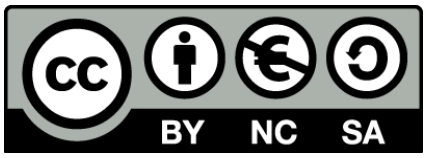
UNIVERSITAT DE BARCELONA

Towards Video Transformers for Automatic Human Analysis

Javier Selva Castelló



DOCTORAL THESIS

Towards Video Transformers for Automatic Human Analysis

Autor: Javier SELVA CASTELLÓ

Director: Prof. Sergio ESCALERA GUERRERO

Co-Director: Dr. Albert CLAPÉS I SINTES

UNIVERSITAT DE BARCELONA

Towards Video Transformers for Automatic Human Analysis

Memòria presentada per optar al grau de doctor per la Universitat de Barcelona

Programa de doctorat en Matemàtiques i Informàtica

Autor: Javier SELVA CASTELLÓ

Director: Prof. Sergio ESCALERA GUERRERO

Co-Director: Dr. Albert CLAPÉS I SINTES

Tutor: Prof. Sergio ESCALERA GUERRERO

UNIVERSITAT DE BARCELONA

Declaration of Authorship

I, Javier SELVA CASTELLÓ, declare that this thesis titled, "Towards Video Transformers for Automatic Human Analysis" and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Javier SELVA CASTELLÓ
September 1st 2023

"It turns out that an eerie type of chaos can lurk just behind a facade of order - and yet, deep inside the chaos lurks an even eerier type of order."

Douglas R. Hofstadter

Abstract

Towards Video Transformers for Automatic Human Analysis

by Javier SELVA CASTELLÓ

With the aim of creating artificial systems capable of mirroring the nuanced understanding and interpretative powers inherent to human cognition, this thesis embarks on an exploration of the intersection between human analysis and Video Transformers. The objective is to harness the potential of Transformers, a promising architectural paradigm, to comprehend the intricacies of human interaction, thus paving the way for the development of empathetic and context-aware intelligent systems. In order to do so, we explore the whole Computer Vision pipeline, from data gathering, to deeply analyzing recent developments, through model design and experimentation.

Central to this study is the creation of UDIVA, an expansive multi-modal, multi-view dataset capturing dyadic face-to-face human interactions. Comprising 147 participants across 188 sessions, UDIVA integrates audio-visual recordings, heart-rate measurements, personality assessments, socio-demographic metadata, and conversational transcripts, establishing itself as the largest dataset for dyadic human interaction analysis up to this date. This dataset provides a rich context for probing the capabilities of Transformers within complex environments. In order to validate its utility, as well as to elucidate Transformers' ability to assimilate diverse contextual cues, we focus on addressing the challenge of personality regression within interaction scenarios. We first adapt an existing Video Transformer to handle multiple contextual sources and conduct rigorous experimentation. We empirically observe a progressive enhancement in model performance as more context is added, reinforcing the potential of Transformers to decode intricate human dynamics. Building upon these findings, the Dyadformer emerges as a novel architecture, adept at long-range modeling of dyadic interactions. By jointly modeling both participants in the interaction, as well as embedding multi-modal integration into the model itself, the Dyadformer surpasses the baseline and other concurrent approaches, underscoring Transformers' aptitude in deciphering multifaceted, noisy, and challenging tasks such as the analysis of human personality in interaction.

Nonetheless, these experiments unveil the ubiquitous challenges when training Transformers, particularly in managing overfitting due to their demand for extensive datasets. Consequently, we conclude this thesis with a comprehensive investigation into Video Transformers, analyzing topics ranging from architectural designs and training strategies, to input embedding and tokenization, traversing through multi-modality and specific applications. Across these, we highlight trends which optimally harness spatio-temporal representations that handle video redundancy and high dimensionality. A culminating performance comparison is conducted in the realm of video action classification, spotlighting strategies that exhibit superior efficacy, even compared to traditional CNN-based methods.

Resum

Towards Video Transformers for Automatic Human Analysis

per Javier SELVA CASTELLÓ

Amb l'objectiu de crear sistemes artificials capaços de reflectir les intrincades habilitats de comprensió i interpretació inherents a la cognició humana, aquesta tesi s'embarca en una exploració de la intersecció entre l'anàlisi humana i els Transformers per a vídeo. L'objectiu és aprofitar el potencial dels Transformers, una prometedora família d'arquitectures, per tal d'entendre les complexitats de la interacció humana, obrint així el camí per al desenvolupament de sistemes intel·ligents, empàtics i conscients del seu entorn. Per aconseguir-ho, explorem totes les branques de la Visió per Computador, des de la recollida de dades fins a l'anàlisi del estat del art, passant pel disseny i l'experimentació amb aquests models.

Una de les pedres angulars d'aquest estudi és la creació d'UDIVA, un ampli conjunt de dades multimodal i multivista que captura interaccions humanes diàdiques cara a cara. Amb 147 participants i 188 sessions, UDIVA integra contingut audiovisual, freqüència cardíaca, perfils de personalitat, dades sociodemogràfiques i transcripcions de les converses, establint-se com el conjunt de dades més gran per a l'anàlisi d'interacció humana diàdica publicat fins ara. Aquestes dades representen un context ric per investigar les capacitats dels Transformers en entorns complexos. Per tal de validar la seva utilitat, així com per verificar aquestes habilitats dels Transformers, ens focalitzem en la regressió de la personalitat dins dels escenaris d'interacció. Primer adaptem un Transformer de vídeo per tal d'integrar múltiples fonts contextuals. Després d'una rigorosa experimentació, obtenim una millora progressiva en els resultats a mesura que s'afegeix més context, validant el potencial dels Transformers per modelar dinàmiques humanes complexes. Arrel d'aquests resultats dissenyem el Dyadformer, una arquitectura per interaccions diàdiques de llarga duració. En modelar conjuntament ambdós participants en la interacció, així com incorporant la multimodalitat al mateix model, el Dyadformer supera la nostra primera proposta i altres treballs concurrents, destacant l'aptitud dels Transformers per resoldre tasques enrevessades, sorolloses i desafiants.

No obstant això, aquests experiments revelen els reptes a l'hora d'entrenar Transformers, especialment relatius al sobreajustament, derivat de la seva demanda de grans conjunts de dades. En conseqüència, concloem la tesi amb una investigació exhaustiva sobre els Transformers per a vídeo, analitzant els dissenys arquitectònics i estratègies d'entrenament, el preprocessament dels vídeos i la tokenització, així com la multimodalitat i aplicacions específiques. Entre aquestes, destaquem tendències que aprofiten de manera òptima les representacions espaciotemporals per tal de gestionar la redundància del vídeo i l'alta dimensionalitat. Per finalitzar, realitzem una comparació de rendiment en l'àmbit de la classificació d'accions a vídeo, destacant estratègies que mostren una eficàcia superior, inclús quan es comparen amb els mètodes tradicionals basats en convolucions.

Acknowledgements

I want to start by thanking all my fellow *altellans*, Alejandro, Carles, Marc, Ciprian, Camilo, Jose, and Sorina, and overall to the HuPBA team, which I've had the honour of being part of. I have greatly enjoyed sharing ideas, long lunch-hour conversations, and worldviews. Spending all those work hours with you has made this office a nicer place to work in. Special thanks go to Alejandro and Carles, who have become true friends. Thanks, Alejandro, for always being there, day and night, for lending me your ears when I needed them the most, for keeping me company during the long working hours, and for teaching me to care. The office was never the same after you left. And thanks, Carles, for helping me see when problems are systemic and not entirely my own fault, for motivating me to keep pushing hard but also reminding me that there is more to life than work. You too, soon, will make it.

I am very grateful to both my supervisors, Sergio and Albert. Sergio, thanks for your guidance and support through all these years, and for being empathetic, patient, and understanding on my hardest moments. I would have not made it without all your optimism and motivation. Albert, I cannot thank you enough for your deep implication in every problem I faced during this thesis, either big or small. In all those long meetings I have not only learned invaluable lessons about machine learning but also about better handling my life.

I must also express my gratitude to my collaborators Marc, Cristina, and Albert. My PhD would have not been possible had I not shared it with you, thanks for taking me in, helping me start, build, and finish this part of my life. Collaborating with you has made this long and difficult journey more bearable, all those long work hours were more enjoyable because you were there with me.

I had the privilege of visiting the VAP lab at Aalborg University, led by Thomas Moeslund. I am really grateful for this opportunity. Anders, Malte, and Gala, I really appreciated you welcoming me there during the months of my research stay. It was really a life-changing experience. Mange tak.

Y por último, aunque en absoluto por ello menos importante, gracias a toda la gente que tengo la suerte de tener en mi vida más allá de la academia. Gracias a mis tres progenitores por criarme, hacerme quien soy, y por todo el apoyo a lo largo de toda mi vida. Gràcies a la meua parella Ana, per més comprensió, afecte i suport emocional del qual seria raonable demanar. T'estime. Gracias a Teo y a Minerva, por los mejores años en Barcelona, llenos de experiencias, viajes y descubrimientos. Gracias a Álvaro, Alberto, Samuel y Heine, por esas conversaciones inabarcables, esas reflexiones sobre la vida, el universo y sobre ser humanos. Y gracias a Ludo y a Enrique, por ayudarme a poner en duda mis preconcepciones del mundo. Estoy realmente agradecido de teneros a todos en mi vida, y espero que sea por muchos años más.

Contents

List of Figures

List of Tables

List of Abbreviations

CV	Computer Vision
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
UDIVA	Understanding Dyadic Interactions from Video and Audio signals
CBQ	Children Behavior Questionnaire
EATQ-R	Early Adolescent Temperament Questionnaire
BFI-2	Big Five Inventory
PEQPN	Post Experimental Questionnaire of Primary Needs
HA	High Angle camera
FC1 and FC2	Frontal Cameras (1 and 2)
GF	General Frontal camera
GB	General Rear camera
MR	Media Recorder
DTW	Dynamic Time Warping
OCR	Optical Character Recognition
SA	Self-Attention
CA	Cross-Attention
MHSA	Multi-Head Self-Attention
MHCA	Multi-Head Cross-Attention
PFFN	Point-wise Feed-Forward Network
STE	SpatioTemporal Encodings
FC	Fully Connected
IoU	Intersection over Union
MSE	Mean Squared Error
VT	Video Transformer
RPN	Region Proposal Network
EF	Encoder Fusion
HEF	Hierarchical Encoder Fusion
CAF	Cross-Attention Fusion
CoAF	Co-Attention Fusion
SSL	Self-Supervised Learning
NLP	Natural Language Processing
MTM	Masked Token Modeling
IP	Interaction Prediction
MTA	Masked Token Alignment

Al meu iaio, per ensenyar-me el valor de l'esforç.

Chapter 1

Introduction

1.1 Motivation

An intricate human world. The world is characterized by an inherent complexity, abounding with stimuli and intricate details, rich in information that can take many forms. Humans have evolved a heterogeneous toolkit of senses, allowing us to relate to our environments through a wide range of signals (visual, olfactive, gustatory, aural, tactile, etc.). Moreover, our brains have adapted over millions of years to properly integrate such varied sensory input by optimizing for survival. We are able to build a coherent and complete representation of the world around us. This allows us to perceive, reason about, understand, and interact with our environments and each other. Moreover, this has allowed humans to build increasingly complex technological spaces and societies. From social media to the ease of global mobility, we live surrounded by a great diversity of viewpoints, cultures, and experiences. This highlights the importance of empathy and understanding of each other.

Building artificial systems that can effectively extract valuable information, integrate diverse sources, and capture meaningful relationships from the complexity of our world is a formidable challenge. Furthermore, bridging the gap between technology and society requires the development of artificial systems capable of understanding our desires, empathizing with us, and engaging in natural, human-like communication. Achieving these goals demands a deep understanding of human cognition as well as the intricacies of human interaction. These technological advances could grant us a new paradigm of interaction between humans and machines, opening the door to unimaginable scientific advances.

Human perception. Crucially, we possess attentional capacity, enabling us to selectively focus on a few pertinent signals while filtering out the overwhelming noise and sensory overload that surrounds us. Moreover, we possess a brain structure that orchestrates the integration of multiple sensory modalities: the thalamus. In this sense, attention allows us to prioritize relevant information, while the thalamus puts everything together, leading to a holistic and meaningful perception of our environment. For instance, in Figure 1.1a multiple cues such as the sounds of birds, smells of flowers, rattling leaves, and bright colors are integrated by the thalamus to construct a single world representation including individual objects. In that setting, attention allows to focus on the hidden tiger, which poses a threat, so the appropriate response can be activated as soon as possible. These abilities are especially

(A) (B)

FIGURE 1.1: The complexity of the world requires proper handling of relevant context while discarding irrelevant information. **(A)**: A monkey in a cluttered environment, multisensory information must be properly integrated to create a unique perceptual representation of the world with part-whole structures. This allows attention to separate irrelevant cues (*e.g.*, flowers or birds) from potential threats (*e.g.*, a tiger lurking in the dark), such that responses to the environment can be taken in time. **(B)**: Depiction of a hungry monkey. In order to get food, it needs to focus on current observations of nuts and rocks, while relating distant memories of the destructive power of rocks and the availability of food inside nuts shells. All this whilst discarding other unrelated observations (*e.g.*, leaves or butterflies) and memories (such as swimming or climbing). Source images to create these figures were generated using Bing Image Creator[1].

relevant when trying to solve a specific task, allowing us to integrate varied sources of context and discard irrelevant information in order to formulate a solution. For instance, discovering the means to crack a nut open requires integration of past memories related to nuts and rocks (knowledge that there is food inside the shell, destructive power of rocks) and recent visual events (nearby nuts and rocks), while discarding unrelated memories and other distracting present elements (see Figure 1.1b).

Vision is our most complex sense [379, 380]. As put by G. Pariente [282]: the human "may be regarded as occupying the summit of visual evolution in the animal kingdom". He was probably referring to vision beyond mere sensing, as it also involves forming assumptions and predictions based on the physical stimuli received by the eyes. Indeed, around 27% of the human cortex is dedicated to tasks concerning visual function [383]. Regarding sensing itself, the aforementioned complexity of the natural world stresses the need for attentional mechanisms to selectively highlight salient elements while filtering out superfluous information. For this, our visual system relies on the interaction of foveal vision and precise eye movements. Foveal vision refers to the ability to see with exceptional detail at the center of our visual field. This ability is crucially complemented by eye movements as a means to actively explore our environment. These eye movements allow us to shift our focus and direct

[1] `https://www.bing.com/images/create`

our attention to different regions of interest within a visual scene. These abilities allow us to scan the environment, track moving objects, and selectively extract relevant information with remarkable flexibility (compared, for instance, to hearing, which is limited to head motion). Crucially, vision does not operate in isolation. The interplay between our internal world models, memories, and recent perceptions participate in building a holistic representation of a given situation. It is this complete context that drives our attention toward specific stimuli or ideas. It is precisely this contextual framework that enables us to discerningly direct our focus within the visual field. For instance, auditory cues can guide our attention toward occluded or out-of-view areas, enabling vision to verify the source of a sound.

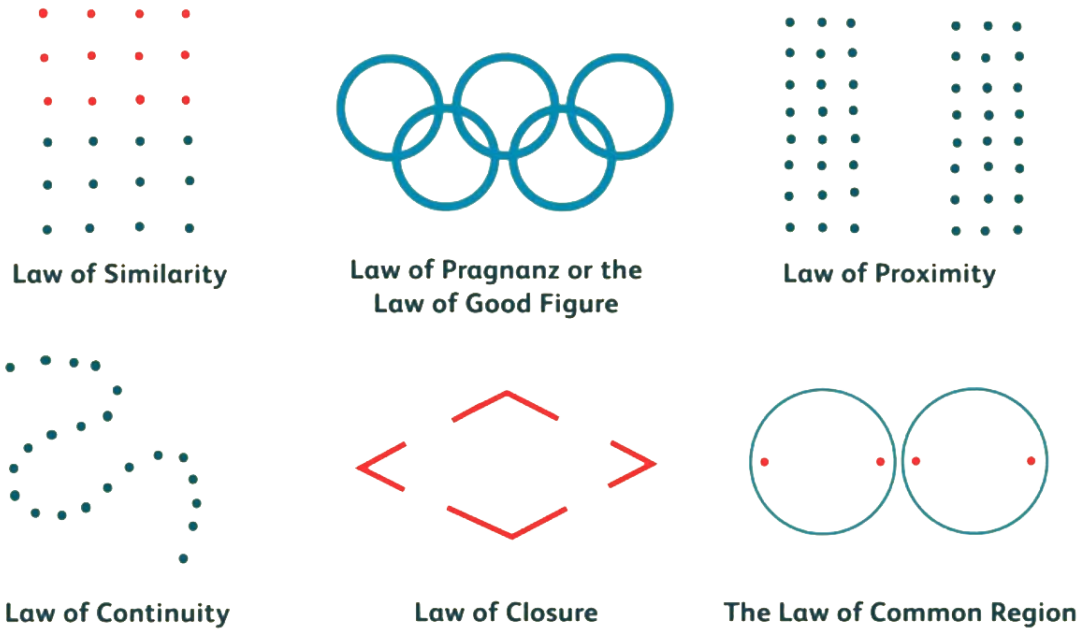


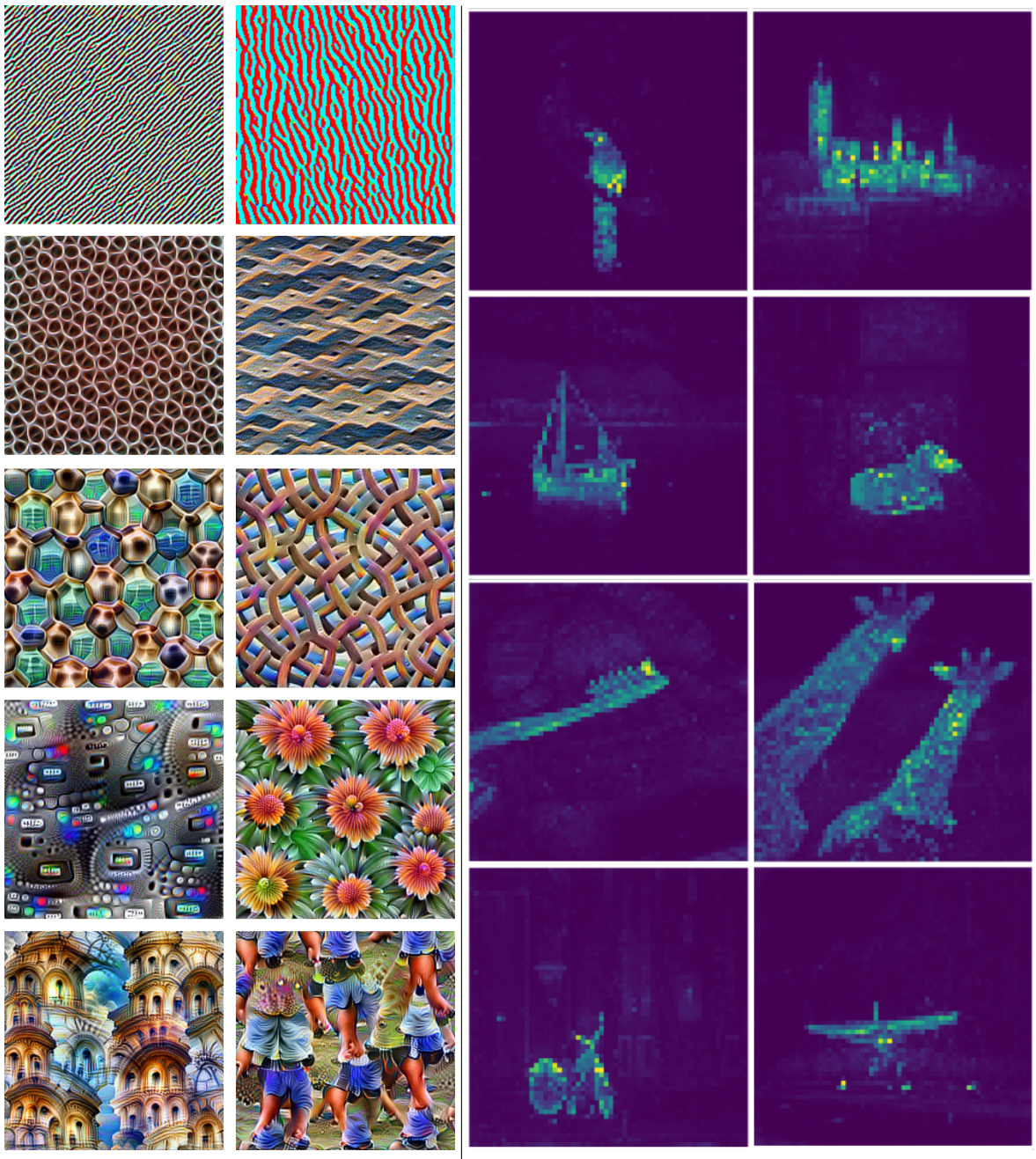
FIGURE 1.2: A few of the Gestalt principles which exemplify how our visual system finds structure within complex patterns. **Similarity**: We tend to group similar entities together (red vs. blue); **Prägnanz**: perception of complex objects as simple as possible (here we perceive 5 complete circles, instead of perceiving partial circles and their intersections as separate entities); **Proximity**: we perceive closer dots as more related than distant ones; **Continuity**: we perceive the points as forming a single curved entity, instead of them being independent; **Closure**: we tend to fill in the gaps, to perceive a rhombus instead of four independent corners; and **Common region**: we perceive each pair of red dots as belonging to a different group. Original figure by James Bascara, from an article by [72].

Patterns and structure in vision. Visual scenes are characterized by dynamic changes, occlusions, cluttered environments, and varying lighting conditions. In order to make sense of such turmoil, our visual system seems to rely on mechanisms to parse the world with an innate structure. In other words, we impose organization on sensory data, unveiling regularity, coherence, and continuity. We are able to detect salient elements and anticipate semantically meaningful relationships. This aligns with the Gestalt laws of perception [410], and notably with the *prägnanz* principle, which dictates that our visual system naturally strives for simplicity and order (see Figure 1.2). Traditionally, and from a neurological point of view, the visual system has been described as a hierarchy of descriptors [171, 384]. However, world

representations based solely on hierarchies of descriptors of surface statistics (*e.g.*, depth, edges, shading...) may be insufficient to explain the phenomenon of structure perception. Elementary surface processing is data-driven, while higher-level perception (like that proposed by Gestalt principles) is goal-driven, guided by semantic models and the expectations of the perceiver. Recognition involves leveraging such models in order to establish correspondences and grouping patterns of scene primitive's attributes into cohesive wholes [414] (*i.e.*, find structure within those hierarchies of descriptors). This can be further exemplified by a recent theory put forward in [58], in which our brains would be using some form of grammar to represent visual concepts which may have been a precursor of the language we humans developed. Indeed, studies have shown that both visual and language-based stimuli activate the same abstract concepts in the brain [40]. This highlights the fantastic ability of our vision to impose structure, categorize, and group complex perceptual patterns as single entities and the relationships between them. Finally, it is noteworthy to mention that temporal structure perception seems to play a pivotal role in comprehending our environment [38]. Not only to perceive and understand motion and dynamic changes but also as a means to parse appearance and static cues, hence suggesting that time perception "can promote spatial grouping" [37].

Limitations of current artificial vision. The more we delve into the intricate mechanisms of human perception, it becomes increasingly evident that our perceptual abilities surpass those of traditional AI methods. While our cognitive system seamlessly integrates multimodal information, directs attention to relevant cues, and forms a rich semantic structure to understand the world, contemporary Computer Vision approaches fall short compared to the depth and complexity of human perception, as we explore next.

In Computer Vision (CV), the ubiquitous Convolutional Neural Network (CNN) has shown great performance and improvement over the years in a myriad of scenarios and applications [221, 32]. CNNs build a hierarchy of representations of a given visual input based on interactions between local descriptors of the previous layer. These hierarchical structures for vision were proposed precisely to mimic the behavior of the early areas of the human visual cortex [115]. In this way, early layers learn simple detectors for edges, whereas later layers detect more complex forms based on those (as can be seen in the work by [272] and in Figure 1.3a). CNNs' success indicates they can indeed extract relevant patterns to form semantically significant representations of the data. Nevertheless, and as discussed above, this may not be enough. CNNs have been shown to prioritize texture over shape cues [125, 24, 124], fixating on surface statistics rather than capturing the underlying structure and higher-level abstractions [182]. This limitation becomes evident in the sensibility exhibited by these networks to high-frequency adversarial attacks [30] (input perturbation through selective color changes to a few individual pixels on an image). This fixation on high-frequency details [2, 438, 394] suggests that these networks may be learning shortcuts to solve tasks based on the wrong reasons [123]. This can be further seen by the tendency of CNNs to misclassify objects when taken out of context [156] (*e.g.*, a pear floating in the ocean). Finally, while there is research analyzing the emergence of structure in vision models (often through the lens of



(A) (B)

FIGURE 1.3: Texture vs. Structure: Comparison between the types of representations learned by CNNs and Transformers. **(A)**: CNNs build representations based on a hierarchy of surface statistics based on learned local descriptors from the previous layer. From top to bottom, increasing depth in the network, linked with increased complexity in the patterns exhibited (figure adapted from [272]). **(B)**: Transformers learn global context interactions and prefer to group semantically significant regions which are input-dependent (figure adapted from [55]). We note that these two forms of visualization are not strictly comparable and are here used for illustrative purposes only, displaying the findings of several works cited in the main text.

Gestalt principles [191, 229]) this has exclusively been studied on image models, yielding inconsistent results [23, 35]. As we have mentioned, time seems to play a crucial role in how these spatial perceptual structures emerge [38], so these findings may require further analysis in the context of video.

Nevertheless, traditional methods dealing with temporal data have limitations of their own. One key issue, especially in video, is that they fundamentally fail to handle long sequences. On the one hand, the Recurrent Neural Network (RNN) family encounters numerical instability due to vanishing and exploding gradients when training on longer sequences [198, 285]. Furthermore, as they collapse sequential observations into unique recurrent states, as more elements are processed, distant information is progressively diluted [21]. On the other hand, CNN-based video models still rely on hierarchies of local descriptors, which are directly extended from image models into the temporal dimension. CNN models excel in tackling appearance-biased tasks, where simply spotting specific objects or backgrounds may suffice (see Section 4.4.3). This capability extends to both static images and, when integrated with recurrent techniques, even videos. Nonetheless, proper modeling of spatiotemporal information in videos requires careful treatment of both space and time. Despite that, existing architectures tend to aggregate information in both dimensions, which impedes some distant temporal relationships to form at the right level of abstraction (see Section 4.3.5). Furthermore, these methods struggle more when solving video benchmarks that require stronger temporal reasoning capabilities, such as Something-Something-v2 [139] (see Section 4.6.2). From this perspective, it may be easier to explain the comparatively limited progress in addressing video-related tasks, as opposed to image-centric tasks, beyond the issue of dimensionality.

Transformers. Consequently, in order to build intelligent systems, capable of handling intricate environments and leveraging the necessary spatio-temporal context to solve a given task, we require novel architectures. Architectures capable of relating distant elements, making sense of a wide range of signals, and focusing on the prevailing structure of the data based on semantically significant relationships while discarding irrelevant information accordingly. It is in this scenario where the Transformer architecture was proposed [385], first intended to better handle sequential data (see Figure 1.4). Transformers use non-local operations (in contrast to traditional, local-based methods like CNNs), thereby enabling them to relate distant components of the input from the first layer. In this sense, they are able to parse the whole context in fewer operations, highlighting fine-grained and semantically meaningful parts of the input. This capability assists in discovering more complex patterns of interactions among temporally remote cues. Transformers use self-attention as their core operation (see Section 3.1.1) which is dynamic and input-dependent (in contrast to the also global but learned fully connected networks). This offers a two-fold benefit: on the one hand, the non-locality of the self-attention allows to associate distant parts of the input; and on the other hand, it is probably thanks to this operation that they group semantically related parts of the input [252, 292]. In this regard, Transformers have been shown to be less sensible to adversarial attacks [33], as they favor more semantic representations [454] (focusing on broader

context and structure instead of on small distracting details) that have even been found to easily transfer to other modalities [238, 363]. In this line are the findings in the work by [55], who showed that, if trained properly, Transformers automatically find salient objects in images (see Figure 1.3b). All this evidence suggests that this novel architecture may be better able to pay attention to structure within the input, focusing on semantically significant content useful to solve the task at hand. Finally, and interestingly, Transformers relax assumptions about the data, giving great flexibility to be adapted to any modality, but also to work seamlessly in multi-modal settings that require information fusion. As we will see in Chapter 3, this will prove to be a very useful feature when in need of leveraging multi-modal cues to solve a complex task.

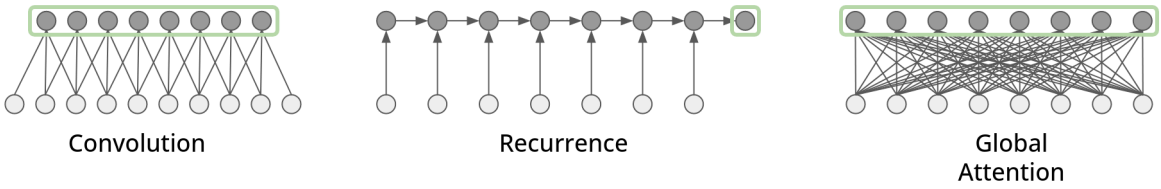


FIGURE 1.4: Difference between the way in which temporal input is processed by different architectures. Light-grey circles depict individual input sequence elements, dark-grey circles illustrate the output sequence representation while green boxes highlight the output that will be forwarded to the next layer. **CNNs** learn representations based on local interactions of the previous layer. **RNNs** sequentially accumulate observations in a single recurrent state, difficulting relationships between distant elements. **Transformers** leverage global contexts through self-attention, allowing to leverage the whole input at each layer. While fully connected layers also use global contexts at each layer, we note that in the case of Transformers, these relationships are input dependent instead of fixed.

Human analysis. However, Transformers are still novel and so far we only possess preliminary evidence. In this sense, we wonder, can they really make sense of complex and multimodal environments, leveraging minutely detailed spatiotemporal context to solve tasks that demand fine-grained perception? To put that to the test, we need a complex environment and a likewise complex task that require such sophisticated perceptual abilities.

Human behavior understanding, with its inherent complexity and the multifaceted nature of humans, presents an intriguing and fertile ground for evaluating the true extent of Transformer capabilities. Humans are very complex creatures with diverse desires and motives: understanding ourselves (why we do what we do) is a task that, in itself, *we* have not fully mastered, and still tend to make false assumptions about others. Human behavior understanding involves analyzing how humans respond to internal and external stimuli, considering the interplay of environmental and psychological factors. The manner in which humans behave is influenced by cultural backgrounds, ethics, interpersonal relationships, politics, and social norms, as well as the behaviors of others we interact with. This has already been clearly stated by the psychology research field: human behavior is much better understood in situations [94]. In other words, context is crucial for understanding ourselves and others. Yet, the integration and interpretation of complex environments to develop a comprehensive understanding of the world, and especially of human behavior, poses an exceptional challenge. Understanding humans requires understanding complex audiovisual cues and the relation

(A) (B) (C) (D)

FIGURE 1.5: Illustration of four potential applications of intelligent empathetic systems capable of humane interaction. **(A)** Personalized assisted living for disabled or elderly people, allowing them for certain degree of independence. **(B)** The integration of these systems has the power to democratize technology and science, enabling people of all backgrounds and ages to effortlessly engage and contribute to advancing our collective knowledge. **(C)** Tailored teaching through custom pedagogical agents that adapt to the needs of each person. **(D)** Hightened human understanding can allow to build emotional companions and even support therapy. Source images to create these figures were generated using Bing Image Creator[1].

between potentially distant events. It involves deciphering facial expressions, body gestures, intonation, and vocalization [11], among many other aspects.

Human interaction. Perceiving all these subtleties required to analyze human behavior demands a scenario that allows humans to freely and naturally express all those behavioral cues. So we ponder, what setting provides such opportunities? Humans are inherently social creatures. It is believed that some intelligent behaviors humans exhibit may have evolved in the context of socializing [298]: we developed language and reason as a means to convince others of our worldview [256]. It may make sense, then, to try to understand human behavior in a context that permeates our everyday lives and may have partly shaped that same behavior: human face-to-face interaction. Moreover, interaction is a setting where the full spectrum of overt behavioral cues can be observed: as we are communicating with others, we reveal part of our emotions through facial expressions or varying the pitch of our voice, we can convey our opinions through speech, display empathy and rapport by mirroring body language, utilize eye contact to convey attention and interest, employ speech pauses to indicate hesitation or thoughtfulness, express assertiveness or engagement through body posture, and use physical contact to convey intimacy or comfort, to name a few. It is also within human interaction that we can study patterns of leader-follower behaviors, as well as the results of cognitive processes within cooperative or competitive task solving.

We hypothesize that, given the abilities exhibited so far by Transformers, they are a promising tool to better understand humans. We are interested in pushing the limit, facing Transformers with a task that can be noisy, subjective, and demanding of abstract representations from challenging environments that require fine-grained spatio-temporal integration of visual and multi-modal cues.

Finally, this particular setting opens the door to many different potential applications (see Figure 1.5). For instance, aiding anthropology or psychology research to further process and analyze great quantities of information from varied sample populations. Another

fascinating application involves the development of empathetic systems capable of comprehending our desires and engaging with us in a more humane manner. Such systems could enable equitable access to technology, transcending barriers based on age or cultural background. This increased technological accessibility not only empowers individuals to satisfy their innate curiosity but also encourages wider participation in technology and science. By lowering access barriers, more people can engage with these fields, fostering research, driving the development of new technologies, and ultimately improving our lives and advancing our species.

1.2 Thesis outline

The objective of this thesis is to work towards a spatiotemporal modeling system capable of extracting useful information for the automatic analysis of humans. In order to advance research in that direction we traverse the entire spectrum of the Machine Learning pipeline. From data gathering to exploring the theoretical potential of the novel Transformer architecture, through modeling the complexity of human face-to-face interactions by means of empirical application of these models. Modeling and understanding humans in interaction is a challenging endeavor, and as more people are involved in an interaction, the behavioral patterns that arise become increasingly complex and nuanced. In order to provide a first incremental step we focus on dyadic face-to-face interactions (*i.e.*, between two participants only). It is in this context that we begin an exploration of human interaction datasets that allow for such intricate analysis. Unfortunately, we find a scarcity of databases that meet all our criteria: dyadic face-to-face interaction, containing detailed annotations that demanded fine-grained task solving, large enough to train a Transformer, as well as with a set of participants and situations that were diverse enough. It is for this reason that we embark on a journey to construct a large dataset capturing human interactions: UDIVA.

The UDIVA dataset (see Chapter 2) is a significant endeavor that involves extensive collaboration with researchers from various fields and institutions. It involves designing an interaction protocol with varied tasks, recruiting participants, recording interaction sessions, processing the gathered data, as well as annotating with varied labels. The initial version of the dataset contains self- and peer-reported personality scores, literal transcriptions of the conversations, as well as other metadata such as gender, ethnicity, or age of the participants. The UDIVA dataset currently continues to be expanded with additional annotations and keeps contributing to the advancement of research as a valuable resource for challenges and workshops. Nonetheless, in its current state, which task could we pose to the Transformer that requires a fine-grained understanding of this interaction environment? Given no dyadic labels are available yet, the challenges we were interested in presenting to the Transformer are best represented by the task of self-reported personality inference.

With this in mind, we build an initial baseline that leverages the Transformer to solve this intricate task (see Chapter 3). The non-local operation at the core of Transformers (see Section 3.1.1) promises great context modeling capabilities. It should allow for picking up and integrating small cues across the input video such that the dyadic interaction can be modeled.

We first benchmark this possibility by extending an existing Transformer architecture [131] to our setting. With it, we are able to successfully predict personality from a target participant given a small 3-second chunk of audiovisual recordings. Motivated by this, we build the Dyadformer, a Transformer-based architecture that effectively integrates multiple information sources through attentional mechanisms from longer videos (up to 30 seconds). This is done in two fashions: on the one hand video and audio representations are fused to construct a comprehensive audiovisual model of the scene; and on the other hand, it enables reciprocal information access among each participant's features. The multi-modal joint modeling of both interactants helps predict individual personality scores, outperforming the baseline and other models from a recent challenge on the UDIVA dataset [281].

During the development of the Dyadformer, we encountered numerous challenges. Factors such as the high dimensionality of video and the requirement for large amounts of data posed significant obstacles. It also became evident that relying solely on supervised training for intricate tasks, such as personality prediction, might not suffice. Given the relative novelty of Transformers, there are still many aspects that remain not fully understood. Consequently, in Chapter 4 we dedicate the remainder of this thesis to conducting an exhaustive analysis of the Video Transformers domain. We dissect them from how embedding networks are used, to self-supervised training strategies, going through architectural design choices specifically tailored to better capture the nuances of video data. Furthermore, we accompany our detailed exploration with a performance comparison of over 20 Video Transformers on the task of video action classification, finding them to outperform CNN counterparts, even at cheaper computational costs in some cases. This effort culminates in an extensive survey that encompasses comprehensive taxonomies and analyses, together with detailed insights and extensive discussions, shedding light on the recent advancements in the field of Video Transformers.

To conclude, given the knowledge gathered during this survey, we are currently extending the Dyadformer to leverage self-supervised losses that can more effectively learn human personality from the UDIVA dataset (see Section 5.1). Our preliminary experiments suggest that this indeed seems to be the case, proving self-supervised learning to be one of the key elements to infuse both general and domain-specific semantics into neural networks. We hope this will provide us with better tools to further our understanding of human behavior, from both individual and dyadic settings, in the context of Video Transformers.

1.3 Summary of Contributions

Next are the main contributions derived from the development of the current thesis.

- We introduce **UDIVA**, a new non-acted **dataset** of face-to-face dyadic interactions, where interlocutors perform competitive and collaborative tasks with different behavior elicitation and cognitive workload. The dataset consists of 90.5 hours of dyadic interactions among 147 participants distributed in 188 sessions, recorded using multiple audiovisual and physiological sensors. Currently, it includes sociodemographics, self- and peer-reported personality, internal state, literal transcriptions and relationship profiling from

participants. To the best of our knowledge, there is no similar publicly available, face-to-face dyadic dataset in the research field in terms of the number of views, participants, tasks, recorded sessions, and context labels.

- As an initial analysis on UDIVA, we propose a **transformer-based method for self-reported personality inference** in dyadic scenarios, which uses audiovisual data and different sources of context from both interlocutors to regress a target person's personality traits. Preliminary results from an incremental study show consistent improvements when using all available context information.

- We then extend this baseline and present the ***Dyadformer***, a novel multi-modal multi-subject Transformer architecture to model individual and interpersonal features in dyadic interactions using variable time windows, thus allowing the capture of long-term interdependencies. Our proposed cross-subject layer allows the network to explicitly model interactions among subjects through attentional operations. Different to the baseline, here the videos from both participants are simultaneously used as target and context. This approach shows how multi-modality and joint modeling of both interactants for longer periods of time helps to predict individual attributes. With Dyadformer, we set new state-of-the-art self-reported personality inference results on individual subjects on the UDIVA dataset.

- Finally, we review over 100 Video Transformer works and present a survey where we analyze the main contributions and trends of works leveraging Transformers to model video. In order to do so, we propose detailed taxonomies of the various design choices throughout the whole Video Transformer pipeline. Specifically, we delve into how videos are handled at input level first. Then, we study the architectural changes made to deal with video more efficiently, reduce redundancy, re-introduce useful inductive biases, and capture long-term temporal dynamics. In addition, we provide an overview of different training regimes and explore effective self-supervised learning strategies for video. Finally, we conduct a performance comparison on the most common benchmark for Video Transformers (i.e., action classification on Kinetics 400 [57] and Something-Something-v2 [139]), finding them to outperform 3D ConvNets even with less computational complexity. We complement every analysis with extensive discussions and thorough insights.

Chapter 2

UDIVA: A Dataset of Human Interaction

2.1 Introduction

People interact with others in many ways, guided by their culture, beliefs, relationships, context, personality, or emotional and health state. Social interactions permeate our lives, beyond our direct relationships with friends and family. They provide the foundation for many of our daily activities, such as work, food access, education, entertainment, exercise, and more. This is one of the reasons why human-human interaction has been widely studied in both sociology and psychology, and more recently, from a machine learning standpoint. Generally speaking, recent technologies are interested in addressing automatic human behavior analysis [4], and/or in the research and development of applications for human-machine interactions [203]. In this new era of ubiquitous intelligent systems becoming more and more seamlessly integrated into our daily lives, the future looks daunting for part of society. "How should we adapt to this new technology? How will the machines learn to interact with us?". While valid, such questions and fears can be tackled from a different perspective, that is, "How can technology providers train these systems to interact in a more humane way while accounting for possible societal biases?". Such questions motivate this work to analyze the simplest way of human-human interactions, *i.e.*, the interaction between two individuals or the so-called *dyadic interaction*.

Current literature in computer vision and machine learning for human behavior understanding has mainly focused on research and development of perception, analysis, and synthesis methods for individual behavior, such as the automated recognition of body movements and actions [291, 56, 248]. Although recent works developed for triadic [184], small groups [121] or multi-party based-interactions [301, 244, 165] can be found in the literature, interpersonal-based tasks such as perception and modeling of the communication flow and the adaptation between communication partners have been largely unexplored from a technological point of view [388]. To advance in such areas, the community is in need of publicly available annotated datasets of non-acted, spontaneous interactions among dyads belonging to different population groups in terms of age, gender, and cultural background. While several acted datasets exist [50], natural interactions are preferred, as they cover the richness and complexity of social communications in real life.

In this chapter, we introduce UDIVA (Understanding Dyadic Interactions from Video and Audio signals), a novel annotated dataset of dyadic face-to-face spontaneous interactions. The purpose is to move beyond automatic individual behavior detection and focus on the development of automatic approaches to study and understand the mechanisms of influence, perception and adaptation to verbal and nonverbal social signals in dyadic interactions, taking into account individual and dyad characteristics as well as other contextual factors. One of our central research question revolves around the feasibility of developing socially-aware systems able to decode complex internal processes and behaviors from individuals by the social signals they convey, and to understand how interaction partners perceive and react to those cues directed to them. UDIVA has been designed and collected with these goals in mind. It is a highly varied multi-modal, multi-view dataset of zero- and previous-acquaintance, non-scripted face-to-face dyadic interactions based on free and structured tasks performed in a lab setup. It consists of 188 interaction sessions, where 147 participants arranged in dyads performed a set of tasks in different circumstances. It has been collected using multiple audiovisual and physiological sensors and currently includes sociodemographic, self- and peer-reported personality, internal state, literal transcriptions of the conversations, and relationship profiling. To the best of our knowledge, there is no similar publicly available, face-to-face dyadic dataset in the research field in terms of the number of views, participants, tasks, recorded sessions, and context labels. The main contribution derived from this chapter is to publicly release UDIVA. By doing so, we aim at fostering multidisciplinary research and the development of new paradigms and technologies for understanding interpersonal behavior.

This chapter is organized as follows: we first introduce human interaction in Section 2.1.1 and personality computing in Section 2.1.2. Then, in Section 2.1.3, we review related dyadic human interaction databases. The data collection process of UDIVA is explained in Section 2.2, with a description of its contents following in Section 2.3. Next, in Section 2.4 we detail the current release of the dataset, UDIVA v0.5, and go through its statistics. We highlight potential applications for good and ethical considerations for the released data in Section 2.5. Finally, Section 2.6 concludes the chapter.

2.1.1 Human Interaction

Human interaction has been a central topic in psychology and social sciences, aiming at explaining the complex underlying mechanisms of communication with respect to cognitive, affective, and behavioral perspectives [48, 43]. In dyadic interactions, we use verbal and nonverbal communication channels to convey our goals and intentions [265, 391] while building a common ground [74]. Both interlocutors influence each other based on the cues we perceive [48]. However, the way in which we handle them depends on a myriad of factors that are crucial to better understanding the nuances of communication and interaction. Such factors, which we refer to as context, may include, but are not limited to: our personal characteristics, either stable (*e.g.*, personality [79], cultural background, and other sociodemographic

information [336]) or transient (*e.g.*, mood [75], physiological or biological factors); the relationship and shared history between both interlocutors; the characteristics of the situation and task at hand [321]; societal norms; and environmental factors (*e.g.*, temperature).

As we have mentioned, social interaction is the basis for most of our daily activities. In this sense, developing tools and systems capable of understanding both, human interaction and behaviour, would open the door to a wide spectrum of applications in many different scenarios. This makes social interactions a complex yet very compelling setting in which to analyze humans: it fosters that all aforementioned sources of context come into play in the form of a wide spectrum of behaviours, demanding AI systems capable of a more comprehensive and complete modeling of human behavior.

Socially intelligent systems are expected to accurately perceive, understand, react, and adapt to the affective and cognitive state of interacting individuals in different contexts, so as to provide a more natural, empathic, tailored communication [83, 102, 280, 270]. To embody this human-likeness into such systems, it is crucial to have a deeper understanding of real human-human interactions first, to allow for the computational modeling of individual and social behaviors, and interpersonal influence [48, 101, 44]. Research in dyadic and small group interactions has enabled the development of automatic approaches for detection, understanding, modeling, and synthesis of individual and interpersonal behaviors, social signals, and dynamics [297, 121, 392, 389]. For measuring interpersonal processes during an interaction (such as non-verbal synchrony [85], rapport [463], or engagement [87]) the joint modeling of all interlocutors and/or other sources of context has been frequently considered. However, for the task of recognizing individual attributes or behaviors in interaction settings, most computational approaches usually consider information from the target interlocutor only. To analyze individual behaviors during a conversation, the joint modeling of both interlocutors is required due to the existing dyadic interdependencies. While these aspects are usually contemplated in non-computational dyadic research [189], context- and interlocutor-aware computational approaches tend to disregard the effect of any sources of context on individual behavior in addition to existing dyadic or group interdependencies [27, 415, 389, 262, 94, 78]. We largely attribute this dficiency to the lack of datasets providing contextual metadata in different situations and populations. In order to foster research on interlocutor- and context-aware approaches for social behavior modeling and understanding, we develop the UDIVA dataset.

2.1.2 Personality

Two of the key labels collected during the recording of the UDIVA dataset are self- and peer-reported personality scores of the participants. Personality is widely defined as the manifestation of individual differences in patterns of thought, feeling, and behavior, that remain relatively stable during time [352]. In recent years, personality psychologists have reached a consensus on the number of basic major dimensions of human personality, which range between five and six [19]. Within the personality computing field [390, 295], personality is usually characterized by the basic Big Five traits [251, 135] – *Open-Mindedness* ("O"), *Conscientiousness* ("C"), *Extraversion* ("E"), *Agreeableness* ("A"), and *Negative Emotionality* ("N"),

often referred to as OCEAN. As described in [390], personality computing often follows the *Brunswik Lens* [45], which is used to describe *externalization* and *attribution* of socially relevant attributes during social interactions. The former is related to personality recognition, which aims to infer self-reported personality traits from observable *distal cues*, i.e., overt behavior. In contrast, the latter is related to personality perception, where the goal is to recognize the *apparent* personality traits of a target person from the perspective of an external observer [176] based on *proximal cues*, i.e., cues that the observer perceives. According to these definitions, we will use the "personality recognition" expression when referring to works that focus on self-reported personality traits (*e.g.*, obtained via self-reported questionnaires), and the "personality perception" expression (or "apparent personality") for referring to works that focus on the personality attributed by external observers. Nevertheless, a method developed for personality recognition could be easily adapted for personality perception and vice versa.

2.1.3 Related work

Human interaction datasets. Research on human behavior and communication understanding has fostered the creation of a plethora of human interaction datasets [196, 300, 94, 355]. Here, we focus on publicly available datasets containing at least audiovisual data, which enable the fusion of multiple modalities and the creation of more complete representations. In the literature, we can find examples of rich, non-acted datasets focused on computer-mediated dyadic settings [51, 201], face-to-face triadic [184, 59], or small group interactions [8]. A number of TV-based datasets with acted interactions also exist [301]. However, in such cases, the interlocutors' internal states are artificially built.

One of the advantages of face-to-face settings is that the full overt behavioral spectrum can be observed and modeled. Existing publicly available face-to-face dyadic interaction datasets are summarized in Table 2.1. Most of them focus exclusively on verbal communication (speech and conversation) or emotion analysis. While some of them (*e.g.*, IEMOCAP [50], NOMCO [279] or MSP-IMPROV [49]) do include annotations that would allow for some form of interaction analysis (*e.g.*, turn-taking), the narrow focus limits the kind of behaviors displayed. MMDB [322] and DAMI-P2C [62] can be highlighted, as they both focus on interaction analysis similar to us, but are limited to adult-infant interaction, missing a broader age representation. Opposed to these, UDIVA has been designed with a multipurpose objective, by including a varied set of tasks that allow for a broader assortment of behaviors to be analyzed, and it includes interactions between a wider range of ages. As can be seen, most of the publicly available interaction datasets are tailor-made for too specific purposes (such as Mimicry, Hirability, or Agreement classification) or limited in the number of participants, recordings (for instance, the second biggest dataset in terms of hours, Spontal [97], lags 30 hours behind the 90.5 hours of UDIVA), views, context annotations, language, or purpose. Hence, there is no big enough general-purpose database in the literature that could allow for an integral analysis of both, the interaction and the participants. For this reason, we next introduce the UDIVA dataset.

Name (Year)	Focus	Interaction	Modality	Annotations	F/M	Sess	Size	#Views	Lang.
IFADV [349] (2007)	Speech & conversation analysis	Non-acted	Audiovisual	Speech features, transcripts	24/10	20	5h	2	Dutch
IEMOCAP [50] (2008)	Emotion recognition	Acted* & Acted	Audiovisual, face & hands MoCap.	Emotions, transcripts, turn-taking	5/5	5	∼12h	2	English
CID [36] (2008)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual	Speech features, transcripts	10/6	8	8h	1	French
Spontal [97] (2010)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual, head & torso MoCap.	Transcripts, speech features	ϕ	120	60h	2	Swedish
NOMCO [279] (2010)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual	Speech & interaction features, gestures, transcripts, emotions	6/6 $^{\phi}$	60	∼6h	3	Danish, Swedish, Finnish
HUMAINE† [92, 93] (2011)	Emotion analysis	Non-acted*	Audiovisual	Emotions	34	18	∼12h	4	English
MMDB [322] (2013)	Adult-infant interaction analysis	Non-acted*	Audiovisual, depth, physiological	Social cues (gaze, vocal affects, gestures...)	121	160	∼13.3h	8 + 1D	English
MAMCO [386] (2014)	Overlap analysis	Non-acted	Audiovisual	Transcripts	6/6	12	∼1h	3	Maltese
4D CCDb [247] (2015)	Speech & conversation analysis	Non-acted	Audiovisual, depth	Facial expressions, head gestures, utterances	2/2	17	∼0.2h	6 + 8M	English
MAHNOB [34] (2015)	Mimicry	Non-acted*	Audiovisual, head MoCap.	Head, face and hand gestures, personality scores (self-reported)	29/31	54	11.6h	2 + 13M	English
MIT Interview [264] (2015)	Hirability analysis	Non-acted*	Audiovisual	Hirability, speech features, social & behavioral traits, transcripts	43/26	138	10.5h	2	English
MPIIEMO [263] (2015)	Bodily emotion analysis	Acted	Audiovisual	Emotions	3/2	8×7×4 (tasks)	∼2.4h	8	German $^{\phi}$
JESTKOD [42] (2015)	Agreement classification	Non-acted*	Audiovisual, body MoCap.	Agreement, emotion	4/6	25	4.3h	1	Turkish
Creative IT [258] (2016)	Emotion recognition	Acted	Audiovisual, body MoCap.	Transcripts, speech features, emotion	9/7	8	∼1h	2	English
MSP-IMPROV [49] (2017)	Emotion recognition	Acted & Non-acted	Audiovisual	Turn-taking, emotion	6/6	6	9h	2	English
NNIME [73] (2017)	Emotion analysis	Non-acted*	Audiovisual, physiological	Emotion, transcripts	22/20	102	∼11h	1	Chinese
RAMAS [293] (2018)	Emotion analysis	Non-acted* & Acted	Audiovisual, depth, body MoCap.	Physiological signals, emotion, interaction traits	5/5	80	∼7h	2 + 1D	Russian
DAMI-P2C [62] (2020)	Adult-infant interaction analysis	Non-acted*	Audiovisual	Emotion, sociodemographics, parenting assessment, child personality (peer-reported)	38/30	65	∼21.6h	1 $^{\phi}$	English
UDIVA (ours) (2020)	Social interaction analysis	$\frac{1}{5}$ Non-acted & $\frac{4}{5}$Non-acted*	Audiovisual, heart rate	Personality scores (self- & peer- reported), sociodemographics, mood, fatigue, relationship type	66/81	188×5 (tasks)	90.5h	6 + 2E	Spanish, Catalan, English

† Here we consider the Green Persuasive and the EmoTABOO [448] datasets together.

TABLE 2.1: Publicly available audiovisual human-human (face-to-face) dyadic interaction datasets. "Interaction", *Acted* (actors improvising and/or following an interaction protocol, *i.e.*, given topics/stimulus/tasks), *Acted** (Scripted), *Non-acted* (natural interactions in a lab environment) or *Non-acted** (non-acted but guided by interaction protocol); "F/M", number of participants per gender (Female/Male) or number of participants if gender is not informed; "Sess", number of sessions; "Size", hours of recordings;"#Views", number of RGB cameras used, and *D* is RGB+D, *E* is Ego, *M* is Monochrome. The ϕ symbol is used to indicate missing/incomplete/unclear information on the source.

2.2 Data collection

In this section we introduce the data collection procedure for the UDIVA dataset. We describe participant recruiment process, the questionaires to assess personality and other internal states, the sensors used (cameras, microphones, etc.), and the tasks that compose the structure of a dyadic session in UDIVA.

2.2.1 Participant recruitment and Questionnaires

Participants were recruited through university and social media ads, as well as word of mouth. Prior to their first dyadic session, participants gave consent to be recorded and to share their collected data for research purposes, in compliance with the EU GDPR[1], and filled in several questionnaires about themselves. A number of individual factors are related to personality and behaviour, such as gender [407] or age [325]. Similarly, cultural and educational indicators have been found to influence interpersonal relationships [336]. Therefore, prior to their first session, each participant filled out a sociodemographic questionnaire, including: age, gender, ethnicity, occupation, maximum level of education, and country of origin. To assess personality and/or temperament, age-dependent standardized questionnaires were administered. In particular, parents of children up to 8 years old completed the Children Behavior Questionnaire (CBQ) [328, 276], participants from 9 to 15 years old completed the Early Adolescent Temperament Questionnaire (EATQ-R) [99, 387], while participants aged 16 and older completed both the Big Five Inventory (BFI-2) [352] and the Honesty-Humility axis of the HEXACO personality inventory [19]. These personality scores were computed on a 1-to-5 scale and were later converted to z-scores using descriptive data from normative samples.

Before and after each interaction session, participants also filled in several questionnaires regarding their current internal state. In particular, all participants (or their parents) completed pre- and post-session mood ([119], as a 1-to-5 rating scale) and fatigue (ad-hoc 1-to-10 rating scale) assessments. The mood assessment contained items drawn from the Post Experimental Questionnaire of Primary Needs (PEQPN [412]). In particular, this included 8 classes: *Good*, *Bad*, *Happy*, *Sad*, *Friendly*, *Tense*, and *Relaxed*. After each session, participants aged 9 and above completed again the previous temperament/personality and mood questionnaires, this time rating the individual they interacted with, to provide their perceived impression. Finally, participants reported the relationship they had with their interaction partner, if any. Participants that did not fill in the pre- and/or post-fatigue questionnaire had their fatigue level set to 0.

2.2.2 Sensors and Recording setting

The setup consisted of 6 HD tripod-mounted cameras recording at a resolution of 1280 $\times$ 720px and 25fps, 1 lapel microphone per participant at 44100 Hz, and an omnidirectional microphone on the table, as depicted in Figure 2.1a. Each participant also wore an egocentric

[1] `https://ec.europa.eu/info/law/law-topic/data-protection_en`.

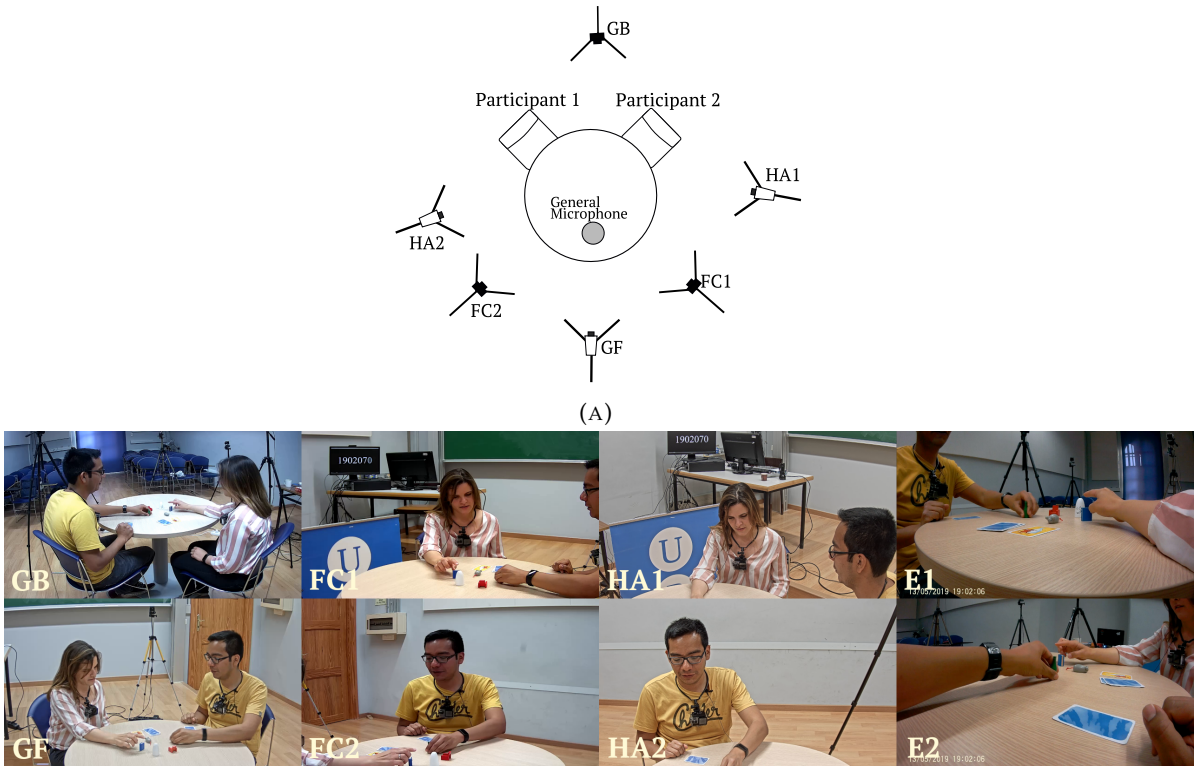


FIGURE 2.1: Recording environment. We used six tripod-mounted cameras, namely **GB**: General Rear camera, **GF**: General Frontal camera, **HA**: individual High Angle cameras and **FC**: individual Frontal Cameras, and two ego cameras **E** (one per participant, placed around their neck). (A) Position of cameras, general microphone, and participants. (B) Example of the time-synchronized 8 views.

camera around their neck recording at 1920 × 1080px and 30fps and a heart rate monitor on their wrist. All the capturing devices are time-synchronized and the tripod-mounted cameras are calibrated. Such setup allows us to collect rich audiovisual data from the participants, in particular with respect to speech, face and upper body. See Figure 2.1b for an example of the camera views. The particular devices used are listed below:

- Tripod-mounted cameras:
  - Two *Revotech i706* (FC1 and FC2) and one *Revotech i712* (GB).
  - Three *Axis M1124* (HA1, HA2 and GF).
- Ego-cameras: Two *Victure AC800*.
- Lapel microphones: Two *RØDE SmartLavPlus*.
- Omnidirectional general microphone: One *Amazon Basics USB Conference Microphone*.
- Heart Rate monitor: Two *Fitbit Charge 3*.

FIGURE 2.2: Examples of the 5 tasks included in the UDIVA dataset from 5 sessions. From left to right: *Talk*, *Lego*, *Animals*, *Ghost*, *Gaze*.

2.2.3 Tasks

Participants were asked to sit at 90° to one another around a table (see Figure 2.1a), to be close enough to perform the administered tasks while facilitating data acquisition. A session consisted of 5 tasks (illustrated in Figure 2.2) eliciting distinct behaviors and cognitive workload:

- **Talk.** Participants were instructed to discuss any subject for approximately 5 minutes. This task allows for analysis of common conversational constructs, such as turn-taking, synchrony, empathy, and quality of interaction, among others. Moreover, it allows to extract further interaction features that can be employed as personality indicators.
- **"Animals" game.** Participants asked 10 *yes*/*no* questions about the animal in the card wore on their own forehead to guess the animal. Animals were classified into 3 difficulty levels. This game elicits cognitive processes (*e.g.*, thinking or gaze signaling events). The duration of this task depends on the participants' ability to find proper questions and the difficulty of the animal.
- **Lego building.** Participants built a Lego together following the instructions leaflet, ranging between 4 difficulty levels. This task fosters collaboration, cooperation, joint attention, and leader-follower behaviors, among others.
- **"Ghost blitz" card game.** Participants played one card per turn, competing with each other to be the first to select the correct figurine from a set of five figurines placed on the table, based on the content of the played card. This task fosters competitive behavior, and allows cognitive processing speed analysis, among others.
- **Gaze events.** Participants followed directions to look *at each other's face*, *at static/moving object*, or *elsewhere*, while moving head and eyes. This task serves as ground truth for gaze gestures and face modeling with varied head poses.

These tasks were selected along with psychologists due to the variety of individual and dyadic behaviors they elicit. In particular, Lego structures have been widely used in observational settings to assess aspects such as communication [1], social skills [228] or teamwork abilities and performance [117]. *Ghost* and *Animals* are examples of board games, proven to be valid assessments of interpersonal skills [147, 381]. All these aspects are, in turn, indicators of personality traits like *Extraversion*, *Agreeableness* or *Conscientiousness* [6]. Cognitive methods, such as the tasks herein used, are routinely used in personality research [18].

Talk was always administered first as a warm-up and *Gaze* was always last, whereas the remaining tasks were administered in randomized order. Prior to the beginning of each task,

it was explained by a lab proctor, who left the recording room while it was taking place. Only for *Gaze*, the proctor gave the instructions while participants performed them. The recording of each task starts when the lab proctor finishes explaining the task to the participants and stops interacting with them, and finishes when the proctor starts interacting with the participants again to deliver the following task. The real given task (*e.g.*, build a Lego building) may finish minutes before the end of the recording. Once they finished the real given task, participants were free to continue playing, talking, or just wait until the proctor entered the recording room and stopped the recording. This extra time allows the emergence of spurious behaviors once the actual task finishes that could prove interesting for analysis (*e.g.*, they play another round, build random legos, talk, stare at the distance in silence, etc.). The specific Lego building and assigned animal cards for each session were selected such that no participants repeated the same Lego or animal twice while forcing a uniform distribution on the number of times each item was used for the total of sessions. To assess their difficulty level, we conducted an anonymous survey among 19 co-researchers.

2.3 Dataset Content

This section introduces the actual contents of the UDIVA dataset. This includes time-synchronized multi-modal, multi-view videos, as well as metadata and transcriptions. We note that egocentric cameras, heart-rate monitors, peer-reported personality, and camera calibration are yet to be processed and synchronized, hence they have not been made publicly available yet.

2.3.1 Synchronization of audiovisual data

The UDIVA dataset currently contains the audiovisual recordings of the 6 HD tripod-mounted cameras (see Figure 2.1a), in *.mp4* format. There is one recording per camera, task, and session, which had to undergo a synchronization process. Cameras were separated into two groups, which are internally synchronized: *Axis* cameras (HA1, HA2 and GF) recorded on one computer using *Noldus' MediaRecorder* software (MR), and *Revotech* cameras (FC1, FC2 and GB) were recorded on a different computer using *CMS* software. Then, the task of synchronization involves aligning the two groups together. To allow for this, a single screen was added in the background, such that it could be seen from one camera of each group (HA1 and FC1), displaying a timestamp with millisecond precision. Furthermore, the CMS group lost 1% of the frames (around 1 frame every 4 seconds). This does not affect the ability to perform analysis on this data, as it is effectively equivalent to having recorded at 24.75 fps, but it does produce synchronization problems. It is for this reason that we used Dynamic Time Warping (DTW) to both align and also suggest where to insert missing frames so the videos would not lose synchrony with time. DTW is an algorithm which compares two sequences and proposes the minimum set of operations (insertion, deletion, or match), to transform one sequence into another. In our case, it is used to detect where, in a sequence of timestamp readings, there is a missing frame. Regarding audio, it was recorded in synchrony with the video through three microphones which were each linked to one camera of the MR group.

One lapel microphone to HA1, another to HA2, and the General Microphone (on the table) to GF. In this sense, auidio does not require any further synchronization.

The process involved the following steps:

1. Manually annotate the timestamps shown in the first and last four frames (8 frames total) of each recorded video from FC1 and HA1.

2. Manually select the region of the video where the screen appears for both FC1 and HA1 (as cameras were only moved between recording days, this was done for just one session of each recording day).

3. For all frames in the video, crop the timestamp region and perform an Optical Character Recognition (OCR) reading of the displayed timestamp[2]). This results in a list of read timestamps in numerical format.

4. For each FC1 video, run DTW with the sequence of timestamp readings compared to an artificially generated sequence (the expected one). The latter was generated through linear interpolation using the manually annotated first and last frames of each video (from step 1). This is needed as OCR detections were not entirely accurate, and while DTW is resilient to the few mistakes, we require an accurate sequence against which to compare the OCR readings. The output of DTW is a list of positions where a frame is missing.

5. Insert missing frames in all three cameras from CMS at those positions by copying the preceding frame.

6. Compute the delay between camera groups by using the manually annotated timestamps.

7. Use the delay to copy the audio from HA1 to FC1, from HA2 to FC2 and from GF to GB.

8. Crop the videos by task using manually annotated task timestamps (see *Others* in Section 2.3.2).

9. Finally, a manual verification process involved assessing whether the six cameras were synchronized at the beginning, middle, and end for each task of each session. This was done by visual inspection of the frames from multiple views close to those three temporal positions and listening whether audio coincides with visual cues.

2.3.2 Metadata

Audiovisual data is accompanied by a set of metadata, described next:

Participant. Metadata about each participant, independent of the session. All information except participant and session IDs, and the total number of sessions done, were reported by the participants themselves. It includes:

[2]For this we used Tesseract-OCR library: `https://github.com/tesseract-ocr/tesseract`

- Anonymized participant ID (3-digit string).
- Gender (F - female, M - male).
- Age (integer).
- Country of origin (string).
- Maximum level of education (string).
- Self-reported personality questionnaire (BFI-2) results, in z-scores for the five OCEAN traits (float, 0-centered).
- Total number of sessions done (integer).
- Session IDs in which the participant has participated in order of occurrence (string).

Session. Metadata about the interaction session, as well as participant metadata specific to a given interaction session. It includes:

- Session ID (string).
- ID of Participant 1, recorded on cameras HA1 and FC1 (string).
- ID of Participant 2, recorded on cameras HA2 and FC2 (string).
- Recording timestamp (date and time).
- Difficulty level of Lego task (integer, from 0 to 3).
- Difficulty level of Animal task (integer, from 0 to 2).
- Language (Spanish, English, or Catalan) (string).
- Self-reported relationship among interaction partners (integer, 0 - zero-acquaintance or 1 - previous-acquaintance).
- Order of administration of the tasks within a session (integer from 1 to 5, one per task).
- Animal assigned to each participant in the Animals task (2 strings).
- Lego assigned for the participants to build (string).
- Self-reported answers of the mood questionnaire (PEQPN) per participant (1 and 2) before and after the session (Good, Bad, Happy, Sad, Friendly, Tense and Relaxed, integers from 1 to 5).
- Self-reported fatigue levels per participant (1 and 2) before and after the session (integers from 0 to 10 - the questionnaire contained values from 1 to 10, participants that did not answer this question have a value of 0 assigned).
- Other relevant notes (string).

Other. Start-end times of the real given task within a task recording (*e.g.*, the time when participants start and end building the requested Lego figurine for the Lego task), and start-end times of segments that do not contain audio and/or video because of privacy or technical issues.

2.3.3 Transcriptions

Literal transcripts of the conversations at utterance level were obtained by a third-party company and manually reviewed for cleanliness and data protection. As illustrated in Figure 2.3, a transcript file is composed of: an utterance number, the start and end times of the utterance synchronized to the videos, the ID of the interaction partner (PART.1/2 for the participant visible from FC1/2, respectively), and the transcribed utterance.

No.	Start	End	Duration	Text
44	00:02:34.590	00:02:39.390	00:00:04.800	PART.1: But I was checking and the six, (39) the other one you included... (29)
45	00:02:42.490	00:02:46.590	00:00:04.100	PART.2: So, at the end, (23) it's eleven images you say? (27)
46	00:02:46.690	00:02:47.690	00:00:01.000	PART.1: Yes. (12)
47	00:02:49.690	00:02:50.690	00:00:01.000	PART.2: Really? (15)
48	00:02:50.690	00:02:52.690	00:00:02.000	PART.1: Do you have access here? (32)
49	00:02:52.690	00:02:54.690	00:00:02.000	PART.2: Yes. (12)
50	00:02:57.290	00:03:02.590	00:00:05.300	PART.1: I think you included one (32) that I did not comment (22) but was a good one. (19)

FIGURE 2.3: Example of a transcript from a short conversational segment included in UDIVA.

2.3.4 Main statistics

The dataset is composed of 90.5h of recordings of dyadic interactions between 147 voluntary participants (55.1% male) from 4 to 84 years old (mean=31.29), coming from 22 countries (68% from Spain). The majority of participants were students (38.8%) and identified themselves as white (84.4%). Participants were distributed into 188 dyadic sessions, with a participation average of 2.5 sessions/participant (max. 5 sessions). To create the dyads, three variables were taken into account: 1) gender (*Female*, *Male*); 2) age group (*Child*: 4-18, *Young*: 19-35, *Adult*: 36-50, and *Senior*: 51-84); and 3) relationship among interlocutors (*Known*, *Unknown*). Participants were matched according to their availability and language while trying to enforce a close-to-uniform distribution among all possible combinations between variables (60 combinations). A minimum age of 4 years and the ability to maintain a conversation in English, Spanish or Catalan were the only inclusion criteria. In the end, the most common interaction group is *Male-Male/Young-Young/Unknown* (15%), with 43% of the interactions happening among known people. Spanish is the majority language of interaction (71.8%), followed by Catalan (19.7%). Half of the sessions include both interlocutors with Spain as the country of origin.

2.4 Release

In this section we describe the UDIVA v0.5 dataset: the preliminary subset of UDIVA which is currently publicly available for research purposes[3]. UDIVA v0.5 contains a subset of the interaction sessions, participants, synchronized camera views, metadata, and transcriptions from UDIVA, in addition to new pose annotations[4]. UDIVA v0.5 dataset includes the two frontal views of the UDIVA dataset, one per participant (FC1 and FC2, see Figure 2.1). There is one video per participant, task, and session. Note that this initial release was motivated by the personality regression task we tackle in Chapter 3. In this sense, "Gaze" task was not included in this release, as very few personality indicators are present in it, and will be made public in the future.

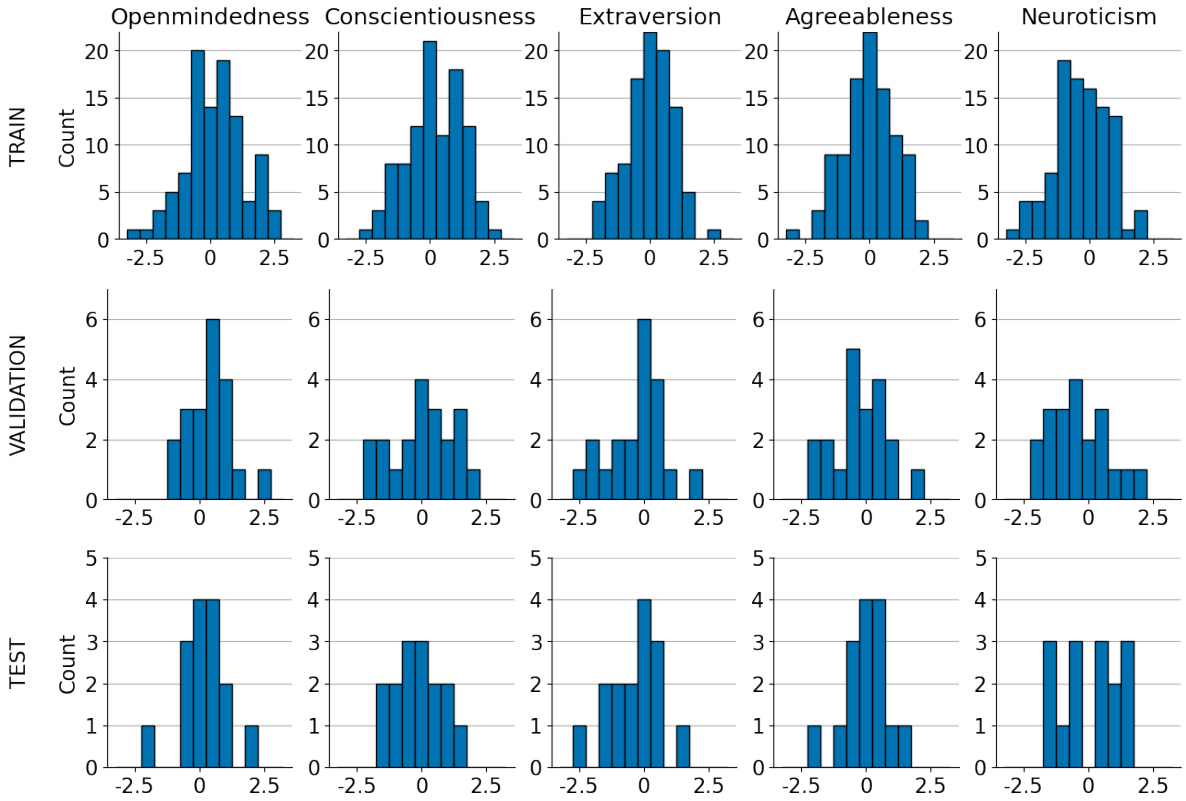


FIGURE 2.4: Distribution of the self-reported personality trait (OCEAN) values across train, validation, and test splits of the UDIVA v0.5 dataset. The x axis refers to z-scores for each personality trait.

2.4.1 Data selection and partition procedure

The interaction sessions included in the UDIVA v0.5 dataset were selected from the complete UDIVA dataset, with the aim of creating subject-independent training, validation, and test splits with similar distribution each in terms of personality traits, age, gender, and relationship among interaction partners. Prior to the data partition process, we first discarded all

[3] `https://chalearnlap.cvc.uab.cat/dataset/41/description/`.

[4] Pose annotations were collected outside the scope of the current dissertation, hence they will not be discussed here any further and are mentioned only for the sake of completeness. See Section 5.1 for a short description and [281] for details.

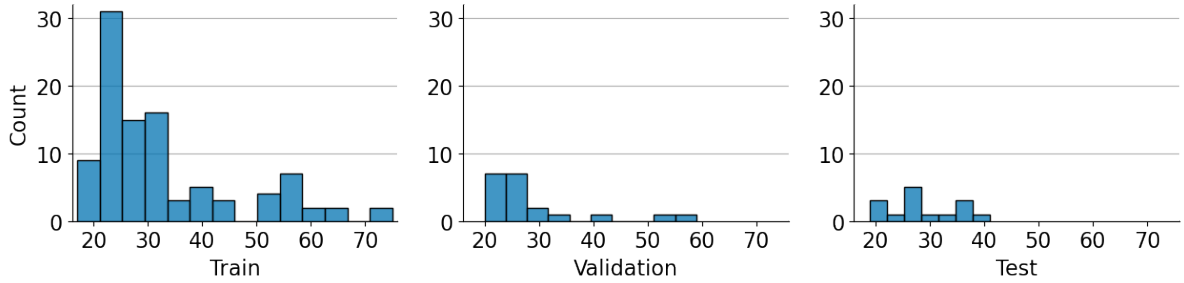


FIGURE 2.5: Age distribution across train, validation, and test splits of the UDIVA v0.5 dataset.

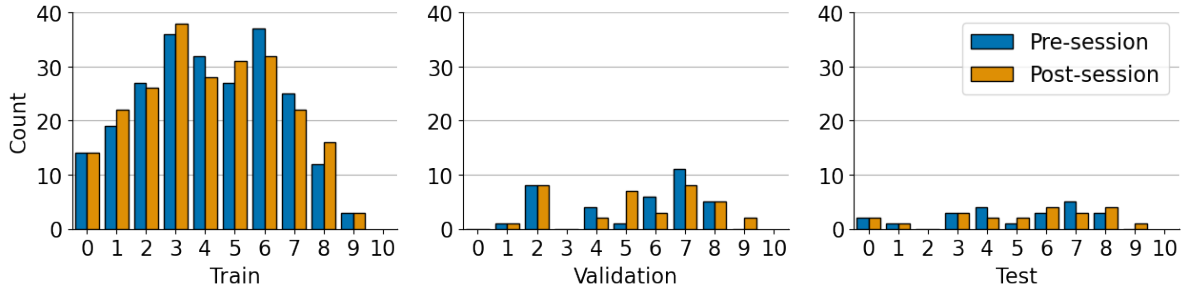


FIGURE 2.6: Pre- and post-session fatigue distribution across train, validation, and test splits of the UDIVA v0.5 dataset.

sessions with participants younger than 16 years old, as such participants filled in personality/temperament questionnaires specific to their age different than the BFI-2 questionnaire (see Section 2.2.1), and not all traits measured with such questionnaires have a one-to-one correspondence to the OCEAN personality traits (see Section 2.1.2 for a description of OCEAN). We also discarded sessions with any major technical issue (*e.g.*, one of the FC views was not available or none of the audio signals were available for a whole recording).

To ensure that no participants appeared in more than one split, some further sessions needed to be discarded. To decide which sessions to keep and how to divide them into the different splits, we followed an iterative procedure. First, we represented the remaining sessions as a graph, where the nodes correspond to participants and the edges correspond to interaction sessions, and assigned weights to sessions based on the number of interactions per participant and the group (*i.e.*, combination of age group, gender, and binary relationship) they belonged to. That is, initially, sessions with participants who interacted in many sessions and/or whose group belonged to a high-density one were assigned a lower weight than those sessions with participants who interacted in fewer sessions and/or belonged to less represented groups. Then, we used a greedy approach that iteratively removed and added sessions based on their importance to find split combinations that fulfilled a split ratio of approximately 3:1:1 with respect to the number of participants and sessions. Such approach started removing those sessions with a lower weight. The weights were updated every time a session was removed or added. Once a sample of all possible split candidates was computed, we calculated a set of costs for each candidate based on:

- the difference in per-trait distributions among each split with respect to the sum of splits by means of the p-value obtained from a Kolmogorov-Smirnov significance test [249];

- the differences in Pearson correlation between each personality trait and self-reported gender among each split;
- the differences in Pearson correlation between each personality trait and self-reported age among each split;
- the differences between age, gender, and relationship distributions with respect to a uniform distribution for validation and test splits.

Finally, we selected the combination that minimized the sum of the costs and that maximized the number of sessions and participants.

The final split contains 116 sessions and 99 participants for training, 18 sessions and 20 participants for validation, and 11 sessions and 15 participants for test. Although the validation split is larger than the test split, the latter contains a better trait balance. The resulting distribution of OCEAN and age values among splits can be observed in Figures 2.4 and 2.5, respectively. Gender ratios were conserved in all splits. In contrast, relationship ratios are significantly different, having 37.9% of *known* people in training, 61.1% in validation, and 81.8% in test. Pre- and post-session mood and fatigue values per split are shown, respectively, in in Figures 2.6 and 2.7. Given that the number of participants in the different splits is low (particularly validation and test), correlations between personality traits and other attributes differed, as expected [335]. Nonetheless, these splits allow for reliable comparability and benchmarking, especially in the context of personality regression, which we focus on in Chapter 3. For other contexts where a higher number of train/test subjects is required, we recommend strategies like leave-one-subject-out or leave-one-dyad-out instead of the provided data splits. We detail per split distribution of personality traits by gender in Table 2.2 and Figure 2.8 and by age in Table 2.3 and Figure 2.9.

Trait	Training	Validation	Test
O	t(97)=-0.07; p=0.94; d=0.01	t(18)=0.62; p=0.54; d=0.3	t(13)=0.85; p=0.41; d=0.47
C	t(97)=2.45; p=0.02; d=0.5	t(18)=0.6; p=0.56; d=0.28	t(13)=2.64; p=0.02; d=1.47
E	t(97)=1.78; p=0.08; d=0.36	t(18)=0.64; p=0.53; d=0.3	t(13)=1.64; p=0.13; d=0.91
A	t(97)=2.65; p=0.01; d=0.54	t(18)=-0.32; p=0.75; d=0.15	t(13)=2.13; p=0.05; d=1.18
N	t(97)=2.71; p=0.01; d=0.55	t(18)=1.09; p=0.29; d=0.52	t(13)=1.65; p=0.12; d=0.92

TABLE 2.2: Gender differences in OCEAN scores on training, validation, and test splits of the UDIVA v0.5 dataset, by means of Student's t-test.

Trait	Training	Validation	Test
O	-0.19 p=0.057 [-0.37;0.01]	-0.34 p=0.146 [-0.68;0.12]	0.01 p=0.981 [-0.51;0.52]
C	0.35 p<0.001 [0.17;0.51]	0.11 p=0.632 [-0.35;0.53]	0.38 p=0.165 [-0.17;0.75]
E	-0.01 p=0.924 [-0.21;0.19]	0.21 p=0.364 [-0.25;0.6]	0.09 p=0.742 [-0.44;0.58]
A	0.26 p=0.01 [0.07;0.43]	0.33 p=0.154 [-0.13;0.67]	-0.21 p=0.46 [-0.65;0.34]
N	-0.13 p=0.218 [-0.31;0.07]	-0.06 p=0.795 [-0.49;0.39]	-0.36 p=0.185 [-0.74;0.18]

TABLE 2.3: Statistical tests and 95% CIs for the correlations between OCEAN scores and age of the UDIVA v.05 splits.

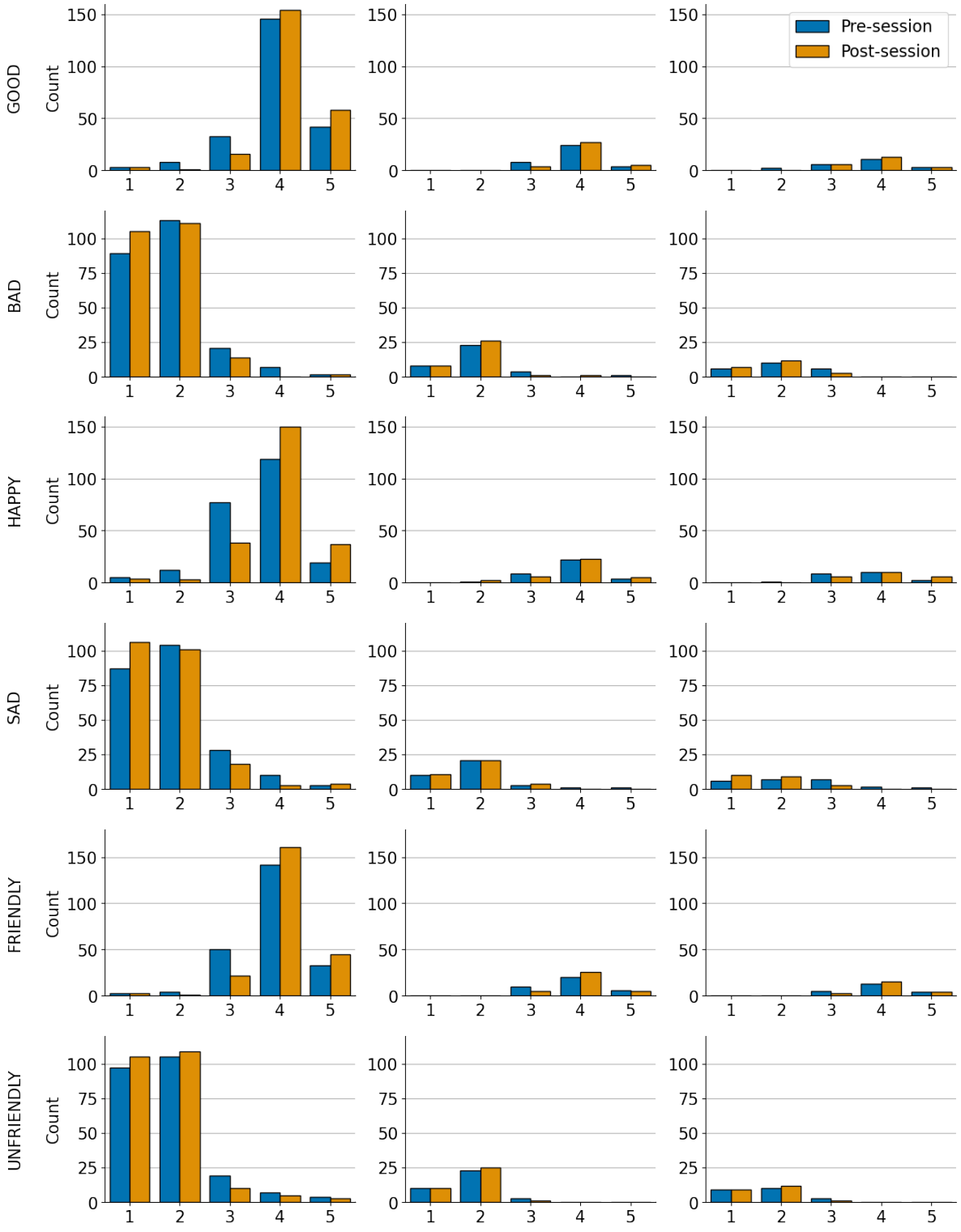


FIGURE 2.7: Pre- and post-session distribution of mood categories across train, validation, and test splits of the UDIVA v0.5 dataset.

2.4.2 UDIVA v0.5 Statistics

UDIVA v0.5 is composed of approximately 80h of recordings of dyadic interactions between 134 voluntary participants (44.78% female) from 17 to 75 years old (mean=31.95, sd=12.57).

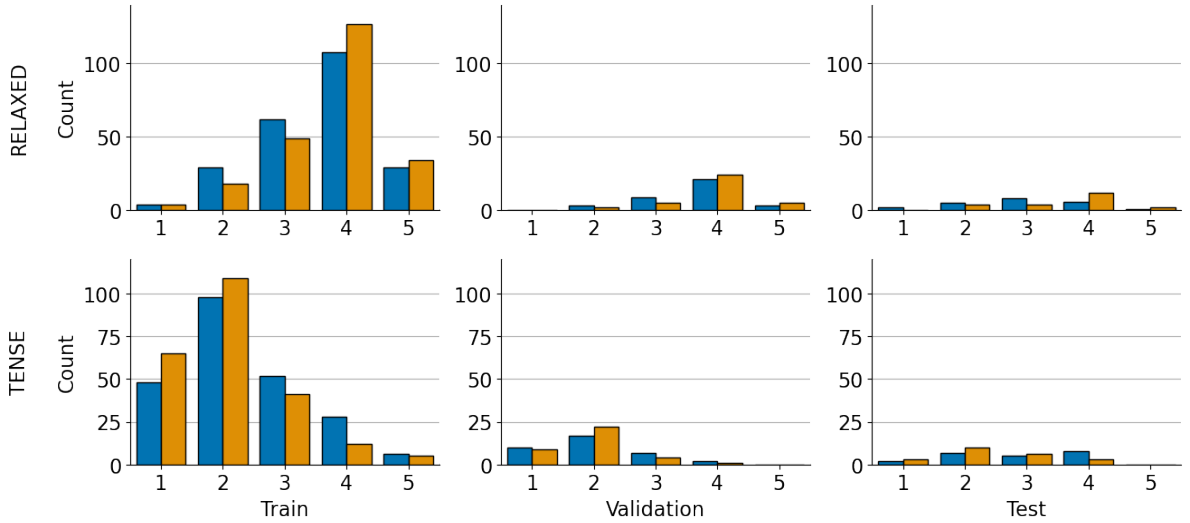


FIGURE 2.7: (Continuation) Pre- and post-session distribution of mood categories across train, validation, and test splits of the UDIVA v0.5 dataset.

Trait	Mean	Std. Dev.	O	C	E	A
O	.21	1.07				
C	.12	1.07	.02 [−.15, .19]			
E	−.12	1.01	.40** [.25, .53]	.24** [.07, .39]		
A	−.04	.97	.08 [−.09, .25]	.26** [.10, .42]	.20* [.03, .36]	
N	−.28	1.07	.01 [−.16, .18]	−.22** [−.38, −.05]	−.10 [−.26, .07]	−.22* [−.38, −.05]

TABLE 2.4: Descriptive statistics (mean and std. deviation, and Pearson's correlation) for self-reported personality OCEAN values of participants from the UDIVA v0.5 dataset. Values in square brackets indicate the 95% confidence interval for each correlation. $^{*}p < .05$, and $^{**}p < .01$.

Participants come from 22 different countries, with 68.66% of them from Spain. In relation to the maximum level of education, participants had mostly a Master's degree (35.82%), followed by a Bachelor's degree (28.36%). Table 2.4 shows the descriptive statistics (*i.e.*, mean, standard deviation, and Pearson's correlation) for self-reported personality OCEAN variables for participants of UDIVA v0.5. Compared to the normative data [120], UDIVA v0.5 sample presented higher levels of *Open-Mindedness* (O) and lower scores in *Negative Emotionality* (N). A low-to-moderate negative correlation between "N" and *Conscientiousness* (C) and *Agreeableness* (A), and low-to-moderate positive correlation between "A" and "C" and *Extraversion* (E) were observed. Finally, "E" correlated moderately with "O" ($r = .40$) and slightly with "C". These findings are comparable to the pattern found in the literature of intercorrelations between personality traits [352, 120].

Participants were distributed into 145 dyadic sessions, with a participation average of 2.16 sessions/participant (max. 5 sessions). 44.14% of the interactions occurred among participants who knew each other before their interaction session (*i.e.*, *known* people). Spanish was the interaction language in most of the dyads (73.10%), followed by Catalan (17.25%), and English (9.65%). Regarding the inner state of the participants, the mean pre- and post-fatigue

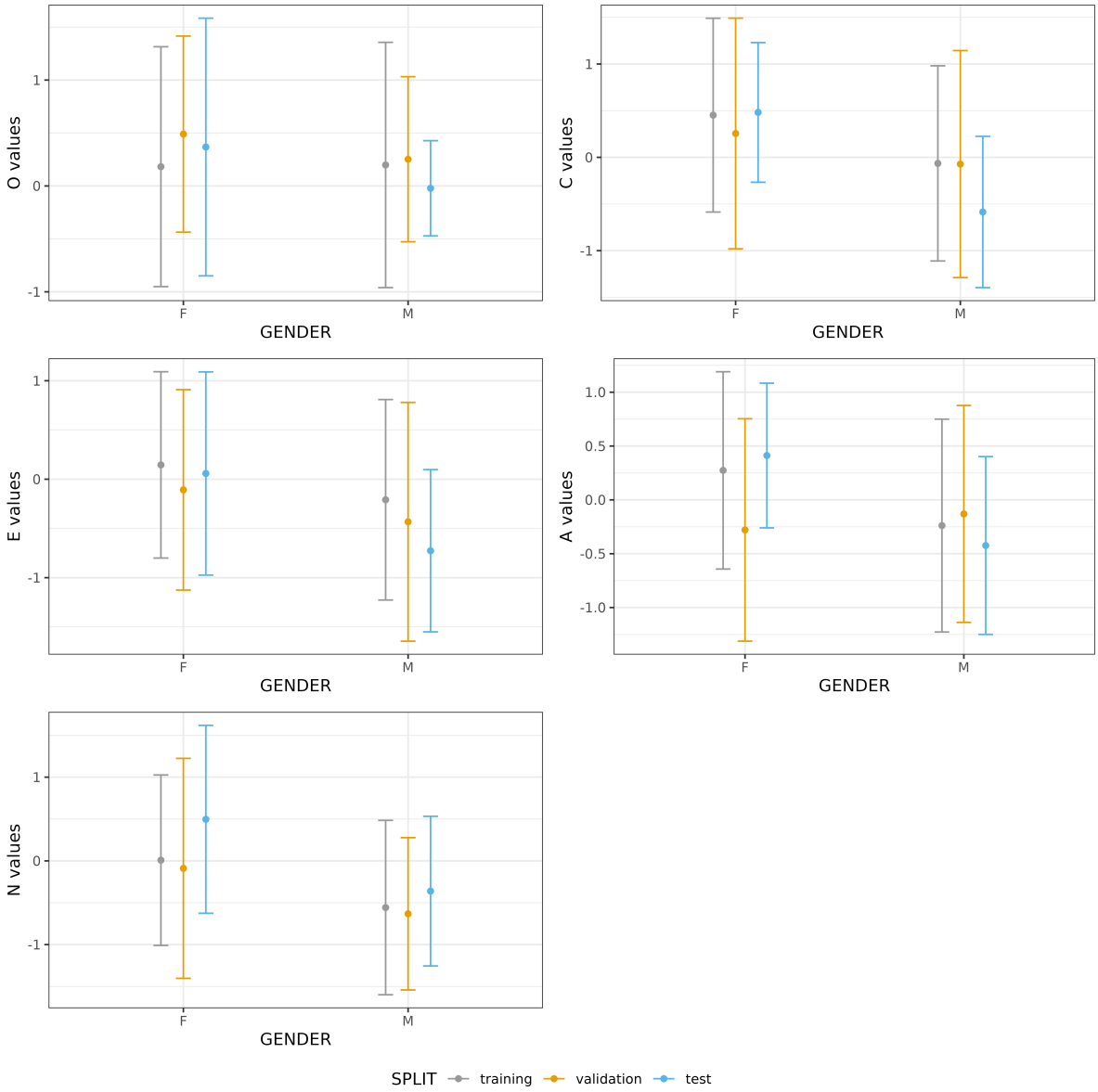


FIGURE 2.8: Gender differences in OCEAN scores on training, validation, and test splits of the UDIVA v0.5 dataset.

	Good	Bad	Happy	Sad	Friendly	Unfriendly	Relaxed	Tense
Pre-session	3.91±0.74	1.83±0.79	3.61±0.78	1.91±0.88	3.86±0.71	1.77±0.81	3.52±0.93	2.34±1.01
Post-session	4.1±0.64	1.67±0.67	3.92±0.72	1.71±0.78	4.04±0.63	1.68±0.73	3.73±0.85	2.07±0.9

TABLE 2.5: Descriptive statistics (mean ± standard deviation) of pre- and post-session mood categories of the UDIVA v0.5 dataset.

value was 4.35 (± 2.32 for pre-session and ± 2.39 for post-session), whereas descriptive statistics of pre- and post-session mood are shown in Table 2.5. As can be seen, interaction sessions slightly improved the mood state of participants (*i.e.*, positive states increased, whereas their respective negative counterparts decreased).

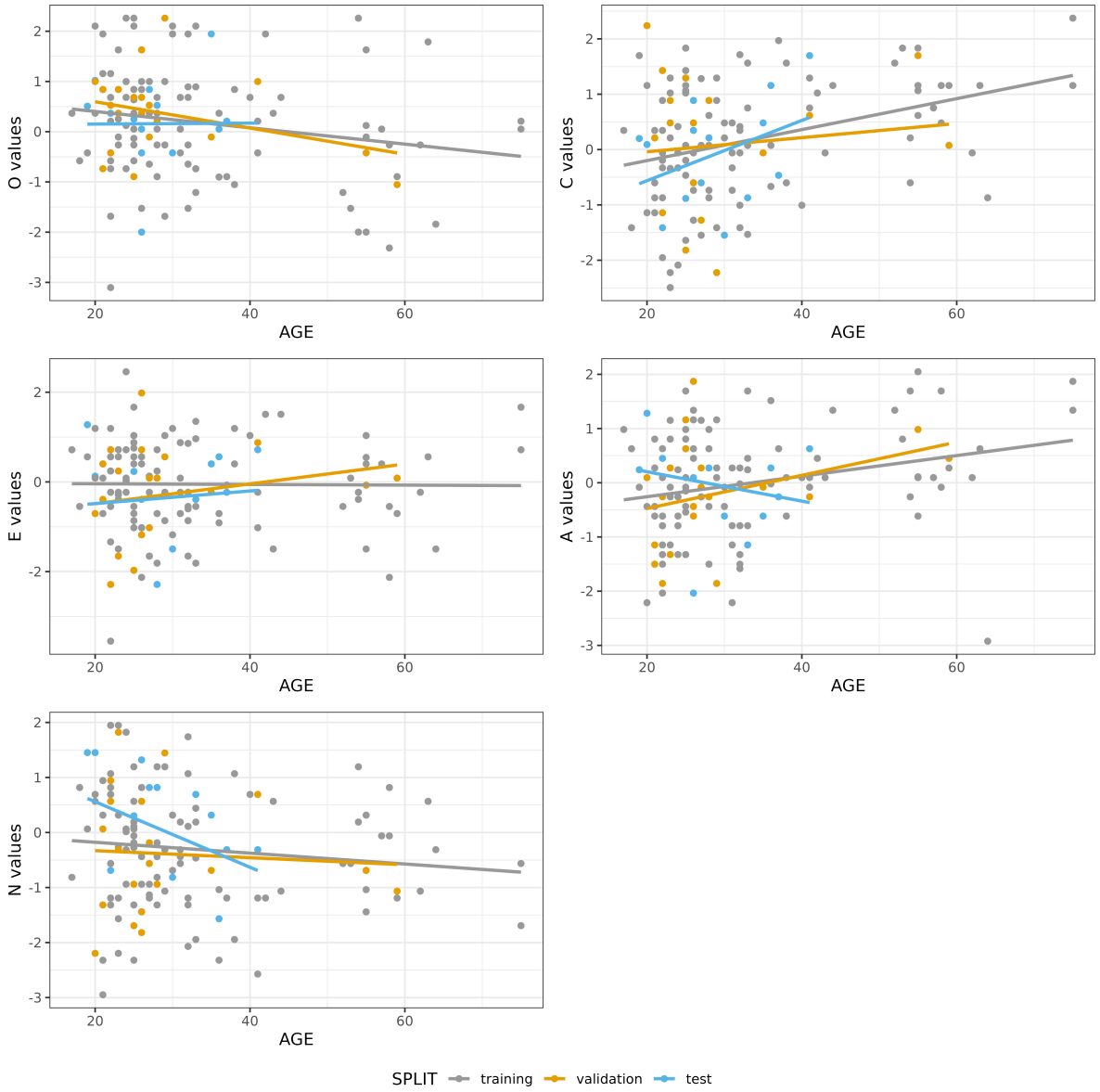


FIGURE 2.9: Relationship between OCEAN scores and age along training, validation, and test splits of the UDIVA v0.5 dataset.

2.5 Applications for good

Understanding dyadic interactions opens the door to a plethora of human-centered application scenarios. Some examples of these include early diagnosis and intervention [98], augmented telepresence [5], personalized agents [103], or human behaviour forecasting [26]. Here we particularly remark potential benefits of advancing personality regression, given that the current release of the UDIVA dataset provides personality scores as one of its key labels, and as it will be the task we tackle next in Chapter 3.

Beyond the basic structure of personality defined by the Big Five traits [251], the assessment of personality still has some limitations. Importantly, personality predicts a number of *consequential outcomes* [278], which have been shown to be consistent and replicable [353].

These consequential outcomes allow a number of applications, which range from clinical applications to human factor engineering. With respect to the former, subjective well-being, risky behaviors, or substance abuse are consistently related to personality [353]. Regarding the latter, personality has been related to performance and fit in the workplace [353], and therefore, it would be plausible that automatic detection of personality traits could allow personalizing working environments and hence increase performance and well-being at the workplace.

Generally speaking, automated video analysis of personality would allow speeding up the process of personality assessment. In the context of personality recognition, it is increasingly showing its potential to improve well-being and mental health through personalized interventions [10]. If successful, we could think of it as a way to overcome all of the caveats related to personality assessment, such as faking in test-takers, especially in high-stakes.

The key use of automatic video processing for the assessment of personality is personalization and customization. That is, it allows for a quick, and relatively cheap form of assessment that could help custom-tailoring environments to make people able to exploit all of their potential (via person-environment fit). For instance, in clinical settings, it would allow to detect people at risk of certain disorders. Furthermore, healthcare delivered by robots is a burgeoning field of research [104], and personalization of robots' personalities seems to have positive impacts on patients' health and social outcomes [13] by means of increasing acceptance of robot's care [368]. Another real-life example is pedagogical agents [84], which are personalized to maximize learners' attention: after an initial video assessment, an AI system could personalize a learning environment for any given student that actually fits their personality characteristics. For instance, extraverted learners take more profit from group activities and positive reinforcement, whereas introverted and conscientious children are more sensitive to negative reinforcement and lonely activities [311].

On a more interactive research field, [161] found they were able to profile personality from a job candidate during a job interview interaction with higher accuracy than self-reports, and [351] found that they were able to predict emotional states of an individual through her facial reactions based on cues displayed by the conversational partner. Also, within work and organizational psychology, automated video analysis coupled with AI could create teams on the spur of the moment and boost their performance for specific tasks. All these benefits rely on the fact that personality is an essential variable in people's interactions with their contexts. The role of personality recognition is to understand the person to the greatest possible degree so as to offer them a personalized version of health, education, or even customer service.

2.5.1 Ethical considerations

Despite all the exciting applications discussed above, research on human behavior understanding and personality recognition also comes with a number of potential ethical pitfalls. It is for this reason that we also describe some of the ethical issues concerning the methodological aspects of this research as well as its potential results and consequences.

Since this research involved the collection and manipulation of sensitive data, several ethical aspects were considered in conducting it. Consent to collect and use the data was asked

with full disclosure of how it would be used, processed, and for how long the data would be available for further processing. Additionally, the data is preserved anonymously and encrypted, all participation was voluntary, and it entailed no degree of harm. Furthermore, the current release is available only for research purposes and does not allow for commercial use. A noteworthy ethical concern in our research is the sample bias towards a WEIRD population [157], since most of the participants in our sample matched the characteristics of a white and highly educated population.

On a different note, there are important ethical concerns in the use of automatic tools for personality recognition. First off, since personality computing may become the new channel for psychological assessment over the next few years [39], researchers and practitioners should ensure data privacy even more as the possibilities of data leakage and misuse have increased [267]. Collected personality data could be misused for personality profiling beyond the scopes of the research purposes, like personnel selection [373], selling strategies [89] or such as it happened in the Cambridge Analytica scandal, when personalized political messages were delivered based on reconstructions of personality profiles from social networks [76].

2.6 Conclusions

This chapter has introduced UDIVA, the largest multiview audiovisual dataset of dyadic face-to-face non-scripted interactions, with 90 hours of recordings, 147 participants, and 188 recording sessions. The UDIVA dataset holds the potential to be used for a plethora of analyses related to individual and dyadic behavior within interaction settings, from both the computational and psychological fields of study. In this sense, we are releasing this data[5] with the purpose of advancing the research and understanding of human communication from a multidisciplinary perspective.

The UDIVA dataset is currently being expanded with further annotations that allow the modeling of more complex interpersonal constructs. Micro-actions, pose or gaze are just a few of them (see Section 5.1). But now, the moment has arrived to assess the capabilities of Transformers on this challenging data. We could try to regress some of the metadata, such as gender or age. However, these are individual traits that could be predicted entirely disregarding the other participant and may not require fine-grained analysis of the scene and interaction environment. For this reason, we believe a more interesting analysis would be that of personality. As we have explored, human behavior in interaction is deeply influenced by our individual personalities. It is for this reason that in the next chapter we explore the abilities of Transformers to model such complex scenarios for the task of personality regression.

[5] https://chalearnlap.cvc.uab.cat/dataset/39/description/

Chapter 3

Modeling Humans through Video Transformers

3.1 Introduction

The way people adapt and react to social signals and behaviors during a conversation depends not only on their individual characteristics (*e.g.*, personality) but also on the specific situation and their shared history [48]. For example, one might behave more relaxed during a conversation with a friend than in a meeting with their foreman. When analyzing social interactions from a computational perspective, all these influential factors should be taken into consideration to truly understand human behavior, even when the focus is on predicting individual attributes such as personality traits [393]. However, this is still not the norm throughout the literature. Our motivation to develop the UDIVA dataset (described in the previous Chapter 2) was to help fill in that gap. With UDIVA, we provide a general-purpose dataset with varied sources of context in order to motivate research that considers all of them when solving interaction and individual-related tasks.

In this chapter, we are interested in benchmarking the abilities of the Transformer architecture [385] to integrate such varied sources of context to solve some challenging task. As an initial analysis of the UDIVA dataset, we focus on automatic self-reported personality recognition [390, 176] (which has plenty of potential applications for good as we have discussed in Section 2.5). This can be considered the first step among the many research lines that can benefit from the designed dataset. To do so, we first design a baseline model inspired by the Video Action Transformer [131]. Motivated by the work in [315], we use a target person's face video and metadata as source information, while the other interlocutor's scene, audio, and further metadata are used as external context. The metadata includes stable and transient characteristics from each interlocutor, as well as session, task, and relationship information. We experimentally evaluate the usefulness of each additional input incrementally, showing consistent improvements when using all the available context sources and modalities.

The obtained baseline scores show that Transformers are indeed capable of integrating multiple sources of context. However, there is a large range of improvement in the regression of personality, as well as several dyadic open challenges to address with the UDIVA dataset.

Overall, despite the growing interest in social interaction understanding, current computational approaches for this area (including our proposed baseline) present various shortcomings. On the one hand, long-term modeling is crucial in interaction settings, as more complex dynamics emerge at different time scales, and an event may unchain effects that take time to be observed [48]. In the case of self-reported personality computing in such scenarios, the need for long-term modeling is heightened, as behavioral manifestations of certain traits may not be fully observed in short periods of time. Hence, more time is needed to find salient patterns arising during the interaction that can be associated with given traits [127]. Most existing works attempting longer-term modeling have generally been limited to single frame descriptors averaged over whole sequences [209], missing to represent the temporal evolution of features. On the other hand, the joint modeling of both interlocutors is another aspect that fails to be properly modeled when assessing individual attributes in dyadic interactions. Despite its importance for triggering individual behaviors that provide insights on individual features [27], most of the works that do model it are focused on analyzing interaction attributes.

In order to address these issues we propose the *Dyadformer*, a novel Transformer architecture to leverage long-term information for joint modeling of both interlocutors in dyadic scenarios. More precisely, we predict the personalities of both interactants by considering not only the audiovisual information and contextual factors (referred to as metadata) independently for each one but also by explicitly modeling their interaction. The proposed model mainly consists of two stages: (1) a *cross-modal* stage where cross-attention encoders fuse multi-modal information, and (2) a *cross-subject* stage which aims to shape the interaction by performing double cross-attention (see Figure 3.3).

This approach presents several advantages over the baseline. First, the baseline regresses participant personality from just 3-second chunks, which may not be enough to properly model long-term interactions. Opposed to this, the Dyadformer inputs longer clips (up to 30 seconds), allowing the model to learn useful longer-range relationships. Second, our first approach does multi-modal fusion by simply concatenating information from video and audio. Differently, the Dyadformer leverages multi-modal Transformers that exploit useful features from each source by looking at interdependencies, and fusing them in a shared representation space. Finally, whereas in the baseline only the personality of the target subject is regressed, the Dyadformer explicitly models the behavior of both individuals simultaneously through our proposed two-stream cross-attentional Transformer. As we will see, thanks to these characteristics, the Dyadformer not only outperforms the baseline, but all participants to a recent *Challenge on Understanding Social Behavior in Dyadic and Small Group Interactions* on UDIVA v0.5 dataset [281], which included a self-reported personality branch.

Our contributions are summarized as follows:

- We validate our hypothesis that Transformers can integrate multiple sources of context thanks to our baseline. Also, we confirm that as they are added results on personality regression are improved.

- To our knowledge, the Dyadformer is the first one to jointly model (and infer) self-reported personality in dyadic interactions using time windows of up to ∼30 seconds of dense audiovisual features.
- Inspired by the classical decoder block of the Transformer network [385], we leverage a cross-attention mechanism to both fuse modalities and allow information to flow between subjects.
- Dyadformer obtained state-of-the-art results on the UDIVA v0.5 dataset for the task of self-reported personality prediction. In particular, we reduce participant-level error by 11.8% compared to the baseline and by 6.1% compared to the best-performing challenge solution.

The rest of the chapter is organized as follows: we first introduce the original Transformer and relevant adaptations for video in Section 3.1.1. Then, we outline related work on social signals, personality computing and the use of multi-modal Transformers for video in Section 3.1.2. We then present, ablate, and discuss the baseline model in Section 3.2. Next, we introduce our proposed Dyadformer in Section 3.3, with extensive experimental results and discussion presented in Section 3.4. Finally, Section 3.5 concludes the chapter.

3.1.1 The Transformer

As the Transformer [385] is the core technology leveraged for this chapter, as well as Chapter 4, next we provide a brief technical introduction to it. Transformers are a recent family of models, originally designed to replace recurrent layers in a machine translation setting. Its purpose was to remedy limitations of sequence modeling architectures by handling whole sequences at once (as opposed to RNNs, which are sequential in nature), allowing further parallelization. Besides, it removes the locality bias of traditional architectures, such as CNNs, and instead learns interactions of non-local contexts of the input.

The Transformer evolves input representations based on interactions among all the sequence elements. These interactions are modulated through a pair-wise affinity function that weighs the contribution that every element should have on any other. This ability to model all-to-all relationships can be especially beneficial to understand motion cues, long-range temporal interactions and dynamic appearance changes in video data. The original Transformer consists of two distinct modules: encoder and decoder, each composed of several stacked Transformer layers (see Figure 3.1). The *encoder* was designed to produce a representation of the source language sentence that will be then attended by the *decoder*, which will eventually translate it into the target language. We first introduce a few necessary concepts, input preprocessing and the self-attention operation, to then follow the flow of the Transformer while explaining its components and functioning.

Input pre-processing: tokenization, linear embedding, and positional encodings. The *tokenization* converts the input source and target language sentences into sequences of words

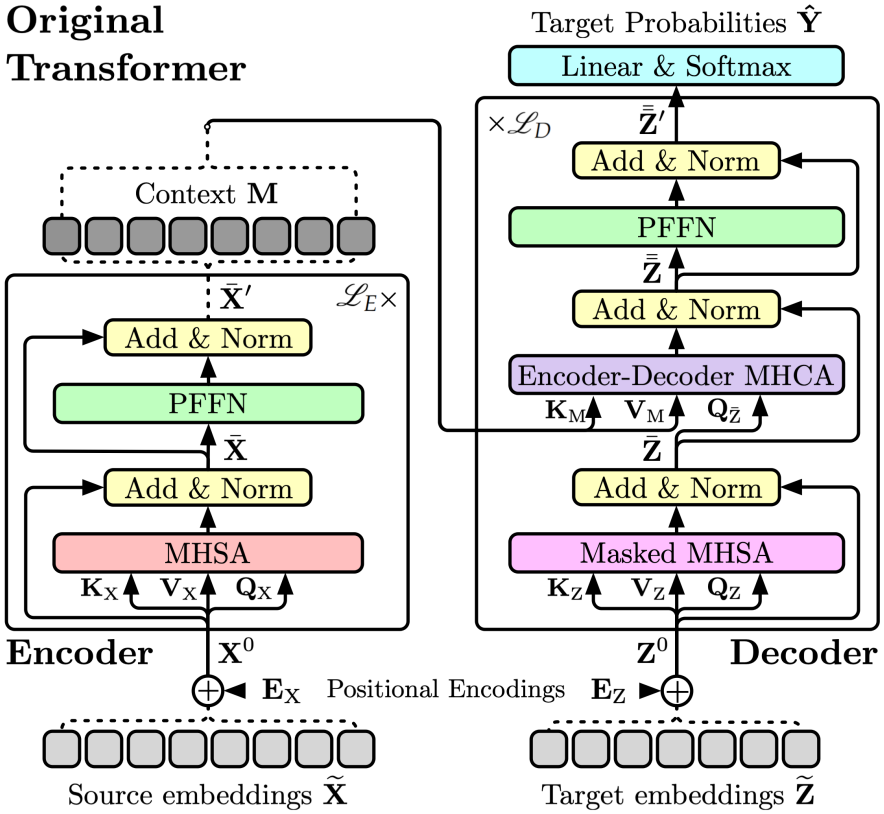


FIGURE 3.1: Visualization of the original Transformer proposed in [385].

(or subwords), namely "tokens". Let $\widetilde{\widetilde{\mathbf{X}}} = (\widetilde{\widetilde{\mathbf{x}}}_1, \ldots, \widetilde{\widetilde{\mathbf{x}}}_{N_\mathrm{x}})$ and $\widetilde{\widetilde{\mathbf{Z}}} = (\widetilde{\widetilde{\mathbf{z}}}_1, \ldots, \widetilde{\widetilde{\mathbf{z}}}_{N_\mathrm{z}})$ be, respectively, the source and target sequences of one-hot encoded tokens over their respective word vocabularies $\mathcal{X}$ and $\mathcal{Z}$ (*i.e.*, $\widetilde{\widetilde{\mathbf{x}}} \in \mathbb{R}^{|\mathcal{X}|}$ and $\widetilde{\widetilde{\mathbf{z}}} \in \mathbb{R}^{|\mathcal{Z}|}$), where $N_\cdot$ represents the number of elements in the sequence. Then, *linear embedding* is simply the step of projecting one-hots to a continuous embedding space via a learnt linear transformation: $f_\mathcal{X} : \mathbb{R}^{|\mathcal{X}|} \mapsto \mathbb{R}^{d_m}$ (analogously $f_\mathcal{Z}$), where d_m will be the dimensionality handled internally by the Transformer. This way, we obtain the source embeddings $\widetilde{\mathbf{X}} = (f_\mathcal{X}(\widetilde{\widetilde{\mathbf{x}}}_1), \ldots, f_\mathcal{X}(\widetilde{\widetilde{\mathbf{x}}}_{N_\mathrm{x}}))$ and target embeddings $\widetilde{\mathbf{Z}} = (f_\mathcal{Z}(\widetilde{\widetilde{\mathbf{z}}}_1), \ldots, f_\mathcal{Z}(\widetilde{\widetilde{\mathbf{z}}}_{N_\mathrm{z}}))$. Finally, *positional encodings* are used to signal the position of the tokens in the sequence to the later (otherwise permutation invariant) attention operations. Defined using a set of (non-learnable) sinusoidal encodings (see [385] for details), these are added to the source/target embeddings before being input to encoder/decoder (as depicted in Figure 3.1): $\mathbf{X}^0 = (\widetilde{\mathbf{x}}_1 + \mathbf{e}^\mathrm{x}_1, \ldots, \widetilde{\mathbf{x}}_{N_\mathrm{x}} + \mathbf{e}^\mathrm{x}_{N_\mathrm{x}})$ and $\mathbf{Z}^0 = (\widetilde{\mathbf{z}}_1 + \mathbf{e}^\mathrm{z}_1, \ldots, \widetilde{\mathbf{x}}_{N_\mathrm{z}} + \mathbf{e}^\mathrm{z}_{N_\mathrm{z}})$, where $\mathbf{e}^\mathrm{x}_\cdot, \mathbf{e}^\mathrm{z}_\cdot \in \mathbb{R}^{d_\mathrm{m}}$.

Self-attention (SA). It is the core operation of the Transformer. Given an arbitrary sequence of token embeddings $\mathbf{X} \in \mathbb{R}^{N_\mathrm{x} \times d_\mathrm{m}}$ (*e.g.*, $\mathbf{X}^0$), it augments (contextualizes) each of the embeddings $\mathbf{x}_i \in \mathbb{R}^{d_\mathrm{m}}$ with information from the rest of embeddings and also itself. For that, the embeddings in $\mathbf{X}$ are linearly mapped to the embedding spaces of *queries* $\mathbf{Q} = \mathbf{X}\mathbf{W}_\mathrm{Q} \in \mathbb{R}^{N_\mathrm{x} \times d_\mathrm{k}}$, *keys* $\mathbf{K} = \mathbf{X}\mathbf{W}_\mathrm{K} \in \mathbb{R}^{N_\mathrm{x} \times d_\mathrm{k}}$, and *values* $\mathbf{V} = \mathbf{X}\mathbf{W}_\mathrm{V} \in \mathbb{R}^{N_\mathrm{x} \times d_\mathrm{k}}$, where $\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{K} \in \mathbb{R}^{d_m \times d_k}$, $\mathbf{W}_\mathrm{V} \in \mathbb{R}^{d_m \times d_v}$, and typically $d_\mathrm{k}, d_\mathrm{v} <= d_m$. Then, self-attention can be computed as follows:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_\text{k}}}\right)\mathbf{V}. \tag{3.1}$$

The dot-product $\mathbf{Q}\mathbf{K}^\top \in \mathbb{R}^{N_\text{x} \times N_\text{x}}$ is a measure of similarity. Intuitively, the larger the similarity between $\mathbf{q}_i \in \mathbf{Q}$ and $\mathbf{k}_j \in \mathbf{K}$ the more relevant the information embedded in $\mathbf{x}_j$ is for $\mathbf{x}_i$. However, this aggregation is not done in the space of $\mathbf{X}$, but in the one of the values. By applying Softmax with temperature $\sqrt{d_k}$, we come up with normalized similarities (the self-attention matrix) that weigh how much each of the values $\mathbf{v}_j$ contributes to the output representation of every other $\mathbf{v}_i$.

Encoder module. It consists of $\mathscr{L}_E$ layers, each including *Multi-Head Self-Attention* (MHSA) and *Position-wise Feed-Forward Network* (PFFN) sub-layers. The MHSA sub-layer performs self-attention through multiple separate heads that map $\mathbf{X}$ to h different representation subspaces (*i.e.*, $\{(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \,|\, 1 \leq i \leq h\}$). $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ are computed via their associated embedding matrices (*i.e.*, $\mathbf{W}_{\text{Q}_i} \in \mathbb{R}^{d_\text{m} \times d_\text{k}}$, $\mathbf{W}_{\text{K}_i} \in \mathbb{R}^{d_\text{m} \times d_\text{k}}$, and $\mathbf{W}_{\text{V}_i} \in \mathbb{R}^{d_\text{m} \times d_\text{v}}$ with $d_\text{k} = d_\text{v} = d_\text{m}/h$). The output of the heads are concatenated and mapped back to a d_m-dimensional space with another linear transformation $\mathbf{W}_\text{O} \in \mathbb{R}^{(h \cdot d_\text{v}) \times d_\text{m}}$:

$$\begin{aligned} \text{MHSA}(\mathbf{X}) &= \text{Concat}(\mathbf{H}_1, ..., \mathbf{H}_h)\mathbf{W}_\text{O}, \\ \text{where } \mathbf{H}_i &= \text{Att}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i), \end{aligned} \tag{3.2}$$

and $\mathbf{H}_i \in \mathbb{R}^{N_\text{x} \times d_\text{v}}$ is the output of the i^th head. We then apply "Add + Norm" to come up with $\bar{\mathbf{X}} = \text{LN}(\mathbf{X} + \text{MSHA}(\mathbf{X}))$, where LN stands for "Layer Normalization" [20]. After this, the following PFFN sub-layer further refines each embedding in $\bar{\mathbf{X}}$ individually (point-wise). This sub-layer is composed of two linear layers and ReLU activation function:

$$\text{PFFN}(\bar{\mathbf{X}}) = \text{ReLU}(\bar{\mathbf{X}}\mathbf{W}_{\text{F}_1})\mathbf{W}_{\text{F}_2}, \tag{3.3}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{N_\text{X} \times d_\text{m}}$, and $\mathbf{W}_{\text{F}_1} \in \mathbb{R}^{d_\text{m} \times (4 * d_\text{m})}$ and $\mathbf{W}_{\text{F}_2} \in \mathbb{R}^{(4 * d_\text{m}) \times d_\text{m}}$. Note, $\mathbf{W}_\cdot$ are weight matrices independent for each encoder layer, but we omit those indices for ease of notation. By applying this, $\bar{\mathbf{X}}' = \text{LN}(\bar{\mathbf{X}} + \text{PFFN}(\bar{\mathbf{X}}))$. In practice, a PFFN sub-layer is equivalent to applying two fully connected layers in a point-wise fashion, *i.e.*, independently to each token (which can also be seen as two 1D convolutional layers with kernel size 1).

Decoder module. Consisting of $\mathscr{L}_D$ layers and fed with $\mathbf{Z}^0$, it substitutes MHSA with two other sub-layers. The first one, *Masked Multi-Head Self-Attention* (Masked MHSA), modifies Att in Equation (3.1) to include a mask, $\mathbf{B} = (b_{ij}), 1 \leq i, j \leq N_\text{z}$, impeding the access to certain tokens. This is added to the result of the dot-product in the numerator (and before the Softmax), as follows: $\mathbf{Q}\mathbf{K}^\top + \mathbf{B} \in \mathbb{R}^{N_\text{Z} \times N_\text{Z}}$, where $b_{ij} = -\infty$ iff $i < j$ (otherwise $b_{ij} = 0$). This draws attention values for the masked attention pairs to 0 when taking exponents in the Softmax. As we will see, such masking is crucial to define the auto-regressive behavior of the decoder module (avoiding tokens to attend to other tokens later in the sequence). The result from this Masked MHSA, $\bar{\mathbf{Z}}$, is now passed to the *Encoder-Decoder Multi-Head Cross-Attention*

(MHCA) sub-layer, which leverages the memory/context produced by the encoder, namely $\mathbf{M}$ (*i.e.*, $\bar{\mathbf{X}}'$ at encoder's $\mathscr{L}_E{}^{\text{th}}$ layer), into $\bar{\mathbf{Z}}$ as follows:

$$\begin{aligned} \text{MHCA}(\bar{\mathbf{Z}}, \mathbf{M}) &= \text{Concat}(\mathbf{J}_1, \dots, \mathbf{J}_h)\mathbf{U}_\text{O}, \\ \text{where } \mathbf{J}_i &= \text{Att}(\bar{\mathbf{Z}}\mathbf{U}_{\text{Q}_i}, \mathbf{M}\mathbf{U}_{\text{K}_i}, \mathbf{M}\mathbf{U}_{\text{V}_i}), \end{aligned} \tag{3.4}$$

and $\mathbf{J}_i \in \mathbb{R}^{N_\text{Z} \times d_v}$ is the output of the i^{th} cross-attention head, $\mathbf{U}_{\text{Q}_i} \in \mathbb{R}^{N_\text{Z} \times d_k}$, $\mathbf{U}_{\text{K}_i} \in \mathbb{R}^{N_\text{x} \times d_k}$, $\mathbf{U}_{\text{V}_i} \in \mathbb{R}^{N_\text{x} \times d_v}$, and $\mathbf{U}_\text{O} \in \mathbb{R}^{(h \cdot d_\text{v}) \times d_\text{m}}$ are weight matrices. Note that $\bar{\mathbf{Z}}\mathbf{U}_{\text{Q}_i}$, $\mathbf{M}\mathbf{U}_{\text{K}_i}$ and $\mathbf{M}\mathbf{U}_{\text{V}_i}$ are, respectively, the queries, keys, and values of MHCA, shown here implicitly to ease notation. Then, $\bar{\bar{\mathbf{Z}}} = \text{LN}(\bar{\mathbf{Z}} + \text{MHCA}(\bar{\mathbf{Z}}, \mathbf{M}))$. Crucially, MHCA is different from MHSA in that, for the former, two distinct sources of information are involved ($\bar{\mathbf{Z}}$ and $\mathbf{M}$), and that the cross-attention (CA) operation effectively augments $\bar{\mathbf{Z}}$ with information from $\mathbf{M}$. The remaining PFFN sub-layer, which is no different from the one in encoder layers, is to produce $\bar{\bar{\mathbf{Z}}}' = \text{LN}(\bar{\bar{\mathbf{Z}}} + \text{PFFN}(\bar{\bar{\mathbf{Z}}}))$. Finally, in the $\mathscr{L}_D{}^{\text{th}}$ layer, the embeddings from the PFFN are each sent through a linear layer followed by softmax to generate the output probabilities over the words in the target vocabulary $\mathcal{Z}$, *i.e.*, $\hat{\mathbf{Y}} \in \mathbb{R}^{N_\text{z} \times |\mathcal{Z}|}$.

Current Transformer trends adopted for video. Many variations to the Transformer have become common in vision and, particularly, video. First, the use of other *special token* besides the ones discussed, such as `[CLS]` (class) or `[MSK]` (mask) tokens. In image/video, these are not strings but directly parameters initialized at random and adapted during the optimization process based on the learning objective. `[CLS]` is used to condensate (into a vector representation) information from the rest of token embeddings in a sequence (representing spatiotemporal patches from the video [17]), and suited for high-level tasks (such as classifying the sequence globally). Using input token embeddings instead of `[CLS]` may cause the model to be biased towards it [402]. Conversely, `[MSK]` is used to replace input embeddings and signal the Transformer to reconstruct those guided by the loss and based on the rest of tokens. This forces the Transformer to learn context of the tokens and how these relate to the masked ones. Thought for language representation learning [88], this has been adopted also for video representation learning [406, 374].

Second, *deviations from the canonical encoder-decoder*: encoder-only or decoder-only Transformer architectures[1]. Encoder-only are suited to produce fixed-size outputs, *i.e.*, augmentations of the input embeddings that can be used for more granular tasks (*e.g.*, per-frame classification) or, when used together with `[CLS]`, to come up with a global representation (*e.g.*, sequence-level classification). For instance, [17, 31, 106] adopted an encoder-only architecture (along with the inclusion `[CLS]`) for video classification following [91]. Instead, decoder-only alternatives enable auto-regressive tasks if the size of the output cannot be determined a priori (thanks to the Masked MHSA) just by knowing the input size. For instance, to predict a series of temporal action detections. Initially proposed by [313] in NLP, these have been

[1]Note we here refer to *encoder* or *decoder* by the role they carry out (*i.e.*, encoding or decoding information), and not to the building blocks that compose them. In this sense, and as an example, the Transformer *encoders* our Dyadformer uses are composed, in part, of MHCA sub-layers (see Section 3.3), different from the canonical encoder defined above which only include MHSA and PFFN.

also followed in the context of video in [259, 367, 456]. Other trends originated in other fields have been followed: swapping the order of the residual connection and layer normalization [17, 106], although no clear general advantage of one over the other has been empirically shown yet; or replacing ReLU in the PFFN by GeLU [130, 31, 400, 186] following [88], with only [130] ablating this decision (finding out that GeLU was slightly outperforming ReLU on their task/data).

Transformer limitations. Transformers have two key limitations: first, given the pair-wise affinity computation in Equation (3.1), they exhibit *quadratic complexity* ($\mathcal{O}(N^2)$), which will be specially problematic for video. In Section 4.3.1, we will explore some works alleviating this issue by reducing the scope of the SA operation. The second limitation is the lack of *inductive biases*. This is a double-edged sword, allowing for a general-purpose architecture that can handle any modality, but severely complicating the learning process. While this can be solved through large quantities of data [91], this further adds to the computational costs of training Transformers. Throughout Chapter 4, we will explore various approaches to solve this issue.

3.1.2 Related work

Next we review relevant related work on social signals and behavriours, automatic personality recognition, the use of long-term modeling for personality recognition, and multi-modal Transformers.

Social signals and behaviors in context. Dyadic and small group interactions are a rich source of overt behavioral cues. They can provide insight into our personal attributes and cognitive/affective inner states dependent upon the context in which they are situated. Context can take many forms, from the interaction partner's attributes and behaviors to spatiotemporal and multi-view information. Joint modeling of both interlocutors and/or other sources of context have been extensively considered when trying to measure interpersonal constructs (*e.g.*, synchrony [85] or rapport [463]), individual social behaviors (*e.g.*, engagement [87]) and impressions (*e.g.*, dominance [462] or empathy [284]). However, for the task of recognizing individual attributes (such as emotion and personality) context has often been misrepresented, in spite of extensive claims on its importance [27, 415, 389, 262].

Regarding utterance- or turn-based emotion recognition in conversation [302] and sentiment analysis, only a handful of recent studies employ interlocutor-aware approaches to incorporate contextual information. Early works were based on handcrafted nonverbal, spatiotemporal dyadic features [204, 257]. Nowadays, most approaches rely on deep learning, using conversation transcripts as input with contextualized word or speaker embeddings [222] and considering past and/or future parts of the conversation as additional context. Richer contexts have been captured by explicitly modeling the temporal dimension. This has been widely achieved via recurrent approaches [243, 424, 129], and more recently with Transformer-like architectures [465, 216]. Some works have further proposed to enrich models with additional modalities, such as raw audiovisual data to enhance the representation of

interlocutors' influences and dynamics [447, 148, 421, 167], or speech cues and personality of the target speaker [213]. Regarding context-aware personality recognition (the focus of this chapter), a similar trend is seen, but the literature is even scarcer, as discussed next.

Automatic personality recognition has been addressed in the literature using different data modalities, *i.e.*, *still images* [60], *image sequences* [359], *audiovisual* [28, 210, 108], *speech and text* [12], and *multi-modal* [14]. Apparent or self-reported personality has also been inferred from gaze behavior [162], mood [348], and even from behavior patterns collected from smartphones [354]. Preliminary studies tended to use handcrafted features representing gestures and speech with standard machine learning techniques [269, 108]. More recent works rely on deep learning approaches (from either handcrafted features [12] or raw data [253]), such as convolutional and/or recurrent architectures to leverage spatiotemporal dynamics [358, 142].

Most works focus on personality recognition or perception from the individual point of view [153, 419], even in dyadic or small group conversational scenarios [15], using only features from the target person. The first works that considered interpersonal dependencies for personality computing in face-to-face interaction scenarios used descriptive statistics of overlapping speech segments, short interjections, backchanneling, or interruptions [296, 382], or percentages of attention given by the target speaker to other participants and attention received by them [212, 16, 359], in addition to other audio/video features of the target speaker. Some works also considered labeled co-occurrent events, such as attention given/received while speaking/not speaking [211]. Okada *et al*. [271] proposed to discover frequent co-occurrent events between multiple modalities and people using graph clustering in a small group scenario. In particular, they used utterance segments, speech, gaze, head and body gestures. In a similar scenario, [108] obtained the highest accuracy when using intrapersonal (speech-, prosodic-, and visual-based), dyadic (speech-based, such as speaking interruptions and backchanneling), and speech-based one-versus-all features.

Most of the aforementioned methods rely on handcrafted interpersonal features. To our knowledge, only a handful of methods propose interlocutor- or context-aware architectures for self-reported personality recognition in small group interactions. Most recent works focus on personality analysis on social media, generally limited to the textual modality (see [245] for a complete review), involving much more people while missing useful cues from face-to-face interactions. The work of Su *et al*. [356] was one of the first for dyadic conversations, but focusing on personality perception (*i.e.*, personality reported by external observers [176]). They relied on a recurrent network to model the relationship between the linguistic features of each speaking turn and personality, and on coupled Hidden Markov Models to then model the long-term turn-taking temporal evolution and cross-speaker contextual information to infer the personality of both individuals. Other works have also focused on modeling transcribed interviews [161], but disregarding the interviewer, hence missing a chance to exploit interpersonal context.

With respect to self-reported personality inference in small group interactions, the work by Lin *et al*. [227] proposed an interlocutor-modulated recurrent attention model with turn-based acoustic features, which models the vocal self and interactive behaviors of a target

speaker during small group interactions. Later, Zhang *et al*. [455] predicted self-reported personality and performance labels by correlation analysis of co-occurrent intrapersonal and interpersonal key action events, extracted from head and hand pose, gaze, and motion intensity features. Regarding context, Principi *et al*. [305] were one of the first to consider person metadata (*e.g.*, gender, age, ethnicity, and perceived attractiveness) with audiovisual data. However, their goal was to better approximate the crowd biases for apparent personality recognition in one-person videos. Concurrently to the work described in this chapter, Shao *et al*. [341] proposed to infer an individual's personality by modeling their cognitive processes. More concretely, the approach first learns a person-specific convolutional network that predicts the target's facial reactions to the other speaker's audiovisual cues (facial landmarks and Mel-frequency cepstral coefficients). Then, personality is inferred from the graph representation of the target's person-specific processor. Since such processor is optimized on all available data of the target, it can use the information from entire videos at once to infer personality, not only short video snippets.

Contrary to previous works, we use different sources of context, including both interlocutors, scene, and task information to infer personality, using for the first time a video-based Transformer adapted to include audio and further context as metadata.

Long-term modeling in personality computing. The need for longer-term modeling in personality regression tasks is highlighted in [350]. The authors proposed a model based on facial features for individual apparent and self-reported personality, but limited to 3-second time windows. Others have attempted long-term modeling of features for personality inference, but most are limited to compute sequence representations by averaging small clips or individual frames features [209, 185], which miss temporal relationships. As far as we know, only one previous work has used up to 1 minute without aggregating across clips [370]. However, they focused on first-impressions regression, which does not benefit from longer temporal windows [11, 413].

In the past years, a new family of architectures has risen to address some limitations of traditional recurrent methods [188], *i.e.*, the Transformer [385], which has shown impressive results for many sequence modeling tasks in a plethora of modalities [88, 91, 31]. As already mentioned, these models are capable of attending to long-range data dependencies with few layers, allowing them to learn very useful representations. Recently, some works have started using Transformer-like architectures to model personality. However, these works tend to focus on the apparent personality of individuals alone [134] by only modeling text features, generally on social media posts [208, 431]. We focus on self-reported personality on real face-to-face dyadic interactions, which our proposed architecture explicitly exploits, and use the Transformer to model video, audio, and metadata modalities altogether. Furthermore, our Dyadformer exploits lonng-term interdependencies up to 30 seconds.

Multi-modal Transformers. Our work is related to the recent use of Transformers [385] to learn multi-modal representations. The most common approach employs contrastive losses to bring paired samples (such as video and caption [130] or subtitles [215]) closer together.

This is generally used for captioning or retrieval tasks, where both modalities provide similar information and the aim is to *translate* between them (see Sections 4.5.2 and 4.5.3). However, in our setting we expect audio and video to convey different complementary information, for which we explore two better suited Transformer families (we describe these alternatives in detail in Section 4.3.4). The first one uses a BERT-like [88] stream which concatenates modalities along the temporal dimension [205, 118] before input, effectively doubling sequence length. Nevertheless, as Transformers scale quadratically with input length, these methods incur in memory efficiency limitations. The second one solves this by using separate cross-attention streams [180, 173], replacing self-attention to allow both modalities to attend and enrich each other, akin to MHCA (see Section 3.1.1), while the separate streams allow for independent modeling and maintain sequence length. This design has generally been used to fuse two modalities, as is our case, but it can be extended further [468]. In Section 3.4.2, we test using a BERT-like approach but, when compared with the latter alternative and in our setting, we find it to underperform. For this reason, we opt for cross-attentional streams to fuse multimodal information and go one step further by also using this technique to model cross-subject interactions (see Section 3.3 for details).

In this regard, our baseline uses audiovisual data and different sources of context from both interlocutors and the situation istelf to regress a target person's personality traits. Multimodal fusion is done by simply concatenating the information from the video and audio modalities, different from the Dyadformer, which exploits cross-attention to fuse modalities within the architecture itself.

3.2 Approaching UDIVA with Transformers

This section provides a first insight into the UDIVA dataset by evaluating it in a personality traits inference task. We present a transformer-based context-aware model to regress self-reported personality traits of a target person during a dyadic interaction. Then, we assess its performance and the effect of adding several sources of context. Method, evaluation protocol and results are described next.

The baseline is a multi-modal attention-based architecture that receives different sources of information and context from both participants in the dyadic interaction to regress the target person's personality traits. It is a re-purposed Video Action Transformer Network [131], the input of which consists of video, audio, and metadata information. The time-synchronized full-length videos of both interlocutors are first split into non-overlapping chunks, and similarly for the audio. The face region from the target person's video is then cropped to attain a face-only video. All these are independently encoded through pre-trained networks. A fundamental characteristic of the Video Action Transformer Network is the selection of the query, keys and values. In the case of the proposed baseline, the query incorporates the face and the metadata of the target person. We consider two types of keys and values: local and extended. Local keys/values include audiovisual embeddings from the target person, while the extended counterparts include audiovisual and metadata embeddings from the other interlocutor. The local and extended key and value embeddings together with the query are

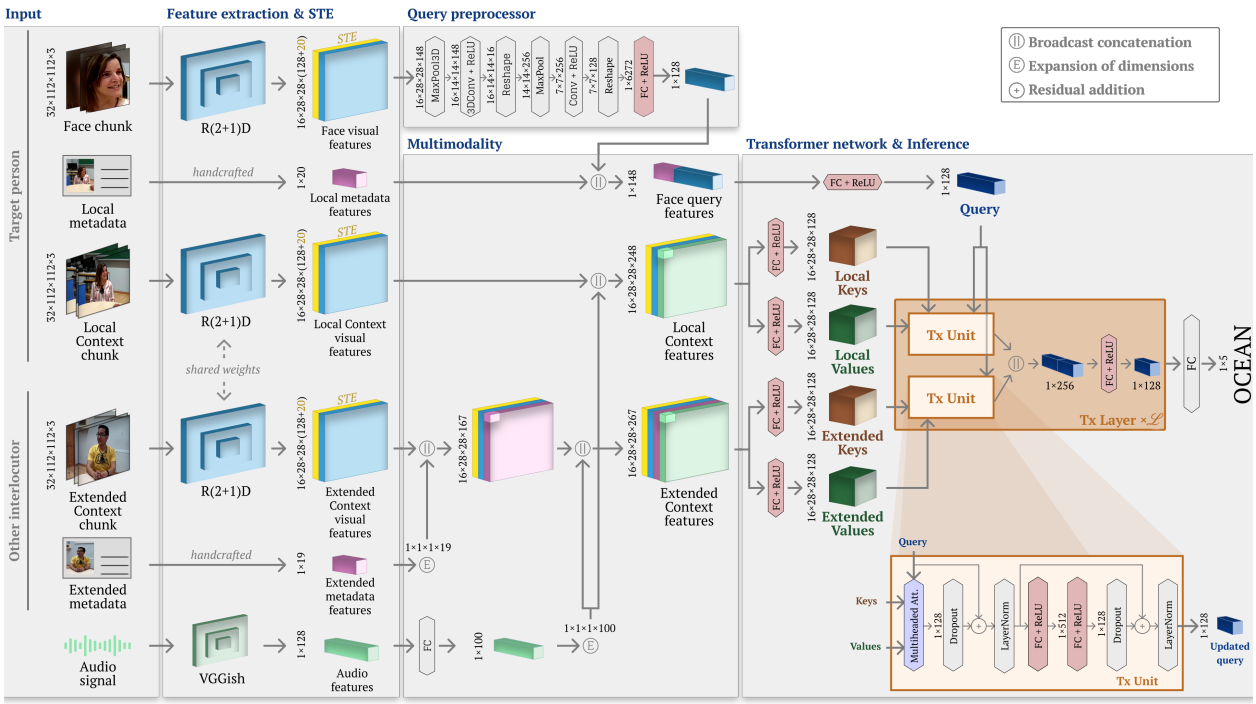


FIGURE 3.2: Pipeline of the proposed baseline methodology to infer self-reported personality (OCEAN) traits from multi-modal synchronized signals and context. Input consists of visual (face, local context, and extended context chunks), audio (raw audio chunk), and metadata (both interlocutors' characteristics, and session and dyadic features). Feature extraction consists of two backbones: a R(2+1)D network for the visual chunks and VGGish for the audio one. The visual features from the R(2+1)D's 3rd residual block are concatenated to spatiotemporal encodings (STE). The VGGish's audio features and handcrafted metadata features are incorporated to visual context/query features and such representations are transformed to the set of Query, Keys, and Values to input to the Transformer network (Tx). The Tx consists of $\mathscr{L}$ Tx layers, each equipped with Local and Extended Context Transformer Tx units. The Tx Units implement multiheaded attention and provide their updated queries, which are combined and fed to the next Tx Layer. Finally, based on the $\mathscr{L}$-th Tx layer output, the network uses a fully-connected (FC) layer to regress per-chunk OCEAN scores.

processed independently in two different units of the transformer layer. They provide two updated queries that are concatenated and linearly projected to produce the final context-updated query. The baseline architecture uses three such transformer layers. Since the model receives chunks of the original videos as input, the predicted personality traits are obtained at chunk level, and later aggregated by particpant using the median from all their chunks.

3.2.1 Baseline method

The attention mechanism of our transformer-based method relates an initial query, in this case the target person's face, to the nonverbal behavior of both interlocutors, the overall scene, and further contextual metadata, and updates it with relevant context. The process is repeated with the updated query in consecutive layers to eventually infer the personality (OCEAN) traits. The proposed method consists of several stages, detailed below. All components and the information flow among them are illustrated in Figure 3.2.

Audiovisual input. We use the pair of RGB time-synchronized full-length videos from both subjects available in UDIVA v0.5 (*i.e.*, FC1 and FC2 camera views). They contain the target person, denoted as *local context*, and the other interlocutor or *extended context*). We divide them into 32-frame non-overlapping chunks and resize each chunk's spatial resolution to 112×112 to obtain, respectively, $\widetilde{\widetilde{\mathbf{X}}}_{\mathrm{V}}^{\mathrm{L}}, \widetilde{\widetilde{\mathbf{X}}}_{\mathrm{V}}^{\mathrm{E}} \in [0, 255]^{32 \times 112 \times 112 \times 3}$. Note that we here use $\mathbf{X}$ to denote any input (different from Section 3.1.1 where $\mathbf{Z}$ was used for decoder inputs). Moreover, we use the superindex to denote context source (*e.g.*, Local or Extended), and drop the layer index to ease notation. The 32 frames of the chunks are sampled with a stride of 2, such that a chunk corresponds to 2.56 seconds of the original videos. Also, we detect the target person's face regions (see below for details) in the original video, crop, and re-scale them to form the face chunk $\widetilde{\widetilde{\mathbf{X}}}_{\mathrm{V}}^{\mathrm{F}} \in [0, 255]^{32 \times 112 \times 112 \times 3}$. Apart from the visual data, we define an audio chunk $\widetilde{\widetilde{\mathbf{x}}}_{\mathrm{A}} \in \mathbb{R}^{132\,300}$ consisting of the raw audio frames acquired at 44.1 KHz from the general microphone (or one of the lapels if the general one was not available for that session), and time-synchronized to its respective video chunk.

Face detection and tracking. In order to detect the faces we use MobileNet-SSD [164], deployed using Tensorflow Object Detection API [170] and pretrained on the Wider Face Dataset [433]. As we consider only frontal cameras, the detection task is not very challenging, therefore, on more than 95% of the videos, the detection ratio is higher than 75%. In case the gap between consecutive detections is lower than 25 frames (1 second), we linearly interpolate the coordinates of the boxes. If the gap was larger than 25 frames, those were discarded. Since there are frames in which the frontal cameras capture both participants, we need to identify the target subject before computing the face chunks. In order to do so, we employ a basic tracking algorithm based on the following 2 steps: (1) *identify* target person's face: given a video, the face of the target person is considered the first detection that has a mean intersection over union (IoU) score higher than 0.2 with respect to all the other faces in the video; (2) *track* target person face throughout the video based on the IoU.

Size of video chunks. The original Video Action Transformer [131] uses an I3D backbone pretrained on Kinetics-400 [57] for spatiotemporal feature extraction. Such backbone uses 64 frames per chunk, which is equivalent to around 3 seconds of video. Instead, we opted for the R(2+1)D backbone [377] pretrained on IG-65M dataset, which has shown to provide significant performance gains [128]. This backbone uses 32 frames per chunk, so by using a stride of 2 we manage to encode approximately the same time window as the original method with half the number of frames while reducing the memory load. This is equivalent to downsampling the original videos from 25 fps to 12.5, that is, 1 frame every 0.08 seconds. Although not frequent, there is a chance to miss some fast-paced facial and body micro-actions in such downsampling process. However, there is also the trade-off we try to balance between losing some of these fast micro-actions and being able to include a larger, and also important, temporal context.

Context type			Source	Value range normalization	Output size
Individual	Stable (across sessions)	Age	Self-reported	$[17, 75] \rightarrow [0, 1]$	1D
		Gender	Self-reported	$\{\text{F}, \text{M}\} \rightarrow \{0, 1\}$	1D
		Cultural background	Self-reported (country of origin)	Recategorization based on cultural differences [255]	6D (one-hot encoding)
	Transient (per session)	Session index	Session info.	$[1, 5] \rightarrow [0, 1]$	1D
		Pre-session mood	Self-reported [119] (8 categories*, Likert scale)	$[1, 5] \rightarrow [0, 1]$ (for each category)	8D
		Pre-session fatigue	Self-reported (Rating scale)	$[0^{\dagger}, 10] \rightarrow [0, 1]$	1D
Session	Order of the task within the session		Session info.	$[1, 4] \rightarrow [0, 1]$	1D
	Task difficulty†		External survey	$[0, 3] \rightarrow [0, 1]$	1D
Dyadic	Interlocutors' relationship		Self-reported	$\{\text{N}, \text{Y}\} \rightarrow \{0, 1\}$	1D

*Categories: *good*, *bad*, *happy*, *sad*, *friendly*, *unfriendly*, *tense*, and *relaxed*.
† Sessions with fatigue data missing were assigned a value of 0.
‡ Tasks with no difficulty level associated were assigned a value of 0.

TABLE 3.1: Description of the different sources of context included as metadata in the proposed personality inference baseline model.

Metadata input. Different sources of context are captured in the form of input metadata. We consider 2 types of metadata (see Table 3.1): (1) *local metadata*, containing *individual* context from the target person and *session* information; and (2) *extended metadata*, with *individual* context from the other interlocutor and *dyadic* features.

Feature extraction. First, we normalize the pixel values of $\{\widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{F}}, \widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{L}}, \widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{E}}\}$ in the range $[0, 1]$, subtracting and dividing them by the mean and standard deviation of the IG-65M dataset [128]. Then, we feed them to the R(2+1)D network [377] backbone, pretrained on that same dataset, and save the rich spatiotemporal features produced by the R(2+1)D's 3rd convolutional residual stack: $\widetilde{\mathbf{X}}_{\text{V}}^{\text{F}} = f_{\text{F}}(\widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{F}}; \mathbf{W}_{\text{F}})$, $\widetilde{\mathbf{X}}_{\text{V}}^{\text{L}} = f_{\text{C}}(\widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{L}}; \mathbf{W}_{\text{C}})$, $\widetilde{\mathbf{X}}_{\text{V}}^{\text{E}} = f_{\text{C}}(\widetilde{\widetilde{\mathbf{X}}}_{\text{V}}^{\text{E}}; \mathbf{W}_{\text{C}})$, where $\mathbf{W}_{\text{F}}$ and $\mathbf{W}_{\text{C}}$ are the weights of the face network $f_{\text{F}}(\cdot)$ and context network $f_{\text{C}}(\cdot)$, respectively (note both local and extended context are processed with the same weights). $\widetilde{\mathbf{X}}_{\text{V}}^{\text{F}}, \widetilde{\mathbf{X}}_{\text{V}}^{\text{L}}, \widetilde{\mathbf{X}}_{\text{V}}^{\text{E}} \in \mathbb{R}^{16 \times 28 \times 28 \times 128}$ denote the *face*, *local context*, and *extended context visual features*, respectively. For the audio feature extraction, we use the VGGish [159] backbone. This VGG-like model, developed specifically for the audio modality and with pre-trained weights $\mathbf{W}_{\text{A}}$ learned on a preliminary version of the YouTube-8M [3], provides a feature vector $\widetilde{\mathbf{x}}_{\text{A}} \in \mathbb{R}^{128}$ encoding information contained in the $\widetilde{\widetilde{\mathbf{x}}}_{\text{A}}$ chunk: $\widetilde{\mathbf{x}}_{\text{A}} = f_A(\widetilde{\widetilde{\mathbf{x}}}_{\text{A}}; \mathbf{W}_{\text{A}})$. Finally, input metadata is normalized according to Table 3.1, and encoded in $\boldsymbol{m}_{\text{L}} \in \mathbb{R}^{20}$ and $\boldsymbol{m}_{\text{E}} \in \mathbb{R}^{19}$ for *local* and *extended metadata features*, respectively.

Spatiotemporal encodings (STE). Following other transformer-like architectures, we need to add positional encodings to our audiovisual feature embeddings $\widetilde{\mathbf{X}}$, which can be either learned or fixed. We opt to learn them end-to-end. Different from the original positional encodings we described in Section 3.1.1, we instead opt to define a vector of zero-centered time

indices $\boldsymbol{t} = \langle -\frac{16}{2}, -\frac{16}{2}+1, \ldots, \frac{16}{2}-1 \rangle$ (as the temporal dimension of the different $\widetilde{\mathbf{X}}$ is 16). The *temporal encodings* are computed by a two-layered network: $\widetilde{\mathbf{E}}_{\mathrm{T}} = \mathrm{ReLU}\left(\mathbf{W}_{\mathrm{T}_1}^{\top}\mathrm{ReLU}\left(\mathbf{W}_{\mathrm{T}_2}^{\top}\boldsymbol{t}\right)\right)$, where $\mathbf{W}_{\mathrm{T}_1} \in \mathbb{R}^{1\times 20}$ and $\mathbf{W}_{\mathrm{T}_2} \in \mathbb{R}^{20\times 10}$ are learned weights. The *spatial encodings* $\widetilde{\mathbf{E}}_{\mathrm{S}}$ are computed by a similar encoding network. Given that 28×28 is the spatial resolution of the features, we feed to the spatial encoding network a tensor of spatially zero-centered position indices $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{28\times 28\times 2}$, where $\boldsymbol{\mathcal{S}}_{i,j} = \langle i - \frac{28}{2}, j - \frac{28}{2} \rangle, \forall i, j \in [0, 28)$ and weights $\mathbf{W}_{\mathrm{S}_1} \in \mathbb{R}^{2\times 20}$ and $\mathbf{W}_{\mathrm{S}_2} \in \mathbb{R}^{20\times 10}$. Then, $\widetilde{\mathbf{E}}_{\mathrm{T}}$ and $\widetilde{\mathbf{E}}_{\mathrm{S}}$ are reshaped to $\mathbf{E}_{\mathrm{T}} \in \mathbb{R}^{16\times 1\times 1\times 10}$ and $\mathbf{E}_{\mathrm{S}} \in \mathbb{R}^{1\times 28\times 28\times 10}$ and concatenated together by broadcasting singleton dimensions, *i.e.*, $\mathbf{E} = \mathbf{E}_{\mathrm{S}} \parallel \mathbf{E}_{\mathrm{T}}$, where $\mathbf{E} \in \mathbb{R}^{16\times 28\times 28\times 20}$.

Multi-modality: fusing visuals with position, audio, and metadata. *Local* and *extended visual context features* ($\widetilde{\mathbf{X}}_{\mathrm{V}}^{\mathrm{L}}$ and $\widetilde{\mathbf{X}}_{\mathrm{V}}^{\mathrm{E}}$) are augmented with positional encodings ($\mathbf{E}$) and audio features. The original 128-dimensional global *audio features* $\widetilde{\mathbf{x}}_{\mathrm{A}}$ are linearly projected to a more compact 100-dimensional representation and reshaped to $\widetilde{\mathbf{X}}_{\mathrm{A}} \in \mathbb{R}^{1\times 1\times 1\times 100}$. Then, the *local context features* are simply $\mathbf{X}^{\mathrm{L}} = \widetilde{\mathbf{X}}_{\mathrm{V}}^{\mathrm{L}} \parallel \mathbf{E} \parallel \widetilde{\mathbf{X}}_{\mathrm{A}}$, where $\mathbf{X}^{\mathrm{L}} \in \mathbb{R}^{16\times 28\times 28\times 248}$. The *extended context features* are similarly augmented, and also include *extended metadata* from the interlocutor. This is achieved by reshaping $\boldsymbol{m}_{\mathrm{E}} \in \mathbb{R}^{19}$ to $\boldsymbol{\mathcal{M}}_{\mathrm{E}} \in \mathbb{R}^{1\times 1\times 1\times 19}$ and applying broadcast concatenation, that is $\mathbf{X}^{\mathrm{E}} = \widetilde{\mathbf{X}}_{\mathrm{V}}^{\mathrm{E}} \parallel \mathbf{E} \parallel \widetilde{\mathbf{X}}_{\mathrm{A}} \parallel \boldsymbol{\mathcal{M}}_{\mathrm{E}}$, resulting in $\mathbf{X}^{\mathrm{E}} \in \mathbb{R}^{16\times 28\times 28\times 267}$. We note that local metadata features $\boldsymbol{m}_{\mathrm{L}}$ are not included with local context because they are concatenated with the query, as we define next.

Query Preprocessor (QP). This small module transforms the positionally augmented *facial features* into a vector form: $\mathbf{x}_{\mathrm{Q}} = \mathrm{QP}(\widetilde{\mathbf{X}}_{\mathrm{V}}^{\mathrm{F}} \parallel \mathbf{E})$, $\mathbf{x}_{\mathrm{Q}} \in \mathbb{R}^{128}$. The QP consists of a 3D max pooling layer of size $(1, 2, 2)$ and stride $(1, 2, 2)$, a 3D conv layer of size $(1, 1, 1)$ and 16 filters, a ReLU activation function layer, a permutation of dimensions and reshaping so that the temporal dimensions and the channels are merged into the same dimension, a 2D max pooling of size $(2, 2)$, a 2D conv layer of size $(1, 1)$, a ReLU activation layer, a flattening, and a fully-connected (FC) layer of size 128, another ReLU, and a dropout layer. Finally, the *face query features* $\mathbf{q} \in \mathbb{R}^{148}$ are built from the combination of the QP output along with the target person's *local metadata*: $\mathbf{q} = \mathbf{x}_{\mathrm{Q}} \parallel \boldsymbol{m}_{\mathrm{L}}$.

Keys, Values, and Query. To obtain the final input to the transformer layers, we first need to transform *local context* and *extended context features* into two different 128-dimensional embeddings (Keys and Values), and also the *face query features* into the query embedding of the same size. The *Local keys* and *Local values* are then $\mathbf{K}_{\mathrm{L}} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{K}_{\mathrm{L}}}^{\top}\mathbf{X}^{\mathrm{L}})$ and $\mathbf{V}_{\mathrm{L}} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{V}_{\mathrm{L}}}^{\top}\mathbf{X}^{\mathrm{L}})$ where $\mathbf{W}_{\mathrm{K}_{\mathrm{L}}}, \mathbf{W}_{\mathrm{V}_{\mathrm{L}}} \in \mathbb{R}^{248\times 128}$, whereas the *Extended keys* and *Extended values* are $\mathbf{K}_{\mathrm{E}} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{K}_{\mathrm{E}}}^{\top}\mathbf{X}^{\mathrm{E}})$ and $\mathbf{V}_{\mathrm{E}} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{V}_{\mathrm{E}}}^{\top}\mathbf{X}^{\mathrm{E}})$, where $\mathbf{W}_{\mathrm{K}_{\mathrm{E}}}, \mathbf{W}_{\mathrm{V}_{\mathrm{E}}} \in \mathbb{R}^{267\times 128}$. The input *Query* representation $\mathbf{q}^0 \in \mathbb{R}^{128}$ is computed as $\mathbf{q}^0 = \mathrm{ReLU}(\mathbf{W}_{\mathrm{Q}^0}^{\top}\mathbf{q})$, where $\mathbf{W}_{\mathrm{Q}^0} \in \mathbb{R}^{148\times 128}$.

Transformer network. Our transformer network (Tx) is composed of $\mathscr{L} = 3$ Tx layers with 2 Tx units each, one for the local context and another one for the extended context. A Tx unit

	Query		**Key** and **Value**			
	Face*	Metadata*	Frame*	Frame‡	Metadata‡	Audio
M	-	-	-	-	-	-
L	✓	-	✓	-	-	-
Lm	✓	✓	✓	-	-	-
LE	✓	-	✓	✓	-	-
LEm	✓	✓	✓	✓	✓	-
LEa	✓	-	✓	✓	-	✓
LEam	✓	✓	✓	✓	✓	✓

* target person and ‡ interlocutor data.

TABLE 3.2: Evaluated scenarios. Mean value prediction (M) obtained from the mean of the per-trait ground truth labels of the training set; and the proposed baseline method with/without Local (L) and Extended (E) context, Metadata (m), and Audio (a) information.

follows the canonical Transformer encoder we defined in Section 3.1.1 (see left side of Figure 3.1) but replaces the MHSA sub-layer by a MHCA one. Intuitively, the Tx unit receives the queries $\mathbf{q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, and iteratively refines $\mathbf{q}$ by cross-attending the extended and local contexts in $\mathbf{K}$ and $\mathbf{V}$. These MHCA have $h = 2$ heads each, which compute a separate $128/h$-dimensional linear projection of the query, the keys, and the values, and apply scaled dot product attention[2]. Then, it concatenates the h outputs, and linearly projects them back to a new 128-dimensional query. After the multi-headed attention, the resulting query follows the rest of the pipeline in the Tx unit (as illustrated in Fig. 3.2) to obtain the *updated query*. Note that each unit in the ℓ-th layer provides its own updated query, denoted as $\mathbf{q}_{\mathrm{L}}^{\ell} \in \mathbb{R}^{128}$ and $\mathbf{q}_{\mathrm{E}}^{\ell} \in \mathbb{R}^{128}$, $0 < \ell \leq \mathscr{L}$. These are next concatenated together and fed to a FC layer to obtain the ℓ-th layer's joint updated query $\mathbf{q}^{\ell} = \mathrm{ReLU}\left(\mathbf{W}_{Q^{\ell}}^{\top}(\mathbf{q}_{\mathrm{L}}^{\ell} \parallel \mathbf{q}_{\mathrm{E}}^{\ell})\right)$, where $\mathbf{W}_{Q^{\ell}} \in \mathbb{R}^{256\times128}$. Finally, $\mathbf{q}^{\ell}$ is fed as input to the next ($\ell + 1$-th) layer.

Inference. The per-chunk OCEAN traits are obtained by applying a FC layer to the updated query from the $\mathscr{L}$-th (last) layer, *i.e.*, $\mathbf{y} = \mathbf{W}_{\mathrm{Y}}^{\top}\mathbf{q}^{\mathscr{L}}$ where $\mathbf{W}_{\mathrm{Y}} \in \mathbb{R}^{128\times5}$. Final per-trait, per-subject predictions are computed as the median of the chunks predictions for each participant.

3.2.2 Ablative experiments

This section describes the experimental setup used to assess the performance of the personality inference baseline model.

Data. We use the UDIVA v0.5, with the subset of sessions and participants, as well as the splits defined in Section 2.4.1. We use frontal camera views (FC1 and FC2, see Fig. 2.1), in line with the proposed methodology. As personality labels, we use the raw OCEAN scores obtained from the self-reported BFI-2 questionnaire, converted into z-scores using descriptive data from normative samples. Since the duration of the videos is not constant throughout sessions and tasks, in order to balance the number of samples we uniformly selected around 120 chunks from each stream, based on the median number of chunks per video. The final

[2]Note that, while in Equation (3.4) MHCA separately maps a single embedding $\mathbf{M}$ into keys and values, we here input distinct embeddings for each of them ($\mathbf{K}$ and $\mathbf{V}$).

sample of chunks contains 94 960 instances for training, 15 350 for validation and 7 870 for test (equivalent to 67.5/10.9/5.6 hours, respectively), distributed among the 4 tasks.

Training strategy. The proposed model was trained using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and a learning rate of $1e-5$. We used a batch size of 2 and the Mean Squared Error (MSE) as the loss function. We compute the validation error approximately 30 times per epoch and select the model that gives the best results considering the mean with its previous and next evaluation scores. The final results were obtained by freezing the layers of the R(2+1)D backbones, as strategies such as finetuning end-to-end or only the last block of the feature extractors led to fast overfitting.

Evaluation protocol. We follow an incremental approach, starting from the *local context*. Six different scenarios are evaluated, summarized in Table 3.2. We train one model per scenario and task, since each of the four tasks can elicit different social signals and behaviors (detailed in Section 2.2.3), which can be correlated to different degrees with distinct aspects of each personality trait. Results are evaluated with respect to the Mean Squared Error between the aggregated personality trait score and associated ground truth label for each individual in the test set. We also compare the results to a mean value prediction, computed as the mean of the per-trait ground truth labels of the training set.

3.2.3 Discussion of results

Obtained per-task results for the different scenarios are shown in Table 3.3. We discuss some of the findings below.

Effect of including extended (E) visual information. The *extended context* contains visual information from the other interlocutor's behaviors and surrounding scene, allowing the model to consider interpersonal influences during a chunk. By comparing "L" vs. "LE" we can observe that, on average, only *Talk* benefits from the addition of the extended visual context. Trait-wise, *Extraversion* improves for all tasks except for *Lego*, which performs worse for all traits. This can be attributed to the fact that the interaction during this type of collaboration is more slow-paced than in other tasks. Therefore, interpersonal influences cannot be properly captured during just one chunk. In contrast, for more natural tasks such as *Talk*, or fast-moving games such as *Ghost*, there are many instant actions-reactions that can be observed during a single chunk, the effect of which is reflected in the improved results for those tasks. This motivates the need to extend the model to capture longer-time interpersonal dependencies, characteristic of human interactions, across a series of ordered chunks along time, to truly benefit from this extended information. The positive effects of including extended visual information are better seen in the presence of metadata. "LEm" obtains lower error than "Lm" on average for all tasks except *Animals*, where it only outperforms the variant without extended context for *Conscientiousness* and *Agreeableness*. This suggests that some features included within the session and dyadic metadata are crucial for the proper integration of the

Arch. \ Trait	O	C	E	A	N	Avg.
Animals						
M	0.731	0.871	0.988	0.672	1.206	0.894
L	0.742	0.879	0.955	0.674	1.133	0.877
Lm	**0.721**	0.874	0.946	0.684	1.154	0.876
LE	0.733	0.832	0.988	0.672	1.221	0.889
LEm	0.736	0.834	0.968	0.669	1.192	0.880
LEa	0.722	0.827	0.954	0.672	1.211	0.877
LEam	0.737	**0.756**	**0.887**	**0.580**	**1.023**	**0.797**
Ghost						
M	0.733	0.887	0.991	0.674	1.220	0.901
L	0.744	0.891	1.010	0.677	1.242	0.913
Lm	0.759	0.859	1.027	**0.642**	1.208	0.899
LE	0.731	0.905	0.956	0.676	1.291	0.912
LEm	0.743	0.944	0.868	0.657	1.153	0.873
LEa	**0.730**	**0.872**	0.950	0.672	1.199	0.885
LEam	0.741	0.893	**0.844**	0.667	**1.139**	**0.857**
Lego						
M	0.738	0.871	0.990	0.676	1.204	0.896
L	**0.723**	0.852	0.917	0.676	1.164	0.866
Lm	0.725	0.798	0.857	0.618	1.101	0.820
LE	0.731	0.885	0.949	0.676	1.230	0.894
LEm	0.727	**0.763**	**0.826**	**0.611**	**1.037**	**0.793**
LEa	0.742	0.867	0.941	0.672	1.229	0.890
LEam	0.745	0.839	0.953	0.659	1.099	0.859
Talk						
M	**0.731**	0.872	0.991	0.673	1.211	0.896
L	0.769	0.769	0.997	0.664	1.177	0.875
Lm	0.743	0.798	0.962	**0.636**	1.168	0.861
LE	0.738	0.793	0.964	0.673	1.094	0.852
LEm	0.825	**0.718**	0.878	0.639	1.047	0.821
LEa	0.757	0.728	0.970	0.664	1.106	0.845
LEam	0.773	0.790	**0.869**	0.670	**0.985**	**0.817**

TABLE 3.3: Obtained results on different tasks. Legend: Mean value prediction (M) obtained from the mean of the per-trait ground truth labels of the training set; and the proposed baseline with/without Local (L) and/or Extended (E) context, Metadata (m), and Audio (a) information.

context provided by the interaction partner. We further validate the importance of metadata next.

Effect of including metadata (m) information. The inclusion of metadata validates our intuition that personal, task, and dyadic details provide relevant information to the model to produce overall better predictions, particularly if the cases "L" vs. "Lm", "LE" vs. "LEm", and "LEa" vs. "LEam" are compared, with the largest improvement observed for *Lego* (11.29%, "LE" vs. "LEm" case). Considering the high heterogeneity and dimensionality of behaviors revealed in an interaction and their multiple meanings, these concise features appear to be beneficial to better guide the model and establish meaningful patterns in the data. Nonetheless, a systematic study would be needed to assess the effect of each feature individually.

Effect of including audio (a) information. From comparing "LE" vs. "LEa" and "LEm" vs. "LEam", we observe that better results are obtained, on average, for all the tasks when audio information is considered. In line with previous literature [390], it is clear that paralinguistic acoustic features are required to better model personality. However, the observed improvement is smaller for *Lego*. One plausible reason would be the noise produced by the Lego pieces while being moved, or by the instructions book while turning its pages close to the microphones, which would interfere with the learning process. In the case of more natural routines like *Talk*, the influence of audio is not as strong as we would have expected. In contrast, *Animals*, another speaking-based task, obtains the best results for almost all traits when audio is considered. There is one salient difference among these two tasks that may explain this pattern. The latter elicits more individual covert thinking and cognitive processes that cannot be entirely observed from the visual modality, so most of the overt information comes from the spoken conversation. In contrast, the former elicits a larger range of visual cues which may be more relevant than acoustic features for certain traits.

Putting everything together. In the last experiment ("LEam"), the model is aware of the overall contextual information. We notice that apart from *Lego*, for which the audio drawbacks were already commented, all the other tasks seem to benefit from the provided knowledge, obtaining the lowest error value on average.

Mean prediction comparison. We observe that *Agreeableness*, followed by *Open-mindedness*, obtain the lowest error among mean value prediction ("M") results, indicating that ground truth labels for such traits are more concentrated. In those cases, none of the models achieve a substantial improvement over the mean prediction, except for *Animals*, where "LEam" obtains an error of 0.58, the lowest overall. At the other end we find *Negative emotionality*, which is the trait with most spread values, but also the one for which we obtain the largest benefits with the evaluated models. In particular, the largest improvement overall (18.66%) is given by "LEam" for *Talk*.

3.2.4 Limitations

The proposed baseline model shows promise in integrating multiple sources of context, enabling it to capture some of the complexity of dyadic interactions. As we have discussed, this can be seen by the consistent improvement in personality regression results as more context is added. Nonetheless, there are noteworthy limitations in its approach.

One significant constraint arises from the fact that the multimodal concatenation occurs outside the Transformer architecture. As they are linearly mapped before being tokenized, this could lead to features from different modalities being mixed. In this sense, the Transformer may fail to selectively attend to different modalities, potentially incurring in the omission of relevant cues that are crucial for understanding the nuances of personality dynamics. Furthermore, the compression of information into a reduced query vector might result in the loss of fine-grained information from the target participant, which can limit the model's potential to accurately capture the intricacies of personality regression in dyadic interactions. Transformers' lack of inductive biases makes them a promising tool for multi-modal integration that could be exploited for our purposes.

Another notable limitation is the reliance on short 3-second chunks as input to the baseline model. Personality is a complex human characteristic that often manifests over longer durations and patterns. Consequently, attempting to infer personality solely from brief interactions can be misleading and excessively noisy. These short segments might not be truly indicative of the individuals' underlying personality traits, thereby introducing a significant level of uncertainty and potential misinterpretation. In order to develop a more robust model, it is imperative to incorporate longer-range interactions to provide a comprehensive view of how personality traits manifest and evolve within a dyadic context.

Finally, this baseline treats one participant as the target and the other as supplementary context. This is inherently equivalent to modeling just one of the participants in the interaction, while the other is bunched together with all remaining contextual factors. In this sense, the baseline overlooks the potential benefits of considering both interactants as independent but equally important entities within the interaction. Dyadic interactions are inherently co-dependent, and the personality dynamics of both participants influence the interaction's overall nature. By treating each participant as an individual with their own personality traits and characteristics, the model could be better equipped to capture the mutual influence, power dynamics, and interplay that contribute to the dyadic interaction, fostering a more accurate prediction of individual behaviors.

Guided by these limitations, we next propose a novel design to better integrate multi-modal cues, leverage longer temporal contexts, and jointly model both participants.

3.3 The Dyadformer

In this section, we present the Dyadformer (depicted in Figure 3.3). The Dyadformer is a multi-subject multi-modal architecture that is composed of a set Transformer layers. The Dyadformer receives as input a sequence of N small, consecutive and temporally aligned

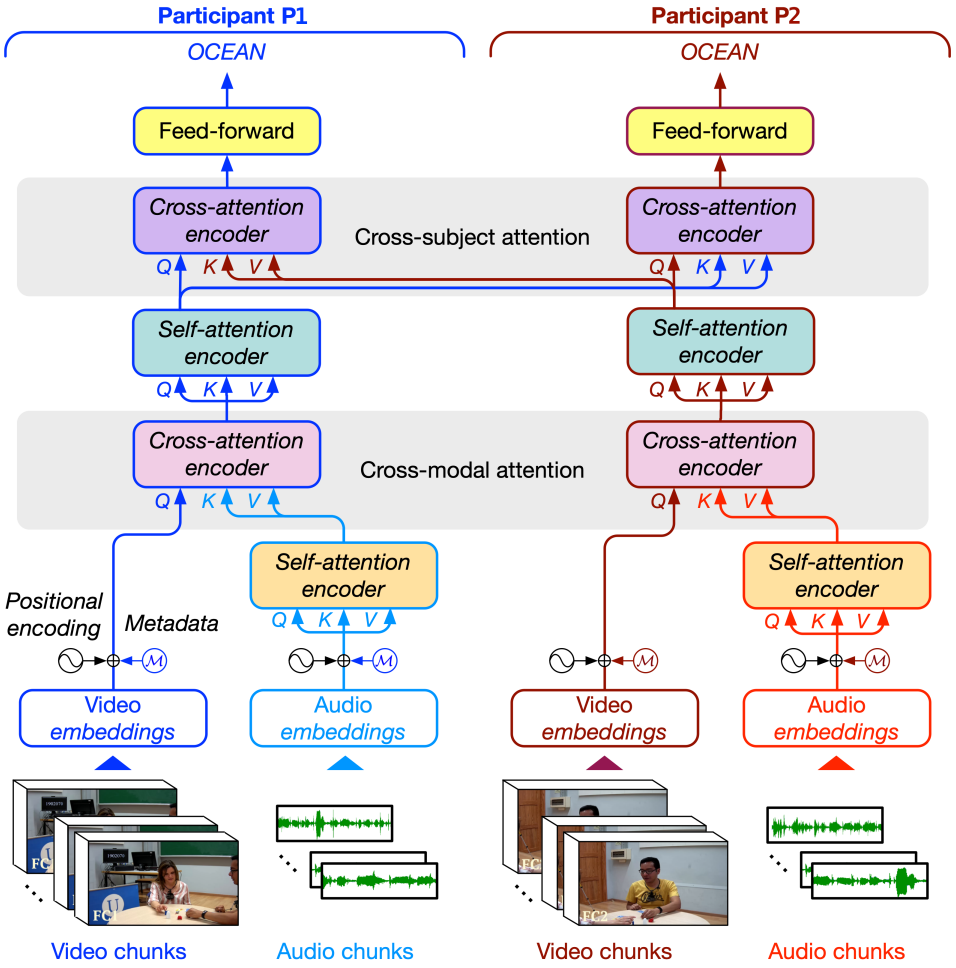


FIGURE 3.3: Proposed Dyadformer including different kinds of attention (self, cross-modal, and cross-subject). The model jointly infers the self-reported personality of both participants (P1 and P2). Model complexity is reduced by sharing weights between parallel encoders (as illustrated by their colors) and across layers within each encoder. $\mathcal{M}$ are the corresponding metadata embeddings of each participant added to both their video and audio embeddings.

video/audio chunks and infers the personality traits for both subjects in a dyadic interaction. It is composed of two main streams, each of which simultaneously processes a single subject.

As discussed in Section 3.1.2, context and interpersonal features are crucial to predict individual features in dyadic and small group interaction scenarios. For this reason, we propose a model which is capable of (a) fusing information from multiple sources (video, audio, and contextual metadata), and (b) allowing per-subject streams to access each other, in order to consider crossed influence during the interaction. To satisfy both, we exploit the cross-attention used in MHCA sub-layers (see Equation (3.4)). For a Transformer focusing on dyadic interactions, the input $\bar{\mathbf{Z}}$ will be from the subject of interest, while the memory $\mathbf{M}$ will be from the other one. The intuition behind this is to allow information from a given subject to *query* for useful information from the other. But first, each stream will create an individual representation for each subject. In order to do so, we draw inspiration from multi-modal Transformer models [468, 180, 173], and use this same cross-attentional mechanism to fuse data coming from video and audio modalities. In this cross-modal module, $\bar{\mathbf{Z}}$ is from the

video modality, while $\mathbf{M}$ is from the audio one, thus enriching video information with the audio signal. Finally, personality scores for both individuals are predicted jointly in a single forward pass.

Input. As done for the baseline, we temporally divide videos and audios into small chunks first. Next, we precompute per-chunk feature representations using pre-trained networks (see Section 3.4.1 for details). Doing so, we can then feed our model with two pairs of sequences $(\dot{\mathbf{X}}^p_{\mathrm{V}}, \dot{\mathbf{X}}^p_{\mathrm{A}})$, where $p \in \{P1, P2\}$ denote the participants, $\dot{\mathbf{X}}^p_{\mathrm{V}} = [\widetilde{\mathbf{X}}^p_{\mathrm{V}_1}, \ldots, \widetilde{\mathbf{X}}^p_{\mathrm{V}_N}]$ is a sequence of precomputed per-chunk video features and $\dot{\mathbf{X}}^p_{\mathrm{A}} = [\widetilde{\mathbf{X}}^p_{\mathrm{A}_1}, \ldots, \widetilde{\mathbf{X}}^p_{\mathrm{A}_N}]$ is the corresponding sequence of precomputed audio features. Note that $\dot{\mathbf{X}}^{P1}_{\mathrm{V}}$, $\dot{\mathbf{X}}^{P1}_{\mathrm{A}}$, $\dot{\mathbf{X}}^{P2}_{\mathrm{V}}$, and $\dot{\mathbf{X}}^{P2}_{\mathrm{A}}$ are all temporally aligned. Apart from these, the model also receives the metadata handcrafted features, namely $\boldsymbol{m}^p$. Then, the precomputed video and audio features, as well as metadata, are linearly projected into d_{m}-dimensional embeddings via three independent linear layers. Next, for each participant, positional encodings and their respective metadata embeddings are summed to their video and audio embeddings. Given $\boldsymbol{m}^p$ has no temporal dimension, before the summation, $\boldsymbol{m}^p$ is replicated N times using the outer product operation: $\boldsymbol{\mathcal{M}}^p = \mathbf{1} \otimes \boldsymbol{m}^p$, where $\mathbf{1}$ is a N-sized vector of ones. Different from the baseline, we here use the positional encodings as defined in the original Transformer [385] (see Section 3.1.1). We only use them to signal temporal position for both video ($\mathbf{E}_{\mathrm{V}} \in \mathbb{R}^{N \times d_{\mathrm{m}}}$) and audio ($\mathbf{E}_{\mathrm{A}} \in \mathbb{R}^{N \times d_{\mathrm{m}}}$). This results in the final features $\mathbf{X}^p_{\mathrm{V}} = \dot{\mathbf{X}}^p_{\mathrm{V}} + \boldsymbol{\mathcal{M}}^p + \mathbf{E}_{\mathrm{V}}$ and $\mathbf{X}^p_{\mathrm{A}} = \dot{\mathbf{X}}^p_{\mathrm{A}} + \boldsymbol{\mathcal{M}}^p + \mathbf{E}_{\mathrm{A}}$, which will be fed as input to the Dyadformer.

Attentional Encoder modules. In our design, we use two main modules to build the complete architecture: the self-attention encoder $\mathrm{Tx}^{\mathrm{SA}}(\mathbf{X})$, which is used to enhance features by attending to themselves, and the cross-attention encoder $\mathrm{Tx}^{\mathrm{CA}}(\mathbf{Z}, \mathbf{M})$, which is used to allow for a set of features to attend to a different source. First, we formalize the sub-layer introduced in Section 3.1.1 as $\mathrm{SubLayer}_{\mathrm{Block}}(x) = \mathrm{LayerNorm}(x + \mathrm{Block}(x))$, where Block can be instantiated by either MHSA (see Equation (3.2)), MHCA (see Equation (3.4)), or PFFN (see Equation (3.3)). Tx modules can be composed of several consecutive layers, and the output from each of them is fed as input to the next. On the one hand, one layer of a self-attention encoder $\mathrm{Tx}^{\mathrm{SA}}(\mathbf{X})$ follows the canonical encoder Transformer and is composed of two consecutive sub-layers: $\bar{\mathbf{X}} = \mathrm{SubLayer}_{\mathrm{MHSA}}(\mathbf{X})$ and $\mathrm{SubLayer}_{\mathrm{PFFN}}(\bar{\mathbf{X}})$. On the other hand, one layer of a cross-attention encoder $\mathrm{Tx}^{\mathrm{CA}}(\mathbf{Z}, \mathbf{M})$ follows the canonical decoder Transformer and is composed of three consecutive sub-layers: $\bar{\mathbf{Z}} = \mathrm{SubLayer}_{\mathrm{MHSA}}(\mathbf{Z})$, $\bar{\bar{\mathbf{Z}}} = \mathrm{SubLayer}_{\mathrm{MHCA}}(\bar{\mathbf{Z}}, \mathbf{M})$, and $\mathrm{SubLayer}_{\mathrm{PFFN}}\left(\bar{\bar{\mathbf{Z}}}\right)$. It is important to note that, when consecutive cross-attention encoder layers are used, $\mathbf{M}$ is the same for all of them, allowing the features in $\mathbf{Z}$ to iteratively attend $\mathbf{M}$ and be progressively augmented.

Cross-modal and cross-subject attention. In order to build the multi-modal representation for each subject, we first feed the audio features $\mathbf{X}^p_{\mathrm{A}}$ to an audio self-attention encoder module $\mathrm{Tx}^{\mathrm{SA}}_{\mathrm{aud}}$ composed by $\mathscr{L}_{\mathrm{aud}}$ layers, such that $\hat{\mathbf{X}}^p_{\mathrm{A}} = \mathrm{Tx}^{\mathrm{SA}}_{\mathrm{aud}}\left(\mathbf{X}^p_{\mathrm{A}}\right)$ where $\hat{\mathbf{X}}^p_{\mathrm{A}} \in \mathbb{R}^{N \times d_{\mathrm{m}}}$. Then, we use a cross-attention encoder $\mathrm{Tx}^{\mathrm{CA}}_{\mathrm{xm}}$ with $\mathscr{L}_{\mathrm{xm}}$ layers to enhance video features $\mathbf{X}^p_{\mathrm{V}}$ with the

new audio features, such that $\hat{\mathbf{X}}^p = \text{Tx}_{\text{xm}}^{\text{CA}}(\mathbf{X}_{\text{V}}^p, \hat{\mathbf{X}}_{\text{A}}^p)$, where $\hat{\mathbf{X}}^p \in \mathbb{R}^{N \times d_{\text{m}}}$ are the multimodal features of subject p.

The audio-enhanced video features of each subject $\hat{\mathbf{X}}^p$ are transformed through a subject encoder $\text{Tx}_{\text{sbj}}^{\text{SA}}$ with $\mathscr{L}_{\text{sbj}}$ layers, such that $\mathbf{S}^p = \text{Tx}_{\text{sbj}}^{\text{SA}}(\hat{\mathbf{X}}^p)$, in order to learn rich relationships within individual subject features. This subject encoder is followed by a cross-attention encoder with $\mathscr{L}_{\text{xs}}$ layers as to allow the features from each subject to draw relevant information from each other, such that $\hat{\mathbf{S}}^{P1} = \text{Tx}_{\text{xs}}^{\text{CA}}(\mathbf{S}^{P1}, \mathbf{S}^{P2})$ and $\hat{\mathbf{S}}^{P2} = \text{Tx}_{\text{xs}}^{\text{CA}}(\mathbf{S}^{P2}, \mathbf{S}^{P1})$, where $\mathbf{S}^p$ and $\hat{\mathbf{S}}^p \in \mathbb{R}^{N \times d_{\text{m}}}$.

Inference. For a given sequence, to infer the personality of the participant p in the dyad, we feed the output subject representations $\hat{\mathbf{S}}^p = \{\hat{\mathbf{s}}_1^p, \ldots, \hat{\mathbf{s}}_N^p\}$ through an average pooling and two FC layers in order to regress the final OCEAN values for p, i.e. $\hat{\mathbf{o}}^p = (\text{ReLU}(\mathbf{y}^p \mathbf{W}_{\text{FC1}})\mathbf{W}_{\text{FC2}})$, where $\mathbf{y}^p = \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{s}}_n^p$, $\mathbf{y}^p \in \mathbb{R}^{d_{\text{m}}}$, $\mathbf{W}_{\text{Y1}} \in \mathbb{R}^{d_{\text{m}} \times 4 * d_{\text{m}}}$ and $\mathbf{W}_{\text{Y2}} \in \mathbb{R}^{4 * d_{\text{m}} \times 5}$.

3.4 Experimental evaluation

Next, we experimentally evaluate a set of variants of the Dyadformer architecture for the task of self-reported personality traits regression and discuss the obtained results. We first describe the experimental setting, then thoroughly ablate the various design choices of the Dyadformer and finally compare against other state-of-the-art models on the UDIVA v0.5 dataset.

3.4.1 Experimental setting

Pre-segmented chunks and feature extraction. For the sake of comparison, we utilize the same set of video and audio chunks of UDIVA v0.5 used for the baseline. As described in Sections 3.2.1 and 3.2.2, chunk availability was limited by the face detection algorithm, such that chunks with no detected face were discarded. Given also the difference in duration throughout sessions and tasks, the final number of chunks per task was uniformly subsampled based on the median. For the Dyadformer we require contiguous sequences of N chunks, resulting in some of them being further discarded. Given these limitations, some tasks do not contain many chunks and, to avoid losing more data, we limited our experiments to $N \leq 12$. As Transformers are known to be data hungry [91], we follow other works [131, 195] in which Transformers have been successfully trained on smaller datasets by leveraging backbones pre-trained on Kinetics [57].

As for the experiments with the baseline, each video chunk is composed of 32 frames at 12.5 fps ($\sim$2.56 seconds), but for the Dyadformer we use a larger spatial resolution of 224×224 pixels (normalized between $[0, 1]$). Audio chunks are 3 seconds long, acquired at 44.1 kHz, and time-synchronized to its respective video chunk (*i.e.*, the centers of corresponding video/audio chunks are aligned). Video, audio, and contextual metadata features are generated for each subject individually. Visual features are computed with R(2+1)D [377] pre-trained on IG-65M [128] and Kinetics [57]. We also fine-tuned its 5th block on the training set of UDIVA v0.5 during 13 epochs (after having replaced the last fully connected layer

by another one of size 5 to predict OCEAN). Once trained, all the pre-segmented chunks of UDIVA v0.5 were reprocessed and the 512-dimensional feature representations output by the second to last layer of R2+1D were saved. Analogously, for audio, we used a VGGish [159] pre-trained on AudioSet [126] to compute a 128-dimensional representation for each audio chunk. Sequences of N such video/audio precomputed features were used as input for each subject in our method.

Parameters and implementation details. Following [88], we fixed $d_{\text{m}} = 768$ and $h = 12$, and hence $d_{\text{k}} = 64$. We set $\mathscr{L}_{\text{aud}} = \mathscr{L}_{\text{xm}}$ and $\mathscr{L}_{\text{sbj}} = \mathscr{L}_{\text{xs}}$ for our experiments. To maximize the number of consecutive N-length training sequences, they were sampled with a stride of 1 chunk. Metadata was included for all the experiments if not otherwise stated (the set of metadata used can be seen in Table 3.1), based on the results from the baseline.

Transformer models quickly grow in number of parameters. In our simplest model (see TF_{v} in Table 3.4) one Transformer layer accounted for ∼7.1M, whereas 8 layers accounted for ∼56.8M parameters (disregarding the backbones and final linear layers). Nevertheless, recent studies on Transformer models in NLP [22, 202], later extended to the audiovisual domain with similar results [205], have shown that weight sharing does not hurt representational power nor performance, while allowing for lighter and faster-to-train models. For this reason, in this work we always shared weights between all equivalent layers of both subject's streams. In other words, both streams were exactly the same. Also, for experiments where layers for any given module $\mathscr{L}_{\cdot} \geq 1$, we shared parameters across them (*e.g.*, all cross-modal Transformer layers share weights).

Our model was trained by minimizing a MSE loss measuring the error of the inferred personality traits at sequence level versus its associated ground truth: $\mathcal{L} = \sum_{p \in \mathcal{P}} \sum_{i=0}^{5} (\mathbf{o}_i^p - \hat{\mathbf{o}}_i^p)^2$, $\mathbf{o}^p$ is the ground truth of self-reported personality and $\mathcal{P} \subseteq \{P1, P2\}$ (depending on the experiment). Model weights were trained by minimizing $\mathcal{L}$ via SGD optimization with weight decay $5e^{-3}$. Training was early stopped after 6 epochs if no improvement was observed on the validation loss. The learning rate was initially set to $5e^{-4}$ and reduced by a factor of 2 after 3 epochs without improvement. The dropout rate throughout all the layers in the architecture was set to 0.2.

Evaluation metrics. For the following experiments, we report the average per-trait MSE at two levels: (a) *sequence-level* (MSE_{seq}), where the error was computed for every N-length sequence by comparing the predictions against ground truth personality of the subject appearing in them. The MSE_{seq} reported is the mean over all the N-length sequences in the test set; and (b) *participant-level* (MSE_{part}), for which we first aggregated the predictions over all the sequences of a given participant by the median, and then compared it to that participant's personality ground truth (akin to how results are reported in Section 3.2.2). In contrast to MSE_{seq}, MSE_{part} removes bias towards participants that appear more in the test set, hence being a more balanced metric for this problem. We choose to report both in this work to compare the effect of the different aggregation mechanisms.

3.4.2 Ablations

Here we include experiments performed to assess the validity of various design choices for the proposed Dyadformer. First, we evaluate an alternative design for the cross-attentional modules, and re-evaluate the usefulness of metadata. Second, we evaluate our two main contributions: (1) the use of multi-modal information and joint modeling of both participants against vanilla self-attention (using only video and one participant at a time); and (2) the inclusion of longer-range temporal context ($N = 6$ and $N = 12$ chunk sequences, corresponding to 15.36 and 30.72 seconds, respectively) with respect to the baseline ($N = 1$, *i.e.*, 2.56 seconds). And finally, we explore the usefulness of the self-attentional modules at different stages of our model. In order to mitigate the stochasticity introduced by the random initialization of the network weights, we repeated each experiment 4 times (or 8 for models with $N = 12$) and report the average of their results.

To assess the cross-attention's contribution we test four variants of our model: (1) a self-attention Transformer (TF_v) on the visual modality only and for each participant separately, *i.e.*, attention is applied within each subject's sequence and neither cross-modal nor cross-subject attention are considered; (2) the Dyadformer with either only cross-modal attention (DF_xm) or (3) cross-subject attention (DF_xs); and (4) the full architecture with both cross-attentions ($\text{DF}_\text{xm,xs}$).

Cross-attention versus bidirectional encoding. Besides CA for both cross-modal and cross-subject interactions, we also tried to follow the approach of bidirectional encoding from BERT [88] (introduced in Section 3.1.2). Instead of using cross-attention, the tokens from the two sequences are concatenated before being fed to canonical Transformer encoder layers[3]. This alternative was implemented through two stages. First, two parallel multi-modal BERT encoders (which share weights among them and within them), each performing video-audio joint attention on its corresponding subjects. Then, their outputs are concatenated and fed to a second stage with one BERT encoder, effectively attending over the two subjects. For a fair comparison with our $\text{DF}_\text{xm, xs}$ with $\mathscr{L}_\text{xm}, \mathscr{L}_\text{xs} \in \{1, 2\}$, we tried with different number of layers for the encoders of this BERT-like architecture such that the number of MHA blocks in both was similar. In particular, BERT with $\mathscr{L}_\text{bm}, \mathscr{L}_\text{bs}$, where $\mathscr{L}_\text{bm}, \mathscr{L}_\text{bs} \in \{3, 6\}$ are, respectively, the number of layers in the multi-modal BERT encoders and the multi-subject one. The BERT configuration $\mathscr{L}_\text{bm} = \mathscr{L}_\text{bs} = 3$ corresponds to the same number of attention layers included in our model with $\mathscr{L}_\text{xm} = L_\text{xs} = 1$ and $\mathscr{L}_\text{bm} = \mathscr{L}_\text{bs} = 6$ corresponds to $\mathscr{L}_\text{xm} = \mathscr{L}_\text{xs} = 2$. Moreover, regardless of the combination of $(\mathscr{L}_\text{bm}, \mathscr{L}_\text{bs})$, the number of parameters of the architecture is 17.1M, which is comparable to either DF_xm or DF_xs (both with 19.4M). We set $N = 12$ for these experiments. We show the results at the bottom of Table 3.4. This variant resulted slightly worse than the equivalent Dyadformer variants ($\text{DF}_\text{xm,xs}$) for all metrics and combinations of layers. These results highlight the effectiveness of the used cross-attentional modules. One possible reason for this to happen is that our cross-attentional design helps

[3]Additional details on various multi-modal fusion mechanisms will be presented in Section 4.3.4. Whilst the Dyadformer employs cross-attention for multi-modal interactions and co-attention for the inter-subject ones, BERT-like approaches use encoder fusion for both.

Arch.	$\mathscr{L}$		MSE_{seq}		MSE_{part}		Params
			$N = 6$	$N = 12$	$N = 6$	$N = 12$	
TF_v	2		0.807	0.771	0.742	0.732	10.0M
	4		0.857	0.792	0.781	0.744	
	6		0.919	0.856	0.837	0.807	
	8		0.948	0.860	0.867	0.804	
	$\mathscr{L}_{\text{xm}}$	$\mathscr{L}_{\text{xs}}$	$N = 6$	$N = 12$	$N = 6$	$N = 12$	
DF_{xm}	1	-	**0.797**	0.767	**0.738**	0.732	19.4M
	2	-	0.845	0.767	0.777	**0.722**	
	3	-	0.880	0.802	0.824	0.762	
DF_{xs}	-	1	0.802	0.768	0.763	0.745	19.4M
	-	2	0.831	0.760	0.778	0.738	
	-	3	0.843	0.767	0.794	0.743	
$\text{DF}_{\text{xm,xs}}$	1	1	0.831	0.760	0.794	0.741	36.0M
	1	2	0.847	0.765	0.802	0.748	
	2	1	0.854	**0.738**	0.809	**0.722**	
	2	2	0.894	0.758	0.842	0.737	
	$\mathscr{L}_{\text{bm}}$	$\mathscr{L}_{\text{bs}}$	$N = 6$	$N = 12$	$N = 6$	$N = 12$	
BERT	3	3	-	0.818	-	0.784	17.1M
	3	6	-	0.820	-	0.780	
	6	3	-	0.814	-	0.766	
	6	6	-	0.800	-	0.761	

TABLE 3.4: Ablation of different architectures and sequence lengths (N chunks) in terms of average sequence- and participant-level mean squared errors: TF_v, a Transformer on each subject's sequence separately; DF_{xm} or DF_{xs}, the Dyadformer with only cross-modal ("xm") or cross-subject ("xs") attention respectively; $\text{DF}_{\text{xm,xs}}$ with both; and BERT, an alternative for multi-modal multi-subject modeling. $\mathscr{L}_\cdot$ are the number of layers in the encoders. Best result per column in bold.

decouple self-attention from accesses to the external memory (through separate MHSA and MHCA operations). The bidirectional encoding, however, emulates accesses to internal and external representations through a single multi-head attention, which may hinder learning to attend differently to one and the other. For these reasons, the BERT variant was discarded and all remaining experiments used the described cross-attentional layers.

Use of metadata. The results obtained with the baseline showed the benefits of using metadata at a marginal computational cost, we here further verify this is also the case for the Dyadformer. As it can be seen in Table 3.5, for the simplest ablated model TF_v ($\mathscr{L} = 2$), that using only video results in very low values for the standard deviation. This *regression to the mean* problem is alleviated by allowing the model to access metadata information. Note that the lack of metadata especially hurts *Extraversion* ("E"), *Agreeableness* ("A"), and *Negative emotionality* ("N"). If we compute the mean of the two sets of standard deviations (with and without metadata, from Table 3.5), we obtain 0.332 versus 0.116, respectively. This indicates the models are more willing to deviate the personality trait predictions from a mean value when incorporating the extra context provided by metadata. This is in line with current state-of-the-art research in personality psychology, which states that personality needs to be expressed in *situations* [321], *i.e.*, taking the interaction context into account. For this reason, all remaining experiments include metadata information.

Cross-modal and cross-subject attentions. As shown in Table 3.4, the two strongest variants were DF_{xm} and $\text{DF}_{\text{xm,xs}}$. Although TF_v was already a strong baseline model, it did not

	O	C	E	A	N
Training (ground truth)	0.255 $\pm$1.136	0.160 $\pm$1.020	$-$0.053 $\pm$0.969	$-$0.006 $\pm$0.957	$-$0.346 $\pm$1.085
TF_v ($\mathscr{L} = 2$) wo/ metadata	$-$0.008 $\pm$0.256	0.057 $\pm$0.112	$-$0.186 $\pm$0.062	$-$0.178 $\pm$0.086	$-$0.431 $\pm$0.064
TF_v ($\mathscr{L} = 2$) w/ metadata	$-$0.053 $\pm$0.323	0.126 $\pm$0.313	$-$0.321 $\pm$0.364	$-$0.134 $\pm$0.345	$-$0.238 $\pm$0.317

TABLE 3.5: Ablation on the regression to the mean problem. Mean and standard deviations of personality trait predictions by one run of the simpler TF_v ($\mathscr{L} = 2$) without and with metadata and the same values over the training ground truth for comparison.

obtain the best result in any metric, suggesting that involving multiple modalities and explicitly modeling interaction among subjects is indeed beneficial for this task. The diminishing trend we observed on the performance of the models when further increasing their depth (number of encoder layers) discouraged us from trying further combinations and/or increasing their capacity with more parameters.

Temporal context. We then evaluated different temporal context lengths, *i.e.*, $N \in \{6, 12\}$, for the aforementioned combinations. As shown in Table 3.4, $N = 12$ achieves better results (lower MSE_{seq} and MSE_{part}) throughout all the ablation. Interestingly, the Dyadformer variants with cross-subject attention, DF_{xs} and $DF_{xm,xs}$, benefited more from longer sequences. This is aligned to the fact that interpersonal dynamics can span very different temporal ranges. That is, the behavior of one interlocutor could be considerably delayed in time. Hence, using $N = 12$ allows such long-term interdependencies to emerge and be further leveraged.

Self-attention before cross-attention. In preliminary experiments, the Dyadformer included self-attention modules before every cross-attention module. However, motivated by the observation of an overfitting trend for overly complex models, we considered discarding all self-attention modules so as to reduce the number of parameters. As a result, for our model in Figure 3.3, we removed the self-attention encoder between the video embedding and the cross-modal encoder, which had no detrimental effect. The self-attention after the audio embeddings was kept to give the audio features a chance to evolve (as video embeddings do during the cross-modal attention), especially given the fact that audio embeddings were extracted from a model not fine-tuned on the personality prediction task – differently from video ones. Regarding the self-attention encoders prior to cross-subject encoders, we experimentally found the impact was negative when removing those layers in our best cross-subject models, *i.e.*, DF_{xs} and $DF_{xm,\,xs}$. Without those layers, MSE_{part} increases from 0.738 and 0.722 (reported in Table 3.4) to, respectively, 0.758 and 0.740. It is for these reasons that we removed the self-attention in the video branch, but kept all other.

3.4.3 Analysis across personality traits and tasks

Here, we analyze the results obtained by the four Dyadformer variants ablated in the previous section. First, we evaluate the results from the four different tasks present in the UDIVA v0.5 dataset, as each of them was designed to elicit different behaviors. Then, we study how different tested variants of the Dyadformer model the different OCEAN traits, given that not all traits are equally expressed nor captured. We compare our results to the two best-performing baseline models "LEm" and "LEam". Note that such models were trained per task, whereas our tested models were trained on all tasks jointly. We first analyze them per-trait and per-task in terms of MSE and later comment on the Pearson correlation results (typically used in personality psychology [39]). For all Dyadformer results reported next, we always use the best combination of N and $\mathscr{L}$. for that particular variant (reported in Table 3.4).

Per-task analysis. As with the baseline, we analyze the performance of the different model variations predicting the OCEAN traits separately depending on the task at hand. The results are shown in Table 3.6. As we can observe, among our models there is not a clear winner regarding MSE metrics. For *Animals*, TF_{v} is the one which provided more accurate results on average ("Avg") both in terms of MSE_{seq} and MSE_{part}, although DF_{xm} did equally well for "A". DF_{xs} outperformed the rest for the "N" trait in this task. Both for *Ghost* and *Lego*, $DF_{\text{xm, xs}}$ and DF_{xm} got the lowest error in terms of MSE_{seq} and MSE_{part}, respectively. Finally, for *Talk*, $DF_{\text{xm, xs}}$ outperformed the rest of the models on average, doing better than the rest for *Open-mindedness* ("O") and *Conscientiousness* ("C") measuring MSE_{seq} and also for "O" and "E" measuring MSE_{part} instead. Some of the findings diverge from the ones we observed with the baseline. For instance, whereas the baseline *Animals* benefited more from audio than *Lego*, we see a contrary trend here. However, note that Dyadformer models are not trained in a task-specific fashion, thus the network has been able to learn from a wider range of behaviors encountered across tasks, which might impact the relative importance of each modality.

Per-trait analysis. Transversely to all tasks except for *Animals*, $DF_{\text{xm, xs}}$ is the most accurate model predicting "O" at participant-level. It is also the best at predicting "E" at participant-level and "C" at sequence-level, whereas DF_{xm} does a better job at participant-level for the latter across all tasks. For "A", $DF_{\text{xm, xs}}$ is a close second after DF_{xs}. Interestingly, for "A", both variants incorporating cross-subject attention improved results."A" is positively correlated with kindness, consideration, and cooperativeness, pro-social behaviors that are more clearly understood when the network attends to both interactants. In contrast, "N" does not usually benefit from cross-subject attention as this trait is more associated to the individual's inner context (*i.e.*, stress, mood changes) [135]. Surprisingly though, we find opposite trends for *Animals*, for which "N" does highly benefit from cross-subject whereas "A" does not.

Per-trait vs. per-task discussion. While, on average, *Talk* is the task obtaining the lowest MSE_{part} error, that is not the case per trait. If we focus on participant-level, the *Talk* scenario does allow to better predict "C", and "E", but *Animals* is more informative for "O" and "A", and *Lego* for "N". At sequence-level, "E" is better predicted with *Lego* and "N" with *Ghost*.

Arch. \ Trait	O	C	E	A	N	Avg
Animals (A)						
Baseline (LEm)	0.736	0.834	0.968	0.669	1.192	0.880
Baseline (LEam)	0.737	**0.756**	**0.887**	0.580	1.023	0.797
TF_v	**0.186**	0.722	**0.659**	**0.049**	1.511	**0.626**
	0.455	1.062	1.283	**0.054**	0.975	**0.766**
	-0.533	0.440	-0.638	**0.894**	0.110	0.055
DF_{xm}	0.206	**0.691**	0.677	0.050	1.658	0.656
	0.515	1.008	1.328	**0.054**	1.041	0.789
	-0.020	0.524	0.458	0.406	0.339	0.342
DF_{xs}	0.242	0.927	0.672	0.123	**1.367**	0.666
	0.628	1.227	1.433	0.134	**0.889**	0.862
	0.267	0.490	0.494	0.353	**0.599**	0.441
$DF_{xm,xs}$	0.263	0.920	0.670	0.115	1.520	0.698
	0.674	1.239	1.448	0.134	0.947	0.888
	0.373	**0.592**	**0.705**	0.341	0.283	**0.459**
Ghost (G)						
Baseline (LEm)	0.743	0.944	0.868	0.657	1.153	0.873
Baseline (LEam)	**0.741**	0.893	0.844	0.667	1.139	0.857
TF_v	1.217	0.609	0.665	0.595	0.783	0.774
	0.858	0.633	0.723	**0.589**	**0.988**	0.758
	-0.535	**0.608**	-0.693	**0.896**	0.137	0.083
DF_{xm}	1.231	**0.563**	**0.629**	0.615	0.778	0.763
	0.889	**0.584**	**0.707**	0.617	0.989	**0.757**
	-0.028	0.565	0.470	0.387	0.343	0.347
DF_{xs}	1.156	0.619	0.778	**0.564**	0.786	0.781
	0.808	0.707	0.781	0.604	1.039	0.788
	0.251	0.517	0.496	0.353	**0.588**	0.441
$DF_{xm,xs}$	**1.122**	0.582	0.733	0.577	**0.775**	**0.758**
	0.771	0.691	0.754	0.616	1.029	0.772
	0.363	0.603	**0.706**	0.334	0.277	**0.457**
Lego (L)						
Baseline (LEm)	**0.727**	0.763	0.826	**0.611**	1.037	0.793
Baseline (LEam)	0.745	0.839	0.953	0.659	1.099	0.859
TF_v	0.925	0.806	0.514	0.614	**0.534**	0.679
	0.808	0.657	0.755	0.710	0.866	0.759
	-0.588	-0.042	-0.741	-0.212	0.193	-0.278
DF_{xm}	0.916	0.753	**0.488**	0.647	0.537	0.668
	0.827	**0.616**	0.743	0.732	**0.844**	**0.752**
	0.103	0.427	0.381	0.382	0.282	0.315
DF_{xs}	0.847	0.801	0.575	0.555	0.567	0.669
	0.749	0.663	0.789	0.709	0.975	0.777
	0.351	0.495	0.512	0.354	**0.511**	0.445
$DF_{xm,xs}$	**0.808**	**0.727**	0.517	**0.527**	0.555	**0.627**
	0.741	0.635	**0.736**	0.747	0.908	0.753
	0.510	**0.580**	**0.714**	**0.388**	0.215	**0.481**
Talk (T)						
Baseline (LEm)	0.825	0.718	0.878	**0.639**	1.047	0.821
Baseline (LEam)	0.773	0.790	0.869	0.670	**0.985**	0.817
TF_v	1.107	0.472	0.561	0.846	1.074	0.812
	0.736	0.513	0.462	0.708	1.076	0.699
	-0.573	0.114	-0.726	-0.020	0.213	-0.198
DF_{xm}	1.117	0.467	**0.526**	0.862	**1.057**	0.806
	0.735	**0.488**	0.440	0.719	1.081	0.693
	0.193	0.452	0.419	**0.404**	0.312	0.356
DF_{xs}	0.896	0.454	0.707	**0.771**	1.095	0.785
	0.632	0.529	0.479	0.671	1.124	0.687
	0.401	0.542	0.529	0.370	**0.525**	0.473
$DF_{xm,xs}$	**0.861**	**0.450**	0.617	0.794	1.082	**0.761**
	0.574	0.504	**0.419**	0.683	1.135	**0.663**
	0.585	**0.597**	**0.743**	0.403	0.229	**0.511**

TABLE 3.6: Results per trait and task. For each model, first row is MSE_{seq}, second row is MSE_{part}, and third row is Pearson Correlation also at participant level (ranging in $[-1, 1]$, closer to 1 is better). The "Avg" column depicts the average performance per row (over all the traits). We also compare to the best baseline results in terms of MSE_{part}. Best result per task, trait, and metric in bold. Also, best result among Dyadformer variants underlined.

These findings are consistent with those from the baseline. This can be useful for psychological research, because it provides evidence that different situations actually enact different traits [116]. Trait-enactment refers to the idea that some situations enact, or activate, certain levels of traits required for this situation [113]. For the case of *Animals*, we can observe a strikingly low error for "A" followed by "O". This suggests that these two traits are likely enacted by this task. This pattern is further confirmed when we look at *Talk*. Extraverted individuals are generally more talkative, but conscientious participants, even though they are not particularly extraverted, will engage in active talking when they are demanded to.

Correlation analysis. We also report the Pearson correlation metric among the per-trait/per-task predictions and the self-reported personality ground truth for the participants in the test partition in Table 3.6. The Pearson correlation represents a normalized measurement of the covariance between the predictions and the ground truth for the test set. By looking at this metric, TF_v displayed the worst average results, mostly correlating negatively with the ground truth. A notable exception is, however, that it obtained the highest correlation (over 0.8) for "A" in *Animals* and *Ghost*. In contrast, it can be observed that all of our Dyadformer variants correlated positively with the ground truth scores (except for DF_{xm} in "O", for which correlation is usually close to zero). DF_{xm} was less accurate for "C", "E" and "N" than DF_{xs} when looking at the Pearson correlation, despite the opposite trend was observed looking at MSE-based metrics. DF_{xs} correlated best with "N", although it showed poor correlation with "A" and "O". $DF_{xm,xs}$ obtained the best "Avg" performance in terms of correlation for all the tasks, followed by DF_{xs}. This shows that explicitly modeling cross-subject interactions helps better approximate the distributions of the traits. The former achieved the highest correlation when predicting "O" and "E", even for *Animals*, where MSE_{part} was very high. More concretely, its strongest correlations were found for the latter trait ($\sim$0.7). $DF_{xm,xs}$ was also the best correlating with "C", except for *Ghost*, where it ranked second. Nevertheless, and opposite to DF_{xs}, it correlated very poorly with "N", while obtaining reasonably good results in "A" for *Lego* and *Talk*. Overall, modeling cross-modal interactions seems to boost results in terms of MSE, both at chunk and participant level, whereas explicit handling of cross-subject interactions seems to provide a better Pearson correlation on average. We hypothesize that multi-modality is guiding the model to make accurate but coarse predictions about individuals by integrating video and audio, and in some way providing relevant cues of their relationship with their environments. Whereas cross-subject attention is providing information about the other participant, aiding with detailed information about that specific part of the context and, in that sense, aiding with refining the prediction. This could be interpreted as multi-modality excelling at modeling the central most populated part of the distribution, while joint modeling of both interlocutors would be helping to better shape it around the periphery. This culminates in the $DF_{xm,xs}$ attaining the best balance across both metrics and establishing itself as the best Dyadformer variant for modeling interaction settings from varied sources of context.

Comparison to the baseline. First of all, it is noteworthy to mention that the results from both the baseline and the Dyadformer confirm that as more sources of context are added,

personality recognition sees improved performance. Apart from this, we can clearly see that the Dyadformer outperforms the baseline by a substantial margin. The best average baseline result is obtained by the "LEam" variant at a MSE_{part} of 0.818 (see Table 3.7). This is largely reduced by the best Dyadformer results by 11.8% (0.722 in Table 3.7). In very few cases (specific task-trait combinations) we find the baseline outperform some Dyadformer variants, but as we discuss in Section 3.4.5 it is expected that different models are better suited for specific traits in specific situations. Nonetheless, despite the baseline being trained separately by task, our different proposed models outperform the two best baseline variations in 15/20 cases, as can be seen in Table 3.6.

The main reason we find for this is the extended temporal context, which we already established in Section 3.4.2. Personality traits are relatively stable over time [79], following specific patterns of change that evolve very gradually, such as the increase of trait "C" throughout life [325]. Yet, when dealing with short temporal segments self-reported personality predictions can be noisy. In fact, within-person variability becomes greater than between-person variability [112], hence aggregation and final personality estimates should be taken with caution. This points towards the possibility that the median from 2.56-second chunks, as employed by the baseline, might be too susceptible to variability. As we have shown with the Dyadformer, the use of longer time windows (up to 30 seconds) capable of capturing long-term interdependencies is very promising to address this problem.

Beyond this, and as we have already discussed, the cross-modal and cross-subject layers are working together to build useful representations that further enhance the performance of the Dyadformer over that of the baseline. On the one hand, the cross-attentional mechanism to fuse video and audio compared to straightforward concatenation provides better chances to enhance visual features with relevant audio cues. On the other hand, the joint modeling of both interlocutors contrasts with the setting of the target participant and additional context. The Dyadformer handles both participants with a shared model, and their respective features are then swapped to be used as context for the other one. Different from the baseline where audio, as well as the video from the other participant and interaction metadata, are bunched together and treated as bulk information, the Dyadformer captures the nuances of the other participant to be used as context, allowing for more nuanced processing of the interaction itself.

3.4.4 Comparison with the state-of-the-art

In [281], we organized a challenge on the UDIVA v0.5 dataset for the task of personality regression. Aside from our baseline, the challenge participants are the only other available results on the UDIVA v0.5 dataset that we can compare against. The scores obtained by the challenge participants for each personality trait and the average scores are reported in Table 3.7. As can be seen, the SMART-SAIR team outperformed the baseline for all the personality traits and was declared the winner of the challenge. However, the FGM Utrecht team can be highlighted for achieving competitive performance by using a straightforward Random Forest regressor trained solely on metadata features (*i.e.*, age, gender, and number of sessions). The table also includes the error obtained by the *mean prediction*, which uses the

Team	MSE error ↓					
	Avg.	**O**	**C**	**E**	**A**	**N**
Mean prediction	0.889	0.725	0.877	0.991	0.673	1.179
Baseline (LEam)	0.818	0.744	0.794	0.886	0.653	1.012
SMART-SAIR	0.769	0.711	0.723	0.867	**0.548**	0.997
FGM Utrecht	0.825	0.752	0.687	0.917	0.671	1.098
TF_v	<u>0.732</u>	0.749	<u>0.653</u>	<u>0.696</u>	0.612	<u>0.949</u>
DF_{xm}	**0.722**	0.763	**0.613**	**0.673**	0.615	**0.944**
DF_{xs}	0.738	<u>0.684</u>	0.687	0.736	<u>0.602</u>	0.982
$DF_{xm,xs}$	**0.722**	**0.646**	0.664	0.699	0.614	0.989

TABLE 3.7: Challenge results compared to Dyadformer on self-reported personality recognition. Per-trait scores are reported by means of Mean Squared Error (MSE) per participant. ↓ indicates the lower the score, the better. *Mean prediction* refers to the performance of a system that returns the average per-trait personality ground truth labels of the training set as the predicted personality. Best results in bold, while second best results are underlined.

mean of the per-trait ground truth personality labels of the training set as the prediction for the individuals on the test set. We observe that both teams outperform this mean prediction on average and for all traits except for "O", for which this mean is only outperformed by the winning team. This metric is useful to quantify how proposed approaches alleviate the regression-to-the-mean effect, which as we have seen, is characteristic of tasks that tackle the inference of attributes with Gaussian-like distributions such as personality. Detailed information about challenge participant's methods can be found in their respective fact sheets: SMART-SAIR[4] and FGM Utrecht[5].

The bottom of Table 3.7 shows how all Dyadformer variants outperform both challenge participants on average. This is also true for "C", "E" and "N". For "O" only the variants including cross-subject attention are capable of surpassing them, while for the "A" trait the SMART-SAIR team outperforms us, leaving our DF_{xs} variant in second place. The winning team uses visual features including facial and body pose landmarks, as well as textual features precomputed from the literal transcriptions (talk-turn duration, content, and sentiment). As we have commented, modeling cross-subject interactions explicitly helps predict "A", and while they model each subject separately, the turn-taking features may have helped them in this regard. Despite this, the SMART-SAIR team was no match for the Dyadformer. Their use of lighter input, compared to raw audio and video, allowed them to use longer contexts but sacrificing fine-grained modeling. Furthermore, their multi-modal fusion is limited to aggregating the prediction of each uni-modal stream. Our model uses raw data as input, allowing it to discover rich fine-grained interactions of multi-modal context while modeling both participants jointly. This highlights that a proper balance between these three factors is key for adequate modeling of self-reported personality in a challenging setting such as interaction.

Finally, it is noteworthy that the Dyadformer outperforms the SMART-SAIR team despite

[4] `https://chalearnlap.cvc.uab.cat/media/results/None/Track-1_top-1_ICCV_Learning_Personalised_Models.pdf`

[5] `https://chalearnlap.cvc.uab.cat/media/results/None/Track_1-Honorable_Mention_Fact_sheet___Challenge.pdf`

them using one model per gender group. Both challenge teams suggested that some correlation between personality and gender exists, which is supported by the literature on the field [407, 246]. Although previous work on personality computing identified the existence of distinct types of gender bias on related datasets [305, 100, 177], most methods found in the personality computing literature are proposing to advance the research on the topic without explicitly taking such correlation into account. The Dyadformer is one of such methods, yet it is still able to outperform both teams. Nonetheless, we believe that further analysis could help assess if the Dyadformer has learned these by-gender differences, which could help explain its improved results.

3.4.5 Discussion

Results have shown that the Dyadformer is better able to model self-reported interactions in dyadic interactions. This is supported by multiple factors. We have seen how the longer temporal windows have a clear effect on prediction, compared to the baseline. The joint modeling of both participants together has proven to be key. This follows observations that understanding the interaction setting provides useful cues to better represent the individuals themselves. Our ablations measured on the Pearson correlation confirm this, as the best results were obtained by $DF_{xm,xs}$, followed on average by DF_{xs}. It can be further seen by the Dyadformer being the only model to explicitly do this on UDIVA v0.5 which has defined a new state-of-the-art. Finding a balance between fine-grained multi-modal information and long-term interactions is key. This is exemplified by comparing our results to the SMART-SAIR team, tied with our ablations showing the DF_{xm} boosting performance. The SMART-SAIR team did use longer temporal contexts, but sacrificing their ability to model fine-grained features, and multi-modality is only used for final agreement. Our multi-modal Transformer provides a better interplay between long contexts and the integration of multiple fine-grained cues, as shown by the results.

Several other lessons can be drawn from these analyses. For instance, challenge participants have shown that vocal behavior (both speech features and transcripts) provides rich cues for personality recognition, which is aligned with previous work on personality perception [142]. While the Dyadformer has shown to be competitive without these, we deem these features to be a potential venue for future research. Results also reveal a recurrent and consistent finding in personality computing [176] that there is no single model that works best for all possible settings and personality traits, suggesting that different traits can be better modeled and recognized by distinct feature representations and methodologies. This can also be observed in our per-trait analysis provided in Section 3.4.3, where we saw some variants outperform others for specific traits.

A final aspect that can be derived from these results is some potential limitations of UDIVA v0.5. On the one hand, it is an unbalanced dataset, at least in some regards. For instance, although the dataset is relatively balanced with respect to the gender attribute, it is extremely unbalanced regarding the spoken language. Besides that, and as mentioned in Section 2.4.2, the dataset is biased towards people with high levels of the *Open-mindedness* trait due to the recruitment procedure. This has been made explicit by the FGM Utrecht team, who

used the number of sessions as one of the three metadata fed into their Random Forest. They suggested that the number of sessions could be correlated with the *Open-mindedness* trait, considering the time investment and social energy required to participate in these sessions. On the other hand, during the experimental phase of both the baseline and the Dyadformer, we encountered severe overfitting issues that were hard to overcome. We hypothesize this may be related to another limitation of the UDIVA and UVIDA v0.5 dataset: the number of subjects, *i.e.*, 147 and 134, respectively. While being one of the face-to-face interaction datasets with the highest number of participants (see Section 2.1.3), correlations are known to achieve their point of stability around 160 subjects for typical scenarios in personality psychology, achieving full stability at 250 [335]. From a psychology standpoint, this means that our findings reveal trends that will not radically change but can vary in terms of size, not in terms of direction of correlation. Thus, the experimental findings of the Dyadformer are reliable. Note that, this limitation refers to the number of participants in the context of regressing personality scores, and is not representative of other features where UDIVA still excels, such as the number of hours of recordings, which could prove useful in setting such as self-supervised learning, or the variety of situations. In this sense, from a machine learning point of view, this limitation may influence supervised training for personality-related tasks, and specially in the case of Transformers, result in overfitting due to the limited number of distinct subjetcs. If we aim to infer personality from overt behavior, the need for further data that covers a wider spectrum of behavior combinations significantly increases. Consequently, our study can motivate the research and design of novel datasets and annotation protocols, in particular large-scale datasets aiming for greater variety of participants for shared, social-related downstream tasks.

3.5 Conclusion

In this chapter, we have started probing the possibilities offered by the UDIVA dataset as a complex scenario of human interaction. In doing so, our exploration also aimed to uncover the abilities of Video Transformers to model such intricate settings, especially for the nuanced and complext task of personality regression.

We have first validated our hypothesis that Transformers can harness multiple contextual cues from the challenging interaction setting offered by UDIVA by adapting an existing Video Transformer. We incrementally combined different sources of context (both interlocutors' scene, acoustic, and task information) finding consistent improvements as they were added, which is consonant with human interaction research in the psychology field. These results already suggest that Transformers are indeed capable of handling intricate environments while retaining semantically meaningful cues. Nonetheless, this method exhibited still some limitations that we tackle with our proposed Dyadformer.

The Dyadformer is a multi-modal multi-subject Transformer for modeling individual and interpersonal features in dyadic interactions with the flexibility to accommodate different time windows, thus allowing the capture of long-term interdependencies. We thoroughly ablate our model in the UDIVA v0.5 dataset for the task of self-reported personality prediction to

demonstrate the contributions of each attentional module, as well as the modeling of longer timesteps. Experimental results demonstrated the reliability of our approach by surpassing previous results in UDIVA v0.5, reducing the error by 11.8% with respect to the baseline and 6.1% with respect to the challenge winner. Results also showed that context (or situations) matters in personality computing. Recently, situations have been put at the forefront of personality research to understand and predict real behavior [321]. In this sense, a promising extension of this work into the psychological realm would be to extract situational perceptions as we compute personality scores, since considering both features would undoubtedly improve behavior understanding.

Finally, we have seen some limitations of UDIVA for personality modeling, opening the door to many future research developments of novel datasets and bias mitigation mechanisms. Nevertheless, the modeling of context-aware long-term interdependencies is challenging in itself, particularly in the case of multi-modal scenarios. In this sense, we observed strong overfitting, which we partially attribute to the reduced number of participants of UDIVA, but also to an inherent feature of Transformers, which seem to be data hungry [91]. This demands a better understanding of architectures and training strategies for modeling human behavior. Especially for addressing the challenges of training Transformers on small to medium datasets, which highlights the need further research on this novel architecture. To fully harness their capabilities, we require a more profound comprehension of this architecture and a deeper analysis of its inner workings. This involves unraveling the architectural and training needs of successful spatio-temporal and multi-modal representation learning within the Transformer family. It is for this reason that in the next chapter, we embark on an in-depth analysis of Video Transformers, in the hopes of gaining the necessary insights to exploit their full potential.

Chapter 4

Video Transformers: A Comprehensive Survey

4.1 Introduction

As we saw in Section 3.1.1, the Transformers [385] are a recent family of models that process sequencial data in a parallel fashion. The two key features that make Transformers special are their non-local core operation and their lack of inductive biases. On the one hand, this lack of inductive biases makes Transformers very versatile, as seen by the quick adoption for modeling many data types [88, 299, 136, 91, 54], including videos [31, 131, 17, 236, 468, 215]. However, and as we have seen with the Dyadformer in Chapter 3, Transformers have a tendency to overfit. As concurrent work has demonstrated for images [91], the lack of inductive biases makes Transformers require large amounts of data. In our particular case we hipothesize this may be related with the limited number of participants, and not the amount of video hours. On the other hand, the use of non-local token mixing causes Transformers to scale quadratically with sequence length N (*i.e.*, $\mathcal{O}(N^2)$, due to the pair-wise affinity computation in Equation (3.1)). Moreover, the video domain further introduces its own challenges, namely a large increase in dimensionality, linked with a high level of information redundancy, and the need to model motion dynamics. This is aggravated when tackling tasks such as personality recognition in interaction, which also demand for long-term modeling. As we will see, Transformers will require several modifications to adapt to the highly redundant spatio-temporal structure of video.

Transformers are still in its infancy, and despite the many claims of their abilities to form semantic representations of the input, and to integrate multi-modal cues into a cohesive representation, much is yet to be understood. Furthermore, the recent surge in *Video Transformer* (VT) works makes it convoluted to keep track of the latest advances and trends. Existing surveys focus on design choices for Transformers in general [226], NLP [187], images [428, 232], or efficient designs [369, 114]. Given the sequential nature of video, as well as the large dimensionality and redundancy introduced by the temporal dimension, directly adopting image-based solutions or NLP-based designs for long-term modeling will not suffice. While other existing surveys include video, they are limited to superficial comments of a few VTs in the broader context of vision Transformers [190, 145, 435], techniques to integrate visual data with other modalities [344, 426], or video-language pre-training [330]. In this sense, they miss

an in-depth analysis that properly captures the challenges of modeling raw image sequences or highly redundant spatiotemporal visual features through Transformers.

In this chapter, we comprehensively survey and analyse advances and limitations of Transformers when considering the particularities of modeling video data. To do so, we review over 100 VT works and provide detailed taxonomies of the various design choices throughout the VT pipeline (namely input, architecture, and training). Finally, we extensively compare performance on the task of video classification based on self-reported results from the state-of-the-art on Kinetics 400 [57] and Something-Something-v2 [241]. With these in depth analysis we aim at better understanding Transformers as a tool, in order to be able to harness their potential to model complex scenarios, such as human interaction.

For an introduction of the original Transformer, we kindly refer the reader to Section 3.1.1. The structure of this chapter is as follows: in Section 4.2 we explore how videos are handled prior to the Transformer; Section 4.3 describes architectural design adaptations to video; Section 4.4 investigates common training strategies; Section 4.5 outlines trends for specific application scenarios; Section 4.6 discusses VTs performance on action classification; and in Section 4.7 we discuss the main trends and limitations. For an extensive list of all VT works surveyed, and details on how each section in this chapter relates to a given work, see Tables 4.1 and 4.2[1].

4.2 Input pre-processing

Here, we review how video is processed before being input to the Transformer. This involves tokenization, embedding, and positioning (see Figure 4.1). Note that, in the context of video, embedding often comes before tokenization: a separate network embeds the raw data to a continuous and compact representation, which can be used directly as a token or be further tokenized into more atomic units.

4.2.1 Embedding

In order to embed video, we find VTs following two main trends: *embedding networks*[2] or *minimal embeddings*. The key difference between the two is size: while minimal embeddings are generally limited to single linear layers, large embedding networks are instantiated as full CNN architectures. Furthermore, while minimal embeddings follow the classic tokenization-then-embedding approach, full embedding networks can be used to embed full input sequences for later tokenization. In the context of video, embedding layers also function as a crucial dimensionality reduction mechanism.

Embedding network. Leveraging an embedding network (such as a CNN), can potentially ease the learning of the Transformer by providing strong initial features thanks to locality

[1] Note that due to its length, the table has been split into two subtables

[2] Note that, in previous chapters, we have refered to embedding networks as "backbones". In this chapter we distinguish between "embedding networks" (the topic of the current section) and Transformer backbones (the core architecture, see Table 4.1)

	Name	Ref.	Yr.	Architecture				Input				Train.
				Arch.	Aggr.	Restr.	Long-t.	Backbone	Embedding Network	Tknz.	Pos.	SSL
Classification	TimeSformer	[31]	'21	E	-	LAS	-	-	Minimal Embedding	P	LA	-
	PE	[205]	'21	E	-	-	-	-	SlowFast[110], RN-50[152]	C	LA	P
	CBT	[360]	'19	E	-	-	-	-	S3D[423]	C	-	P
	ViViT	[17]	'21	E	H	A	-	ViT [91]	Minimal Embedding	P	LA	-
	ELR	[308]	'19	E	-	-	-	-	I3D[57]	P	-	-
	FAST	[441]	'21	E	-	-	-	-	Minimal Embedding	P	LA	-
	VATNet	[131]	'19	E	Q	-	-	-	I3D[57], Faster R-CNN (RP only)[323]	P + I	FA	-
	VATT	[7]	'21	E	-	S	-	-	Minimal Embedding	P	LA	P
	MViT	[106]	'21	E	H	-	-	-	Minimal Embedding	P	LA	-
	SCT	[451]	'21	E	H	L	-	-	Minimal Embedding	P	LA	-
	CATE	[362]	'21	E	-	-	-	-	SlowFast[110] (Slow br.)	C	-	P
	LapFormer	[199]	'20	E	-	-	-	-	RN-50[152]	P	FA	-
	TRX	[294]	'21	E	-	-	-	-	RN-50[152]	F	FA	-
	LTT	[186]	'20	E	-	-	-	-	R(2+1)D[377]	F	LA	-
	Actor-T	[122]	'20	E	-	-	-	-	I3D[57], HRNet[395]	I	FA	-
	STiCA	[290]	'21	E	-	-	-	-	R(2+1)D-18[377], RN-9[152]	F	LA	A
	GroupFormer	[217]	'21	ED	Q	L	-	-	I3D[57]	I + F	LA	-
	Video Swin	[236]	'21	E	H	L	-	-	Minimal Embedding	P	LR	-
	VTN	[268]	'21	E	H	L	-	ViT [91]	Minimal Embedding	P	LA	-
	Video-Swin-V2	[234]	'22	E	H	L	-	-	RN-50[152]	P	LR	P
	MTV	[430]	'22	E	H	-	-	ViT [91]	Minimal Embedding	P	LA	-
	Motionformer	[288]	'21	E	-	-	-	-	Minimal Embedding	P	LA	-
	X-ViT	[47]	'21	E	-	L	-	ViT [91]	Minimal Embedding	P	LA	-
	ObjTr	[416]	'21	E	-	-	-	-	Faster R-CNN [323], RN-101 [152]	I	FA + LA	-
	MViTv2	[219]	'22	E	H	-	-	-	Minimal Embedding	P	LR	-
	MaskFeat	[406]	'22	E	H	-	-	MViT [219]	Minimal Embedding	P	LR	P
	LSTCL	[396]	'22	E	-	-	-	Swin [236]	Minimal Embedding	P	LA	P
	RViT	[432]	'22	E	-	-	R	ViT [91]	Minimal Embedding	P	LA	-
	Direcformer	[378]	'22	E	-	A	-	TimeSformer [31]	Minimal Embedding	P	LA	-
	VideoMAE	[374]	'22	E	-	S*	-	ViT [91]	Minimal Embedding	P	LA	P
	BEVT	[398]	'22	E	H	L	-	Swin [236]	Minimal Embedding	P	LA	P
	TIME	[446]	'22	E	-	-	-	Motionformer [288]	Minimal Embedding	P	LA	A
	TokenLearner	[331]	'21	E	H	-	-	ViT [91]	Minimal Embedding	P	LA	-
	SVT	[319]	'22	E	-	A	-	TimeSformer [31]	Minimal Embedding	P	LA	P
	UniFormer	[214]	'22	E	H	L	-	-	Minimal Embedding	P	LA + LR *	-
Captioning	ActBERT	[468]	'20	E	-	-	-	-	R(2+1)D[377], Faster R-CNN [323]	I + C	LA	P
	HERO	[215]	'20	E	H	-	-	-	RN-101[152], SlowFast[110]	F	FA	P
	MART	[207]	'20	ED	-	-	R	-	RN-200[152], BN Inception[175]	F	FR	-
	VideoBERT	[361]	'19	E	-	-	-	-	S3D[423]	C	LA	P
	E2E-DC	[467]	'19	ED	-	-	-	-	RN-200[152], BN Inception[175]	F	FA	-
	BMT	[173]	'20	ED	-	-	-	-	I3D[57]	F	FA	-
	AMT	[443]	'21	ED	-	-	-	-	RN-200[152], BN-Inception[175]	F	FA	-
	MDVC	[174]	'20	ED	-	-	-	-	I3D[57]	F	FA	-
	RLM	[220]	'20	D	-	-	-	-	I3D[57]	C	FA	-
Retrieval	HiT	[231]	'21	E	-	-	-	-	S3D[423], SENet-154[166]	F + C	LA	T
	COOT	[130]	'20	E	H	-	-	-	RN-152[152]; ResNext-101[422]; I3D[57]	F	-	T
	MMT	[118]	'20	E	-	-	-	-	S3D[423], DenseNet-101[169], RN-50[152], SENet-154[166]	P + F	FA	T
	Support-set	[289]	'21	E	-	-	-	-	RN-152[152], R(2+1)D-34	F	-	T
	TCA	[340]	'21	E	-	-	-	-	iMAC[137], L-3-iRMAC[200]	F	-	T
	MDMMT	[96]	'21	E	-	-	-	-	CLIP[312]	F	LA	T
	Fast and Slow	[259]	'21	D	-	-	-	-	TSM RN-50[223]	P	-	T
	ClipBERT	[206]	'21	E	-	S*	-	-	RN-50[152]	P	LA	-
	CACL	[144]	'22	E	-	-	-	-	RN-50[152]	F	LA	P
Tracking	Hopper	[466]	'21	ED	-	-	-	-	ResNeXt-101[422], DETR[54]	I + F	LA	-
	DTT	[440]	'21	ED	-	-	-	-	RN-50[152]	P	LA	-
	TrDIMP	[397]	'21	ED	-	-	-	-	RN-50[152]	P	-	-
	TransT	[68]	'21	E	-	-	-	-	RN-50[152]	P	FA	-
	STARK	[429]	'21	ED	-	-	-	-	RN-50[152]	P	FA	-
	Trackformer	[254]	'22	ED	Q	-	MR	-	RN-50[152]	P	FA	-
	VDRFormer	[464]	'22	ED	Q	-	MR	-	RN-101[152]	P	FA	-

*: Non-attentional sparsity (e.g., input level).

TABLE 4.1: General overview of relevant Video Transformers surveyed. In *Architecture*, "Arch.": architecture, that is Encoder (E), Decoder (D), or Encoder-Decoder (ED); "Aggr.", aggregation strategy, either Hierarchical (H) or Query-driven compression (Q); "Restriction", can be Local (L), Axial (A), Sparse (S), or a mix. "Long-t.": long-term temporal modeling, Memory (M), Recurrence (R), or a both. In *Input*, "Backbone" refers to Transformer backbone; "Tknz', the tokenization strategy, patch- (P), instance- (I), frame- (F), or clip-wise (C); and "Pos.", the positional embedding, can be Fixed Absolute (FA), Fixed Relative (FR), Learned Absolute (LA), Learned Relative (LR), or a combination. (Continuation in Table 4.2)

	Name	Ref.	Yr.	Architecture				Input				Train.
				Arch.	Aggr.	Restr.	Long-t.	Backbone	Embedding Network	Tknz.	Pos.	SSL
Low-level	ET-Net	[409]	'21	ED	-	-	-	-	ConvLSTM[343]	P	FA	T
	STTN	[450]	'20	ED	-	-	-	-	2D CNN (custom)	P	-	T
	FuseFormer	[230]	'21	ED	-	-	-	-	I3D[57]	P	-	T
	SAVM	[408]	'20	ED	-	L	-	-	Minimal Embeddings	P	LR	T
	VLT	[316]	'20	ED	-	-	-	-	VQ-VAE[274]	P	FR	T
	TransformerFusion	[41]	'21	E	-	S*	M	-	RN-18[152]	F	LA	-
Segmentation	VisTR	[405]	'21	ED	-	-	-	-	RN-50[152]	P	FA	-
	MFN	[404]	'21	E	-	-	-	-	3D CNN (custom)	P	FA	-
	CMSANet	[437]	'21	E	-	-	-	-	DeepLab-101[64]	P	FA	-
	IFC	[172]	'22	ED	Q	-	-	-	RN-101[152]	P	FA	-
	TeViT	[434]	'22	ED	Q	-	-	MsgShifT [434, 399]	Minimal Embedding	P	FA	-
	AOT	[436]	'21	E	-	L	MR	Swin [236]	MobileNet-V2[332]	P	FA + RL	-
O.D.	PCSA	[141]	'20	E	-	L	-	-	MobileNet-V3[163]	P	-	-
	TCTR	[445]	'21	ED	-	-	-	-	RN-50[152]	P	FA	-
	PMPNet	[439]	'20	ED	-	-	-	-	GraphCNN (custom)	P	-	-
	ORViT	[160]	'22	ED	-	-	-	-	Faster R-CNN[323], RN-50 [152]	P + I	LR	-
Summ.	H-MAN	[233]	'19	E	-	-	-	-	VAE-GAN[240]	F	-	-
	VasNet	[105]	'19	E	-	-	-	-	GoogLeNet[366]	F	FA	-
	BiDAVS	[225]	'20	E	-	-	-	-	GoogLeNet[366]	F	LR	-
	VMTN	[337]	'19	E	Q	-	-	-	ResNet-18[152], SENet-101[166]	P	FA	-
Localiz.	HISAN	[304]	'19	E	-	-	-	-	Faster R-CNN[323]	I + F	-	-
	STVGBert	[357]	'21	E	Q	-	-	-	RN-101[152]	P	-	-
	MeMViT	[417]	'22	E	H	-	M	MViTv2 [219]	Minimal Embedding	P	LR	-
	MSAT	[456]	'21	E	-	-	-	-	C3D[375]	C	FA	-
	RTD-Net	[367]	'21	D	-	-	-	-	I3D[57]	F	LR	-
	LSTR	[425]	'21	ED	Q	-	M	-	RN-50[152]	F	FA	-
Others	SiaSamRea	[442]	'21	E	-	S*	-	ClipBERT [206]	RN-50[152]	P	LA	A
	Perceiver	[178]	'21	E	Q	-	-	-	Minimal Embedding	P	LA	-
	AVT	[132]	'21	E	H	-	-	ViT [91]	Minimal Embedding	P	LA	A
	OadTR	[400]	'21	ED	-	-	-	-	RN-200[152], BN-Inception[175]	F	LA	-
	STTran	[77]	'21	ED	-	L	-	-	RN-101 F R-CNN[323]	I + F	LA	-
	E.T.	[286]	'21	E	-	-	-	-	Faster R-CNN[323], Mask R-CNN[151]	F	FA	-
	SMT	[107]	'19	ED	Q	S*	M	-	RN-18[152]	F	FA	-
	JSLT	[53]	'20	ED	-	-	-	-	InceptionV4[365]	F	FA	-
	MSLT	[52]	'20	ED	-	-	-	-	InceptionV4[365]	F	FA	-
	SBL	[239]	'20	ED	-	-	-	-	RN-18[152]	F	-	-
	MDAM	[194]	'19	E	Q	-	-	-	RN-152[152]	F	FA	-
	PSAC	[460]	'21	E	-	-	-	-	Minimal Embedding	P	FA	-
	BTH	[218]	'21	E	-	-	-	-	VGG-16[346]	F	FA	P
	BERT4SessRec	[70]	'20	E	-	-	-	-	GoogLeNet[366]	C	FA	P
	Dyadformer	[80]	'21	E	-	-	-	-	R(2+1)D-152[377]	C	FA	-
	MM-Transformer	[329]	'22	ED	-	L	-	-	Mask R-CNN [151]	I	FA	-

*: Non-attentional sparsity (e.g., input level)

TABLE 4.2: (Continuation of Table 4.1)

inductive biases. We can roughly categorize the choice of embedding network by the types of relationships they encode into spatial and spatiotemporal. Within *spatial embeddings*, we find 2D CNN networks, typically ResNet variants [152, 422], pre-trained on large image corpora (most commonly ImageNet [86, 324]) to learn general filters that can extract meaningful representations of individual frames. This has been shown to work effectively in the context of video [207, 308, 158, 141, 199, 183, 220, 230]. However, 2D convolutions lack the ability to model temporal information. For this reason, we also find the use of *spatiotemporal embedding* networks (*e.g.*, in [401, 308, 360, 131, 220]). These are generally instantiated as 3D CNNs (such as I3D [57] and S3D [423]), commonly pre-trained on large video datasets, such as Kinetics [57, 56] or HowTo100M [261], to produce features involving temporal relationships. Alternatively, LSTMs [233] or a hybrid ConvLSTM [343, 404, 409], are also leveraged to embed local temporal information. While spatial embeddings produced by spatial embedding networks are limited to per-token spatial interactions, spatiotemporal counterparts help provide initial locally-based temporal interactions.

Minimal embeddings. Inspired by the success of ViT [91], some works [17, 230, 31, 441, 178, 91] omit deep embedding networks and subdivide the input (*i.e.*, tokenize) and then perform embedding with only few linear projections or convolutions. In this sense, they are guaranteed to not share information between tokens, leaving the learning of interactions between them entirely to the Transformer. Empirical studies like [31, 178], show that *standalone Transformers* (*i.e.*, without complex CNN embedding networks) are as performant as CNN counterparts, although the resulting model becomes data-hungry and computationally expensive. Given that, training and deploying VTs with minimal embeddings may benefit from architectural modifications inducing necessary biases (see Section 4.3).

4.2.2 Tokenization

When dividing a video into smaller tokens to form the input sequence to the Transformer, we find several categories depending on the token input receptive field (*i.e.*, the extent of

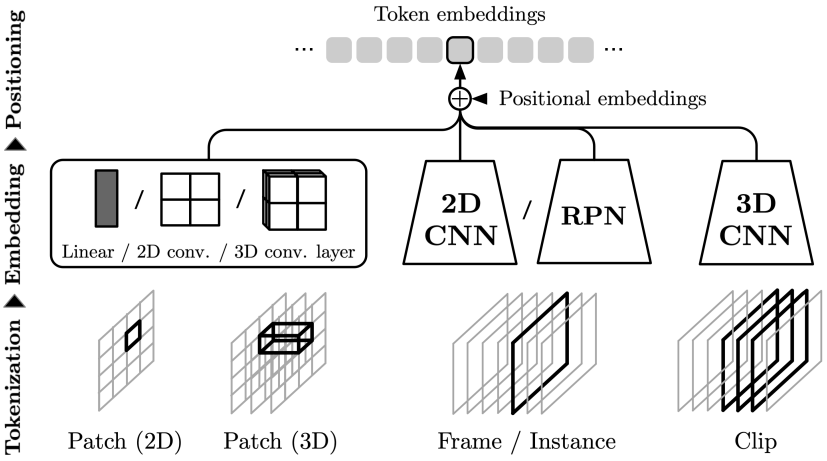


FIGURE 4.1: Overview of the input-preprocessing step, showing tokenization, embedding strategies, and positioning (inclusion of positional information).

the original input covered by a given token before being processed by the Transformer). We distinguish between patch, instance, frame, and clip tokenization (see Figure 4.1).

Patch-wise tokenization. Most VTs follow ViT [91] and employ a 2D-based patch tokenization [460, 31, 441, 450], dividing the input video frames into regions of fixed spatial size [31, 441, 460] or even multi-scale patch sizes [450]. Others propose using 3D patches instead [236, 17, 7, 106, 374] (also regarded as *cubes*), allowing to consider local motion features within the tokens themselves. While non-overlapping patches are the most common, a few works propose using overlapping 2D [230] or 3D [106] patches for smoother information flow between neighbouring patches. Due to their access to neighbouring information in the input, we also regard positions of intermediate feature maps from CNN embedding networks as patches (*e.g.*, 2D in [230, 357, 397, 405] or 3D in [122, 308]), as their exact receptive field will depend on the specific setting in which they are produced. Overall, patch-based tokenization provides the finer granularity, allowing to properly model spatiotemporal interactions in the VT.

Instance-wise tokenization. We refer to instances as semantically meaningful (foreground) regions that extend their reach beyond small patches but still smaller than whole frames [416, 286, 468, 131]. On the one hand, a *Region Proposal Network* (RPN in Figure 4.1), such as a Faster R-CNN [323], can be used to generate region proposals and their corresponding embeddings [416]. Thus, they allow to reason about foreground objects or region interactions. Alternatively, in [468, 131, 217], this kind of tokenization is combined with other coarser tokenizations (frame- and clip-wise tokenization) allowing to form instance-context relationships. Instance-based tokenization can be regarded as a form of sparse sampling (*e.g.*, [160, 329]), potentially reducing redundancy and allowing to input relatively large temporal sequences of per-frame instance representations to the VT without running into efficiency limitations.

Frame-wise tokenization. In this case, the learning of per-frame local spatial interactions is entirely left to the embedding network, while the Transformer focuses on modeling the temporal interactions among the resulting frame tokens (*e.g.*, [289, 450, 286, 215, 467, 443, 400, 294]). This allows longer videos to be modeled, as the input sequence length will be shorter for a given input video, specially compared to patch tokenization. Nevertheless, the Transformer may have a hard time modeling fine-grained spatial interactions. Still, some tasks focusing on frame-level predictions (such as video summarization [105]) may not require them.

Clip-wise tokenization. Condensing the information of several frames (clip) into each individual token allows further reducing the input sequence length (*e.g.*, in [118, 468, 205, 361, 360, 130]). This way, the Transformer can effectively consume more frames to cover longer temporal spans. This makes clip tokenization very suitable for long-term modeling tasks. Given the high dimensionality of clips, it is necessary to embed them into single token representations through large embedding networks: for instance [456] with C3D, [468] with 3D ResNet-50,

[361] with S3D, [186] with R(2+1)D, or [205] with SlowFast, to mention a few. This tokenization could also be suitable for retrieval tasks, where a high-level representation of the video is required [118, 468]. Clip-based tokenization exacerbates pros and cons of frame-based tokenization where fine-grained information may be lost or mixed, preventing the Transformer from disentangling it later, in favor of efficiency when handling longer videos.

4.2.3 Positional Embeddings (PE)

Given that SA is an operation on sets, signaling positional information is necessary in order to exploit the spatiotemporal structure of videos. This is done via positional embeddings (PE), which can be either *fixed* or *learned* and then *absolute* or *relative*: fixed absolute [404, 131, 106], learned absolute [178, 468, 205], fixed relative [207, 316], or learned relative [225, 408, 236]. Absolute variants are summed to the input embeddings but can also be concatenated [405, 178, 437], while for the relative ones the positional information is introduced directly in the multi-head attention [418].

Absolute. These positional embeddings are generally 1D. This naturally fits frame or clip tokenization to indicate position in the only remaining (temporal) dimension. However, when dealing with patch-wise tokenization, fixed 1D in raster order may seem counter-intuitive, as the last patch i-th from row j, will be regarded as closer to the first patch in the next row $j+1$, than to patch i at row $j-1$ (or $j+1$). For this reason, 2D absolute PE [132, 122] accounting for joint space wh and time t dimensions, and 3D absolute PE [178, 441, 404, 405] for width w, height h, and t have also been proposed, disregarding [91] who found 1D learned absolute PE to suffice – at least for images.

Relative. The idea behind relative PEs is that the positional information added when computing attention between token i and j depends on their relative position, making them translation equivariant. In other words, 1D relative PE added when computing attention between token at position i and $j = i + k$ will be the same regardless of the value for i (*i.e.*, $-k$). Relative PEs are generally added as an additional bias term (as in [207, 342, 82, 234]) in the dot-product between $\mathbf{Q}$ and $\mathbf{K}$ (modifying Equation (3.1)). We find different variants of relative PEs applied to VTs, for instance [408, 236, 367] are based on decomposable attention [283], whereas [225] follows the approach of relation-aware attention [342]. Still, to the best of our knowledge, there is yet no study providing insights on their functional differences and effects on performance.

4.2.4 Discussion on input pre-processing

Most VTs employ large CNN embeddings to reduce input dimensionality (aiding with data redundancy) and to exploit their ability to produce strong representations (thanks to local inductive biases). This significantly alleviates complexity and simplifies training when employing Transformers for video tasks. The success of these methods is clearly visible by the

amount of works which utilize large embedding networks as opposed to minimal embeddings (see Table 4.1). While minimal embeddings are indeed lighter than large CNN counterparts, they do result in overall more costly models if used naively. As they do not provide the necessary inductive biases, these will have to be provided elsewhere (such as in the Transformer design – see Section 4.3 –, or during training, through large-scale (self-)supervised pre-training – see Section 4.4). Regarding tokenization, it has an impact on two main factors: (1) it will affect the granularity at which information is modeled by the VT (longer temporal spans by using frame- or clip-based tokenization, and more fine-grained spatiotemporal modeling when employing patches); (2) it will impact the input sequence length, and consequently the computational complexity of the model. For these reasons, most works use a patch-based approach accompanied by some efficient design, or frame-based tokenization, as it provides better long-term modeling scalability.

We find that the interactions between embedding and tokenization play a crucial role in defining the abstraction level and granularity at which the Transformer can model interactions. On the one hand, large embedding networks allow to produce tokens sharing information between them, guided by interactions defined by the CNN's inductive biases. In this regard, it may be desirable to leverage 3D CNNs that provide local interactions among spatiotemporally neighboring positions. On the other hand, some tokenization strategies (such as 3D patches or clips) allow the formation of fine-grained temporal interactions within the token itself. This can be further motivated by most state-of-the-art VTs employing 3D patches. In this sense, the choices of embedding network and tokenization need to be carefully considered, as they will affect the level at which spatial and temporal interactions can be formed.

Finally, the fixed absolute PEs proposed in [385] require less parameters than the learned counterpart. However, the latter could be learning relevant positional relations that Fourier-like approaches are unable to capture (similarly to how learned convolutional filters replaced the handcrafted ones). The vast majority of VTs employ these absolute variants while the use of relative counterparts is still marginal (see Table 4.1). We believe, however, that the translation equivariance the latter provide could prove beneficial for generalizing to unseen lengths (see Section 4.7). This is highly valuable for the video domain as it is much more prone to display inconsistent temporal length (and cannot be re-scaled as easily as spatial dimensions, without harming fine-grained motion modeling – see Section 4.3.5).

4.3 Architecture

In this section we overview Transformer designs. The different alternatives focus on specific limitations of VTs or on better exploiting the abundant information in videos. In Section 4.3.1 we analyse approaches to reduce the number of tokens accessible in a single attention operation, aiming to reduce quadratic complexity. Then, in Section 4.3.2 we describe proposals to enhance the long-range temporal modeling capabilities of VTs. Next, in Section 4.3.3 we explore specialized designs to separately capture fine-grained and coarse-level features. Finally, and for the sake of completeness, we comment on multi-modal fusion designs to accomodate multiple modalities together with video in Section 4.3.4.

4.3.1 Efficient designs

Given the high dimensionality of video it may be challenging to represent long time spans without potentially incurring in information loss or stumbling upon the quadratic attention matrix problem. For this reason, many works decompose full attention into multiple smaller SA operations. This has a two-fold benefit, as it will reduce the size of individual attention matrices while infusing different inductive biases. Two main trends are observed: (1) *restricted* approaches, which limit the scope of a single SA operation, but maintain the sequence length throughout the network; and (2) *aggregation* approaches, which focus on progressively condensing information into smaller sets of tokens. A complete overview of our proposed taxonomy for efficient video designs can be seen in Figure 4.3.

Restricted approaches

In order to approximate full receptive field (*i.e.*, the whole input sequence), restriction relies on stacking multiple such smaller SA (similar to local filters in CNNs). We categorize restricted approaches by how subsets of tokens are selected for each SA. It can be by attending *local* token neighborhoods, specific *axis* (*i.e.*, height, width or time) or *sparsely* sampled subsets of tokens (see Figure 4.2a).

Local approaches. These are defined as the restriction by limiting attention to specific neighborhoods. Similar to *sliding* filters in CNNs, the works in [268, 141, 31, 77, 436] define the neighborhoods by sampling nearby tokens given a query. Instead, other works [236, 234, 408, 451] proposed limiting SA to small *fixed* windows, performing full SA separately in each of them. Relaxing the locality constraint only to time, in [434, 47] the fixed windows span all patches of a given frame. While sliding window local attention allows for more flexible learning (as each query has an independent local neighborhood), it has been shown to be cumbersome to implement [29]. Let S and T be the number of tokens in space and time respectively (*i.e.*, $S \cdot T = N$), local approaches reduce the computational complexity of VTs from $\mathcal{O}((S \cdot T)^2)$ down to $\mathcal{O}(S \cdot T)$ assuming a small (and constant) spatiotemporal neighborhood size. These approaches gain locality biases and linear complexity at the expense of non-local receptive fields, and will require depth to account for it. For this reason, in order to allow *information to flow between windows*, we find different neighborhood sizes for each head in [141, 408], shifting the fixed windows on every layer in [236, 234] and swapping groups of features or neighborhood aggregation tokens between windows in [434, 47]. Instead, the use of global tokens is seen in [31, 451] (alternating between local and sparsely global attention), in [268] (where the `[CLS]` token attends and is attended by all tokens, acting as a bottleneck for non-local information) and in [47] (which includes a global Transformer layer at the end).

Axial approaches. Different from local approaches, axial ones define the restriction to attention by specific axes (*i.e.*, *height*, *width*, or *time*). These can only be applied in patch-based tokenization models, where the underlying structure of the data along the different axes is

kept. *Full axial attention* decomposition has been tested for VTs, either by attending over individual axes in three consecutive MHSA sub-layers [31], or in a single one where each query token attends to all tokens that share with it the position in at least two axis [95]. However, it is more common to decompose attention into spatial and temporal, for modeling intra-frame and inter-frame interactions respectively. *Spatiotemporal* decomposition reduces computational complexity from $\mathcal{O}(S^2 \cdot T^2)$ to $\mathcal{O}(S^2 \cdot T + S \cdot T^2)$. The way in which spatial and temporal attention are related in the architecture will define the granularity at which spatial, temporal, and spatiotemporal interactions of the input tokens are learned. On the one hand, allowing attention to both axes at each Transformer layer allows for spatiotemporal relationships to form throughout layers. This can be done sequentially, through two MHSA sub-layers, as in [31, 17] (and subsequent work [378, 446, 319]) or in parallel for latter combination, seen in [17] through independent spatial and temporal heads and in [217] through separate streams for each axis. On the other hand, entirely *separating spatial from temporal* attention into consecutive modules as explored in [77, 132]. In this sense, it is not until the latter layers that temporal modeling occurs, where it may be too late for certain spatial relationships to form.

Sparse approaches. Sparse restrictions do not limit the scope of attended tokens, opposed to local and axial approaches. Instead, given the high redundancy in video data [461], sparse models provide a way to reduce unnecessary computation while maintaining a global receptive field at each layer. Sparsity can be *embedded in the SA* operation by restricting it to fixed strided patterns for each query [31, 95]. In other words, a given query is only allowed to attend (at most) to every other token on each axis. These are generally used to complement dense local attention. Other approaches involve some form of *clustering*. This can be done through a hard assignment, where tokens get separated into groups (*e.g.*, by k-means), allowing only attention within each of them. Intuitively, as SA contextualizes token representations through their relationships, these groupings allow to attend directly to the most relevant ones for each token, discarding the ones that will contribute less. In order to allow inter-group flow of information, [217] employs centroid SA, broadcasting contextualized cluster representations to each token within, whereas [451] uses an aggregation mechanism for later global modeling. Alternatively, in [288] $\mathbf{Q}$ and $\mathbf{K}$ are softly clustered into a subset of maximally orthogonal prototypes sampled from $\mathbf{Q}$ and $\mathbf{K}$ themselves, performing SA in that reduced space. This can also be seen as computing SA between a reduced (sparse) set of queries with the full K and V, followed by the full Q attending to the (sparse) result from the previous operation.

Aggregation approaches

Aggregation can take many forms, which we roughly categorize into transformer-based and pooling-based. Most commonly, thanks to *Transformers*' ability to encode highly contextual representations, they themselves can be used as global-based aggregation techniques, either through the `[CLS]` token (for instance in [17, 132]) or a small set of queries (*e.g.*, [131, 217], see below). Nevertheless, we also find works employing *learned pooling*, which can be applied

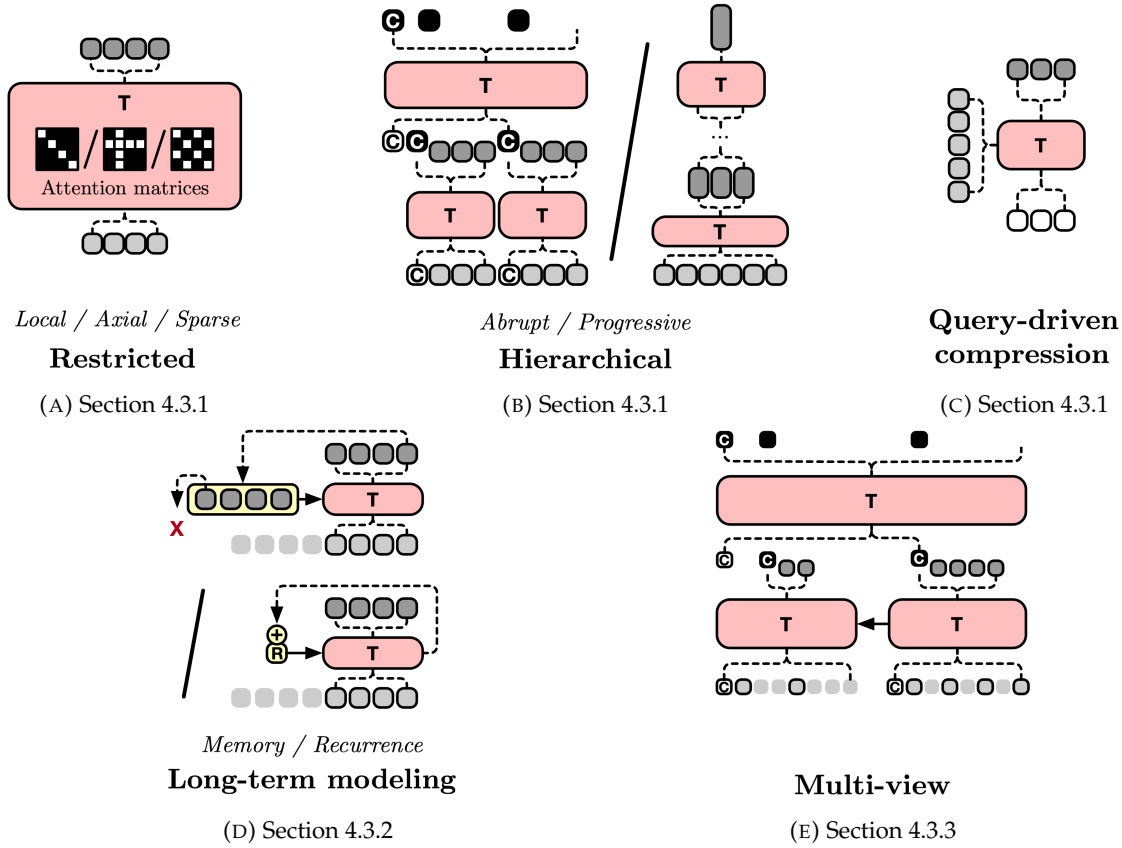


FIGURE 4.2: Visualization of the different design choices for VTs. Data tokens are shown in light gray (and black stroke if the token is used), whereas augmented tokens are in darker gray; those in white are initialized learnable tokens; and [CLS] tokens are indicated with "C" (filled black after being augmented). Data flowing into the (T)ransformer from the side is used for cross-attention.

globally or locally. For the former, a fully connected layer (*e.g.*, [130, 337]) or a set of convolutional layers ([331]) map the token sequence into an attention-like matrix, used to perform a weighted pooling over the whole sequence. For the latter, individual neighborhoods of tokens are aggregated by concatenating their embeddings into a single vector and mapping it to lower dimensionality through a linear layer (*e.g.*, [236, 417]). This is similar but not equivalent to strided convolutions, where each element of the kernel would weight a different token of the neighborhood. Although this method only aggregates in local neighborhoods, it is generally used in works which enrich the tokens with non-local context prior to aggregation (*e.g.*, [236, 417]), and as we discuss in Section 4.3.5 this may prove to be a crucial feature.

We note that it is not clear whether any of these is an absolute better option, as each has proven more valuable than others on different areas of application. However, in some cases it can make more intuitive sense to use one over the others, for instance in MViT [417] or Swin [236] where aggregation happens at many local levels, pooling is more straightforward, whereas `[CLS]` tokens may be easy to use when predefined bigger neighborhoods are used instead. And although not necessarily efficient on its own (for instance, when used as final representation for a downstream task, such as computing a classification score [17, 91, 70, 361, 468] or measuring similarity for retrieval [118, 289, 96]), these ideas can be used to build efficient models by progressively condensing information in a smaller set of tokens throughout

the network.

Aggregation-based VTs can be roughly categorized into *hierarchical* and *query-driven compression*. The key distinction is whether the input sequence length is reduced for all **Q**, **K** and **V**, or if a small set of tokens (**Q**) is used to condense information from the full input sequence (**K** and **V**).

Hierarchical. These designs can be further divided into abrupt or progressive hierarchy. The former employ bigger neighborhoods (*e.g.*, whole frames) and perform a single aggregation step, whereas the latter tend to work on smaller neighborhoods and involve multiple such steps (see Figure 4.2b). In both cases, the improvement to efficiency comes from the fact that deeper layers will have to process a smaller sequence length.

Abrupt approaches divide the input tokens into separate groups which are independently processed by a Transformer, to learn intra-group relationships. Then, information from each subset is aggregated, generally through a `[CLS]` token (*e.g.*, [17, 132]), although some use learneable global pooling in the form of linear [130] or convolutional layers [331]. The aggregated representations are then fed into the next stage, modelling inter-group relationships. We only find one work leveraging pure temporal hierarchy [130], which models frame-then-clip interactions. It is more common to employ spatiotemporal hierarchical models. These works ([17, 132, 268, 47, 439, 451, 172]) are the aggregation equivalent of spatiotemporal axial methods: a first module (generally a ViT [91] or Swin [235] architecture), learns spatial patch-wise interactions, and a second one models frame-level temporal interactions. Interestingly, in [331] multiple aggregation tokens are used for each frame, containing different features. As we discuss later in Section 4.3.5, these approaches may loose the ability to model fine-grained features after aggregation, potentially missing on relevant temporal cues.

Progressive approaches, tackle this limitation by learning spatiotemporal interactions at all levels. In works such as Video Swin [236] and MViT [106] (as well as their followups [219, 234, 417, 406, 396, 398, 160]) non-local interactions are learned at each level, whereas in [214] the first layers are limited to local interactions. In both cases, sequence length is progressively aggregated by local neighborhoods (*i.e.*, through learneable local pooling) while expanding the tokens' dimensionality. While this increased model capacity for deeper layers will require more parameters (weight matrices **W** quadratically grow with the number of feature channels), it is generally compensated by smaller dimensionality in shallower layers. Interestingly, the work of [451] combines both types of hierarchy, by progressively downsampling in the spatial module, for latter aggregation and high-level temporal modeling.

Query-driven compression. Another aggregation-based approach consists in defining the set of queries **Q**, such that $N^{(\mathrm{Q})} \ll N$. Then, the computations are reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \cdot N^{(\mathrm{Q})})$. In these, SA is performed only on the tokens that correspond to **Q**, while **K** and **V** will be attended over via CA. With this, the $N^{(\mathbf{Q})}$ queries will iteratively access the whole input to distill the most useful information and aggregate it in the token embeddings corresponding to the queries. The intuition behind this is similar to how the input tokens to

the decoder get refined by repeatedly cross-attending to the encoder's memory **M** (see Figure 3.1). In the context of VTs, these queries are generally defined to be either an aggregated or sub-sampled version of the input data, or they are an independent set of tokens.

Aggregating the input into queries (*e.g.*, through global pooling) can be used to build global streams while maintaining access to a broader low-level context within **K** and **V**. This may be useful for tasks that require a high level representation of the input clip (*e.g.*, video retrieval [130], scene or action classification [337] or group activity recognition [217]). Interestingly, in [357] this idea is developed by forming a reduced set of queries at each layer. In particular, T and S embeddings resulting from spatial and temporal average pooling respectively, are concatenated and used to attend the full set of keys and values. Another alternative to produce a reduced set of query tokens results from *sub-sampling* the input. In this manner, the queries can be used to reason about specific regions or objects (*e.g.*, by extracting a small set of boxes from the input clip to be used as queries [131, 466]).

Using a fixed set of *learnable queries* to cross-attend the input was first explored in [178] to build a global stream, where latent embeddings are used to progressively gather information from the raw high-dimensional input. In VT literature it is more common to use these learneable queries in an object-centric fashion, extending on DETR [54] (used to detect objects at each frame) and propagating detection tokens to build recurrent Transformers (*e.g.*, [254, 464], see Section 4.5.4). Alternatively, a set of independent *text-based queries* can be defined from the text modality to aggregate relevant visual information for video question answering [194]. This idea naturally extends the original Transformer, replacing the textual encoder by a video one while maintaining the auto-regressive text decoder, for video captioning [207, 215, 173, 259] or dense captioning [173, 467, 443] (through further event sampling).

4.3.2 Long-term (temporal) modeling

Capturing capturing long-term dynamics might be crucial for video tasks, as events observed at a given moment could potentially be only understood by looking far away in time. We saw a clear example of this in Chapter 3, where some human behavior can only be understood by events occurred long ago. We here focus on works that propose dealing with long-term temporal modeling. We roughly categorize them into memory- (*e.g.*, [417, 207]) and recurrence-based approaches (*e.g.*, [432, 254]). Whereas recurrent ones aggregate information into fixed-size representations, memory-based are variable-size and allow selective attention. In both, portions (*i.e.*, frames/clips) of the videos are processed sequentially in a sliding window fashion to keep manageable compute and GPU-memory but still making sure relevant information from past windows is at reach.

Memory

Naively caching many past raw (high-dimensional) input frames quickly becomes prohibitive. Instead, one can store global frame features [425, 107] or convolutional maps late in the embedding network [436], intermediate embeddings across Transformer layers (*e.g.*, those from patches [417]), or the Transformer's output embeddings [41]. In particular, when dealing with

patch embeddings, aggregation might be needed before storing them [417]. On top of that, some works maintain several memories with different temporal reach (long/short) [425, 436], abstraction level [417], or granularity (fine/coarse) [417, 41].

Memory access. Memories are mostly accessed via either cross-attention [107, 425, 436] or self-attention [417, 41]. By concatenating input and memory tokens sequence-wise to perform self-attention, the cost of the operation is $\mathcal{O}((N_{\mathrm{M}} + N_{\mathrm{X}})^2)$. Although manageable with small memories, cross-attention turns out to be much more affordable, with cost $\mathcal{O}(N_{\mathrm{M}} \cdot N_{\mathrm{X}})$ if we assume $N_{\mathrm{X}} \ll N_{\mathrm{M}}$. Either way, if $N^{(M)}$ happens to be too large, one can reduce the number on the tokens on-the-fly when accessing them [107, 425, 417] by either query-driven compression [107, 425] or progressive aggregation [417] – both seen in Section 4.3.1. On the one hand, existing works using *query-driven compression* for efficient memory access follow a two-stage bottleneck compression: a first Transformer compresses the memory into a smaller set of tokens, whereas a second one "decompresses" the output of the former into a larger set but still much smaller than the original memory. In the case of [425] the second Transformer is also deeper than the one in the first stage. It also uses two separate sets of learnable tokens to perform the aggregation in both stages, while [107] uses a hard selection of memory tokens in the first stage (obtained via *Farthest Point Sampling* [309]). Besides the efficiency gained from such two-stage factorisations, we intuit distinguished underlying roles of each stage. While the first focuses on rough selection/compression, the second tries to recover as much information as possible, aggregating and further refining embeddings. On the other hand, *progressive memory aggregation* throughout the Transformer layers provides later access to finer-to-coarser details. For instance, [417] keeps spatially aggregated $K^{\ell}_{(t-M^{\ell}):(t-2)}$ from previous timesteps after a learnable pooling and concatenates with lastly cached memory that is to be compressed in this iteration (*i.e.*, K^{ℓ}_{t-1}), and the current input's K^{ℓ}_{t} to be used in the ℓ-th MHSA sub-layer (and analogously for V embeddings).

Multiple memories. Multiple memories (*e.g.*, short- and long-term) can be separately accessed and their respective memory-enhanced embeddings fused [436]. Alternatively, a short-term memory (with fewer tokens) driving the compression of the long-term one [425]. In multi-layer memories [417], the ones in later Transformer layers implicitly access information provided by earlier ones, effectively allowing local memory accesses to approximate the full receptive field of the memory in deeper layers.

Memory update. As we move forward in time, memories are discarded in a First-in First-out (FIFO) fashion [107, 425, 417, 436]. A notable exception is [41], which leverages the self-attention weights to discard memory token(s) that are less attended by the rest.

Recurrence

Drawing inspiration from RNNs/LSTMs, recurrence mechanisms have also been proposed to deal with long video sequences. Here we distinguish between recurrence applied between intermediate layers in the VT [432, 207] and outside of it [254, 464].

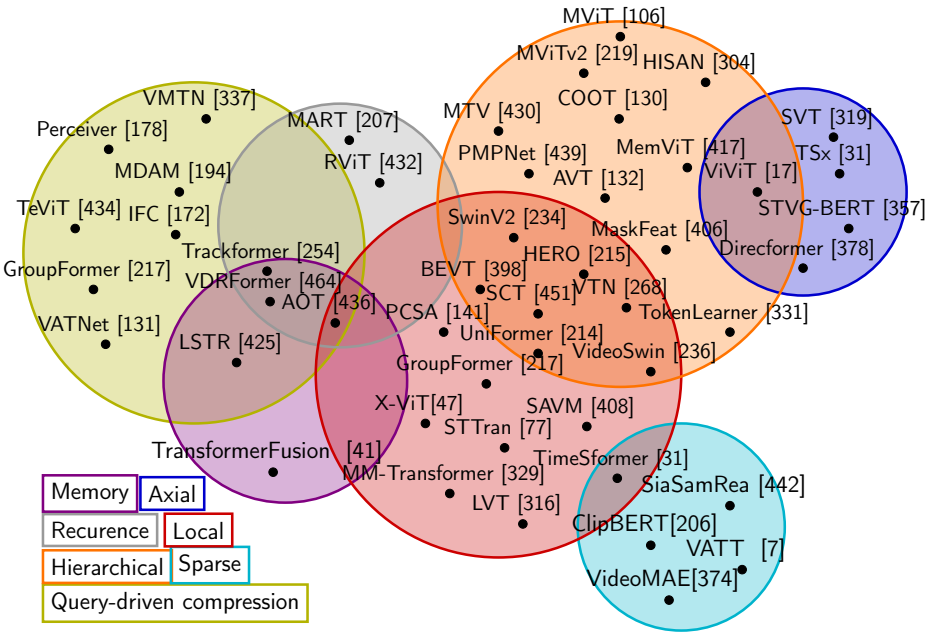


FIGURE 4.3: Venn diagram displaying our proposed taxonomy of efficient VT designs (best viewed in color). We describe Local, Axial and Sparse approaches in Section 4.3.1, and Hierarchical and Query-driven compression in Section 4.3.1. While not efficient on their own, we have also considered Memory and Recurrence (see Section 4.3.2), as they could be seen as mechanisms to efficiently handling long sequences.

Recurrence between layers. Within the first category, we find RViT [432] and MART [207]. RVIT [432] is essentially a ViT-like spatial Transformer that propagates the output of every self-attention sub-layer forward in time. Acting as recurrent states, these are added to the embeddings from the current time step after projecting both to its own **Q**, **K**, and **V**. Instead, MART [207] leverages the embeddings alone to form **Q** whereas a sequence-wise concatenation of those with the recurrent state is used to derive **K** and **V**. Differently from RViT, the recurrent state is not the output of SA, but the result of a gating mechanism between the previous state and the current input embedding.

Recurrence outside layers. Recurrence can also be established outside the Video Transformer. In other words, the output embeddings from the Transformer at time $t-1$, namely $\hat{\mathbf{X}}^D_{t-1}$, can be propagated to its own input at t. In the context of object detection, the works of [254, 464] propose an encoder-decoder architecture for object tracking where the decoder augments a set of learnable tokens while attending to the encoder's representation of the current frame. At time $t = 0$, the decoder augments an initial set of learnable tokens that will become recurrent tokens. At $t > 0$, the decoder augments the sequence-wise concatenation of the recurrent tokens at $t-1$ and added learnable tokens at t to capture newly appeared objects. Trained using pairs of frames, these can still deal with long sequences during inference. One may argue that having a token for each object could be regarded as a form of memory, but from the point of view of time the information is being recurrently aggregated into a fixed-size representation.

4.3.3 Multi-view approaches

Opposed to dense sampling of single views, a few VTs define multiple views of a given video to solve the task at hand in a cooperative fashion. Note that the use of multiple views can be seen in various scenarios. For example, instance-based contrastive approaches employ multiple views but to drive the loss (see Section 4.4.2). Note that the multi-view approaches we discuss here are related to multi-view sampling at inference (see Section 4.6.1), where multiple views of the input are independently processed and the outputs averaged. Crucially, here this technique is leveraged also during training. A clear example of this parallel is [206], which defines *sparse views* by uniformly sampling video frames with a fixed stride but varying starting positions. Then, separate streams process each view and the final classification is reached by averaging predictions in a late fusion manner. This work could be seen as the sparse equivalent to fixed window local restriction. In this sense, it only incurs $\mathcal{O}(R^2k)$, where k is the number of sparse sequences (*i.e.*, $R \cdot k = N$). As weights are shared across streams, no parameter increase is incurred.

Interestingly, many approaches define views by *varying the resolution* of a given clip, while allowing interactions between them to form throughout the network (*i.e.*, early fusion). This was first explored for video in [450] by using patches of different spatial size at each head, and later extended to time in [430] by using 3D patches instead. In the latter case, a multi-stream network is used where each stream models the same video but tokenizes with different temporal resolution (inspired by the SlowFast Network [110]), allowing information flow between views through CA and a final global stream (in an abrupt hierarchical fashion). In [409] a similar architectural setting is used, but the views are sampled from the output of progressively deeper layers of a ConvLSTM embedding network. In this sense, each view holds smaller spatial resolution, but bigger temporal context. Intuitively, these methods use redundancy to their advantage, helping the network become robust to missing information in single views, while each stream models a coherent representation of the full input.

4.3.4 Multi-modal fusion

The human experience of the world is inherently multi-modal. Studies in both Psychology [347] and Computer Vision [277, 9] have consistently found that multi-modality provides useful cues for learning without the need for supervision [260]. This, tied with the versatility of Transformers for handling any type of data, leads into many VTs to be deployed in multi-modal settings.

One very generalized multi-modal approach is to leverage *multi-modal fusion* strategies to combine the embeddings of different modalities into a joint multi-modal representation. When done properly, it allows to exploit complementary cues across modalities and reinforce cross-modal information. In this section we discuss these task-agnostic architectural changes to accomodate multi-modal input. In Section 4.5.2 we discuss modifications for particular tasks, particularly those regarding *multi-modal translation*, where the objective is to autoregressively produce the output by incorporating contextual video (or multi-modal) information. Beyond this, we also find *multi-modal alignment*. Alignment, differently from the previous

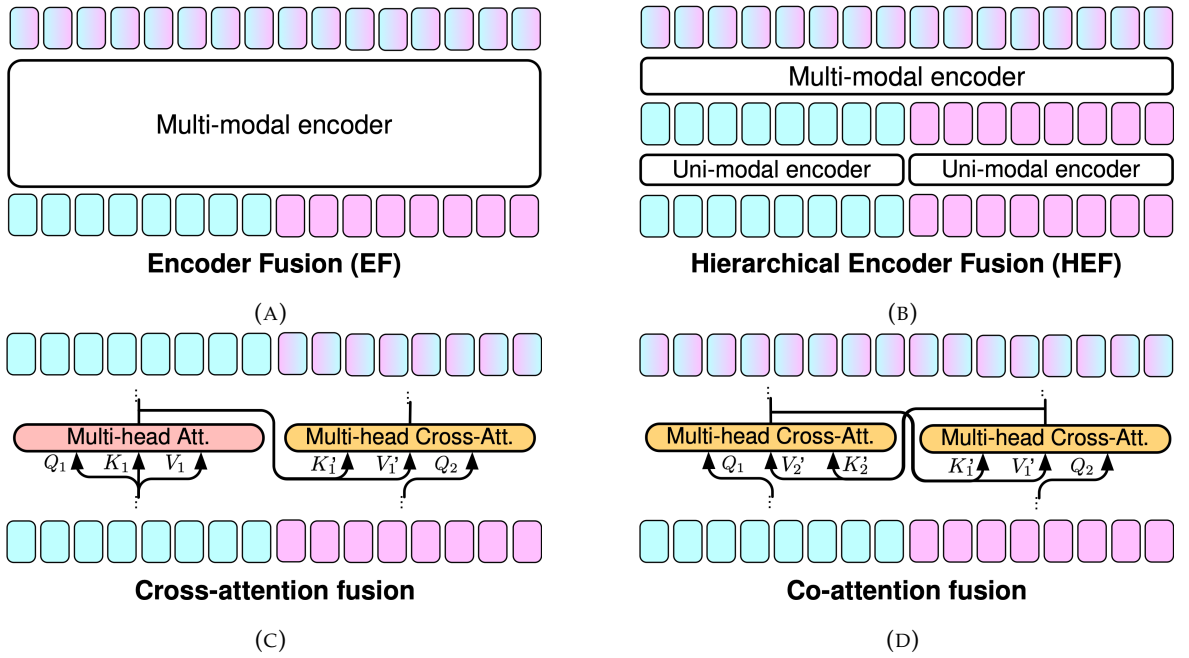


FIGURE 4.4: Four main trends to arrange encoder modules when performing multi-modal fusion (best viewed in color). See Section 4.3.4 for the details.

two, does not require architectural changes, but is performed during training via contrastive learning. We refer the reader to Section 4.4.2 for an introduction of contrastive losses in general, and Section 4.5.3 for a brief overview on how these are used for video retrieval in the context of multi-modality.

Video can be fused with many modalities, but most often with audio [173, 178], text [361, 207, 468], and optical flow derived from the video itself [186, 122, 308]. Some challenges when fusing modalities is difference in size, sampling rate, or the redundancy of their contained semantics. Text is much lower dimensional than video, but at the same time it provides useful high-level semantics that could help pinpoint salient parts of the video or establish priors about its relationships. In contrast, audio is sampled at much higher frequencies than text and hence SA among video and raw audio frames is impractical without embedding audio first into a more compact representation. Commonly, log-mel spectrograms are used to represent relatively long-term temporal audio chunks into a 2D image-like representation that can be input to an off-the-shelf 2D-CNN [159].

To fuse modalities, Transformer-based strategies typically rely on the concatenation of input sequences from all input modalities or, alternatively, on some form of cross-attention (see Figure 4.4). Within the first alternative, we distinguish *encoder fusion* and *hierarchical encoder fusion*. Among cross-attention based strategies, we find *cross-attention fusion*, which is one-sided attention of one stream over the other, and *co-attention fusion* where two streams attend to each other simultaneously. The latter two are typically implemented as encoder-like Transformer layers but modified to include MHCA sub-layers (as we did for the Dyadformer in Chapter 3), *i.e.*, decoder-like layers but without the masking (and occasionally dropping the initial SA sublayer). Moreover, we not only discuss "how" but also "where" the fusion takes place in the architecture, *i.e.*, early, middle, or late stages.

Encoder fusion (EF). Before being input to the encoder, the token embeddings of different modalities are concatenated either sequence-wise [361, 118, 231, 220, 215] (see Figure 4.4a) or channel-wise [107]. One can think of the former as how BERT [88] deals with pairs of language sentences. Encoder fusion considerably increases the computation cost of the SA operation up to $\mathcal{O}((N_1 + \ldots + N_M)^2)$, where M is the number of modalities and M_m the number of tokens in the m-th modality.

In order to identify which tokens belong to each modality, most VTs incorporate the so-called modality embeddings (*e.g.*, [205, 468, 118, 289, 231]). These are similar to positional embeddings, but signaling the source modality of the tokens in the multi-modal input sequence, so the Transformer can treat them accordingly. These learned *modality embeddings* are summed (or concatenated) with the feature embeddings and the positional embeddings altogether. Exceptionally, there are works that handle this differently. One example is the work of [361], in which a separator token is used in a similar fashion to `[SEP]` in BERT [88], originally used to indicate that a new sentence is starting, adapted here to indicate that the following tokens are from a different modality. On a different note, [178] comments on the limitation that all input modalities ought to have the same dimensionality in order to concatenate them. To solve this, the authors concatenated fixed modality embeddings of different sizes for each modality so tokens from all modalities end up having the same dimensionality.

Hierarchical encoder fusion (HEF). Encoder fusion can also be done hierarchically by augmenting modality-specific token embeddings on individual encoders first, concatenating their outputs, and sending those to a multi-modal encoder[205, 286, 360] (see Figure 4.4b). This kind of fusion allows intra-modal information to be handled before modeling the inter-modal patterns, which can be beneficial when dealing with modalities that are not highly correlated or precisely aligned at the input level. The point at which to do so has to be determined experimentally. Also, computational cost is reduced with respect to the previous encoder fusion by a constant factor, depending on the number of layers in both the uni-modal and the multi-modal encoders. Although utilizing multiple encoders increases the amount of parameters, this can be alleviated through weight sharing [205].

Despite being more rigid, encoder fusion allows for an unbounded number of modalities. Cross-attention, even when limited to two modalities only, allows for more flexible modality fusion. We distinguish between cross-attention (one-sided) and co-attention (two-sided), the first being from one modality over another and the second being mutually conducted in parallel.

Cross-attention fusion (CAF). When fusing modalities through CA, one modality will query information from an auxiliary modality that will provide context (see Figure 4.4c). The simplicity of this idea together with the flexibility it provides, causes many works to use it very differently. In [194, 174] separate encoder-decoders for each modality are proposed, each cross-attending to text embeddings, before the outputs of the different encoders are combined. The work of [178] proposes only one stream that keeps augmenting a small set of latent embeddings by repeatedly cross-attending to the same very long multi-modal input sequence of minimal embeddings. For this case, CA layers are interleaved with SA layers that refine the cross-attended information. In [468], a three-stream Transformer is proposed

where the central one cross-attends to the other two, and these will then attend to the embeddings generated by their respective opposite previous cross-attentions. The work of [52] also uses three streams, one per modality. The fusion, however, is achieved within a master stream which substitutes its SA by asymmetric cross attention over the other two at the same time (concatenating both sets of keys and values). CAF only involves two modalities at once, but has a reduced cost compared to encoder fusion, *i.e.*, $\mathcal{O}(N_1 \cdot N_2)$. Notably, asymmetric attention allows the master stream to discard information from a specific modality when it is not relevant.

Co-attention fusion (CoAF). Differently from cross-attention fusion, the two modalities in co-attention are augmented in parallel by attending to each other's embeddings. The MHCA sub-layer in each of the streams computes the queries from its own embeddings, whereas keys and values come from the other stream (see Figure 4.4d). It was originally proposed for images and language in ViLBERT [237] and followed by a few video works [80, 357, 173]. In [80, 357], SA is entirely replaced by CA, but it can also be kept [173]. Indeed, [456] has their modalities co-attending to each other and self-attending to themselves, claiming this keeps intra-modal and inter-modal dynamics separate, up to some degree. In contrast to encoder fusion, the computational cost is reduced to $\mathcal{O}((N_1 \cdot N_2)^2)$.

All of these fusion strategies allow for different early and middle fusion strategies. However, another alternative is to simply late fuse modalities. In order to do that, the different modalities are run through parallel encoders and, then, their outputs are combined. These outputs could be class score distributions [122], as typically done for Two-stream ConvNets [345] in classification problems. Although suboptimal for Transformers, this strategy might still be beneficial when training data is scarce. Alternatively, late fusion could also be regarded as combining the augmented aggregation token (or pooling over all output token embeddings) by concatenation [174, 186, 194] (or summation [303]) and then using this for classification. In either case, the modalities do not share information explicitly.

4.3.5 Discussion on architecture

VT designs focus on reducing computational complexity and handling the redundancy of videos without compromising spatiotemporal modeling capabilities. Furthermore, restrictions imposed on VTs to make them more efficient will bias them towards favoring certain kinds of relationships. Some examples of this are abrupt hierarchy, which learns temporal translation equivariance in spatial layers by modeling each frame independently, or local approaches, which enforce locality biases.

However, efficient designs and inductive biases do not explicitly handle redundancy. Video redundancy can be mostly attributed to appearance-based semantics varying slowly through time, even when small variations in specific pixels occur [461]. However, the extended information provided by these subtle changes in many consecutive frames may be crucial to properly model fine-grained motion features [110]. In order to learn spatiotemporal relationships from video, this must be taken into account. Reducing spatial redundancy may be desirable, as it will allow to focus on more relevant parts of the video (*e.g.*, through

aggregation or sub-sampling of tokens). However, this requires careful consideration: removing certain information too early into the network may limit the formation of crucial temporal interactions later on. Prior works on modeling CNN with video have shown this to be the case: early aggregation of spatial features hinders the formation of fine-grained motion features [339, 375, 224], and temporal pooling seems to hurt spatiotemporal representation learning [61, 109]. With Transformers, tackling this may involve taking into account non-local neighborhoods before deciding which information is to be discarded.

Motivated by this, we derive three crucial aspects for spatiotemporal modeling: (1) explicit spatial redundancy reduction while (2) allowing to model temporal features at all levels in (3) high-fidelity temporal contexts. Different VTs exhibit varying degrees of capabilities in these three aspects. *Restricted approaches* allow for low-level temporal modeling and, due to the lack of aggregation, always maintain temporal fidelity. Given their potential to overlap low- and high-level features they can be suitable for both low-level (*e.g.*, segmentation [95]) and high-level (*e.g.*, classification [31]) tasks, but with certain limitations. Hierarchical approaches effectively exhibit (1) and (3) through aggregation on the spatial dimensions only (except in [214]). Particularly, for *progressive hierarchy* (*e.g.*, [219, 234]), the gradual increase in channel dimensionality provides deeper layers with larger capacity to represent high-level concepts while further limiting the modeling of redundant low-level features. Furthermore, by leveraging different levels of spatiotemporal non-local contexts (*e.g.*, [236, 106]) at least in deeper layers (e.g, [214]), they guarantee that extended temporal fidelity is exploited before aggregation. In contrast, the *abrupt counterparts* (*e.g.*, [132]) may be suffering from early aggregation. While training end-to-end may infuse temporal feedback into spatial layers (which may be sufficient for appearance-biased video benchmarks, see Section 4.4.3), they may lack proper motion modeling. This can be addressed by allowing to form spatiotemporal interactions before aggregation, either locally (by explicitly sharing information between neighborhoods [47, 434, 451, 172] as well as by using 3D patches [17, 430]) or globally [331]. *Query-driven compression* approaches reduce redundancy through aggregation when used to derive global streams [130, 337, 217], or through sparsity when reasoning about individual objects or regions [217, 131]. In both cases, the small set of queries form high-level representations of (parts of) the input, while maintaining temporal fidelity in keys and values. However, they may exhibit a limited capability to form low-level temporal features. While iterative accesses may alleviate the dangers of early aggregation for high-level tasks (*e.g.*, classification [337, 178, 131]), low-level tasks may require to also evolve the fine-grained input representations [217] or to infuse them back with high-level features from the queries [357] (akin to clustering-based sparse approaches). This is similar to the behaviour exhibited by *recurrent* VTs. As temporal information is collapsed into the recurrent state, they may suffer from early aggregation (losing temporal fidelity), which may be specially detrimental for high-level tasks [432]. However, these approaches may excel on applications that only require low-level reasoning of the current observation, enhanced with the forwarded high-level past context (such as for tracking [254], segmentation [436] or dense video captioning [207]). *Memory*-based approaches exhibit great capabilities for preserving temporal resolution of the

input. They can tackle redundancy through aggregation (*e.g.*, upon storing [417] or dynamically on access [425]) or sparsity (either by storing only some past observations [436], by dropping elements in the memory according to their relevance [41], or only attending to a small subset of memory tokens [107]). Finally, *multi-view* approaches working at different input resolutions explicitly allow the formation of separate coarse- and fine-level features while allowing interactions between them [409, 450]. However, as redundancy is not explicitly managed, the success of these methods may be limited to computationally heavy models [430]. Sparse counterparts heavily downsample the input sequence, hurting temporal fidelity and requiring to compensate with other modalities [442, 206, 7].

Regarding multi-modality, concatenation-based strategies trade extended sequence lengths for efficiency. While more computationally expensive, EF does not require experimentally validating the fusion point as with HEF. Cross-attention fusion methods also grapple with finding optimal fusion points, yet middle fusion could help with asynchronous input modalities. They are more efficient than concatenation counterparts, given they do not extend sequence length at any point. As cross-attention variants collapse multiple uni-modal streams into single sequences, we hypothesize they could be better suited to enforce unified multi-modal representation spaces, potentially reducing redundancy. On the contrary, concatenation-based methods might be more approapriate for facilitating independent modality-wise learning, but some features could be redundantly represented. Finally, although CAF and CoAF are inherently restricted to two modalities, cumbersome approaches can be found that leverage this kind of fusion for more than two streams [468].

4.4 Training a Transformer

The two main limitations of Transformers will heavily influence the way in which they are trained. On the one hand, large-scale pre-taining aids Transformers overcome their lack of inductive biases [91, 67, 88], but recent studies suggest that self-supervised pre-training (see Section 4.4.2) alleviates the need for large supervised datasets [374, 406]. On the other hand, some solutions to the lack of inductive biases aggravate computational costs. CNN embedding networks add to the memory footprint and potentially overflow GPU memory when training, specially if done end-to-end. Avoiding overfitting big models requires strong regularization [427] and lots of data [452], which is further problematic when handling several stages of training that require more time and compute. Finally, leveraging self-supervised tasks is computationally heavy, specially for video.

4.4.1 Training regime

We next explore how VTs are trained, from a lens of embedding networks and pre-training. Pre-training involves one or more training stages before transferring the network to a downstream task, for which the model is either fine-tuned or linearly probed (training a few linear layers on top of the frozen Transformer).

End-to-End training with minimal embeddings. End-to-end training of deep neural networks has proven to outperform multiple stage algorithms. To ease memory limitations while allowing for end-to-end training of the Transformer, it is common to use minimal embeddings. Some train in a supervised fashion [31, 178, 17, 408, 106, 451], directly for a downstream task on large datasets, such as Kinetics-700 [56]. However, all these works leveraged efficient architectures and thanks to the inductive biases these designs provide, the network will pick up on relevant patterns faster, and more capacity can be given to Transformer layers. Other works aiming for smaller datasets train aided by some data augmentation [445, 409, 441] or self-supervised losses [450, 230, 132, 259, 7] on medium to large datasets. Stand-alone Transformers seem to be able to learn without large CNN embeddings if aided by the inductive biases that efficient designs, data augmentation or self-supervised losses provide. Still, most of these require multiple stages of training either through large datasets or computationally heavy self-supervised techniques.

End-to-End with embedding networks. Other works train Transformer and deep CNN embedding layers end-to-end either with a pre-trained embedding network [357, 68, 194], fine-tuning just the later layers [360, 337], or training end-to-end from scratch [404, 239, 397, 290]. Some were able to train end-to-end by capping Transformers to 1$\sim$4 layers [460, 186, 439], suggesting that just a few Transformer layers after a large embedding network may be enough to boost performance. Some others' success is attributable to leveraging efficient designs (*e.g.*, local SA [141]) or weight sharing [205] – that reduces the effective number of parameters to be stored in memory (discussed later in Section 4.7). Finally, [405, 131] report having substantial computational resources available, which allowed them to fit in memory both, a large embedding network and a big Transformer. Empirical studies on both image [317, 81] and video [214] Transformers have consistently found improvements when training Transformers and CNN embedding layers end-to-end. This may further be seen in works reporting improved CNN-based results alone after being trained as the embedding net of a Transformer [205, 360, 362], pointing towards CNNs benefiting from long-term temporal feedback provided by the Transformer layers.

Frozen embedding networks. The most common approach by far for VTs is leveraging some large pre-trained and frozen CNN embedding network. These are then used for feature extraction, which further boosts cost-effectiveness, as these features can be pre-computed. Transformer layers are then trained for a downstream task on those features. Compared to end-to-end training from scratch, it is often cheaper and more efficient to employ state-of-the-art models, which have been carefully tuned to perform well on some supervised task. While it is definitely common to use medium to large datasets (as in [207, 467, 286, 173, 294, 456, 443, 174]), with this approach, many video works [308, 304, 53, 70, 307, 122, 52, 367, 400, 220] are still able to train the Transformer on small datasets ($<$10k training samples). Nevertheless, these approaches are limited by the quality of the pre-trained features, and could be biased towards the task they were trained on (which are generally supervised).

Pre-trained Transformers. Video-based pre-training has proven to work best for video classification tasks [406, 396], maybe due to the distribution gap, as pre-training only on images does not provide any motion cues. Nevertheless, image-based pre-training may provide stronger spatial features, given the higher variability of appearance and number of categories (providing better semantics regarding objects) compared to video datasets (where many consecutive frames contain similar appearance statistics). It is for this reason that we find many VTs leveraging image pre-trained Transformers, commonly on some ImageNet variant [86, 324]. This is generally done in one of two fashions. On the one hand, some works [17, 132, 430, 268, 436, 47] leverage a pre-trained image Transformer (generally ViT [91] or Swin [235]) as the spatial stage of an abrupt hierarchical VT, training the later temporal layers from scratch. On the other hand, a pre-trained image Transformer can be directly adapted by using 3D patches to factor time in (as well as inflating linear embeddings and positional biases to account for this change) before fine-tuning for video [236, 234, 106, 406]. Finally, some object-centric approaches (*e.g.*, [466, 405, 254]) leverage a pre-trained Transformer-based object detectors (*e.g.*, DETR [54]) as initialization for a tracking Transformer.

4.4.2 Self-supervised pretext tasks

Harvesting large annotated datasets incurs in additional labeling costs, and may further influence towards human-induced annotation biases [326, 71]. *Self-supervised learning* (SSL) has been recently shown to alleviate data needs for an equivalent supervision-based pre-training (*e.g.*, [374, 406]), while providing more robust [155] and general features [192, 140, 69]. Despite the great success of SSL in both NLP [88] and Image Transformers [149], they are not as widespread in the video domain, which could be attributed to the large costs involved in such process. Therefore, we next analyse benefits and limitations of SSL for VTs, so as to motivate further research on this area.

Traditional time-related pretext tasks (*e.g.*, arrow of time or playback speed, see [334] for a complete review) are rarely used in the context of VTs. They are generally limited to shuffling the input sequence, and training the network to correctly reorder it [215, 446, 144, 378]. The task can be formulated either token-wise (by classifying correct position of each token [215, 446]) or sequence-wise (by predicting the permutation from a predefined set [378] or the edit distance from the original sequence [144]). Intuitively, by solving such a task, the network learns what coherent temporal dynamics look like. Alternatively, we find a few works leveraging generative losses, akin to traditional inpainting [287]. These are only used in VTs where they themselves are the task being solved for downstream. In these cases, it is only natural to use SSL, given that the objective is the data itself. These models are trained through a reconstruction similarity loss [316] (e.g. L_2 or LPIPS [457]), sometimes accompanied by an adversarial setting [230, 450] or other techniques to enforce temporal consistency [409].

However, these have not found as much success [334] compared to (1) *Instance-based learning* and (2) *Masked Token Modeling* (MTM), which we explore next. The former learns sequence-level representations that are invariant to different spatiotemporal perturbations, whereas the latter masks individual token representations of the input and tries to reconstruct them.

Instance-based learning

Instance-based approaches for VTs leverage contrastive losses (generally *InfoNCE* [273]) to make representations of whole sequences invariant to certain augmentations. These approaches define one anchor $\mathbf{x}$, a positive sample $\mathbf{x}^+$ and a set of G negative samples to contrast against $\{\mathbf{x}_g^-\}$, where $1 \leq g \leq G$. These tasks force representations for the positive pair to be similar, while it drives apart representations for the negative (dissimilar) pairs. Minimising InfoNCE can be seen as maximising a lower bound on the mutual information between $\mathbf{x}$ and $\mathbf{x}^+$ [273]. It does so by means of a similarity function $sim(\cdot)$, which measures the affinity between two elements. This function assigns a high score to positively correlated pairs $(\mathbf{x}, \mathbf{x}^+)$ and a low score to negatively correlated ones $(\mathbf{x}, \mathbf{x}^-)$. These losses have also been used in the context of cross-modal matching for VTs, but this has already been explored in [330]. Hence, we only briefly discuss them in the context of video retrieval (see Section 4.5.3), and focus here on their uses for video only.

View mining. Positive pairs tend to be differently augmented versions (generally regarded as *views*) of the same video sample. In VTs, it is customary to apply spatial augmentations (*e.g.*, random cropping, color jittering, horizontal flips or Gaussian blur) consistently through time (*i.e.*, applying the same augmentation to all frames [310]). By aligning multiple views' representations, the model learns to be invariant to such perturbations. However, spatial augmentations alone are not enough for video SSL [334], and generating temporal views needs to be done carefully. For instance, reversing or randomly shuffling a clip may make the model invariant to temporal causality. In VTs (similar to other video literature [111]), it is common to use multiple temporal [416, 362, 396, 144] or spatiotemporal [319, 290] crops of a given video to form the positive pairs, with varying temporal spans [396, 290, 319] and frame-rates (i.e, speed) [319, 396], whereas negatives are sampled among all other training videos. Learning invariance to such changes may be useful for high-level tasks where a wide abstract understanding of the video is enough. Nevertheless, this could disregard local view-dependent information in favour of redundant cross-view information [306], favouring the formation of appearance-biased features (see Section 4.4.3). We hypothesize this is related to the level of features learned by enforcing this type of invariance, *i.e.*, appearance features that describe the overall video clip, but not low-level motion cues that evolve over time and are different among the various views. By forcing the network to similarly represent all positive pairs (clips from the same video), it is left only with shared information among them, *i.e.*, overall appearance of the video. In this sense, we regard proper positive and negative sampling to be crucial to ensure that both general abstract features as well as fine-grained view-dependent local ones are preserved. For instance, balancing the modeling of local and global representations [444], or sampling positives nearby in time to the anchor with higher probability [310], so the network is not forced to depend entirely on high-level appearance features. In VTs we find several approaches to tackle this. Some works use multiple global and local potentially overlapping views as positives [319, 290, 396], which may allow for better modeling of part-whole relationships. Intuitively, this forces global views to preserve information in the local ones, while maintaining global context awareness in local views. Alternatively, the alignment

task can be relaxed, skewing away from learning absolute invariance to changes between views. One example is seen in [362], which conditions alignment on the temporal shift between crops. Another solution found in VT works consists in introducing asymmetries in the networks computing the different views' representations: using additional predictors [396], momentum encoders [319, 396] (originally proposed in [150]) and even CNNs [144] (probably helping infuse some locality bias from CNN representations into the Transformer). Introducing some of these asymmetries has indeed been found to boost downstream performance on image [372] and video tasks [111]. Intuitively, they may be relaxing the alignment task into a more predictive setting, allowing features to be aware of context, not so much invariant to it. In other words, it enforces one view to contain enough information so that when projecting it, it can be aligned with the other view, making the representation of a given crop predictable from another without directly sharing all the information. On a different note, we also find some works combining instance-level learning with token-level learning through MTM [360, 362, 416] (see Section 4.4.2), which as we will discuss, seem to better preserve local token-wise information of different parts of the video.

Negative sampling. One crucial limitation of contrastive approaches is their need for large negative sets [66]. These are generally mined from the batch, which can be very limiting in the context of full video representations, as it may not always be possible to hold enough different instances in a batch. VTs tackle this through large *memory banks* that store representations of past batches [396, 340, 231] (which may further serve as regularizers, due to storing sample representations from past iterations produced by the same model with slightly different weights) or through *hard negative mining* (forcing the network to learn small nuances in the views by trying to separate somewhat similar samples, measured by feature representation distances [205, 218]). Finally, we also find works *dropping negatives* altogether. One example is seen in [319], which formulates learning as instance-based classification, where every positive view has to be classified in the same pseudo-class. Another example is the work in [442], where, during training, multiple sparse views of the input are independently processed and the aggregated prediction is used to distill the consensus into single view streams (which are the ones later being deployed).

Masked Token Modeling

MTM draws inspiration from the *Masked Token Prediction* loss proposed in BERT [88]. It randomly replaces some input tokens with a learnable `[MSK]` token and the network is trained to predict (classify) the original tokens. This forces the Transformer to learn contextualized representations by trying to retrieve the masked part of the input based on all remaining tokens. However, different from language tokens, visual tokens cannot be easily mapped to a discrete and limited-size vocabulary so as to pose MTM as a classification task. For perspective, a pixel codebook would require $256^3 \approx 16$M distinct elements, whereas BERT employed a vocabulary of 30K. Furthermore, posing it as a classification task would disregard the distance of the prediction to the actual ground truth value, distracting the network with high-frequency details of the data which could be irrelevant. To solve this issue in the context of VTs, we

roughly find three families of approaches, categorized by the type of target: (1) working at *feature* level either through regression [215, 63, 218, 406] or contrasting [215, 205, 360, 70], as well as (2) *quantization* of visual tokens [361, 398]. Interestingly, some works have actually found success (3) regressing the original token in *pixel* space [374, 234]. We also find VTs classifying token contents [468, 416], but as these require manual annotations, we do not delve into them here.

Feature-based MTM. These works regress a feature-based representation of the masked tokens. This can be posed as a prediction (*e.g.*, using an MSE loss) [215, 63, 218] or as a contrastive task [215, 205, 420, 70]. The target token representation is obtained from the input embedding network (*e.g.*, [215, 205, 420]) or from an external encoder [360]. In this sense, by requiring an additional network, these models potentially incur in additional compute and memory costs. In order to avert this, [406] proposes using the HOG features of the masked region, which are cheaper to compute and can be pre-computed. Interestingly, the work of [132] uses masked causal self-attention instead of replacing tokens with `[MSK]`. In this sense, all tokens are tasked with solving feature-based MTM by trying to predict the next token's representation (similar to a predictive coding setting [320, 273]).

Quantization-based MTM. Employing this technique involves discretizing video tokens to a limited codebook, generally requiring some pre-trained network to define it. For instance, in [361] a S3D [423] followed by hierarchical k-means is used for both, embedding the tokens prior to the Transformer, and the discrete (cluster assignment) pseudo-label for the prediction, whereas in [398] a VQ-VAE [318] is used instead, but only to generate the ground truth for the masked tokens. The use of quantization makes it possible for these models to optimize the network with a classification objective, akin to NLP counterparts. However, similar to many feature-based MTM, these approaches also require an additional pre-trained network.

Pixel-based MTM. Opposed to previous approaches, pixel-based alternatives directly regress the pixel space for masked regions [374, 234]. They do not require any further networks nor computing additional features, making them very simple to implement. However, pixels as targets have been argued to focus on irrelevant high-frequency details of data, which could be detrimental for high-level tasks [146]. Nevertheless, this may be more nuanced and require further research, as we discuss next in Section 4.4.3.

4.4.3 Discussion on training strategies

Training stand-alone VTs requires balancing solutions to the lack of inductive biases with potentially limited computational budgets. This implies factoring in large datasets, SSL and efficient designs while accounting for the large dimensionality of videos, properly sized clips, batches and architectures. VTs are dominated by fully supervised training aided by *large frozen CNN embeddings* (which are not so common in other fields, such as NLP), and disregard pre-training of Transformer layers. The benefits of the CNN-Transformer tandem are twofold. On the one hand, long-range temporal interactions provided by Transformers boosts

CNN's performance in many application scenarios[308, 286, 173, 304, 53, 199, 122, 367, 225]. On the other hand, the embedding network provides initial representations and dimensionality reduction, alleviating Transformers' training limitations. Nevertheless, this approach caps the potential of Transformers to model spatiotemporal interactions (specially long-term ones) and depends on the transferability of the pre-trained embedding features (*e.g.*, problems with distribution or task shift).

The canonical pre-training then fine-tunning paradigm acts as a *smart form of initialization*. We hypothesize that skewing from it may allow to avoid catastrophic forgetting [250] while achieving more generalizable features. For example, by incorporating self-supervised auxiliary losses during fine-tuning, as done by some VTs [132, 446]. In [215] a training schedule is proposed which samples a different (self-)supervised task at each batch, showing improved results for video retrieval as more tasks are added. Alternatively, recent works (*e.g.*, [398, 133, 453]) deviate from the trend of image-based pre-training, and achieve promising results by optimizing for image and video tasks in a joint manner.

SSL is not as widespread for VTs when compared to supervised or image-based initialization. However, we believe VTs could greatly benefit from large-scale unlabeled videos, as well as from the inductive biases SSL provides. In this sense, we see great promise on the current developments on SSL that are better suited to train visual Transformers. MTM could be seen from the lense of generative-based pre-training as it bears great resemblance with CNN-based inpainting [287]. However, the tendency of CNNs to overfixate on high-frequency features [2, 182] (which does not seem to be the case in vision Transformers [30]), may explain why generative-based approaches have not gained as much popularity for pre-training CNNs [334, 181]. We believe that the success of MTM may be attributable to Transformers providing explicit granularity through tokenization. In order to *conquer* the complex global task of inpainting large missing areas of the input, MTM *divides* it into smaller local predictions. This is specially true in both 2D- and 3D-based patch tokenization approaches [406, 374, 398]. Intuitively, the model needs understanding of both global appearance and motion semantics as well as low-level local patterns to properly gather the necessary context to solve token-wise predictions. This may allow VTs to learn more holistic representations (*i.e.*, better learning of part-whole relationships). Nevertheless, given the high redundancy of videos it could be trivial for the network to find shortcuts, borrowing information from neighboring spatiotemporal positions instead. It has been found that high masking ratios (*e.g.*, 40%-60% in MaskFeat [406] or even 75%-90% in VideoMAE [374]), specially compared to NLP (15%-20% in BERT [88]) or images (20%-50% in MAE [149]), indeed force the network to capitalize on global relationships of the data, as seen by improved performance on high-level semantic tasks (see Section 4.6.2). Furthermore, ablations in [406, 374] suggest that the masking strategy can also impact the learning of such shortcuts, showing that masking blocks of tokens in space consistently through time helps to avoid them. Regarding the choice of target for MTM, quantized and feature-based seem to work best for video [406] (albeit requiring an additional pre-trained network). Pixel based provide the cheapest target, but are generally discarded arguing they may fixate on irrelevant high-frequency details of data. However, the generally used MSE loss has been shown to disregard such details [458, 275, 287], so further

research may be needed. Finally, we highlight HOG features, which provide the best compute/performance trade-off (see Section 4.6.2), as they are cheap to compute while providing partial invariance to various deformations.

Despite requiring large batches for negative mining, instance-based contrastive approaches have consistently shown potential for high-level video tasks [334]. By contrasting differently spatiotemporal augmented views, the network learns invariance to appearance perturbations, spatial scale and occlusions, as well as changes of perspective or illumination naturally present in video [138, 403]. However, the model may also become invariant to temporal translation and deformation, effectively disregarding fine-grained motion dynamics and biasing it towards appearance-based static cues (which is enough for appearance-biased datasets – *e.g.*, UCF101 or Kinetics – where the presence of certain objects may suffice to predict an action class [168, 46]). As we discussed in Section 4.4.2, re-introducing motion modeling requires relaxing the alignment task through network asymmetries (*e.g.*, [144, 319]) or careful sampling techniques (*e.g.*, [290, 396]) to balance part-whole relationship learning. However, compared to MTM, it is easier for these approaches to overlook low-level view-dependent temporal information, crucial for proper motion modeling [444, 310].

In this context, we see promise in combining instance-based contrastive learning and MTM, both in multi-task scenarios [360, 362, 416] as well as feature-based contrastive MTM [215, 205, 420, 360] (as opposed to direct regression). These latter could combine the holistic feature learning of MTM while potentially accounting for the uncertainty of modeling missing information through contrastive losses (as the model is not tasked with explicit hard prediction [146]). For instance, in [215], this alternative is found to outperform L2 feature regression in the context of video-moment retrieval. These approaches remain, to the best of our knowledge, unexplored in the context of patch-based tokenization, where the cardinality of the negative set would be much larger than for instance-based approaches (allowing for many hard negatives from the same sequence as well as easy negatives from all other sequences in the batch). Nevertheless, it is still unclear what these models are actually learning, so future research is needed for proper interpretation of SSL features, which currently are mostly evaluated based on their success on downstream performance [334, 181].

4.5 Task-specific designs

In this section, four major subsections review specific designs of the most popular video tasks (see Table 4.1): Action classification in Section 4.5.1, Video translation (*e.g.*, captioning) in Section 4.5.2, Retrieval in Section 4.5.3, and Object-centric tasks (*e.g.*, detection and tracking) in Section 4.5.4. For the sake of completeness, these are followed by short summary subsections regarding the remaining tasks: Low-level in Section 4.5.5, Segmentation in Section 4.5.6, and Summarization in Section 4.5.7.

Transformers have also been applied for action anticipation [132, 400], sign-language translation [53, 52], visual-question answering [194, 460], autonomous driving [303], robot navigation [107], visual-language navigation [286], personality recognition [80], lip reading [239], dynamic scene graph generation [77], and multimedia recomendation [70]. As not

many video Transformers have tackled this, it is too early to ascertain specific trends, so we have simply listed them here for completeness.

4.5.1 Classification

Regarding video classification, most works rely on pure Transformers [17, 236, 31, 451, 7] that for the most part focus on efficiency: both [17] and [31] test various space-time decompositions, whereas [17] also tests tokenization strategies (2D vs 3D patches). They found that a pre-trained ViT [91] encoding 2D patches with a temporal encoder on top performed the best. The works of [236] and [451] propose different types of restricted attention: the former restricts locally in shifting windows and the latter by only attending to previous frame's patches after having exchanged information with another efficient attention mechanism [197]. In [106] they opt for 3D patches whose receptive field is enlarged across stages by subsequently merging token embeddings. Others pursue building very deep Transformers by maintaining a very compact latent representation [178]. These larger Transformers for classification require large labeled datasets for fully-supervised training [106, 236] or heavily rely on self-supervised pre-training [361, 205]. For multi-modal datasets, encoder fusion [361] or hierarchical encoder fusion is utilized [360].

Several other works rely on larger (usually CNN-based) embedding networks [361, 205, 360, 294, 186, 122, 290, 362], facilitating the training on smaller datasets. When equipped with these embedding networks, shallow encoders can serve as mere pooling operators [186, 122, 290, 308]. For detection, Transformers are also a natural way to fuse information among detections initially made by these embedding networks [217] or to allow them to attend over a larger visual context [131]. Although mostly used in pure Transformers, efficient designs have been explored for this kind of works as well, for instance through weight sharing [205]. We note that these architectures can also be used for regression tasks, by simply replacing the final output head to produce continous values.

4.5.2 Video translation

The translation task intends to map the raw input video to an output signal of a different nature and with an arbitrary length. In the context of video literature, translation can be considered a multi-modal task: given video (and possibly additional modalities) it is translated into a set of (generally) non-video outputs. Although the output could also be video, it is often a signal in another modality (*e.g.*, language) or simply a sequence of discrete symbols. The most popular instantiation of translation is *video captioning* [207, 215, 173, 259] that consists in producing natural language descriptions of what is globally going on in the video. When producing separate captions for different video subparts independently, this is referred to as *dense video captioning* [467, 443]. A more specialized type of video captioning is *sign-language translation* [53, 52]. Additional other forms of translation are: *video reasoning* [466], that extends the task of captioning by allowing a natural language prompt along with the video; *video-language dialogue* systems, which add to reasoning the requirement of back and forth communication with an external agent while reasoning about the visuals [220]; *temporal (or*

spatiotemporal) action localization [367] to produce a list of, respectively, temporal begin and end times or a "tube" of bounding boxes containing the human actions in the video; or *robot video-based navigation* [107], in which the video – and perhaps other sensory inputs – are translated to the next action (a sequence of next actions) to take. Most commonly exploited input modalities for the translation task along with video are text [207, 220], audio [173, 220], and optical flow [367], but others are also used (*e.g.*, human pose [52] or depth [107]).

VTs tackling translation typically leverage encoder-decoder architectures, in which video is passed through the encoder and served as context to the decoder – similarly to the original Transformer (see Section 3.1.1), only that the encoder is a video encoder instead of a language one. Task-specific modifications of this design are found for dense video captioning [173, 467, 443], where a temporal proposal generator is attached after the encoder to tell the decoder where/when in the sequence it has to focus. While most methods first divide the video in regions and then re-encode them independently to produce per-clip captions [173, 467], the work in [443] eliminates the re-forwarding by making the most of local self-attention, which limits the leakage of information across the encoded proposals. [367] tackled temporal action detection by relying only on a Transformer decoder. Inspired by DETR [54], proposals are sourced from a set of learnable token embeddings input to the decoder. The decoder augments these tokens which are later classified into actions and used to regress their temporal position and length.

Multi-modal translation generally uses encoder-decoder or decoder-only schemes. When multi-modal input is provided, multi-modal fusion (see Section 4.3.4) can be adopted before the translation. Yet some works completely skip the prior fusion [220, 259] or opt for a hybrid module that fuses and decodes at the same time [207]. Except for the latter approach, the decoder roughly maintains its canonical form, although there are works who propose small variations. For instance, the MHSA sub-layer can be removed to slim the models [220] or substituted by a moving average [443]. The decoder layers can also cross-attend to the encoder layers of equivalent depth [467, 173], or simultaneously to the output embeddings of different modalities separately [52]. There are also designs that go without a Transformer encoder, replacing it by an external non-Transformer module [367] or relying entirely on the decoder [220, 207]. The work in [220] follows the prompt-based input of GPT-2 [314] and feds n video features as the first tokens in the decoding sequence and decodes the caption starting at the $n+1$-th input embedding. In particular, [207] prompts not only the visual features but the current language sentence features to generate the next sentence in a paragraph. All in all, prompting is the generalization of the original shifting operation in the original Transformer [385], where the decoding starts at a shifted position to account for the start token.

4.5.3 Video retrieval

The task of retrieval consists in recovering a piece of information associated to a particular query. Those associations can be video-video pairs [340] or pairs composed of different modalities (video with, most often, language [130, 231, 259, 289] or language plus audio [118, 96]). Retrieval relies on a distance metric among the representations of the queries and the retrieval candidates. The representations are learnt during training using the ground truth

to minimize the distances between the representations of the corresponding pairs while repelling those from the query the non-corresponding candidates' representations in a joint space. This can be done through classification, by extending BERT's *Next Sentence Prediction* (see [88]) to a cross-modal matching task, forcing the network to find co-occurrent information in both modalities [205, 361, 468]. Alternatively, this can naturally extended into a contrastive setting. In retrieval, it is common to use two anchors (which form the positive pair) and two negative sets, one from each modality. In VT literature we find these losses instantiated through a combined hinge loss [215, 130] or *Bi-directional Max-Margin* [118, 96, 289], which enforce similarity for true pairs to be higher than that of negative pairs, by at least a given margin. Alternatively, InfoNCE is also used [231, 290], normalizing the similarity score of positive pairs by that of a set of negative pairs, effectively forcing the network to learn similar representations for correctly paired samples and viceversa for negative ones. While the most common approach is to align final output representations, some works leverage hierarchical contrastive losses, which also align intermediate feature representations [130, 231]. During inference, the aligned representations are fixed, so the task simply becomes a search (*e.g.*, K-Nearest Neighbors) to find the top-k examples most similar to a given query within the database of candidates' pre-computed representations.

One interesting variation of this pipeline is [259], in which the alignment is performed on the outputs of a siamese two-stream video-and-language CNN for faster retrieval instead. Then, a decoder-only Transformer fed with the text as input and CNN-based video features as context re-ranks the previously top-k retrieved elements using the decoding likelihood score. In a similar spirit, [289] also leverages the likelihood of a language-based decoder-only Transformer during training as a loss that measures how well the query language caption can be reconstructed from the weighted combination of the features from all the non-corresponding videos in the batch. Those weights are based on the similarity of the query caption with the captions of those other videos. Finally, [118] aligns at the same time video, audio, and recognized speech with a language caption. The language-video, language-audio, and language-speech similarities are aggregated before contrastive alignment with a mixture of weights governed by the content of the caption (*i.e.*, the language-video similarity is given more weight if the caption refers to something that is more salient in the video than in the other modalities).

4.5.4 Object-centric tasks

Tasks such as object detection, tracking, and segmentation are inherently object-centric in nature, and recent works [160, 172, 434, 254, 464] within these tasks have begun to leverage temporally coherent object representations. These designs tend to focus on per object outputs. A large part of the information within a given frame is redundant (as mentioned in Section 4.3.1), therefore leveraging known and relevant content from previous frames can be used to focus the global attention to object relevant cues. As such, these approaches typically leverage memory or recurrency (see Section 4.3.2) to correlate object information temporally.

Recent works [172, 434] leverage a set of "messenger" tokens to relay contextual information between frames within abrupt hierarchical architectures (see Section 4.3.1), resulting in hierarchical-like approaches for temporal information sharing. IFC-transformer [172] processes the relationship of frame-specific aggregation tokens (akin to `[CLS]`) by interleaving isolated encoder layers. The temporally enhanced tokens are then brought back to their neighborhoods to communicate temporal information. Instead, TeViT [434] shifts individual tokens between consecutive frames to achieve object specific aggregation, effectively accumulating temporal information across different steps sequentially. The work of [436], performs both long-term and short-term information sharing in parallel, being later concatenated. Due to varied framerate and inter-frame changes in content, smoothness cannot be guaranteed through long-term alone, thus short-term attention is computed on a smaller spatiotemporal neighbourhood to ensure smooth and continuous predictions between frames. With regards to tracking, most approaches leverage recurrent architectures. On the one hand, some use object specific tokens derived from the output bounding box of previous time-steps. Both spatial and size information can then be reccurently propogated[254, 464] or used to produce region-specific attention for each bounding box concurently between frames[160]. On the other hand, and inspired by recent works in vision Transformers (particularly DETR [54] and its variants), some works propagate "detection tokens" to the next frames in a recurrent fashion. DETR is an object detection tool which enhances learned input query tokens by attending to image features through a Transformer decoder (similar to query-driven compression, see Section 4.3.1). The augmented query tokens are then each decoded to either a detected object or background. The works of [464, 254] extend this design to video by concatenating detection tokens from previous frames to the existing learned query tokens, in addition to storing each detection in memory for increased robustness to occlusions in the video sequence.

Different to these, some architectures leverage the object features to aggregate contextual information such as [304, 160], which attempt to enhance existing representations with more focus on object centric information. ORViT [160] leverages auxiliary bounding boxes at each Transformer layer, whereas the GroupFormer [217] uses them to isolate objects into specific action classification branches. Simply put, these approaches try to reduce redundancy by focusing on a few object-centric tokens that attend to the whole input. Unlike recurrent and memory based designs, these types of approaches do not aim for a computationally efficient design, but rather efficient in the sense of information-rich representation, leveraging object-centric information in addition to global context information.

4.5.5 Low-level tasks

Given the high dimensionality of video data, video generation tasks are quite challenging, and not many video Transformers try to address them. In particular, [408, 316] tackle future frame prediction, [409] generates grayscale video from event-based videos and [450, 230] perform video inpainting. Most of these propose to embed a Transformer in between some form of convolutional auto-encoder, in order to evolve representations between encoder and decoder [409, 450, 230]. The only exception is [408], which performs local attention and generate

video autoregressively one pixel-channel at a time. Interestingly, [230] outperforms [450] in all tested benchmarks for inpainting by using an overlapping patch tokenization strategy.

4.5.6 Segmentation

Most works in segmentation leverage temporal relations to refine intermediate feature representations[404, 405, 441]. Most notably, [405] leverages the Transformer's ability to view the entire sequence to include an auxiliary loss where representations of individuals are matched temporally. In this way, the network effectively learns to implicitly track objects and leverage temporal fine-grained information. Another example is [437], that leverages a word-visual attention mechanism allowing a textual query to attend to specific content in multiple spatial scales, performing segmentation based on said query.

4.5.7 Summarization

Few works have used Transformers for the task of video summarization by predicting frame-wise importance scores. We find two key trends when solving this task through VTs: the use of RNNs as an intial step [233, 364] and using individual frames to attend to aggregated subsets of the video, either from a GRU [364] or by using a masked Transformer [225].

4.6 Performance on video classification

The task of video classification has attracted the most research in Transformers for video, given the generality of the task and availability of large datasets for training and evaluation, things that allow for a more comprehensive performance analysis. Next, we overview the particularities of video classification (Section 4.6.1) and then analyse VT state-of-the-art performance on it (Section 4.6.2).

4.6.1 Video classification

Video classification aims to predict the class of a given input sequence of frames. For the task, a VT will encode descriptive high-level global representations of a given sample. Then, some linear layers followed by a softmax provide with a class-score probability distribution. The category with maximum probability should match the ground-truth class. VTs competing to become state-of-the-art in classification tend to be standalone (*i.e.*, use minimal embedding), and thus will be the ones we cover. Next, we present the benchmarking datasets, experimental protocols, and details on clip sampling.

Evaluation datasets. The most popular dataset is the large-scale *Kinetics-400* (K400) [57], consisting of 306K 10-second clips and 400 manually annotated human actions categories with at least 400 examples per class. Kinetics-600 (K600) – an extension of K400 with 495K clips and 600 classes – is only used for pre-training, but not for evaluation. K400/K600 are however known to be appearance-biased [446]. To better assess the modeling of more complex temporal dynamics, most works are also evaluated on *Something-Something v2* (SSv2) [139,

241]. SSv2 is an egocentric human action dataset where some of the categories can only be distinguished by having an understanding of the arrow of time (*e.g.*, "Moving [sth] away from [sth]" vs "Moving [sth] closer to [sth]"). SSv2 consists of 220K videos of duration ranging from 2 to 6 seconds and 174 fine-grained categories.

Experimental protocols. We find two learning protocols being followed: training from scratch or pre-training the model first. *Training from scratch* is rarer because of the size of the models (especially, their larger variants). When following *pre-training*, the weights learnt during a first stage are used to initialize the model that is to be trained in the downstream dataset/task. Common pre-training strategies for video classification are (a) image-based pre-training on ImageNet, and either (b) supervised or (c) a self-supervised video pre-training (generally on video datasets larger than the downstram one, *e.g.*, K600 for evaluating K400 and K400/K600 for SSv2). After initialization, the models are trained on the downstream dataset, fine-tunning existing weights and adapting new ones.

Clip sampling. Models are fed with trimmed video clips. These are relatively short, with a number of frames T' typically 8 to 64 frames and fixed spatial resolution $S' = H' \times W'$ pixels (often $H' = W' = 224$, hence shortened to $S' = 224^2$, see "Input" in Tables 4.3 and 4.4). However, to make sense of these numbers, and specially T', it is crucial to consider the temporal stride τ, *i.e.*, the step between frames when sampling them from the original video. A larger τ extends the temporal span of the clip w.r.t. the video without incurring in extra computation costs, while also reducing the redundancy among otherwise nearby sampled frames. For instance, with $\tau = 4$ and $T' = 64$, a clip covers a temporal span equivalent to a densely ($\tau = 1$) sampled clip of 256 frames. Importantly, τ must be chosen factoring in the temporal resolution of videos (*e.g.*, $\sim$25 FPS in K400) considering fine-grained motion modeling will be sacrificed in favour of context.

Views. The clips generated can be regarded as *temporal views* (related to the views described in Section 4.4.2, which are used for some methods during pre-training). During training, one temporal view per video is gathered at a random temporal position. These are constructed with fixed size $T' \times S'$ and stride τ. For inference, most models follow a multi-view approach: the classification decision for the video is achieved by averaging the prediction obtained from different spatiotemporal crop views.

4.6.2 Comparison among state-of-the-art models

To draw comparisons we consider the factors defined by the columns of Tables 4.3 and 4.4. Among those, the most interesting one to study is perhaps the pre-training strategy, which will drive the rest of the section, separately analysing K400 and SSv2.

K400: training from scratch. Doing so, we only find MViTv2 [219] and its predecessor MViT [106]. The main difference between the two is the inclusion of an extra residual pooling connection and the use of relative positional encoding. With these, "MViTv2-B 32@3" (82.9%)

performs better than its older counterpart "MViT-B 32@3" by +2.7%. In fact, it also surpasses "MViT-B 64@4" – which has an increased temporal receptive field (2.6x) – by +1.7%. Later in [406], the same authors explored different initialization strategies and showed overfitting of the larger variants of MViTv2 when not using effective initialisation. This can be seen for "MViTv2-L↑", with increased spatial resolution (from 224 to 312) and compute (from 51 MP to 218 MP), performing worse (-0.7%) than "MViTv2-B 32@3'. Although this is to be expected, the smaller variants are still able to learn from scratch successfully – as we will see later, even better than 3D ConvNets. Given the need of pre-training for larger models, we next discuss the two most popular strategies in the context of K400 and demonstrate its large positive effect (*e.g.*, "MViTv2↑ 32@3" boosts its results from 82.2% to 85.3% by leveraging image-based pre-training).

K400: image pre-training. The majority of VTs pre-train on either ImageNet-1K ("IN"), ImageNet-21K ("IN21"), or JFT-300M ("JFT"). IN and IN21 consist of 1K and 21K classes and over 1.2M and 14M examples respectively, whereas JFT is a non-public dataset with 300M multi-label images and 18,291 non-mutually-exclusive labels. Other works have been using their own image datasets or extending public ones. For instance, "Video-SwinV2-G" [234] (86.8%), being the best performing model, extended IN21 (14M images) with a private dataset of images ("P" in Tab. 4.3), totalling 70M samples. Close second is "MViTv2-L↑ 40@3" [219] (86.1%), with weights pre-trained exclusively on IN21 while only dropping by -0.7% with respect to the first one, but with 14x less parameters. Third is "MTV-H"[430] (85.8%), this one pre-trained on JFT with 300M images. Unfortunately, in this work, the authors used JFT to pretrain their largest models ("MTV-L" and "MTV-H") and IN21 to train "MTV-B"/"MTV-B (320)", therefore not validating the actual contribution of JFT w.r.t. IN-21K for any of the variants, making it difficult to discern the actual contribution of the model scaling. Also TokenLearner [331] completely relies on JFT for all the experiments, with its best model "TokenLearner 16at18 (L/10)" (85.4%) coming fourth. It was ViViT [17] that showed how the same model variant trained on JFT, "ViViT-L (JFT)" (83.5%), was considerably improving upon the same variant pre-trained on IN-21K ("ViViT-L"), by +1.8%. It is, hence, of great merit that "MViTv2-L↑ 40@3" (86.1%) still surpasses, respectively by +0.3% and +0.7%, the results of "MTV-H" and "TokenLearner 16at18 (L/10)". It is true that compared to those, MViTv2 variant utilizes larger spatial (312^2, versus 224^2 and 256^2 pixels) and temporal receptive field (120 vs 64 frames), but the number of TFLOPs and the amount of pre-training data to process are still both lower: 14 MP versus 300 MP in JFT for MTV and TokenLearner, and 42 TFLOPs versus 47 and 48 TFLOPs.

In terms of cost-effectiveness, we find "UniFormer-B" [214] (83.0%), "SCT-L" (83.0%) [451], "Direcformer" [378] (82.8%) – this one based on [31]-, and "MViTv2-S" (82.6%) [406]. These models only suffer a drop between -3.1% and -3.3% of accuracy but between 10x and 70x less FLOPs w.r.t. "MViTv2-L↑ 40@3".

K400: video (self-supervised) pre-training. An emerging trend in the literature is to perform all SSL pre-training, fine-tunning and evaluation on the same dataset [406, 374, 396,

	Pretrain	Name	Ref.	Input	TF × v_t × v_s	MP.	Acc.
ConvNets	-	SlowFast (R101+NL)	[110]	16@8 × 256^2	0.23 × 10 × 3	60	79.8
		X3D-XXL	[109]	16@5 × 312^2	0.19 × 10 × 3	20	80.4
	IG65 (video)	R(2+1)D-152	[128]	32@1 × 128^2	0.25 × 10 × 1	118	81.3
		ir-CSN-152	[376]	32@2 × 224^2	0.10 × 10 × 3	NA	**82.6**
Scratch	-	MViT-S	[106]	16@4 × 224^2	0.03 × 5 × 1	26	76.0
		MViT-B		32@3 × 224^2	0.17 × 5 × 1	37	80.2
		MViT-B		64@4 × 224^2	0.46 × 3 × 3	37	81.2
		MViTv2-S	[219]	16@4 × 224^2	0.06 × 5 × 1	35	81.0
		MViTv2-B		32@3 × 224^2	0.23 × 5 × 1	51	**82.9**
		MViTv2-L↑	[406]	32@3 × 312^2	2.06 × 5 × 3	218	82.2
Image pretr. (I)	IN	UniFormer-B	[214]	16@4 × 224^2	0.10 × 4 × 1	50	82.0
		UniFormer-B		32@4 × 224^2	0.26 × 4 × 3	50	83.0
		Swin-B	[236]	32@2 × 224^2	0.28 × 4 × 3	88	80.6
	IN21	SCT-L	[451]	24@10 × 224^2	0.34 × 4 × 3	60	83.0
		Swin-B	[236]	32@2 × 224^2	0.28 × 4 × 3	88	82.7
		Swin-L↑		32@2 × 384^2	2.11 × 10 × 5	200	84.9
		TS	[31]	8@16 × 224^2	0.20 × 1 × 3	121	78.0
		ViViT-L-FE	[17]	32@2 × 224^2	3.98 × 1 × 3	352	81.7
		VTN-3 (Aug)	[268]	250@1 × 224^2	4.22 × 1 × 1	114	79.8
		DirecFormer	[378]	8@32 × 224^2	0.20 × 1 × 3	124	82.8
		Mformer	[288]	96@3 × 224^2	0.96 × 10 × 3	NA	81.1
		Mformer↑		64@4 × 336^2	1.19 × 10 × 3	NA	80.2
		X-ViT	[47]	16@1 × 224^2	0.28 × 1 × 3	92	80.2
		X-ViT		16@1 × 224^2	0.28 × 2 × 3	92	80.7
		MTV-B	[430]	32@2 × 224^2	0.4 × 4 × 3	310	81.8
		MTV-B↑		32@2 × 320^2	0.96 × 4 × 3	310	82.4
		RViT-XL	[432]	64@NA × 224^2	11.90 × 3 × 3	108	81.5
		MViTv2-S	[406]	16@4 × 224^2	0.07 × 10 × 1	36	82.6
		MViTv2-L↑		32@3 × 312^2	2.06 × 5 × 3	218	85.3
		MViTv2-L↑	[219]	40@3 × 312^2	2.83 × 5 × 3	218	86.1
	(IN-21 + P) (SSL)	SwinV2-G↑	[234]	8@NA × 384^2	NA × 4 × 3	3 K	**86.8**
	JFT	ViViT-L-FE	[17]	32@2 × 224^2	3.98 × 1 × 3	352	83.5
		ViViT-H		32@2 × 224^2	3.98 × 4 × 3	352	84.9
		MTV-L	[430]	32@2 × 224^2	1.50 × 4 × 3	NA	84.3
		MTV-H		32@2 × 224^2	3.71 × 4 × 3	NA	85.8
		TokenLearner	[331]	64@1 × 256^2	4.08 × 4 × 3	450	85.4
Video pretr. (V)	K400 (SSL)	LSTCL (Swin-B*)	[396]	16@8 × 224^2	0.36 × 5 × 1	88	81.5
		MaskFeat-S	[406]	16@4 × 224^2	0.07 × 10 × 1	36	82.2
		MaskFeat-L↑		32@3 × 312^2	2.06 × 5 × 3	218	86.3
		MaskFeat-L↑		40@3 × 312^2	2.83 × 4 × 3	218	86.4
		MaskFeat-L↑↑		40@3 × 352^2	3.79 × 4 × 3	218	86.7
		VideoMAE (ViT-B)	[374]	16@4 × 224^2	0.18 × 5 × 3	87	80.9
		VideoMAE (ViT-L)		16@4 × 224^2	0.60 × 5 × 3	305	84.7
		VideoMAE↑ (ViT-L)		32@4 × 320^2	3.96 × 5 × 3	305	85.8
	K600 (SSL)	MaskFeat-L	[406]	16@4 × 224^2	0.34 × 10 × 1	218	85.1
		MaskFeat-L↑↑		40@3 × 352^2	3.79 × 4 × 3	218	**87.0**
I+V	IN + K400 (SSL)	SVT (TS)	[319]	8@NA × 224^2 + 64@NA × 96^2	0.20 × 1 × 3	121	78.1
	IN (SSL) + K400 (SSL)	BEVT	[398]	16@NA × 224^2	0.28 × 4 × 3	88	80.6
		BEVT (Dall-E tknzr.)		16@NA × 224^2	0.28 × 4 × 3	88	81.1

↑: increased spatial resolution.
"IN21 + P": extension of IN21 with a private dataset (70M images in total).

TABLE 4.3: Accuracy (top-1) on Kinetics-400. "Pretrain": pre-training strategy; "Input": temporal and spatial size of the views; "TF": TFLOPs; v_t and v_s: the number of temporal and spatial views; and "MP": parameters ($\times 10^6$).

398]. "MaskFeat-L↑ 40@3" [406] reaches 86.4%, thus showing the contribution of MaskFeat (SSL) pre-training compared with supervised training on the same architecture, *i.e.*, MViTv2, by +0.3%. That result of MaskFeat is also only -0.2% behind the best image-based pre-trained model (*i.e.*, "Video-SwinV2-G"). Then, "MaskFeat-L↑↑ 40@30" by switching K400 with K600 and slightly increasing the spatial resolution from 312 to 352 (still lower than 382 of "Video-SwinV2-G"), the model obtains state-of-the-art results (87%), outperforming any of the image pre-trainings. VideoMAE [374] comes second in this category consisting of a ViT backbone with 3D inflation of the patch embeddings. This outperforms all image-based pre-trained models, except for "Video-SwinV2-G". Thus it seems learning motion priors during pre-training has a very positive effect on performance when targeting video classification.

K400: ConvNets. For the sake of completeness, we compare VTs to 3D ConvNets, which were state-of-the-art right until VTs managed to surpass them. See how "MViTv2-S" (81.1%) trained from scratch, exceeds the performance of comparable ConvNets: "SlowFast R101+NL" (79.8%) and "X3D-XXL" (80.4%). This might be attributable to the higher temporal fidelity of MViTv2 being more profitable than extra context – at least on short videos. The number of views for testing were also higher for both (30 versus 5 in " MViTv2-S"). Nonetheless, it also consumes 18x - 22x less TFLOPs and works on a smaller spatial resolution (224 only, versus 256 or 312). Switching to "MViTv2-B 32@4" (82.9%), we see how trained from scratch this model does better than ConvNets pre-trained on the very-large weakly-annotated video dataset IG65 (*i.e.*, 'R(2+1)D-152" [128] (81.3%) and "ir-CSN-152" [376] (82.6%)), even when using half the views.

Moving to the study of SSv2, we found none of the works train from scratch. Another thing to note is the number of temporal views used because of the shorter duration of SSv2 videos compared to Kinetics. Despite that, the temporal dynamics are harder to capture as we will see next.

SSv2: image pre-training. Although less common than for K400, there are works that pre-train on image datasets. Among the ones using IN, "DirecFormer" [378] (64.9%) is the one performing the best. It surpasses its own backbone ("TS" [31]) in all the variants by forcing the learning of temporal order of shuffled input frames via auxiliary SSL. "TIME" [446] is another one using auxiliary SSL ablated with different VT backbones. This one, not so much competing in performance with larger model variants, still points out the effectiveness of temporal guidance for image-based pre-trained models when transferred to the downstream video task. Finally, trained on IN-21K, "X-ViT" (66.4%) [47] is the absolute winner in this category. Unfortunately, by focusing on efficiency alone, it is not able to compete with heavier models that are supervisedly pre-trained on K400.

SSv2: video (supervised) pre-training. It is quite common reusing supervisedly trained checkpoints on Kinetics by transferring them to SSv2 for fine-tuning. These have often also been pre-trained on IN-1K/IN-21 before Kinetics. However, to better disentangle video and image contributions, we focus first on video-only pre-training models, and particularly on

	Pretrain	Name	Ref.	Input	TF × v_t × v_s	MP.	Acc.
CN	IN	TDN (R101)	[110]	8@1 × 256^2 + 16@1 × 256^2	0.2 × 1 × 3	198	**69.6**
Image pretr. (I)	IN	TS*	[446]	8@NA × 224^2	NA × 1 × 3	121	62.1
		Mformer*		8@NA × 224^2	NA × 1 × 3	NA	63.8
		TIME (TS*)		8@NA × 224^2	NA × 1 × 3	121	63.7
		TIME (Mformer*)		8@NA × 224^2	NA × 1 × 3	NA	64.7
		DirecFormer	[378]	8@32 × 224^2	0.20 × 1 × 3	124	64.9
	IN21	TS	[31]	8@16 × 224^2	0.20 × 1 × 3	121	59.5
		TS-HR		16@16 × 448^2	1.70 × 1 × 3	121	62.2
		TS-L		96@4 × 224^2	2.38 × 1 × 3	121	62.4
		ViViT-L	[17]	32@2 × 224^2	3.98 × 1 × 3	352	**65.9**
		X-ViT	[47]	16@NA × 224^2	0.28 × 1 × 3	92	66.2
		X-ViT		32@NA × 224^2	0.42 × 1 × 3	92	66.4
Video pretr. (V)	K400	MViT-B	[106]	32@3 × 224^2	0.17 × 1 × 3	37	67.1
		MViT-B		64@4 × 224^2	0.46 × 1 × 3	37	67.7
		MViTv2-B	[219]	32@3 × 224^2	0.23 × 1 × 3	51	70.5
	K600	MViT-B	[106]	32@3 × 224^2	0.17 × 1 × 3	37	67.8
		MViT-B-24		32@3 × 224^2	0.24 × 1 × 3	53	68.7
	K400 (SSL)	LSTCL (Swin-B*)	[396]	16@8 × 224^2	0.36 × 5 × 1	88	67.0
		MaskFeat-L↑	[406]	40@3 × 312^2	2.83 × 4 × 3	218	74.4
	K600 (SSL)	MaskFeat-L↑	[406]	40@3 × 312^2	2.83 × 1 × 3	218	75.0
	SSv2 (SSL)	VideoMAE (ViT-B)	[374]	16@2 × 224^2	0.18 × 2 × 3	87	70.6
		VideoMAE (ViT-L)		16@2 × 224^2	0.60 × 2 × 3	305	74.2
		VideoMAE (ViT-L)		32@2 × 320^2	1.44 × 1 × 3	305	**75.3**
Image + video pretr. (I + V)	IN + K400	UniFormer-B	[214]	16@4 × 224^2	96.67 × 1 × 3	50	70.4
		UniFormer-B		32@4 × 224^2	259.00 × 1 × 3	50	71.2
	IN21 + K400	Swin-B	[236]	32@2 × 224^2	0.28 × 1 × 3	88	69.6
		X-ViT	[47]	16@1 × 224^2	0.28 × 1 × 3	92	67.2
		MViTv2-B	[219]	32@3 × 224^2	0.23 × 1 × 3	51	72.1
		MViTv2-L↑		40@3 × 312^2	2.83 × 1 × 3	218	**73.3**
		Mformer	[288]	96@3 × 224^2	0.96 × 1 × 3	NA	67.1
		Mformer↑		64@4 × 336^2	1.19 × 1 × 3	NA	68.1
		MTV-B	[430]	32@2 × 224^2	0.40 × 4 × 3	310	67.6
		MTV-B↑		32@2 × 320^2	0.96 × 4 × 3	310	68.5
		RViT-XL	[432]	64@NA × 224^2	35.70 × 1 × 3	108	67.9
		MViTv2-S	[219]	16@4 × 224^2	0.06 × 1 × 3	35	68.2
		ORViT MF-L	[160]	32@4 × NA	1.26 × 1 × 3	148	69.5
	IN + K600	UniFormer-B	[214]	16@4 × 224^2	96.67 × 1 × 3	50	70.2
		UniFormer-B		32@4 × 224^2	259.00 × 1 × 3	50	71.2
	IN21 + K600	SCT-L	[451]	24@10 × 224^2	0.34 × 4 × 3	60	68.1
	IN + K400 (SSL)	SVT (TS)	[319]	8@NA × 224^2 + 64@NA × 96^2	0.20 × 1 × 3	121	59.2
	IN21 + K400 (SSL)	MaskFeat-L	[406]	40@3 × 224^2	2.83 × 1 × 3	218	**73.3**
	IN (SSL) + K400 (SSL)	BEVT	[398]	16@NA × 224^2	0.32 × 1 × 3	88	70.6
		BEVT (Dall-E tknzr.)		16@NA × 224^2	0.32 × 1 × 3	88	71.4

*: re-implementation.

↑: increased spatial resolution.

TABLE 4.4: Accuracy (top-1) in Something-Something v2. "Input": temporal and spatial size of the views; "TF": TFLOPs; v_t and v_s: the number of temporal and spatial views; "MP": parameters ($\times 10^6$); and "Pretrain": pre-training strategy.

those relying on K600. Looking at "MViT-B 32@3" (67.8% pretrained on K600) and "MViT-B 64@4" (67.7% pretrained on K400), with temporal receptive field of 96 to 128 frames respectively, we see there is no improvement in SSv2 by extending temporal context, but slightly better performance when keeping finer temporal resolution (stride 3 instead of 4). Even more interesting is that the deeper "MViT-B-24 32@3" (with 24 layers) outperforms +0.9% upon the 12-layered 32@3 variant. This suggests more complex temporal dynamics might require not necessarily increasing temporal resolution, but higher abstract spatiotemporal semantics being modelled. That or advancements on architectural designs to better model those without going deeper, as done by "MViTv2-B 32@3" (70.5%) also with 12 layers. Finally, if we have a look at models that leverage image-based pre-training before Kinetics, we find further improvement (*e.g.*, "MViTv2-B 32@3" from 70.5% to 72.1%). What does not seem to be as useful, according to "UniFormer" variants, is to switch from K400 to K600.

SSv2: video (self-supervised) pre-training. The only model pre-training on SSv2 is VideoMAE ("VMAE"), which turns out to be the best performing one. In particular, "VMAE (ViT-L) 32@2" (75.3%) slightly improves upon "MaskFeat-L↑ 40@3", despite it being self-supervisedly pre-trained on K400 (74.4%) or K600 (75.0%). It does so with almost half the temporal context, half the FLOPs, and – importantly – with less data. All in all, VideoMAE and MaskFeat seem to point out pixel- and feature-based MTM approaches compare favourably with "SVT" (instance-based invariance learning) or "BEVT" (quantization-based MTM) despite the latter are also using image-based pre-training.

4.6.3 Discussion on performance

We have introduced the task of video classification and analysed the performance of state-of-the-art models on Kinetics-400 and Something-Something v2. Our main finding was that the pre-training strategy was the biggest factor influencing the performance of VTs for video classification, thus the following discussion will address three questions related to this: (1) *Can Video Transformers be trained from scratch?*, (2) *Which is the best pre-training strategy?*, and (3) *How can we effectively model stronger spatiotemporal dynamics?*.

For the smallest models, *training from scratch* seems to be doable. In particular, MViT [106] and MViTv2 [219] are able to, respectively, compete with and slightly surpass 3D ConvNets trained from scratch. In fact, MViTv2 even outperforms those pre-trained on very large weakly-annotated video datasets (*e.g.*, IG-65M). In particular, we attribute the success of those to the locality bias they infused (via the local pooling-based progressive aggregation – discussed in Section 4.3.1) –, which allow these models to go deeper without exploding in computational complexity while still keeping their self-attention operation global. However, training from scratch seems to be the least desirable strategy to follow.

Among *pre-training strategies*, video-based ones, either supervised (*e.g.*, on K400/K600 before fine-tuning SSv2) or self-supervised, are superior to image-based pre-training alone. Image-based supervised pre-trained models seem to be able to partially compensate the lack of temporal modeling with appearance diversity by leveraging huge – often non-public –

image datasets (*e.g.*, JFT [430, 331, 17] or extensions of IN21 [234]). Alternatively, image-based self-supervised learning only competes with video pre-training when leveraging prohibitevely large models [234]. However, starting from very diverse and general appearance features will not harm the modeling of time in later stages, but serve as a good initialization for subsequent video-based pre-training (*e.g.*, all those works that combine IN/IN21 and K400/K600 before fine-tuning on SSv2) or fine-tuning stages with temporal SSL auxiliary losses (*e.g.*, [378, 446]). Nevertheless, we can see how self-supervised video pre-training surpasses supervised regimes. In particular, MaskFeat [406] (with MViTv2 [219] backbone) and VideoMAE [374] (with a plain ViT [17]) outperform those pre-trained on video in a supervised way.

For the *successful modeling of spatiotemporal patterns*, Masked Token Modeling stands out (see Section 4.4.3). Concretely, MaskFeat (feature-based MTM) obtains the best results on K400 and is second best on SSv2, which is dominated by VideoMAE (pixel-based MTM). Interestingly, these models do not require extra data or manual annotations to surpass all other models, being able to self-supervisedly pre-train on the evaluation dataset itself. Unfortunately, instance-based invariance learning (*e.g.*, [319, 396]), being that popular for image representation learning, heavily underperforms compared to MTM for video classification.

Apart from the importance of pre-training, other findings in Section 4.6.2 we want to highlight are: first, that the modeling of the complex spatiotemporal dynamics seem to benefit more from deeper models and temporal fidelity than extended temporal spans; second, that naive adoptions of image and NLP models (*e.g.*, VTN [268], which leverages the image-based ViT [29] to model space and the language-based Longformer [29] to model time) might not work that well; and third, that although joint self-supervised learning on image and video (*i.e.*, BEVT [398]) is promising, still has a long way to go.

4.7 Final Discussion

In this chapter we have comprehensively analysed trends and advances on leveraging Transformers to model video.

Complexity. Given the inherent complexity of Transformers and the great dimensionality of videos, most changes focus on handling the computational burden. This is done transversally across the various stages of the VT pipeline. We find this is most generally addressed with frozen embedding networks, easing Transformer learning through the provided inductive biases and reducing input dimensionality. The Transformer in this context is used to enhance these representations through long-range interactions, which seems enough to boost performance in many areas of application. However, this trend alone may be limiting the potential of Transformers to learn non-local low-level motion cues. We are excited to see novel VT designs (*e.g.*, MViT [106]) which greatly reduce complexity thanks to the inductive biases embedded in the Transformer itself (sometimes becoming lighter than CNN counterparts, see Section 4.6.2). We also see great promise in MTM when separating the representation

learning from the reconstruction which is done by an additional decoder discarded after pre-training [374]. This separations allows the (deeper) encoder to only leverage unmasked tokens, which greatly alleviates training complexity when using large masking ratios. Crucially, this sacrifices the possibility to leverage certain designs for the VT, as input structure is lost (*e.g.*, local or hierarchical approaches may not find enough tokens in a given neighborhood to learn valuable representations).

Spatial redundancy and temporal fidelity. Modeling temporal interactions requires special considerations not present when only modeling appearance (*i.e.*, with image Transformers). On the one hand, the highly redundant appearance information in videos [461, 374] makes it difficult to model information-rich representations that avoid repeatedly representing similar or same sub-representations. It has been proven that pure attentional models loose expressivity with depth, collapsing towards uniform attention in deeper layers [90, 91, 178]. It further seems that this smoothing of the attention matrix is accompanied by highly uniform token representations and even redundant weight matrices [65]. Proper handling of video redundancy is crucial in VTs, where we hypothesize these observations may get exacerbated. On the other hand, few exceptions aside, many current designs and SSL approaches directly inherit from image approaches without careful consideration of the nuances that come with time, making them strongly biased to learn appearance features. As we have seen, allowing temporal features to form at both low- and high-level while accounting for the necessary temporal fidelity is also critical. In this sense, reducing redundancy for video should be mostly targeting appearance features.

Key advancements on VTs. Regarding *architectural choices*, we find progressive hierarchical approaches to stand out. They carefully consider non-local temporal contexts before spatial aggregation. This effectively tackles the redundancy problem while avoiding early aggregation problems that hinder learning of fine-grained motion features. However, to properly handle long-range interactions without loosing temporal fidelity memory-based approaches with adequate sampling or aggregation techniques may be crucial. Regarding *self-supervised learning*, MTM forces to leverage global spatiotemporal semantic contexts through high masking ratios when solving local token-wise predictions. By doing so, it is driven to learn both motion and appearance cues necessary to solve the task. Nevertheless, we look forward to further developments in sampling techniques for instance-based contrastive approaches that skew from appearance biases towards motion-specific features.

Inductive biases. As we have seen, inductive biases are a pivotal aspect for all facets of VTs. They alleviate the need for data by providing stronger cues that the Transformer can pick up faster. Frozen *embedding networks* could be regarded as infusing task-specific biases, as the Transformer is bounded to learn on the provided representations, which in turn are dependant on the pre-training auxiliary task. Some examples include detected bounding boxes of objects [132, 160], higher-level (action) features [468], or scene, motion, OCR and facial features, among others [118]. We have also seen how most *architectural designs* infuse

some inductive biases to aid training the Transformer. However, in this regard, VT literature so far is limited when considering infusing motion specific biases that help the network to pick up relevant spatiotemporal cues. Just two works deviate from this trend. Motionformer [288] proposes trajectory attention to reason about aggregated object or region representations through implicit motion paths in both time and space. Differently, OrViT [160] leverages separate motion and appearance streams. The former learns trajectories of individual objects or regions which get later added to patch-wise token representations of the whole video appearance, effectively infusing motion into it. Finally, beside locality biases or invariance to perturbations induced by different *training losses*, we deem interesting to highlight works infusing causality biases by training the network to sort shuffled video sequences [378, 446]. Or in a different vein, the work in [144], which combines the benefits of both CNNs and Transformers for video learning through a siamese distillation setting, effectively inducing CNN locality biases into the Transformer.

4.7.1 Generalization

It has been shown that vision Transformers are robust to various perturbations [33, 242], suggesting they may be better able to form abstract semantic representations [454], probably due to their ability to leverage non-local contexts [292]. These findings point towards Transformers favouring out-of-distribution (OOD) generalization [154]. A few VTs have studied this on OOD data [468, 289, 409, 230, 340, 233] or evaluated the learned features in other settings [130, 361, 441, 367], showing consistent results. Nevertheless the issue of generalization of video may entail studying other aspects that are still under-researched. For instance, we hypothesize that generalizing to varied frame sampling rates may require further training or conditioning the network to said rates such that it may become robust. We do observe, however, that some existing work may display capabilities to generalise to unseen sequence lengths, as we discuss next.

Unseen sequence length. One issue to account for when processing sequences of unseen length is positional encodings. While we expect them to generalize to shorter sequences, they may have trouble when dealing with longer ones (which may be desirable to provide with extended temporal fidelity during deployment), as no positional information is present to account for them. We find few VTs showing that PEs can easily be extended by fine-tunning the model on longer sequence lengths [7, 106]. Recent VT works have also seen promising results when leveraging input conditioned RPEs [214] or by learning a small network that computes log-scale relative positional biases [234], which pose a great potential to easily generalize to unseen sequence lengths. Similarly, long-range modeling architectures could also handle sequences of any given length, as they process inputs sequentially within fixed windows, but they may require RPEs [417].

Multi-modality. Video is inherently multi-modal (*i.e.*, contains visual and aural information), which could be leveraged to learn more general representations. As we have repeatedly explored, the lack of inductive biases makes Transformers very versatile tools to handle

this multi-modality. It has been found that high-level semantic features learned by language-based Transformers generalize to other modalities [363, 238]. In the context of VTs we find VideoBERT [361], where a pre-trained language BERT [88] model is used as initialization for the video stream, showing promising results in this direction. Lately, there has been a great interest in using these architectures to solve multi-modal tasks (see [330] for a complete review). We hypothesize that the lack of inductive biases may allow Transformers to learn shared multi-modal representation spaces that exhibit better generalization capabilities. When targeting video-only tasks (*e.g.*, tracking, segmentation, classification) we see potential in multi-modal SSL to learn such spaces. We find a few VTs leveraging instance-based multi-modal learning approaches [130, 205, 360, 215] to align representations from various modalities. For instance, [7] successfully performs heavy downsampling of video by aligning it with audio and textual modalities, the model in [290] learns to attend to the spatial sources of audio within the video by aligning audio with visual crops. Interestingly, this alignment is further enforced in some works by sharing weights between Transformer streams modeling different modalities [7], sometimes even showing improved results compared to not sharing [205]. As pointed out in [334], alignment has proven to be very useful for video (at least in the context of classification) outside of Transformers, specially when pairing video with audio or text.

4.8 Conclusion

In this chapter, we have presented an overview of Video Transformer across all elements: input pre-processing, architectural designs, and training strategies. We have also detailed multi-modal fusion mechanisms, per-application trends, and provided an in-depth comparison of performance on action classification. Despite its recent appearance, VTs have already seen an explosion in attempts to capture the nuances of spatiotemporal video representation learning, but we believe there is still a long way to go. We hope that with the release of this study, we can motivate further research on better architectures, more motion-focused SSL, and more careful integration of Transformers with CNNs to better exploit the advantages each of them provides.

With regard to our exploration of Video Transformers for human analysis, we have gained a deeper understanding of the type of changes that would be required to capture fine-grained long-term relationships that could benefit human interaction modeling. In this sense, we believe that the Dyadformer could benefit from more granularity in the tokens used, as well as some form of progressive aggregation to integrate it. Currently, we have already started to test out some of the observations from this chapter, namely, end-to-end training of the embedding layers, as well as some potentially useful SSL losses tailored for our case. We detail these next in Section 5.1.

Chapter 5

Conclusions

The current dissertation has presented the work done in the past years to advance research toward more powerful and humane technologies. We have worked in the direction of two long-term objectives: 1) developing systems that can interact with us in a more humane way, for which we require 2) building automatic methods capable of understanding and dealing with complex environments. In order to pave the way towards these objectives, our multi-faceted contributions cover a wide range of aspects of the machine learning pipeline: from data gathering and curation, to literature search and review, and through novel model developments. These endeavors, which have been thoroughly documented in this dissertation, have culminated in significant contributions to the field of video understanding. In line with objective 1), we have released UDIVA, a pioneering large-scale human interaction dataset that will be a valuable resource for future research in this area. Intersecting both objetives 1) and 2), we have developed the Dyadformer, an innovative architecture able to learn fine-grained long-range spatiotemporal dependencies between humans in interaction. Finally, as an in-deph exploration of objective 2), we have conducted a meticulous analysis of the recent advances in Video Transformers, which provides a comprehensive overview of the state-of-the-art in this field. In the following, we review the main contributions and findings derived from the development of this thesis, as well as ongoing work to extend it, and exciting venues for future research that steers navigation of the road ahead.

On the path to building artificial perceptual systems capable of making sense of the world and interacting with us in a natural way, two crucial components are necessary: extensive data and advanced algorithms. First of all, in Chapter 2 we have detailed the collection and initial annotation of the UDIVA dataset, the largest multiview audiovisual dataset of dyadic face-to-face non-scripted interactions up to this date. It involved the recruitment of 147 participants, who were distributed among 188 interaction sessions, resulting in 90.5 hours of audiovisual recordings of human interaction. The participants were involved in a variety of tasks each eliciting, and hence allowing to study, different patterns of behavior. The data includes 8 camera views (two of which are first-person views), audio, heart rate, transcriptions of the spoken words, self- and peer-reported personality as well as extensive individual sociodemographic metadata. The UDIVA dataset is an ongoing effort of synchronization and annotation. Despite this, and as we saw in Chapter 3, the current release[1] already provides interesting challenges for human interaction modeling. Finally, we have seen that the UDIVA

[1] `https://chalearnlap.cvc.uab.cat/dataset/41/description/`.

dataset still exhibits biases and limitations, at least when aiming for personality recognition tasks. Particularly, the dataset is slightly biased towards people with high levels of the *Openness* trait, and it may fall a bit short on the number of participants. These and other possible sources of bias make the problem of human modeling harder, but also stimulate research on bias identification and mitigation methods and related areas, in addition to the multi-modal personality computing problem by itself. Nevertheless, we believe UDIVA will help advance research in human understanding and interaction, as well as incentivize the development of new algorithms for automatic perception that can exploit relevant cues in complex environments. It can also prove useful as a large-scale pool of 90.5 hours of data for self-supervised pre-training, and to be combined with other interaction datasets, in scenarios such as dataset fusion or cross-dataset evaluation. We hope that its limitations will motivate further research on bias mitigation, the collection of larger datasets (specially in terms of the number of participants), and the design of novel annotation protocols.

In Chapter 3 we showed how the available subset of UDIVA (v0.5) is useful to start benchmarking Transformers on the challenging task of personality recognition. In this second part of the thesis, we have analyzed the abilities of Transformers to deal with an intricate task in the complex environment of human interaction. We started by evaluating this ability through direct extension of an existing Video Transformer architecture, proving the feasibility of tackling personality regression through these models. We incrementally combined different sources of context (both interlocutors' scene, acoustic, and task information) finding consistent improvements as they were added, which is consonant with human interaction research in the psychology field. We then develop the Dyadformer, a Transformer-based design specifically tailored for multi-modal human interaction settings. We compare several forms of aggregation, either by plain concatenation as done by the baseline, or in a BERT-like fashion by extending sequence length. Both approaches have demonstrated to be worse (both in terms of performance and computation) compared to the cross-attentional mechanism used by the Dyadformer. Thanks to these, our model is able to integrate information from different sources and interactants successfully to form a complete dyadic representation. Furthermore, the use of longer time windows (up to 30 seconds) was a clear key point, as personality predictions from shorter periods of time can be very noisy, as they lack enough temporal context to contain longer-term behavioral patterns. Even if not a single variant of our model is best for predicting all traits, as we have discussed, this is consistent with previous literature. Nonetheless, it is clear that joint modeling of both interlocutors, building multi-modal representations through our cross-subject layer, and extending temporal context improves performance. This is further reinforced by the Dyadformer outperforming not only the baseline, but also all challenge participants. This chapter validates two key aspects of the thesis: first, that the UDIVA offers a challenging environment to study the capabilities of neural networks to model complex environments when tackling noisy and subjective tasks; and second, that Transformers are capable of successfully doing so. These two points have been demonstrated by comparing to the baseline and challenge participant results, where 1) some participants failed to improve upon the baseline and 2) the Dyadformer was established as the new state-of-the-art for personality recognition on the UDIVA dataset, by a substantial

margin.

Despite these promising results, Transformers still pose several challenges. In particular, and as we have seen in Chapter 4, they require large quantities of data and cumbersome training regimes. When dealing with video, its large dimensionality will aggravate Transformer's quadratic complexity. While this further highlights the need for larger datasets such as UDIVA, it also signals to the current limitations of the novel Transformer family. It is for this reason that in this final chapter, we have embarked on an in-depth exploration of video Transformers. We found that most works still leverage CNNs as embedding networks to reduce the dimensionality of the input. But this should be done carefully. As we have detailed, if dimensionality is reduced too much it could hinder learning of some crucial relationships later on. This is especially seen for long-term dependencies if either space or time are irreversibly compressed before Transformer's non-local operations come into play. Alternatively, we find many Video Transformers opting to re-introduce relevant inductive biases into the architecture itself. This has a double effect: on the one hand, it helps alleviate some training limitations of Transformers (*e.g.*, it can mitigate the need for larger datasets), as certain relationships can be learned faster, as well as reducing complexity; but on the other hand, this generally implies sacrificing globality. As we also saw, this can also be achieved through self-supervised learning, which has an impact on the kinds of relationships the model learns, as well as using data augmentation. Nevertheless, these latter methods still pose a computational challenge by themselves. In addition, we saw how novel Transformer models are starting to outperform CNN counterparts in both Kinetics 400 [57] and Something-Something v2 [139], and in some cases even with reduced parameter count and fewer FLOPs. As we have explored, the improved results may be partially attributable to Transformers' abilities to learn more semantically abstract representations. On the one hand, the architecture itself (and particularly the non-local self-attention operation), as well as the use of patches as the minimum building blocks, offer greater abilities to disregard high-frequency details and focus on larger semantically relevant patterns. On the other hand, the training objective also plays a crucial role in shaping the kinds of relationships that the Transformer learns. In this sense, MTM seems to provide a better part-whole balance, hence skewing learning towards more holistic representations. This can be seen by the works using such objectives while preserving non-locality in the earlier layers before reducing input dimensionality, which outperform other approaches (see Section 4.6.2). We hope that with the release of this survey, we contribute to the advancement of Video Transformers and to a deeper understanding of the intricate ways in which they operate.

5.1 Future work

As we have previously mentioned, the UDIVA dataset is an ongoing effort. It has already been extended with landmark and gaze annotations using automatic methods. In particular, 68 face fiducials were regressed by the 3DDFA_v2 algorithm presented by Guo et al. [143]; 24 full-body joints and detection confidence were retrieved by using the MeTRAbs method [333]; 21 hand landmarks were retrieved with the hand estimator module from FrankMocap [327];

and 3D eye gaze direction vector was computed with the ETH-XGaze baseline method [459]. Furthermore, a challenge and associated workshop on "Social dyadic interactions" was organized around this dataset, aiming to solve the task of personality recognition, as well as one of behavior forecasting based on the aforementioned landmarks. We kindly refer the interested reader to [281] for a detailed description of the process to extract all of these annotations, as well as the challenge organization and results. Regarding next steps, there is still much more work pending. First of all, further processing and synchronization of ego-cameras and heart-rate monitors, as well as pair-wise camera calibration, are yet to be completed. Also, we are currently planning on extending it with further annotations, such as facial emotion expressions, continuous action/intention for human-object-human interaction, as well as high-level behavior labels and perceived personality states over time. We hope that these will allow for a more holistic analysis of human interaction from both individual and dyadic perspectives and to further evaluate learned patterns of behavior. Furthermore, regarding the aforementioned limitations and biases for personality recognition, if we want the next generation of intelligent systems for personality computing to be fair with respect to different contexts, setups, and demographics, we need to address the fairness problem in some way, either through the design and development of new datasets and annotation protocols or through new methodologies capable of mitigating different types of bias. A future research question to be addressed on this domain could be: *is it possible to build generic and "bias-free" features for different kinds of modalities (e.g., audio, visual) and attributes (e.g., gender, age)?* Finally, we plan on releasing the remaining sessions and camera views, still not publicly available.

Regarding our work surveying advances on Video Transformers, we find that, despite seeing clear trends, VTs are still in their infancy and much more research is needed. First of all, we find a severe lack of explainability tools that properly asses the kind of spatiotemporal representations that different designs and self-supervised losses provide. Overlaying head-specific attention heat-maps of the first layer over a given input may provide some ad-hoc explanations on what the model deems relevant [338, 179, 411]. Even if some VTs have explored this direction (e.g., [286, 289, 178, 268, 290, 41, 132, 215]), this technique may prove overly cumbersome for video, as it requires inspecting such per-sample activations for multiple full video sequences. Possible future venues could analyze the learned patterns of attention preferred by different heads (as in [266]), which may clue on relevant design choices that favor such patterns, or leveraging the aforementioned versatility of Transformers to probe the model through textual descriptions (as done for images in [371]). Furthermore, we see an interesting future direction in analyzing whether video-based features would also generalize to other modalities. For instance by following a similar approach as in [363] and tuning a few adapter layers to map other modalities into the video representation space. Beyond current MTM approaches, other traditional losses could be adapted to the token granularity, such as 3D jigsaw puzzles [193]. Regarding instance-based methods, adapting recent developments to images such as Barlow Twins [449] or VicReg [25] which focus on preserving views-dependent information, may prove beneficial to video modeling. Nevertheless, further research is still needed to alleviate the computational burden of self-supervision in video. Finally, key advancements in architectural choices and training techniques for VTs are mostly

limited to high-level tasks, hindering analysis of the contributions they provide for general video representation learning. In these lines, VTs have barely tackled generative tasks such as frame prediction [408, 316] or inpainting [230, 450]. We believe that token granularity and long-range modeling capabilities of Transformers could benefit these tasks. However, the high dimensionality and the complex interactions within video data, as well as the tendency of Transformers to disregard high-frequency details may pose severe challenges to solving these tasks. In this sense, combining the strenghts of CNNs and Transformers could be key for generative tasks.

Finally, regarding the Dyadformer, and after the lessons learned during the development of Chapter 4 we have already started working on extending our model in several directions. First of all, we have skewed away from leaving the video embedding network fixed, and it is now trained end-to-end with the rest of the Dyadformer, while the audio one is still frozen. But more exciting are our advances on the training objective for the Dyadformer. In particular, we are working on two self-supervised losses: Interaction Prediction (IP) and Masked Token Alignment (MTA). Preliminary results deem these pre-training losses to be very promising.

Interaction Prediction is inspired by BERT's *Next Sentence Prediction* [88]. Simply put, via their corresponding [CLS] tokens, we predict if a pair of sampled sequences of the two participants temporally correspond or not. Intuitively, this forces the network to leverage useful interaction information when learning to represent the input. We define $\mathcal{L}_{IP} = -\mathbf{y} \log \tilde{\mathbf{y}} + (1 - \mathbf{y}) \log (1 - \tilde{\mathbf{y}})$ as the binary cross-entropy, where $\mathbf{y}$ is the true label and $\tilde{\mathbf{y}}$ is the prediction. Positive interactions ($y = 1$) are from the same session and aligned in time. Negative interactions ($y = 0$) are either sampled randomly from different sessions or within the same session but with a random time offset (see Figure 5.1).

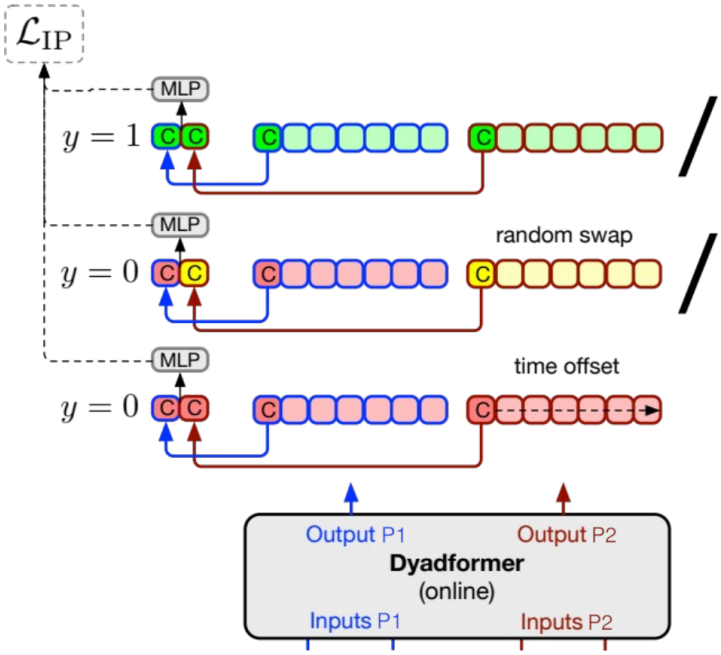


FIGURE 5.1: Illustration of the Interaction Prediction SSL loss. Binary classification of correspondence between the sequences of Subject P1 and Subject P2. Positive interactions ($y = 1$) are used 50% of the time. Within negative interactions, 10% come from another random session, 80% from the same session with a time offset and 10% are kept unchanged.

Masked Token Alignment combines contrastive learning with Masked Token Modeling. As we discussed in Section 4.4.3, we see great promise in these combined approaches, as working with tokens allows to build large negative pools for contrastive learning while MTM strives for a holistic part-whole based representation of the input. In particular, for the contrastive part, we leverage a siamese setting, akin to the one used by BYOL [140] (see Figure 5.2). On the one hand, the weights of the online network $\boldsymbol{\omega}^o$ are trained normally by optimizing the $\mathcal{L}_{MTA}$ loss (see Equation (5.1)). On the other hand, the momentum network's weights are updated by a moving average $\boldsymbol{\omega}^m_t \leftarrow \beta\boldsymbol{\omega}^m_{t-1} + (1-\beta)\boldsymbol{\omega}^o_t$, where β is a decay rate controlling the speed at which the momentum network weights become the online ones, and t is a given training iteration. Regarding the MTM part of MTA, we randomly mask 20% input tokens when used as input to the online network, while the momentum one receives the equivalent unmasked sequence. The objective for the online network is to learn token interactions such that the masked tokens' representation $\mathbf{x}^o_i$ matches the one of that same token from the momentum network $\mathbf{x}^m_i$ more than it does any other token $\mathbf{x}^m_j$ (where $i \neq j$). In doing so, we force the online network to leverage the necessary context from unmasked tokens. The $\mathcal{L}_{MTA}$ loss can be formalised as an InfoNCE [273] loss:

$$\mathcal{L}_{MTA} = -\sum_{i=1}^{M} log \frac{exp(f(\mathbf{x}^o_i)^T) f(\mathbf{x}^m_i)}{exp(f(\mathbf{x}^o_i)^T) f(\mathbf{x}^m_i) + \sum_{j=1, i\neq j}^{M} exp(f(\mathbf{x}^o_i)^T) f(\mathbf{x}^m_j)}, \tag{5.1}$$

where i, j represent the position of a given token, whereas o, m indicate whether the token belongs to the *online* or *momentum* networks, respectively, and M is the total number of distinct tokens in the batch.

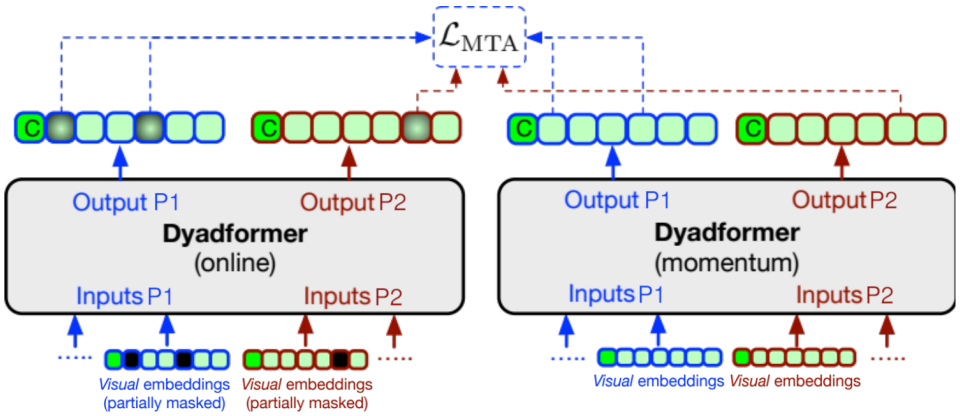


FIGURE 5.2: Illustration of the Masked Token Alignment SSL losse. Randomly masked tokens are contrastively aligned with their representation from a siamese network (trained as a momentum encoder) using InfoNCE. Following BERT [88], within masked tokens, 80% are replaced by a learned `[MSK]` token, 10% are replaced by a random token and 10% are kept unchanged.

With this preliminary approach, we are seeing very promising results, which are summarized in Table 5.1. As it can be observed, pre-training with either MTA or IP achieves lower error compared to the baseline of fully supervised training. This is further improved when both SSL losses are combined, achieving the lowest error when fine-tuning.

In the future, we will continue to pursue this research direction, in order to fully exploit Transformers' abilities to capture semantic representations of complex environments. We are

Pre-training	Training	Min(MSE)↓ #(pre-training epochs)					
		0	10	20	30	40	50
None	Fine-tuning (40 epochs)	0,8044	-	-	-	-	-
MTA	Probing (20 epochs)	-	0,8272	**0,7906**	0,8370	0,8352	0,8541
IP	Probing (20 epochs)	-	0,8164	0,8817	0,8608	0,8676	**0,7903**
MTA + IP	Probing (20 epochs)	-	0,8133	0,8298	**0,7687**	0,8127	0,7994
MTA + IP	Fine-tuning (20 epochs)	-	0,8865	0,8419	0,7932	**0,7592**	0,8473

TABLE 5.1: MSE validation error for variants of the proposed SSL losses to pre-train the Dyadformer. Best results are highlighted **row-wise** and column-wise. Note that these values have been computed on a different data fold of the UDIVA dataset than the one used in Chapter 3, thus they are not comparable to the results provided there.

currently experimenting with masking strategies to maximize results as well as the proper balance between the two losses. Moreover, in addition to audio/video-based personality computing, our model allows for straightforward adaptations to other modalities, as well as extending our analysis to other individual and dyadic features, beyond personality recognition. On the one hand, seeing that the use of language features proved to be beneficial for other models in the challenge (as we discuss in Section 3.4.5) we plan to extend the multimodality layers of the Dyadformer to also leverage the transcriptions as an additional cue to be combined with the audio. Also, future work will include the validation of the architecture for longer time windows, both through increased sequence length, but also by recomputing the chunks without the face detection limitations, which will surely boost the Dyadformer human modeling capabilities. Finally, we believe that poses could be used as an additional cue for supervision, forcing the network to include necessary information in its input representation so that pose sequences could be recovered. On the other hand, the Dyadformer could be easily extended to model other individual (*e.g.*, emotion or engagement) and interaction constructs (*e.g.*, rapport or synchrony) as it exploits both by design, as well as to other interaction datasets (*e.g.*, [97, 62]).

5.2 Final remarks

The remarkable cognitive abilities we humans possess have gifted us with an increasingly profound understanding of physics, enabling us to unravel its mysteries through the pursuit of scientific discovery. This knowledge has not only empowered us to craft tools and refine technology, but it has also deepened our comprehension of one another, nourishing meaningful connections and collaborative efforts that have led to the development of more advanced societies. Science, technology, and society seem to form a symbiotic relationship where advancements in science and technology empower us to refine our interactions and behaviors, which in turn further encourage collaboration to deepen our understanding of science and to develop novel technologies.

We aim for a future where the boundaries between computational systems and human dynamics blur, yielding not just sophisticated technologies, but also profound insights into the universe and our own nature. Through this work, we hope to have laid some of the necessary groundwork to make us one step closer to not only refining the capabilities of AI but

also to fostering a deeper empathy between machines and humans. By providing a novel large-scale dataset, deep insights on the new Transformer technology, and a novel architecture for automatic human analysis, we hope our work can help to develop more humane computational systems that are able to better understand and interact with humans. There is still a long road ahead towards more human-centered technology. We believe that the contributions laid here collectively mark a significant milestone in the evolution of computational systems in that direction. As we conclude this dissertation, hope that our work has the power to inspire new avenues of research, ignite innovative applications, and propel the development of more humane computational systems. These avenues will necessarily require the involvement of experts from fields as diverse as machine learning, psychology, and human-computer interaction to collaborate in unison, unraveling the nuances of human behavior modeling and paving the way for the next generation of empathetic machines. It will also require strengthening the cornerstones of trustworthy AI through unwavering commitments to responsibility, rigorous bias mitigation, and a vigilant regulatory framework that ensures that this remarkable technology remains a force for good.

Through this, we hope humanity will achieve democratization of technology, making it more accessible to people of all ages and cultural backgrounds. This open access can then form the foundation for stimulating collaboration, giving rise to an environment where collective creativity drives advancements in science and overall human development. The horizon ahead holds promise for enriching our lives, amplifying our collective wisdom, and creating a more positive future through unity, innovation, and shared purpose.

Appendix A

Publications

- *M. Oliu*, *J. Selva*, *S. Escalera* (2018). **Folded Recurrent Neural Networks for Future Video Prediction**. Proceedings of the European Conference on Computer Vision (ECCV).
- *C. Palmero*, *J. Selva*, *M. A. Bagheri*, *S. Escalera* (2018). **Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues**. Proceedings of British Machine Vision Conference (BMVC).
- *C. Palmero**, *J. Selva**, *S. Smeureanu**, *J. C. S. Jacques Junior*, *A. Clapés*, *A. Moseguí*, *Z. Zhang*, *D. Gallardo-Pujol*, *G. Guilera*, *D. Leiva* and *S. Escalera*. (2021). **Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset**. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops.
- *C. Palmero*, *G. Barquero*, *J. C. S. Jacques Junior*, *A. Clapés*, *J. Núñez*, *D. Curto*, *S. Smeureanu*, *J. Selva*, *Z. Zhang*, *D. Saeteros*, *D. Gallardo-Pujol*, *G. Guilera*, *D. Leiva*, *F. Han*, *X. Feng*, *J. He*, *W. Tu*, *T. B. Moeslund*, *I. Guyon*, *S. Escalera* (2022). **ChaLearn LAP Challenges on Self-Reported Personality Recognition and Non-Verbal Behavior Forecasting During Social Dyadic Interactions: Dataset, Design, and Results**. Understanding Social Behavior in Dyadic and Small Group Interactions, Proceedings of Machine Learning Research (PMLR).
- *D. Curto**, *A. Clapés**, *J. Selva**, *S. Smeureanu*, *J. C. S. Jacques Junior*, *D. Gallardo-Pujol*, *G. Guilera*, *D. Leiva*, *T. B. Moeslund*, *S. Escalera* and *C. Palmero* (2021). **Dyadformer: A Multi-Modal Transformer for Long-Range Modeling of Dyadic Interactions**. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops.
- *J. Selva*, *A. S. Johansen*, *S. Escalera*, *K. Nasrollahi*, *T. B. Moeslund* and *A. Clapés*. (2023). **Video Transformers: A Survey**. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Bibliography

[1] Jennifer Abel and Molly Babel. "Cognitive load reduces perceived linguistic convergence between dyads". In: *Language and Speech* 60.3 (2017), pp. 479–502 (cit. on p. 20).

[2] Antonio A. Abello, Roberto Hirata, and Zhangyang Wang. "Dissecting the High-Frequency Bias in Convolutional Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 863–871 (cit. on pp. 4, 95).

[3] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "YouTube-8M: A Large-Scale Video Classification Benchmark". In: *CoRR* abs/1609.08675 (2016). arXiv: 1609.08675. URL: http://arxiv.org/abs/1609.08675 (cit. on p. 47).

[4] Palwasha Afsar, Paulo Cortez, and Henrique Santos. "Automatic visual detection of human behavior: A review from 2000 to 2014". In: *Expert Systems with Applications* 42.20 (2015), pp. 6935 –6956. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2015.05.023. URL: http://www.sciencedirect.com/science/article/pii/S0957417415003516 (cit. on p. 13).

[5] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. "To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations". In: *2019 International Conference on Multimodal Interaction*. 2019, pp. 74–84 (cit. on p. 31).

[6] Icek Ajzen. "Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology". In: *Advances in experimental social psychology*. Vol. 20. Elsevier, 1987, pp. 1–63 (cit. on p. 20).

[7] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text". In: *NeurIPS* (2021) (cit. on pp. 71, 74, 83, 89, 90, 97, 110, 111).

[8] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. "Salsa: A novel dataset for multimodal group behavior analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2015), pp. 1707–1720 (cit. on p. 16).

[9] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. "Self-Supervised MultiModal Versatile Networks". In: *NeurIPS*. 2020 (cit. on p. 84).

[10] Leo Alexander III, Evan Mulfinger, and Frederick L Oswald. "Using big data and machine learning in personality measurement: Opportunities and challenges". In: *European Journal of Personality* 34.5 (2020), pp. 632–648 (cit. on p. 32).

[11] Nalini Ambady and Robert Rosenthal. "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." In: *Psychological bulletin* 111.2 (1992), p. 256 (cit. on pp. 8, 43).

[12] Guozhen An and Rivka Levitan. "Lexical and Acoustic Deep Learning Model for Personality Recognition". In: *Proc. Interspeech 2018*. 2018, pp. 1761–1765 (cit. on p. 42).

[13] Sean Andrist, Bilge Mutlu, and Adriana Tapus. "Look like me: matching robot personality via gaze to increase motivation". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 3603–3612 (cit. on p. 32).

[14] Salvatore Anzalone, Giovanna Varni, Serena Ivaldi, and Mohamed Chetouani. "Automated Prediction of Extraversion During Human-Humanoid Interaction". In: *Int. Journal of Social Robotics* 9.3 (2017), pp. 385–399 (cit. on p. 42).

[15] Oya Aran and Daniel Gatica-Perez. "Cross-domain personality prediction: from video blogs to small group meetings". In: *International Conference on Multimodal Interaction*. 2013, pp. 127–130 (cit. on p. 42).

[16] Oya Aran and Daniel Gatica-Perez. "One of a kind: Inferring personality impressions in meetings". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 11–18 (cit. on p. 42).

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. "ViViT: A Video Vision Transformer". In: *ICCV*. 2021 (cit. on pp. 40, 41, 69, 71, 73, 74, 78–80, 83, 88, 90, 91, 97, 103, 104, 106, 108).

[18] Damaris Aschwanden, Mathias Allemand, and Patrick L Hill. "Cognitive methods in personality research". In: *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment* (2020), pp. 49–54 (cit. on p. 20).

[19] Michael Ashton and Kibeom Lee. "The HEXACO-60: A short measure of the major dimensions of personality". In: *Journal of personality assessment* 91 (July 2009), pp. 340–5. DOI: 10.1080/00223890902935878 (cit. on pp. 15, 18).

[20] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *NeurIPS* (2016) (cit. on p. 39).

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2015 (cit. on p. 6).

[22] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. "Deep Equilibrium Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/01386bd6d8e091c2ab4c7c7de644d37b-Paper.pdf (cit. on p. 57).

[23] Nicholas Baker and James H Elder. "Deep learning models fail to capture the configural nature of human shape perception". In: *iScience* 25.9 (Aug. 2022), p. 104913 (cit. on p. 6).

[24] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. "Deep convolutional networks do not classify based on global object shape". In: *PLOS Computational Biology* 14.12 (Dec. 2018), pp. 1–43. DOI: 10.1371/journal.pcbi.1006613. URL: https://doi.org/10.1371/journal.pcbi.1006613 (cit. on p. 4).

[25] Adrien Bardes, Jean Ponce, and Yann Lecun. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". In: *ICLR*. 2022 (cit. on p. 116).

[26] German Barquero, Johnny Núñez, Zhen Xu, Sergio Escalera, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. "Comparison of Spatio-Temporal Models for Human Motion and Pose Forecasting in Face-to-Face Interaction Scenarios". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Proceedings of Machine Learning Research. 2022 (cit. on p. 31).

[27] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. "Context in emotion perception". In: *Current Directions in Psychological Science* 20.5 (2011), pp. 286–290 (cit. on pp. 15, 36, 41).

[28] Ligia Batrinca, Bruno Lepri, and Fabio Pianesi. "Multimodal recognition of personality during short self-presentations". In: *ACM Workshop on Human Gesture and Behavior Understanding*. 2011, pp. 27–28 (cit. on p. 42).

[29] Iz Beltagy, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer". In: *arXiv* (2020) (cit. on pp. 77, 108).

[30] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. "Adversarial robustness comparison of vision transformer and mlp-mixer to cnns". In: *Proceedings of the 32nd British Machine Vision Conference (BMVC)*. 2021 (cit. on pp. 4, 95).

[31] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?" In: *ICML*. 2021 (cit. on pp. 40, 41, 43, 69, 71, 73, 74, 77, 78, 83, 88, 90, 97, 103–106).

[32] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope". In: *Electronics* 10.20 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10202470. URL: https://www.mdpi.com/2079-9292/10/20/2470 (cit. on p. 4).

[33] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. "Understanding Robustness of Transformers for Image Classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10231–10241 (cit. on pp. 6, 110).

[34] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. "The MAHNOB Mimicry Database: A database of naturalistic human interactions". In: *Pattern recognition letters* 66 (2015), pp. 52–61 (cit. on p. 17).

[35] Valerio Biscione and Jeffrey S Bowers. "Mixed Evidence for Gestalt Grouping in Deep Neural Networks". In: *Computational Brain & Behavior* (2023) (cit. on p. 6).

[36] Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. "The corpus of interactional data: A large multimodal annotated resource". In: *Handbook of linguistic annotation*. Springer, 2017, pp. 1323–1356 (cit. on p. 17).

[37] Randolph Blake and Sang-Hun Lee. "Temporal structure in the input to vision can promote spatial grouping". In: *Biologically Motivated Computer Vision: First IEEE International Workshop, BMCV 2000 Seoul, Korea, May 15–17, 2000 Proceedings 1*. Springer. 2000, pp. 635–653 (cit. on p. 4).

[38] Randolph Blake and Sang-Hun Lee. "The Role of Temporal Structure in Human Vision". In: *Behavioral and Cognitive Neuroscience Reviews* 4.1 (2005). PMID: 15886401, pp. 21–42. DOI: 10.1177/1534582305276839. eprint: https://doi.org/10.1177/1534582305276839. URL: https://doi.org/10.1177/1534582305276839 (cit. on pp. 4, 6).

[39] Wiebke Bleidorn and Christopher James Hopwood. "Using machine learning to advance personality assessment and theory". In: *Personality and Social Psychology Review* 23.2 (2019), pp. 190–203 (cit. on pp. 33, 61).

[40] Michael F. Bonner and Russell A. Epstein. "Object representations in the human brain reflect the co-occurrence statistics of vision and language". In: *Nature Communications* 12.1 (2021), pp. 4081–. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24368-2. URL: https://doi.org/10.1038/s41467-021-24368-2 (cit. on p. 4).

[41] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. "Transformerfusion: Monocular rgb scene reconstruction using transformers". In: *NeurIPS* (2021) (cit. on pp. 72, 81–83, 89, 116).

[42] Elif Bozkurt, Hossein Khaki, Sinan Keçeci, B Berker Türker, Yücel Yemez, and Engin Erzin. "JESTKOD database: Dyadic interaction analysis". In: *2015 23nd Signal Processing and Communications Applications Conference (SIU)*. IEEE. 2015, pp. 1374–1377 (cit. on p. 17).

[43] Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Sage Publications, 2014 (cit. on p. 14).

[44] Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Routledge, 2021 (cit. on p. 15).

[45] Egon Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 1956 (cit. on p. 16).

[46] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. "Revisiting the "Video" in Video-Language Understanding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022 (cit. on p. 96).

[47] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. "Space-time mixing attention for video transformer". In: *NeurIPS* (2021) (cit. on pp. 71, 77, 80, 83, 88, 91, 104–106).

[48] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007 (cit. on pp. 14, 15, 35, 36).

[49] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception". In: *IEEE Transactions on Affective Computing* 8.1 (Feb. 2017), pp. 67–80. DOI: 10.1109/TAFFC.2016.2515617 (cit. on pp. 16, 17).

[50] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42.4 (Nov. 2008), p. 335. ISSN: 1574-0218. DOI: 10.1007/s10579-008-9076-6 (cit. on pp. 13, 16, 17).

[51] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. "The NoXi database: multimodal recordings of mediated novice-expert interactions". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, pp. 350–359 (cit. on p. 16).

[52] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. "Multi-channel transformers for multi-articulatory sign language translation". In: *ECCV*. 2020 (cit. on pp. 72, 87, 90, 96–98).

[53] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. "Sign language transformers: Joint end-to-end sign language recognition and translation". In: *CVPR*. 2020 (cit. on pp. 72, 90, 95–97).

[54] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers". In: *ECCV*. 2020 (cit. on pp. 69, 71, 81, 91, 98, 100).

[55] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9650–9660 (cit. on pp. 5, 7).

[56] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. "A short note on the kinetics-700 human action dataset". In: *arXiv* (2019) (cit. on pp. 13, 73, 90).

[57] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6299–6308 (cit. on pp. 11, 46, 56, 70–73, 101, 115).

[58] Patrick Cavanagh. "The Language of Vision*". In: *Perception* 50.3 (2021). PMID: 33583254, pp. 195–215. DOI: 10.1177/0301006621991491. eprint: https://doi.org/10.1177/0301006621991491. URL: https://doi.org/10.1177/0301006621991491 (cit. on p. 4).

[59] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement". In: *IEEE Transactions on Affective Computing* (2017) (cit. on p. 16).

[60] Fabio Celli, Elia Bruni, and Bruno Lepri. "Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures". In: *International Conference on Multimedia*. 2014, pp. 1101–1104 (cit. on p. 42).

[61] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. "Deep analysis of cnn-based spatio-temporal representations for action recognition". In: *CVPR*. 2021 (cit. on p. 88).

[62] Huili Chen, Yue Zhang, Felix Weninger, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. "Dyadic Speech-based Affect Recognition using DAMI-P2C Parent-child Multimodal Interaction Dataset". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 97–106 (cit. on pp. 16, 17, 119).

[63] Jingwen Chen and Hongyang Chao. "VideoTRM: Pre-training for Video Captioning Challenge 2020". In: *ACM-MM*. 2020 (cit. on p. 94).

[64] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *TPAMI* (2017) (cit. on p. 72).

[65] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. "The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy". In: *CVPR*. 2022 (cit. on p. 109).

[66] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *ICML*. 2020 (cit. on p. 93).

[67] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. "When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations". In: *arXiv* (2021) (cit. on p. 89).

[68] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. "Transformer tracking". In: *CVPR*. 2021 (cit. on pp. 71, 90).

[69] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning". In: *CVPR*. 2021 (cit. on p. 91).

[70] Xusong Chen, Dong Liu, Chenyi Lei, Rui Li, Zheng-Jun Zha, and Zhiwei Xiong. "Bert4sessrec: Content-based video relevance prediction with bidirectional encoder representations from transformer". In: *ACM-MM*. 2019 (cit. on pp. 72, 79, 90, 94, 96).

[71] Yunliang Chen and Jungseock Joo. "Understanding and mitigating annotation bias in facial expression recognition". In: *ICCV*. 2021 (cit. on p. 91).

[72] Kendra Cherry. *What Are the Gestalt Principles?* (Cit. on p. 3).

[73] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus". In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2017, pp. 292–298 (cit. on p. 17).

[74] Herbert H Clark, Robert Schreuder, and Samuel Buttrick. "Common ground at the understanding of demonstrative reference". In: *Journal of verbal learning and verbal behavior* 22.2 (1983), pp. 245–258 (cit. on p. 14).

[75] Gerald L Clore and Jeffrey R Huntsinger. "How emotions inform judgment and regulate thought". In: *Trends in cognitive sciences* 11.9 (2007), pp. 393–399 (cit. on p. 15).

[76] Nicholas Confessore. *Cambridge Analytica and Facebook: The scandal and the fallout so far*. Published at The New York Times 2018, April 4. 2018 (cit. on p. 33).

[77] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. "Spatial-Temporal Transformer for Dynamic Scene Graph Generation". In: *ICCV*. 2021 (cit. on pp. 72, 77, 78, 83, 96).

[78] Enric Corona, Albert Pumarola, and Guillem Alenyà. "Context-aware Human Motion Prediction". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6992–7001 (cit. on p. 15).

[79] Ronen Cuperman and William Ickes. "Big Five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables"." In: *Journal of personality and social psychology* 97.4 (2009), p. 667 (cit. on pp. 14, 64).

[80] David Curto, Albert Clapes, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B. Moeslund, Sergio Escalera, and Cristina Palmero. "Dyadformer: A Multi-Modal Transformer for Long-Range Modeling of Dyadic Interactions". In: *ICCV-W*. 2021 (cit. on pp. 72, 87, 96).

[81] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. "CoAtNet: Marrying Convolution and Attention for All Data Sizes". In: *NeurIPS*. 2021 (cit. on p. 90).

[82] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context". In: *ACL*. 2019 (cit. on p. 75).

[83] Kerstin Dautenhahn. "Socially intelligent robots: dimensions of human–robot interaction". In: *Philosophical transactions of the royal society B: Biological sciences* 362.1480 (2007), pp. 679–704 (cit. on p. 15).

[84] Robert O. Davis. "The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis". In: *Educational Research Review* 24 (2018), pp. 193–219 (cit. on p. 32).

[85] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. "Interpersonal synchrony: A survey of evaluation methods across disciplines". In: *IEEE Transactions on Affective Computing* 3.3 (2012), pp. 349–365 (cit. on pp. 15, 41).

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009 (cit. on pp. 73, 91).

[87] Soumia Dermouche and Catherine Pelachaud. "Engagement Modeling in Dyadic Interaction". In: *2019 International Conference on Multimodal Interaction*. 2019, pp. 440–445 (cit. on pp. 15, 41).

[88] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Computational Linguistics NAACL-HLT*. 2019, pp. 4171–4186. URL: https://aclweb.org/anthology/papers/N/N19/N19-1423/ (cit. on pp. 40, 41, 43, 44, 57, 58, 69, 86, 89, 91, 93, 95, 99, 111, 117, 118).

[89] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. "A survey on personality-aware recommendation systems". In: *Artificial Intelligence Review* (2021), pp. 1–46 (cit. on p. 33).

[90] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. "Attention is not all you need: Pure attention loses rank doubly exponentially with depth". In: *ICML*. 2021 (cit. on p. 109).

[91] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021 (cit. on pp. 40, 41, 43, 56, 68, 69, 71–75, 79, 80, 89, 91, 97, 109).

[92] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data". In: *International conference on affective computing and intelligent interaction*. Springer. 2007, pp. 488–500 (cit. on p. 17).

[93] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. "The HUMAINE database". In: *Emotion-Oriented Systems*. Springer, 2011, pp. 243–284 (cit. on p. 17).

[94] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. "Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 206–212 (cit. on pp. 7, 15, 16).

[95] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. "Sstvos: Sparse spatiotemporal transformers for video object segmentation". In: *CVPR*. 2021 (cit. on pp. 78, 88).

[96] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. "Mdmmt: Multidomain multimodal transformer for video retrieval". In: *CVPR*. 2021 (cit. on pp. 71, 79, 98, 99).

[97] Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House. "Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture." In: *LREC*. 2010, pp. 2992–2995 (cit. on pp. 16, 17, 119).

[98] Rana El Kaliouby, Rosalind Picard, and Simon Baron-Cohen. "Affective computing and autism". In: *Annals of the New York Academy of Sciences* 1093.1 (2006), pp. 228–248 (cit. on p. 31).

[99] Lesa Ellis and Mary Rothbart. "Revision of the Early Adolescent Temperament Questionnaire". In: *Poster presented at the 2001 Biennal Meeting of the Society for Research in Child Development* (Jan. 2001). DOI: 10.1037/t07624-000 (cit. on p. 18).

[100] Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Stéphane Ayache, Evelyne Viegas, Furkan Gürpinar, Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem, Marcel A. J. van Gerven, and Rob van Lier. "Modeling, recognizing, and explaining apparent personality from videos". In: *IEEE Transactions on Affective Computing* (2020) (cit. on p. 66).

[101] Valentín Escudero, Minsun Lee, and Myrna L. Friedlander. "Dyadic Interaction Analysis". In: *The Cambridge Handbook of Group Interaction Analysis*. Ed. by Elisabeth Brauner, Margarete Boos, and MichaelaEditors Kolbe. Cambridge Handbooks in Psychology. Cambridge University Press, 2018, 45–67 (cit. on p. 15).

[102] Anna Esposito, Leopoldina Fortunati, and Giuseppe Lugano. "Modeling emotion, behavior and context in socially believable robots and ICT interfaces". In: *Cognitive Computation* 6.4 (2014), pp. 623–627 (cit. on p. 15).

[103] Anna Esposito and Lakhmi C Jain. "Modeling emotions in robotic socially believable behaving systems". In: *Toward Robotic Socially Believable Behaving Systems-Volume I*. Springer, 2016, pp. 9–14 (cit. on p. 31).

[104] Connor Esterwood and Lionel P Robert. "A Systematic Review of Human and Robot Personality in Health Care Human-Robot Interaction". In: *Frontiers in Robotics and AI* (2021), p. 306 (cit. on p. 32).

[105] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. "Summarizing videos with attention". In: *ACCV*. 2018 (cit. on pp. 72, 74).

[106] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale Vision Transformers". In: *ICCV*. 2021 (cit. on pp. 40, 41, 71, 74, 75, 80, 83, 88, 90, 91, 97, 102, 104, 106–108, 110).

[107] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. "Scene Memory Transformer for Embodied Agents in Long-Horizon Tasks". In: *CVPR*. 2019 (cit. on pp. 72, 81, 82, 86, 89, 96, 98).

[108] Sheng Fang, Catherine Achard, and Séverine Dubuisson. "Personality classification and behaviour interpretation: An approach based on feature categories". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 225–232 (cit. on p. 42).

[109] Christoph Feichtenhofer. "X3d: Expanding architectures for efficient video recognition". In: *CVPR*. 2020 (cit. on pp. 88, 104).

[110] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. "SlowFast Networks for Video Recognition". In: *ICCV*. 2019 (cit. on pp. 71, 84, 87, 104, 106).

[111] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. "A large-scale study on unsupervised spatiotemporal representation learning". In: *CVPR*. 2021 (cit. on pp. 92, 93).

[112] William Fleeson and Eranda Jayawickreme. "Whole trait theory". In: *Journal of research in personality* 56 (2015), pp. 82–92 (cit. on p. 64).

[113] William Fleeson and Mary Kate Law. "Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability." In: *Journal of personality and social psychology* 109.6 (2015), p. 1090 (cit. on p. 63).

[114] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. "A Practical Survey on Faster and Lighter Transformers". In: *arXiv* (2021) (cit. on p. 69).

[115] Kunihiko Fukushima. "Neocognitron: A hierarchical neural network capable of visual pattern recognition". In: *Neural networks* 1.2 (1988), pp. 119–130 (cit. on p. 4).

[116] David C Funder. "Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort". In: *Current Directions in Psychological Science* 25.3 (2016), pp. 203–208 (cit. on p. 63).

[117] Riccardo Fusaroli, Johanne S Bjørndahl, Andreas Roepstorff, and Kristian Tylén. "A heart for interaction: Shared physiological dynamics and behavioral coordination in a collective, creative construction task." In: *Journal of Experimental Psychology: Human Perception and Performance* 42.9 (2016), p. 1297 (cit. on p. 20).

[118] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. "Multi-modal Transformer for Video Retrieval". In: *ECCV*. 2020 (cit. on pp. 44, 71, 74, 75, 79, 86, 98, 99, 109).

[119] David Gallardo-Pujol, Antonio Andrés-Pueyo, and Alberto Maydeu-Olivares. "MAOA genotype, social exclusion and aggression: An experimental test of a gene–environment interaction". In: *Genes, Brain and Behavior* 12.1 (Oct. 2012), pp. 140–145. DOI: 10.1111/j.1601-183X.2012.00868.x (cit. on pp. 18, 47).

[120] David Gallardo-Pujol, Victor Rouco, Anna Cortijos-Bernabeu, Luis Oceja, Christopher J Soto, and Oliver P John. "Factor structure, gender invariance, measurement properties and short forms of the Spanish adaptation of the Big Five Inventory-2 (BFI-2)". In: (2021) (cit. on p. 29).

[121] Daniel Gatica-Perez. "Automatic nonverbal analysis of social interaction in small groups: A review". In: *Image and Vision Computing* 27.12 (2009), pp. 1775–1787 (cit. on pp. 13, 15).

[122] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. "Actor-transformers for group activity recognition". In: *CVPR*. 2020 (cit. on pp. 71, 74, 75, 85, 87, 90, 95, 97).

[123] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2 (Nov. 2020), pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: https://doi.org/10.1038/s42256-020-00257-z (cit. on p. 4).

[124] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "Partial success in closing the gap between human and machine vision". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 23885–23899. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf (cit. on p. 4).

[125] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Bygh9j09KX (cit. on p. 4).

[126] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio Set: An ontology and human-labeled dataset for audio events". In: *Proc. IEEE ICASSP 2017*. New Orleans, LA, 2017 (cit. on p. 57).

[127] Katharina Geukes, Simon M Breil, Roos Hutteman, Steffen Nestler, Albrecht CP Küfner, and Mitja D Back. "Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The PILS and the CONNECT study". In: *PloS one* 14.1 (2019), e0210424 (cit. on p. 36).

[128] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. "Large-scale weakly-supervised pre-training for video action recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, May 2019, pp. 12046–12055 (cit. on pp. 46, 47, 56, 104, 105).

[129] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. "COSMIC: COmmonSense knowledge for eMotion Identification in Conversations". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020, pp. 2470–2481 (cit. on p. 41).

[130] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. "COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning". In: *NeurIPS*. 2020 (cit. on pp. 41, 43, 71, 74, 79–81, 83, 88, 98, 99, 110, 111).

[131] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. "Video Action Transformer Network". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 10, 35, 44, 46, 56, 69, 71, 73–75, 78, 81, 83, 88, 90, 97).

[132] Rohit Girdhar and Kristen Grauman. "Anticipative Video Transformer". In: *ICCV*. 2021 (cit. on pp. 72, 75, 78, 80, 83, 88, 90, 91, 94–96, 109, 116).

[133] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. "Omnivore: A Single Model for Many Visual Modalities". In: *CVPR*. 2022 (cit. on p. 95).

[134] Dersu Giritlioğlu, Burak Mandira, Selim Firat Yilmaz, Can Ufuk Ertenli, Berhan Faruk Akgür, Merve Kınıklıoğlu, Aslı Gül Kurt, Emre Mutlu, Şeref Can Gürel, and Hamdi Dibeklioğlu. "Multimodal analysis of personality traits on videos of self-presentation and induced behavior". In: *Journal on Multimodal User Interfaces* (2020), pp. 1–22 (cit. on p. 43).

[135] Lewis R Goldberg. "The structure of phenotypic personality traits." In: *American psychologist* 48.1 (1993), p. 26 (cit. on pp. 15, 61).

[136] Yuan Gong, Yu-An Chung, and James Glass. "AST: Audio Spectrogram Transformer". In: *Interspeech*. 2021 (cit. on p. 69).

[137] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. "End-to-end learning of deep visual representations for image retrieval". In: *ICCV* (2017) (cit. on p. 71).

[138] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. "Watching the World Go By: Representation Learning from Unlabeled Videos". In: *arXiv* (2020) (cit. on p. 96).

[139] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. "The "something something" video database for learning and evaluating visual common sense". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5842–5850 (cit. on pp. 6, 11, 101, 115).

[140] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *NeurIPS* (2020) (cit. on pp. 91, 118).

[141] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. "Pyramid constrained self-attention network for fast video salient object detection". In: *AAAI*. 2020 (cit. on pp. 72, 73, 77, 83, 90).

[142] Yağmur Güçlütürk, Umut Güçlü, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel AJ Van Gerven, and Rob Van Lier. "Multimodal First Impression Analysis with Deep Residual Networks". In: *IEEE Transactions on Affective Computing* 9.3 (2017), pp. 316–329 (cit. on pp. 42, 66).

[143] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. "Towards fast, accurate and stable 3d dense face alignment". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer. 2020, pp. 152–168 (cit. on p. 115).

[144] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. "Cross-Architecture Self-supervised Video Representation Learning". In: *CVPR*. 2022 (cit. on pp. 71, 91–93, 96, 110).

[145] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. "A Survey on Vision Transformer". In: *IEEE TPAMI*. 2020 (cit. on p. 69).

[146] Tengda Han, Weidi Xie, and Andrew Zisserman. "Video representation learning by dense predictive coding". In: *ICCV-W*. 2019 (cit. on pp. 94, 96).

[147] Willard W Hartup, Doran C French, Brett Laursen, Mary Kathleen Johnston, and John R Ogawa. "Conflict and friendship relations in middle childhood: Behavior in a closed-field situation". In: *Child Development* 64.2 (1993), pp. 445–454 (cit. on p. 20).

[148] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018, pp. 2122–2132 (cit. on p. 42).

[149] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners". In: *CVPR*. 2022 (cit. on pp. 91, 95).

[150] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CVPR*. 2020 (cit. on p. 93).

[151] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn". In: *ICCV*. 2017 (cit. on p. 72).

[152] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 71–73).

[153] Daniel Helm and Martin Kampel. "Single-Modal Video Analysis of Personality Traits using Low-Level Visual Features". In: *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2020, pp. 1–6 (cit. on p. 42).

[154] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. "Pretrained Transformers Improve Out-of-Distribution Robustness". In: *ACL*. 2020 (cit. on p. 110).

[155] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. "Using self-supervised learning can improve model robustness and uncertainty". In: *NeurIPS* (2019) (cit. on p. 91).

[156] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. "Natural Adversarial Examples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 15262–15271 (cit. on p. 4).

[157] Joseph Henrich. *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK, 2020 (cit. on p. 33).

[158] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. "Deepfake Detection Scheme Based on Vision Transformer and Distillation". In: *arXiv* (2021) (cit. on p. 73).

[159] Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. "CNN Architectures for Large-Scale Audio Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017 (cit. on pp. 47, 57, 85).

[160] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. "Object-region video transformers". In: *CVPR*. 2022 (cit. on pp. 72, 74, 80, 99, 100, 106, 109, 110).

[161] Louis Hickman, Rachel Saef, Vincent Ng, Sang Eun Woo, Louis Tay, and Nigel Bosch. "Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews". In: *Human Resource Management Journal* (2021) (cit. on pp. 32, 42).

[162] Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. "Eye movements during everyday behavior predict personality traits". In: *Frontiers in human neuroscience* (2018), p. 105 (cit. on p. 42).

[163] Andrew Howard, Mark Sandler, Grace Chu, and Liang-Chieh et al. Chen. "Searching for mobilenetv3". In: *CVPR*. 2019 (cit. on p. 72).

[164] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV] (cit. on p. 46).

[165] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. "EmotionLines: An Emotion Corpus of Multi-Party Conversations". In: *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association, 2018 (cit. on p. 13).

[166] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *CVPR*. 2018 (cit. on pp. 71, 72).

[167] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. "MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5666–5675. DOI: 10.18653/v1/2021.acl-long.440. URL: https://aclanthology.org/2021.acl-long.440 (cit. on p. 42).

[168] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. "What makes a video a video: Analyzing temporal information in video understanding models and datasets". In: *CVPR*. 2018 (cit. on p. 96).

[169] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *CVPR*. 2017 (cit. on p. 71).

[170] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3296–3297. DOI: 10.1109/CVPR.2017.351 (cit. on p. 46).

[171] David H Hubel and Torsten N Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of physiology* 160.1 (1962), p. 106 (cit. on p. 3).

[172] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. "Video instance segmentation using inter-frame communication transformers". In: *NeurIPS* (2021) (cit. on pp. 72, 80, 83, 88, 99, 100).

[173] Vladimir Iashin and Esa Rahtu. "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer". In: *BMVC*. 2020 (cit. on pp. 44, 54, 71, 81, 85, 87, 90, 95, 97, 98).

[174] Vladimir Iashin and Esa Rahtu. "Multi-modal dense video captioning". In: *CVPR*. 2020 (cit. on pp. 71, 86, 87, 90).

[175] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *ICML*. 2015 (cit. on pp. 71, 72).

[176] Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel AJ Van Gerven, Rob Van Lier, et al. "First impressions: A survey on vision-based apparent personality trait analysis". In: *IEEE Transactions on Affective Computing* (2019) (cit. on pp. 16, 35, 42, 66).

[177] Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. "Person Perception Biases Exposed: Revisiting the First Impressions Dataset". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. Jan. 2021, pp. 13–21 (cit. on p. 66).

[178] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. "Perceiver: General Perception with Iterative Attention". In: *ICML*. 2021 (cit. on pp. 72, 73, 75, 81, 83, 85, 86, 88, 90, 97, 109, 116).

[179] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *NAACL-HLT*. 2019 (cit. on p. 116).

[180] Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. "SBAT: Video Captioning with Sparse Boundary-Aware Transformer". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 630–636. DOI: 10.24963/ijcai.2020/88. URL: https://doi.org/10.24963/ijcai.2020/88 (cit. on pp. 44, 54).

[181] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE TPAMI* (2020) (cit. on pp. 95, 96).

[182] Jason Jo and Yoshua Bengio. *Measuring the tendency of CNNs to Learn Surface Statistical Regularities*. 2017. arXiv: 1711.11561 [cs.LG] (cit. on pp. 4, 95).

[183] Adrian Johnston and Gustavo Carneiro. "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume". In: *CVPR*. 2020 (cit. on p. 73).

[184] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. "Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in a Triadic Interaction". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 13, 16).

[185] Jyoti Joshi, Hatice Gunes, and Roland Goecke. "Automatic prediction of perceived traits using visual cues under varied situational context". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 2855–2860 (cit. on p. 43).

[186] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition". In: *ECCV*. 2020 (cit. on pp. 41, 71, 75, 85, 87, 90, 97).

[187] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing". In: *arXiv* (2021) (cit. on p. 69).

[188] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. "A comparative study on transformer vs rnn in speech applications". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2019, pp. 449–456 (cit. on p. 43).

[189] David A Kenny. "Models of non-independence in dyadic research". In: *Journal of Social and Personal Relationships* 13.2 (1996), pp. 279–294 (cit. on p. 15).

[190] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in vision: A survey". In: *ACM CSUR* (2022) (cit. on p. 69).

[191] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. "Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure". In: *Computational Brain & Behavior* 4.3 (Sept. 2021), pp. 251–263. ISSN: 2522-087X. DOI: 10.1007/s42113-021-00100-7. URL: https://doi.org/10.1007/s42113-021-00100-7 (cit. on p. 6).

[192] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. "Selfreg: Self-supervised contrastive regularization for domain generalization". In: *ICCV*. 2021 (cit. on p. 91).

[193] Dahun Kim, Donghyeon Cho, and In So Kweon. "Self-supervised video representation learning with space-time cubic puzzles". In: *AAAI*. 2019 (cit. on p. 116).

[194] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. "Multimodal dual attention memory for video story question answering". In: *ECCV*. 2018 (cit. on pp. 72, 81, 83, 86, 87, 90, 96).

[195] Myeongjun Kim, Taehun Kim, and Daijin Kim. "Spatio-Temporal Slowfast Self-Attention Network For Action Recognition". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 2206–2210 (cit. on p. 56).

[196] Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen. *Multimodal corpora: from models of natural interaction to systems and applications*. Vol. 5509. Springer, 2009 (cit. on p. 16).

[197] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer". In: *ICLR*. 2019 (cit. on p. 97).

[198] John F. Kolen and Stefan C. Kremer. "Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies". In: *A Field Guide to Dynamical Recurrent Networks*. 2001, pp. 237–243. DOI: 10.1109/9780470544037.ch14 (cit. on p. 6).

[199] Satoshi Kondo. "LapFormer: surgical tool detection in laparoscopic surgical video using transformer architecture". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* (2020) (cit. on pp. 71, 73, 95).

[200] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. "Visil: Fine-grained spatio-temporal video similarity learning". In: *ICCV*. 2019 (cit. on p. 71).

[201] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cit. on p. 16).

[202] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS (cit. on p. 57).

[203] P. A. Lasota, T. Song, and J. A. Shah. *A Survey of Methods for Safe Human-Robot Interaction*. 2017. URL: https://ieeexplore.ieee.org/document/8186877 (cit. on p. 13).

[204] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan. "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions". In: *Tenth Annual Conference of the International Speech Communication Association*. 2009 (cit. on p. 41).

[205] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. "Parameter Efficient Multimodal Transformers for Video Representation Learning". In: *ICLR*. 2021 (cit. on pp. 44, 57, 71, 74, 75, 86, 90, 93, 94, 96, 97, 99, 111).

[206] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. "Less is more: Clipbert for video-and-language learning via sparse sampling". In: *CVPR*. 2021 (cit. on pp. 71, 72, 83, 84, 89).

[207] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. "MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning". In: *ACL*. 2020 (cit. on pp. 71, 73, 75, 81–83, 85, 88, 90, 97, 98).

[208] Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. "Multilingual Transformer-Based Personality Traits Estimation". In: *Information* 11.4 (2020), p. 179 (cit. on p. 43).

[209] Bruno Lepri, Nadia Mana, Alessandro Cappelletti, Fabio Pianesi, and Massimo Zancanaro. "Modeling the personality of participants during group interactions". In: *International Conference on User Modeling, Adaptation, and Personalization*. Springer. 2009, pp. 114–125 (cit. on pp. 36, 43).

[210] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. "Connecting meeting behavior with extraversion—A systematic study". In: *IEEE Transactions on Affective Computing* 3.4 (2012), pp. 443–455 (cit. on p. 42).

[211] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. "Connecting meeting behavior with extraversion—A systematic study". In: *IEEE Transactions on Affective Computing* 3.4 (2012), pp. 443–455 (cit. on p. 42).

[212] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. "Employing social gaze and speaking activity for automatic determination of the extraversion trait". In: *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*. 2010, pp. 1–8 (cit. on p. 42).

[213] Jeng-Lin Li and Chi-Chun Lee. "Attention Learning with Retrievable Acoustic Embedding of Personality for Emotion Recognition". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 171–177 (cit. on p. 42).

[214] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. "UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning". In: *ICLR*. 2022 (cit. on pp. 71, 80, 83, 88, 90, 103, 104, 106, 110).

[215] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. "HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training". In: *EMNLP*. 2020 (cit. on pp. 43, 69, 71, 74, 81, 83, 86, 91, 94–97, 99, 111, 116).

[216] Qingbiao Li, Chunhua Wu, Zhe Wang, and Kangfeng Zheng. "Hierarchical Transformer Network for Utterance-Level Emotion Recognition". In: *Applied Sciences* 10.13 (2020), p. 4447 (cit. on p. 41).

[217] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. "GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer". In: *ICCV*. 2021 (cit. on pp. 71, 74, 78, 81, 83, 88, 97, 100).

[218] Shuyan Li, Xiu Li, Jiwen Lu, and Jie Zhou. "Self-Supervised Video Hashing via Bidirectional Transformers". In: *CVPR*. 2021 (cit. on pp. 72, 93, 94).

[219] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection". In: *CVPR*. 2022 (cit. on pp. 71, 72, 80, 83, 88, 102–104, 106–108).

[220] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. "Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog". In: *IEEE/ACM TASLP* (2021) (cit. on pp. 71, 73, 86, 90, 97, 98).

[221] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2022), pp. 6999–7019. DOI: 10.1109/TNNLS.2021.3084827 (cit. on p. 4).

[222] Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. "Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition". In: *Proc. Interspeech 2020* (2020), pp. 394–398 (cit. on p. 41).

[223] Ji Lin, Chuang Gan, and Song Han. "TSM: Temporal Shift Module for Efficient Video Understanding". In: *ICCV*. 2019 (cit. on p. 71).

[224] Ji Lin, Chuang Gan, and Song Han. "TSM: Temporal Shift Module for Efficient Video Understanding". In: *ICCV*. 2019 (cit. on p. 88).

[225] Jingxu Lin and Sheng-hua Zhong. "Bi-Directional Self-Attention with Relative Positional Encoding for Video Summarization". In: *ICTAI*. 2020 (cit. on pp. 72, 75, 95, 101).

[226] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. "A Survey of Transformers". In: *AI Open* (2022) (cit. on p. 69).

[227] Yun-Shao Lin and Chi-Chun Lee. "Using interlocutor-modulated attention blstm to predict personality traits in small group interaction". In: *International Conference on Multimodal Interaction*. 2018, pp. 163–169 (cit. on p. 42).

[228] Sally Lindsay, Kara Grace Hounsell, and Celia Cassiani. "A scoping review of the role of LEGO® therapy for improving inclusion and social skills among children and youth with autism". In: *Disability and health journal* 10.2 (2017), pp. 173–182 (cit. on p. 20).

[229] Hanyuan Liu, Chengze Li, Xueting Liu, and Tien-Tsin Wong. "Neural Recognition of Dashed Curves With Gestalt Law of Continuity". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 1373–1382 (cit. on p. 6).

[230] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. "FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting". In: *ICCV*. 2021 (cit. on pp. 72–74, 90, 91, 100, 101, 110, 117).

[231] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. "HiT: Hierarchical Transformer With Momentum Contrast for Video-Text Retrieval". In: *ICCV*. 2021 (cit. on pp. 71, 86, 93, 98, 99).

[232] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. "A Survey of Visual Transformers". In: *arXiv* (2021) (cit. on p. 69).

[233] Yen-Ting Liu, Yu-Jhe Li, Fu-En Yang, Shang-Fu Chen, and Yu-Chiang Frank Wang. "Learning hierarchical self-attention for video summarization". In: *ICIP*. 2019 (cit. on pp. 72, 73, 101, 110).

[234] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. "Swin transformer v2: Scaling up capacity and resolution". In: *CVPR*. 2022 (cit. on pp. 71, 75, 77, 80, 83, 88, 91, 94, 103, 104, 108, 110).

[235] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: (2021) (cit. on pp. 80, 91).

[236] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video Swin Transformer". In: *CVPR*. 2022 (cit. on pp. 69, 71, 72, 74, 75, 77, 79, 80, 83, 88, 91, 97, 104, 106).

[237] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks". In: *NeurIPS*. 2019 (cit. on p. 87).

[238] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. "Frozen Pretrained Transformers as Universal Computation Engines". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.7 (June 2022), pp. 7628–7636. DOI: 10.1609/aaai.v36i7.20729. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20729 (cit. on pp. 7, 111).

[239] Mingshuang Luo, Shuang Yang, Xilin Chen, Zitao Liu, and Shiguang Shan. "Synchronous Bidirectional Learning for Multilingual Lip Reading". In: *BMVC*. 2020 (cit. on pp. 72, 90, 96).

[240] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. "Unsupervised video summarization with adversarial lstm networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 202–211 (cit. on p. 72).

[241] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. "On the effectiveness of task granularity for transfer learning". In: *arXiv* (2018) (cit. on pp. 70, 102).

[242] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. "On the Robustness of Vision Transformers to Adversarial Examples". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Aug. 2021, pp. 7838–7847 (cit. on p. 110).

[243] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. "Dialoguernn: An attentive rnn for emotion detection in conversations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 6818–6825 (cit. on p. 41).

[244] Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro. "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection". In: *Workshop on Tagging, mining and retrieval of human related activity information*. 2007, pp. 9–14 (cit. on p. 13).

[245] Davide Marengo, Christian Montag, et al. "Digital phenotyping of big five personality via facebook data mining: a meta-analysis". In: *Digital Psychology* 1.1 (2020), pp. 52–64 (cit. on p. 42).

[246] Herb Marsh, Benjamin Nagengast, and Alexandre Morin. "Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects." In: *Developmental psychology* 49 (Jan. 2012) (cit. on p. 66).

[247] Andrew David Marshall, Paul L Rosin, Jason Vandeventer, and Andrew Aubrey. "4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations". In: *Auditory-Visual Speech Processing,{AVSP} 2015* (2015), pp. 157–162 (cit. on p. 17).

[248] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. "Actions in context". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 2929–2936 (cit. on p. 13).

[249] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78 (cit. on p. 26).

[250] Michael McCloskey and Neal J Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation*. 1989 (cit. on p. 95).

[251] Robert R. McCrae and Oliver P. John. "An Introduction to the Five-Factor Model and Its Applications". In: *Journal of Personality* 60.2 (1992), pp. 175–215 (cit. on pp. 15, 31).

[252] Paria Mehrani and John K. Tsotsos. "Self-attention in vision transformers performs perceptual grouping, not attention". In: *Frontiers in Computer Science* 5 (2023). ISSN: 2624-9898. DOI: 10.3389/fcomp.2023.1178450. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1178450 (cit. on p. 6).

[253] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. "Recent trends in deep learning based personality detection". In: *Artificial Intelligence Review* (2019), pp. 1–27 (cit. on p. 42).

[254] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. "Trackformer: Multi-object tracking with transformers". In: *CVPR*. 2022 (cit. on pp. 71, 81–83, 88, 91, 99, 100).

[255] Yaw M Mensah and Hsiao-Yin Chen. "Global clustering of countries by culture–an extension of the GLOBE study". In: *Available at SSRN 2189904* (2013) (cit. on p. 47).

[256] Hugo Mercier and Dan Sperber. *The enigma of reason*. Harvard University Press, 2017 (cit. on p. 8).

[257] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information". In: *Image and Vision Computing* 31.2 (2013), pp. 137–152 (cit. on p. 41).

[258] Angeliki Metallinou, Zhaojun Yang, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. "The USC CreativeIT Database of Multimodal Dyadic Interactions: From Speech and Full Body Motion Capture to Continuous Emotional Annotations". In: *Lang. Resour. Eval.* 50.3 (Sept. 2016), pp. 497–521. ISSN: 1574-020X. DOI: 10.1007/s10579-015-9300-0. URL: https://doi.org/10.1007/s10579-015-9300-0 (cit. on p. 17).

[259] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. "Thinking Fast and Slow: Efficient Text-to-Visual Retrieval With Transformers". In: *CVPR*. 2021 (cit. on pp. 41, 71, 81, 90, 97–99).

[260] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. "End-to-End Learning of Visual Representations From Uncurated Instructional Videos". In: *CVPR*. 2020 (cit. on p. 84).

[261] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips". In: *ICCV*. 2019 (cit. on p. 73).

[262] Philip Moore. "Do We Understand the Relationship between Affective Computing, Emotion and Context-Awareness?" In: *Machines* 5.3 (2017), p. 16 (cit. on pp. 15, 41).

[263] Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. "Emotion recognition from embedded bodily expressions and speech during dyadic interactions". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 663–669 (cit. on p. 17).

[264] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. "Automated prediction and analysis of job interview performance: The role of what you say and how you say it". In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, pp. 1–6 (cit. on p. 17).

[265] Shrikanth Narayanan and Panayiotis G Georgiou. "Behavioral signal processing: Deriving human behavioral informatics from speech and language". In: *Proceedings of the IEEE* 101.5 (2013), pp. 1203–1233 (cit. on p. 14).

[266] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers". In: *NeurIPS* (2021) (cit. on p. 116).

[267] Suvendu Kumar Nayak and Ananta Charan Ojha. "Data leakage detection and prevention: Review and research directions". In: *Machine Learning and Information Processing* (2020), pp. 203–212 (cit. on p. 33).

[268] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. "Video Transformer Network". In: *ICCV*. 2021 (cit. on pp. 71, 77, 80, 83, 91, 104, 108, 116).

[269] Laurent Son Nguyen, Alvaro Marcos-Ramiro, Martha Marrón Romera, and Daniel Gatica-Perez. "Multimodal Analysis of Body Communication Cues in Employment Interviews". In: *International Conference on Multimodal Interaction (ICMI)*. 2013, pp. 437–444 (cit. on p. 42).

[270] Olivia Nocentini, Laura Fiorini, Giorgia Acerbi, Alessandra Sorrentino, Gianmaria Mancioppi, and Filippo Cavallo. "A survey of behavioral models for social robots". In: *Robotics* 8.3 (2019), p. 54 (cit. on p. 15).

[271] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. "Personality trait classification via co-occurrent multiparty multimodal event discovery". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015, pp. 15–22 (cit. on p. 42).

[272] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization". In: *Distill* (2017) (cit. on pp. 4, 5).

[273] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv* (2018) (cit. on pp. 92, 94, 118).

[274] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural discrete representation learning". In: *arXiv* (2017) (cit. on p. 72).

[275] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. "A review on deep learning techniques for video prediction". In: *IEEE TPAMI* (2020) (cit. on p. 95).

[276] Nuria Osa, Roser Granero, Eva Penelo, Josep Domènech, and Lourdes Ezpeleta. "The Short and Very Short Forms of the Children's Behavior Questionnaire in a Community Sample of Preschoolers". In: *Assessment* 21 (Nov. 2013). DOI: 10.1177/1073191113508809 (cit. on p. 18).

[277] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning". In: *ECCV*. 2016 (cit. on p. 84).

[278] Daniel Ozer and Veronica Benet-Martínez. "Personality and the Prediction of Consequential Outcomes". In: *Annual review of psychology* 57 (Feb. 2006), pp. 401–21. DOI: 10.1146/annurev.psych.57.102904.190127 (cit. on p. 31).

[279] P Paggio, J Allwood, Jokinen Ahlsén, and K Jokinen. "The NOMCO Multimodal Nordic Resource-Goals and Characteristics". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 10) Valletta, Malta. May*. European Language Resources Association (ELRA), 2010, pp. 19–21 (cit. on pp. 16, 17).

[280] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. "Empathy in virtual agents and robots: A survey". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.3 (2017), pp. 1–40 (cit. on p. 15).

[281] Cristina Palmero, German Barquero, Julio CS Jacques Junior, Albert Clapés, Johnny Núnez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, et al. "Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR. 2022, pp. 4–52 (cit. on pp. 10, 25, 36, 64, 116).

[282] Georges Pariente. "The role of vision in prosimian behavior". In: *The study of prosimian behavior*. Ed. by Gerald A. Doyle and Robert D. Martin. Academic Press New York, 1979. Chap. 10, pp. 411–459 (cit. on p. 2).

[283] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. "A Decomposable Attention Model for Natural Language Inference". In: *Empirical Methods in Natural Language Processing*. 2016 (cit. on p. 75).

[284] Hye Jeong Park and Jae Hwa Lee. "Looking into the Personality Traits to Enhance Empathy Ability: A Review of Literature". In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 173–180 (cit. on p. 41).

[285] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *International conference on machine learning*. Pmlr. 2013, pp. 1310–1318 (cit. on p. 6).

[286] Alexander Pashevich, Cordelia Schmid, and Chen Sun. "Episodic Transformer for Vision-and-Language Navigation". In: *ICCV*. 2021 (cit. on pp. 72, 74, 86, 90, 95, 96, 116).

[287] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. "Context encoders: Feature learning by inpainting". In: *CVPR*. 2016 (cit. on pp. 91, 95).

[288] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. "Keeping your eye on the ball: Trajectory attention in video transformers". In: *NeurIPS* (2021) (cit. on pp. 71, 78, 104, 106, 110).

[289] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. "Support-set bottlenecks for video-text representation learning". In: *ICLR*. 2021 (cit. on pp. 71, 74, 79, 86, 98, 99, 110, 116).

[290] Mandela Patrick, Po-Yao Huang, Ishan Misra, Florian Metze, Andrea Vedaldi, Yuki M. Asano, and João F. Henriques. "Space-Time Crop & Attend: Improving Cross-Modal Video Representation Learning". In: *ICCV*. 2021 (cit. on pp. 71, 90, 92, 96, 97, 99, 111, 116).

[291] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman. "Structured learning of human interactions in TV shows". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.12 (2012), pp. 2441–2453 (cit. on p. 13).

[292] Sayak Paul and Pin-Yu Chen. "Vision transformers are robust learners". In: *Proceedings of the AAAI Conference on Artificial Intelligence (CAI)*. Vol. 36. 2. 2022, pp. 2071–2081 (cit. on pp. 6, 110).

[293] Olga Perepelkina, Evdokia Kazimirova, and Maria Konstantinova. "RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing". In: *International Conference on Speech and Computer*. Springer. 2018, pp. 501–510 (cit. on p. 17).

[294] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. "Temporal-Relational CrossTransformers for Few-Shot Action Recognition". In: *CVPR*. 2021 (cit. on pp. 71, 74, 90, 97).

[295] Le Vy Phan and John F Rauthmann. "Personality computing: New frontiers in personality assessment". In: *Social and Personality Psychology Compass* 15.7 (2021), e12624 (cit. on p. 15).

[296] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. "Multimodal recognition of personality traits in social interactions". In: *Proceedings of the 10th international conference on Multimodal interfaces*. 2008, pp. 53–60 (cit. on p. 42).

[297] Rosalind W Picard. *Affective computing*. MIT press, 2000 (cit. on p. 15).

[298] Steven Pinker. "The cognitive niche: Coevolution of intelligence, sociality, and language". In: *Proceedings of the National Academy of Sciences* 107.supplement_2 (2010), pp. 8993–8999 (cit. on p. 8).

[299] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. "Skeleton-based action recognition via spatial and temporal transformer networks". In: *CVIU* (2021) (cit. on p. 69).

[300] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information Fusion* 37 (2017), pp. 98–125 (cit. on p. 16).

[301] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 527–536 (cit. on pp. 13, 16).

[302] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. "Emotion recognition in conversation: Research challenges, datasets, and recent advances". In: *IEEE Access* 7 (2019), pp. 100943–100953 (cit. on p. 41).

[303] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving". In: *CVPR*. 2021 (cit. on pp. 87, 96).

[304] Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. "Hierarchical self-attention network for action localization in videos". In: *CVPR*. 2019 (cit. on pp. 72, 83, 90, 95, 100).

[305] Ricardo Darío Pérez Principi, Cristina Palmero, Julio C. S. Jacques Junior, and Sergio Escalera. "On the Effect of Observed Subject Biases in Apparent Personality Analysis from Audio-visual Signals". In: *IEEE Transactions on Affective Computing* 12.3 (2019), pp. 607–621 (cit. on pp. 43, 66).

[306] Senthil Purushwalkam and Abhinav Gupta. "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases". In: *NeurIPS* (2020) (cit. on p. 92).

[307] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. "Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos". In: *IEEE Signal Processing Letters* (2019) (cit. on p. 90).

[308] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. "Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation". In: *CVPR*. 2019 (cit. on pp. 71, 73, 74, 85, 90, 95, 97).

[309] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space". In: *NeurIPS* (2017) (cit. on p. 82).

[310] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. "Spatiotemporal contrastive video representation learning". In: *CVPR*. 2021 (cit. on pp. 92, 96).

[311] Boele Raad and Henri Schouwenburg. "Personality in learning and education: A review". In: *European Journal of Personality* 10 (Dec. 1996), pp. 303 –336 (cit. on p. 32).

[312] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *arXiv* (2021) (cit. on p. 71).

[313] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training". In: *OpenAI Preprint*. 2018 (cit. on p. 40).

[314] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners". In: *OpenAI Blog*. 2019 (cit. on p. 98).

[315] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. "Integrating Multimodal Information in Large Pretrained Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 2359–2369. DOI: 10.18653/v1/2020.acl-main.214. URL: https://www.aclweb.org/anthology/2020.acl-main.214 (cit. on p. 35).

[316] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. "Latent Video Transformer". In: *arXiv* (2020) (cit. on pp. 72, 75, 83, 91, 100, 117).

[317] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. "Stand-Alone Self-Attention in Vision Models". In: *NeurIPS*. 2019 (cit. on p. 90).

[318] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-shot text-to-image generation". In: *ICML*. 2021 (cit. on p. 94).

[319] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. "Self-supervised video transformer". In: *CVPR*. 2022 (cit. on pp. 71, 78, 83, 92, 93, 96, 104, 106, 108).

[320] Rajesh PN Rao and Dana H Ballard. "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature Neuroscience* (1999) (cit. on p. 94).

[321] John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. "The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics." In: *Journal of Personality and Social Psychology* 107.4 (2014), p. 677 (cit. on pp. 15, 59, 68).

[322] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. "Decoding children's social behavior". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 3414–3421 (cit. on pp. 16, 17).

[323] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *NeurIPS*. 2015 (cit. on pp. 71, 72, 74).

[324] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. "ImageNet-21K Pretraining for the Masses". In: *NeurIPS*. 2021 (cit. on pp. 73, 91).

[325] Brent W Roberts, Kate E Walton, and Wolfgang Viechtbauer. "Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies." In: *Psychological bulletin* 132.1 (2006), p. 1 (cit. on pp. 18, 64).

[326] Filipe Rodrigues and Francisco Pereira. "Deep learning from crowds". In: *AAAI Conference on AI*. 2018 (cit. on p. 91).

[327] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. "FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration". In: *IEEE International Conference on Computer Vision Workshops*. 2021 (cit. on p. 115).

[328] Mary Rothbart, Stephan Ahadi, Karen Hershey, and Philip Fisher. "Investigations of Temperament at Three to Seven Years: The Children's Behavior Questionnaire". In: *Child development* 72 (Sept. 2001), pp. 1394–408. DOI: 10.1111/1467-8624.00355 (cit. on p. 18).

[329] Debaditya Roy and Basura Fernando. "Action anticipation using pairwise human-object interactions and transformers". In: *IEEE TIP* (2021) (cit. on pp. 72, 74, 83).

[330] Ludan Ruan and Qin Jin. "Survey: Transformer based Video-Language Pre-training". In: *AI Open* (2022) (cit. on pp. 69, 92, 111).

[331] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. "Tokenlearner: Adaptive space-time tokenization for videos". In: *NeurIPS* (2021) (cit. on pp. 71, 79, 80, 83, 88, 103, 104, 108).

[332] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *CVPR*. 2018 (cit. on p. 72).

[333] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. "MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2020) (cit. on p. 115).

[334] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. "Self-supervised learning for videos: A survey". In: *arXiv* (2022) (cit. on pp. 91, 92, 95, 96, 111).

[335] Felix D Schönbrodt and Marco Perugini. "At what sample size do correlations stabilize?" In: *Journal of Research in Personality* 47.5 (2013), pp. 609–612 (cit. on pp. 27, 67).

[336] Ullica Segerstrale and Peter Molnár. *Nonverbal communication: where nature meets culture*. Routledge, 2018 (cit. on pp. 15, 18).

[337] Hongje Seong, Junhyuk Hyun, and Euntai Kim. "Video multitask transformer network". In: *ICCV*. 2019 (cit. on pp. 72, 79, 81, 83, 88, 90).

[338] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *ACL*. 2019 (cit. on p. 116).

[339] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. "Only time can tell: Discovering temporal data for temporal modeling". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021 (cit. on p. 88).

[340] Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. "Temporal Context Aggregation for Video Retrieval with Contrastive Learning". In: *WACV*. 2021 (cit. on pp. 71, 93, 98, 110).

[341] Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. "Personality Recognition by Modelling Person-specific Cognitive Processes using Graph Representation". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 357–366 (cit. on p. 43).

[342] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. "Self-Attention with Relative Position Representations". In: *NAACL*. 2018 (cit. on p. 75).

[343] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wangchun Woo. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in Neural Information Processing Systems (NeurIPS)* 28 (2015) (cit. on pp. 72, 73).

[344] Andrew Shin, Masato Ishii, and Takuya Narihira. "Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision". In: *IJCV* (2022) (cit. on p. 69).

[345] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *NeurIPS*. 2014 (cit. on p. 87).

[346] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv* (2014) (cit. on p. 72).

[347] Linda Smith and Michael Gasser. "The development of embodied cognition: Six lessons from babies". In: *Artificial Life* (2005) (cit. on p. 84).

[348] Gizem Sogancioglu, Heysem Kaya, and Albert Ali Salah. "Can mood primitives predict apparent personality?" In: *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2021, pp. 1–8 (cit. on p. 42).

[349] Eric Sanders Rob van Son Wieneke Wesseling and Henk van den Heuvel. "The IFADV Corpus: a Free Dialog Video Corpus". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias. http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. ISBN: 2-9517408-4-0 (cit. on p. 17).

[350] Siyang Song, Shashank Jaiswal, Enrique Sanchez, Georgios Tzimiropoulos, Linlin Shen, and Michel Valstar. "Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition". In: *IEEE Transactions on Affective Computing* (2021) (cit. on p. 43).

[351] Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. "Learning Graph Representation of Person-specific Cognitive Processes from Audio-visual Behaviours for Automatic Personality Recognition". In: *arXiv preprint arXiv:2110.13570* (2021) (cit. on p. 32).

[352] Christopher Soto and Oliver John. "The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model With 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power". In: *Journal of Personality and Social Psychology* 113 (July 2017), pp. 117–143. DOI: 10.1037/pspp0000096 (cit. on pp. 15, 18, 29).

[353] Christopher J. Soto. "How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project". In: *Psychological Science* 30.5 (2019), pp. 711–727. DOI: 10.1177/0956797619831612 (cit. on pp. 31, 32).

[354] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. "Predicting personality from patterns of behavior collected with smartphones". In: *Proceedings of the National Academy of Sciences* 117.30 (2020), pp. 17680–17687 (cit. on p. 42).

[355] Alexandros Stergiou and Ronald Poppe. "Analyzing human–human interactions: A survey". In: *Computer Vision and Image Understanding* 188 (2019), p. 102799 (cit. on p. 16).

[356] Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. "Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations". In: *Transactions on Audio, Speech, and Language Processing* 24.4 (2016), pp. 733–744 (cit. on p. 42).

[357] Rui Su, Qian Yu, and Dong Xu. "STVGBert: A Visual-Linguistic Transformer Based Framework for Spatio-Temporal Video Grounding". In: *ICCV*. 2021 (cit. on pp. 72, 74, 81, 83, 87, 88, 90).

[358] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal. "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features". In: *ECCVW*. 2016 (cit. on p. 42).

[359] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions". In: *ICMI*. 2013, pp. 3–10 (cit. on p. 42).

[360] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. "Contrastive Bidirectional Transformer for Temporal Representation Learning". In: *arXiv* (2019) (cit. on pp. 71, 73, 74, 86, 90, 93, 94, 96, 97, 111).

[361] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. "Videobert: A joint model for video and language representation learning". In: *ICCV*. 2019 (cit. on pp. 71, 74, 75, 79, 85, 86, 94, 97, 99, 110, 111).

[362] Chen Sun, Arsha Nagrani, Yonglong Tian, and Cordelia Schmid. "Composable Augmentation Encoding for Video Representation Learning". In: *ICCV*. 2021 (cit. on pp. 71, 90, 92, 93, 96, 97).

[363] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5227–5237 (cit. on pp. 7, 111, 116).

[364] Yi-Lin Sung, Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. "Video Summarization with Anchors and Multi-Head Attention". In: *ICIP*. 2020 (cit. on p. 101).

[365] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *AAAI*. 2017 (cit. on p. 72).

[366] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *CVPR*. 2015 (cit. on p. 72).

[367] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. "Relaxed Transformer Decoders for Direct Action Proposal Generation". In: *ICCV*. 2021 (cit. on pp. 41, 72, 75, 90, 95, 98, 110).

[368] Adriana Tapus and Maja J Matarić. "User personality matching with a hands-off robot for post-stroke rehabilitation therapy". In: *Experimental robotics*. Springer. 2008, pp. 165–175 (cit. on p. 32).

[369] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. "Efficient transformers: A survey". In: *ACM CSUR* (2020) (cit. on p. 69).

[370] Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. "What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube". In: *IEEE Transactions on Affective Computing* 6.2 (2014), pp. 193–205 (cit. on p. 43).

[371] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality". In: *CVPR*. 2022 (cit. on p. 116).

[372] Yuandong Tian, Xinlei Chen, and Surya Ganguli. "Understanding self-supervised learning dynamics without contrastive pairs". In: *ICML*. 2021 (cit. on p. 93).

[373] Nancy T Tippins, Frederick L Oswald, and S Morton McPhail. "Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action". In: *Personnel Assessment and Decisions* 7.2 (2021), p. 1 (cit. on p. 33).

[374] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *NeurIPS*. 2022 (cit. on pp. 40, 71, 74, 83, 89, 91, 94, 95, 103–106, 108, 109).

[375] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: *ICCV*. 2015 (cit. on pp. 72, 88).

[376] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. "Video classification with channel-separated convolutional networks". In: *CVPR*. 2019 (cit. on pp. 104, 105).

[377] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. "A closer look at spatiotemporal convolutions for action recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Nov. 2018, pp. 6450–6459 (cit. on pp. 46, 47, 56, 71, 72).

[378] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. "Direcformer: A directed attention in transformer approach to robust action recognition". In: *CVPR*. 2022 (cit. on pp. 71, 78, 83, 91, 103–106, 108, 110).

[379] John K Tsotsos. "Analyzing vision at the complexity level". In: *Behavioral and brain sciences* 13.3 (1990), pp. 423–445 (cit. on p. 2).

[380] John K. Tsotsos. "Complexity Level Analysis Revisited: What Can 30 Years of Hindsight Tell Us about How the Brain Might Represent Visual Information?" In: *Frontiers in Psychology* 8 (2017). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01216. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01216 (cit. on p. 2).

[381] Marion K Underwood, Bertrina L Scott, Mikal B Galperin, Gretchen J Bjornstad, and Alicia M Sexton. "An observational study of social exclusion under varied conditions: Gender and developmental differences". In: *Child Development* 75.5 (2004), pp. 1538–1555 (cit. on p. 20).

[382] Fabio Valente, Samuel Kim, and Petr Motlicek. "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus". In: *Thirteenth annual conference of the international speech communication association*. 2012 (cit. on p. 42).

[383] David C Van Essen. "Organization of visual areas in macaque and human cerebral cortex". In: *The visual neurosciences* 1 (2003), pp. 507–521 (cit. on p. 2).

[384] David C Van Essen and John HR Maunsell. "Hierarchical organization and functional streams in the visual cortex". In: *Trends in neurosciences* 6 (1983), pp. 370–375 (cit. on p. 3).

[385] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 5998–6008. URL: `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf` (cit. on pp. 6, 35, 37, 38, 43, 55, 69, 76, 98).

[386] Alexandra Vella and Patrizia Paggio. "Overlaps in Maltese: a comparison between map task dialogues and multimodal conversational data". In: *NEALT Proceedings. Northern European Association for Language and Technology; 4th Nordic Symposium on Multimodal Communication; November 15-16; Gothenburg; Sweden*. 093. Linköping University Electronic Press. 2013, pp. 21–29 (cit. on p. 17).

[387] Ferran Viñas Poch, Monica González Carrasco, Eugenia Gras Pérez, Claustre Jane Ballabriga, and Ferran Casas Aznar. "Psychometric properties of the EATQ-R among a sample of Catalan-speaking Spanish adolescents". In: *Universitas Psychologica* 14.2 (2015), pp. 747–758 (cit. on p. 18).

[388] Alessandro Vinciarelli, Paraskevi Chatziioannou, and Anna Esposito. "When the Words are Not Everything: The Use of Laughter, Fillers, Back-Channel, Silence, and Overlapping Speech in Phone Calls". In: *Frontiers in ICT* 2 (2015), p. 4 (cit. on p. 13).

[389] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. "Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions". In: *Cognitive Computation* 7.4 (2015), pp. 397–413 (cit. on pp. 15, 41).

[390] Alessandro Vinciarelli and Gelareh Mohammadi. "A Survey of Personality Computing". In: *IEEE Transaction on Affective Computing* 5.3 (2014), pp. 273–291 (cit. on pp. 15, 16, 35, 52).

[391] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. "Social signal processing: Survey of an emerging domain". In: *Image and Vision Computing* 27.12 (2009), pp. 1743 –1759. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2008.11.007. URL: http://www.sciencedirect.com/science/article/pii/S0262885608002485 (cit. on p. 14).

[392] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. "Bridging the gap between social animal and unsocial machine: A survey of social signal processing". In: *IEEE Transactions on Affective Computing* 3.1 (2011), pp. 69–87 (cit. on p. 15).

[393] Seth A. Wagerman and David C. Funder. "Personality psychology of situations". In: *The Cambridge Handbook of Personality Psychology*. Ed. by Philip J. Corr and GeraldEditors Matthews. Cambridge Handbooks in Psychology. Cambridge University Press, 2009, 27–42 (cit. on p. 35).

[394] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. "High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 4).

[395] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. "Deep high-resolution representation learning for visual recognition". In: *TPAMI* (2020) (cit. on p. 71).

[396] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. "Long-short temporal contrastive learning of video transformers". In: *CVPR*. 2022 (cit. on pp. 71, 80, 91–93, 96, 103, 104, 106, 108).

[397] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. "Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking". In: *CVPR*. 2021 (cit. on pp. 71, 74, 90).

[398] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. "Bevt: Bert pretraining of video transformers". In: *CVPR*. 2022 (cit. on pp. 71, 80, 83, 94, 95, 104–106, 108).

[399] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pvt v2: Improved baselines with pyramid vision transformer". In: *Computational Visual Media* 8 (2022) (cit. on p. 72).

[400] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. "OadTR: Online Action Detection With Transformers". In: *ICCV*. 2021 (cit. on pp. 41, 72, 74, 90, 96).

[401] Xiaofang Wang, Xuehan Xiong, Maxim Neumann, AJ Piergiovanni, Michael S Ryoo, Anelia Angelova, Kris M Kitani, and Wei Hua. "Attentionnas: Spatiotemporal attention cell search for video classification". In: *ECCV*. 2020 (cit. on p. 73).

[402] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks". In: *CVPR*. 2018 (cit. on p. 40).

[403] Xiaolong Wang and Abhinav Gupta. "Unsupervised learning of visual representations using videos". In: *ICCV*. 2015 (cit. on p. 96).

[404] Yuhao Wang, Zhuoran Liu, Yibo Xia, Chunbo Zhu, and Danpei Zhao. "Spatiotemporal module for video saliency prediction based on self-attention". In: *Image and Vision Computing* (2021) (cit. on pp. 72, 73, 75, 90, 101).

[405] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. "End-to-End Video Instance Segmentation With Transformers". In: *CVPR*. 2021 (cit. on pp. 72, 74, 75, 90, 91, 101).

[406] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. "Masked feature prediction for self-supervised visual pre-training". In: *CVPR*. 2022 (cit. on pp. 40, 71, 80, 83, 89, 91, 94, 95, 103–106, 108).

[407] Yanna Weisberg, Colin Deyoung, and Jacob Hirsh. "Gender Differences in Personality across the Ten Aspects of the Big Five". In: *Frontiers in psychology* 2 (Aug. 2011), p. 178 (cit. on pp. 18, 66).

[408] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. "Scaling Autoregressive Video Models". In: *ICLR*. 2020 (cit. on pp. 72, 75, 77, 83, 90, 100, 117).

[409] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. "Event-Based Video Reconstruction Using Transformer". In: *ICCV*. 2021 (cit. on pp. 72, 73, 84, 89–91, 100, 110).

[410] Max Wertheimer. "A brief introduction to gestalt, identifying key theories and principles". In: *Psychol Forsch* 4.1 (1923), pp. 301–350 (cit. on p. 3).

[411] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation". In: *EMNLP-IJCNLP*. 2019 (cit. on p. 116).

[412] Kipling D Williams, Cassandra L Govan, Vanessa Croker, Daniel Tynan, Maggie Cruickshank, and Albert Lam. "Investigations into differences between social-and cyberostracism." In: *Group dynamics: Theory, research, and practice* 6.1 (2002), p. 65 (cit. on p. 18).

[413] Janine Willis and Alexander Todorov. "First impressions: Making up your mind after a 100-ms exposure to a face". In: *Psychological science* 17.7 (2006), pp. 592–598 (cit. on p. 43).

[414] Andrew P Witkin and Jay M Tenenbaum. "On the role of structure in vision". In: *Human and machine vision*. Elsevier, 1983, pp. 481–543 (cit. on p. 4).

[415] Aidan GC Wright. "Current directions in personality science and the potential for advances through computing". In: *IEEE Transactions on Affective Computing* 5.3 (2014), pp. 292–296 (cit. on pp. 15, 41).

[416] Chao-Yuan Wu and Philipp Krahenbuhl. "Towards long-form video understanding". In: *CVPR*. 2021 (cit. on pp. 71, 74, 92–94, 96).

[417] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. "Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition". In: *CVPR*. 2022 (cit. on pp. 72, 79–83, 89, 110).

[418] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. "Rethinking and improving relative position encoding for vision transformer". In: *ICCV*. 2021 (cit. on p. 75).

[419] Liangqing Wu, Dong Zhang, Qiyuan Liu, Shoushan Li, and Guodong Zhou. "Speaker personality recognition with multimodal explicit many2many interactions". In: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2020, pp. 1–6 (cit. on p. 42).

[420] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. "Hierarchical Self-Supervised Representation Learning for Movie Understanding". In: *CVPR*. 2022 (cit. on pp. 94, 96).

[421] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. "Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion". In: *Sensors* 21.14 (2021), p. 4913 (cit. on p. 42).

[422] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks". In: *CVPR*. 2017 (cit. on pp. 71, 73).

[423] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. "Rethinking spatiotemporal feature learning for video understanding". In: *ECCV* (2017) (cit. on pp. 71, 73, 94).

[424] Songlong Xing, Sijie Mai, and Haifeng Hu. "Adapted dynamic memory network for emotion recognition in conversation". In: *IEEE Transactions on Affective Computing* (2020) (cit. on p. 41).

[425] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. "Long short-term transformer for online action detection". In: *NeurIPS* (2021) (cit. on pp. 72, 81–83, 89).

[426] Peng Xu, Xiatian Zhu, and David A Clifton. "Multimodal learning with transformers: a survey". In: *arXiv* (2022) (cit. on p. 69).

[427] Qi Xu, Ming Zhang, Zonghua Gu, and Gang Pan. "Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs". In: *Neurocomputing* (2019) (cit. on p. 89).

[428] Yifan Xu, Huapeng Wei, Minxuan Lin, Yingying Deng, Kekai Sheng, Mengdan Zhang, Fan Tang, Weiming Dong, Feiyue Huang, and Changsheng Xu. "Transformers in computational visual media: A survey". In: *Computational Visual Media* (2022) (cit. on p. 69).

[429] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. "Learning spatio-temporal transformer for visual tracking". In: *ICCV*. 2021 (cit. on p. 71).

[430] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. "Multiview transformers for video recognition". In: *CVPR*. 2022 (cit. on pp. 71, 83, 84, 88, 89, 91, 103, 104, 106, 108).

[431] Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. "Multi-Document Transformer for Personality Detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 16. 2021, pp. 14221–14229 (cit. on p. 43).

[432] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. "Recurring the Transformer for Video Action Recognition". In: *CVPR*. 2022 (cit. on pp. 71, 81–83, 88, 104, 106).

[433] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. "WIDER FACE: A Face Detection Benchmark". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on p. 46).

[434] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. "Temporally Efficient Vision Transformer for Video Instance Segmentation". In: *CVPR*. 2022 (cit. on pp. 72, 77, 83, 88, 99, 100).

[435] Yuting Yang, Licheng Jiao, Xu Liu, Fang Liu, Shuyuan Yang, Zhixi Feng, and Xu Tang. "Transformers Meet Visual Learning Understanding: A Comprehensive Review". In: *arXiv* (2022) (cit. on p. 69).

[436] Zongxin Yang, Yunchao Wei, and Yi Yang. "Associating objects with transformers for video object segmentation". In: *NeurIPS* (2021) (cit. on pp. 72, 77, 81–83, 88, 89, 91, 100).

[437] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. "Referring segmentation in images and videos with cross-modal self-attention network". In: *TPAMI* (2021) (cit. on pp. 72, 75, 101).

[438] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. "A Fourier Perspective on Model Robustness in Computer Vision". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/b05b57f6add810d3b7490866d74c0` `Paper.pdf` (cit. on p. 4).

[439] Junbo Yin, Jianbing Shen, Chenye Guan, Dingfu Zhou, and Ruigang Yang. "Lidar-based online 3d video object detection with graph-based message passing and spatiotemporal transformer attention". In: *CVPR*. 2020 (cit. on pp. 72, 80, 83, 90).

[440] Bin Yu, Ming Tang, Linyu Zheng, Guibo Zhu, Jinqiao Wang, Hao Feng, Xuetao Feng, and Hanqing Lu. "High-Performance Discriminative Tracking With Transformers". In: *ICCV*. 2021 (cit. on p. 71).

[441] Bingyao Yu, Wanhua Li, Xiu Li, Jiwen Lu, and Jie Zhou. "Frequency-Aware Spatiotemporal Transformers for Video Inpainting Detection". In: *ICCV*. 2021 (cit. on pp. 71, 73–75, 90, 101, 110).

[442] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. "Learning from inside: Self-driven siamese sampling and reasoning for video question answering". In: *NeurIPS* (2021) (cit. on pp. 72, 83, 89, 93).

[443] Zhou Yu and Nanjia Han. "Accelerated masked transformer for dense video captioning". In: *Neurocomputing* (2021) (cit. on pp. 71, 74, 81, 90, 97, 98).

[444] Liangzhe Yuan, Rui Qian, Yin Cui, Boqing Gong, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. "Contextualized Spatio-Temporal Contrastive Learning with Self-Supervision". In: *CVPR*. 2022 (cit. on pp. 92, 96).

[445] Zhenxun Yuan, Xiao Song, Lei Bai, Zhe Wang, and Wanli Ouyang. "Temporal-Channel Transformer for 3D Lidar-Based Video Object Detection for Autonomous Driving". In: *Transactions on Circuits and Systems for Video Technology* (2021) (cit. on pp. 72, 90).

[446] Sukmin Yun, Jaehyung Kim, Dongyoon Han, Hwanjun Song, Jung-Woo Ha, and Jinwoo Shin. "Time Is MattEr: Temporal Self-supervision for Video Transformers". In: *ICML*. 2022 (cit. on pp. 71, 78, 91, 95, 101, 105, 106, 108, 110).

[447] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. "Multi-attention recurrent network for human communication comprehension". In: *AAAI Conference on Artificial Intelligence*. Vol. 2018. NIH Public Access. 2018, p. 5642 (cit. on p. 42).

[448] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. "Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer. 2007, pp. 464–475 (cit. on p. 17).

[449] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow twins: Self-supervised learning via redundancy reduction". In: *ICML*. 2021 (cit. on p. 116).

[450] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. "Learning Joint Spatial-Temporal Transformations for Video Inpainting". In: *ECCV*. 2020 (cit. on pp. 72, 74, 84, 89–91, 100, 101, 117).

[451] Xuefan Zha, Wentao Zhu, Lv Xun, Sen Yang, and Ji Liu. "Shifted Chunk Transformer for Spatio-Temporal Representational Learning". In: *NeurIPS* (2021) (cit. on pp. 71, 77, 78, 80, 83, 88, 90, 97, 103, 104, 106).

[452] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. "Scaling vision transformers". In: (2022) (cit. on p. 89).

[453] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M Dai, Ruoming Pang, and Fei Sha. "Co-training Transformer with Videos and Images Improves Action Recognition". In: *arXiv* (2021) (cit. on p. 95).

[454] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. "Delving Deep Into the Generalization of Vision Transformers Under Distribution Shifts". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 7277–7286 (cit. on pp. 6, 110).

[455] Lingyu Zhang, Indrani Bhattacharya, Mallory Morgan, Michael Foley, Christoph Riedl, Brooke Welles, and Richard Radke. "Multiparty Visual Co-Occurrences for Estimating Personality Traits in Group Meetings". In: *The IEEE Winter Conference on Applications of Computer Vision*. 2020, pp. 2085–2094 (cit. on p. 43).

[456] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. "Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos". In: *CVPR*. 2021 (cit. on pp. 41, 72, 74, 87, 90).

[457] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *CVPR*. 2018 (cit. on p. 91).

[458] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric". In: *CVPR*. 2018 (cit. on p. 95).

[459] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation". In: *European Conference on Computer Vision*. Springer. 2020, pp. 365–381 (cit. on p. 116).

[460] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. "Vidtr: Video transformer without convolutions". In: *ICCV*. 2021 (cit. on pp. 72, 74, 90, 96).

[461] Zhang Zhang and Dacheng Tao. "Slow Feature Analysis for Human Action Recognition". In: *IEEE TPAMI* (2012) (cit. on pp. 78, 87, 109).

[462] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. "From facial expression recognition to interpersonal relation prediction". In: *International Journal of Computer Vision* 126.5 (2018), pp. 550–569 (cit. on p. 41).

[463] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior". In: *International conference on intelligent virtual agents*. Springer. 2016, pp. 218–233 (cit. on pp. 15, 41).

[464] Sipeng Zheng, Shizhe Chen, and Qin Jin. "VRDFormer: End-to-End Video Visual Relation Detection With Transformers". In: *CVPR*. 2022 (cit. on pp. 71, 81–83, 99, 100).

[465] Peixiang Zhong, Di Wang, and Chunyan Miao. "Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 165–176 (cit. on p. 41).

[466] Honglu Zhou, Asim Kadav, Farley Lai, Alexandru Niculescu-Mizil, Martin Renqiang Min, Mubbasir Kapadia, and Hans Peter Graf. "Hopper: Multi-hop Transformer for Spatiotemporal Reasoning". In: *ICLR*. 2021 (cit. on pp. 71, 81, 91, 97).

[467] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. "End-to-end dense video captioning with masked transformer". In: *CVPR*. 2018 (cit. on pp. 71, 74, 81, 90, 97, 98).

[468] Linchao Zhu and Yi Yang. "Actbert: Learning global-local video-text representations". In: *CVPR*. 2020 (cit. on pp. 44, 54, 69, 71, 74, 75, 79, 85, 86, 89, 94, 99, 109, 110).