Deep Spatio-Temporal Neural Network for Facial Analysis

Decky Aspandi Latif

TESI DOCTORAL UPF / year 2020

THESIS SUPERVISOR
Xavier Binefa Valls
Department of Technology Information and Communication

To my mom, father, sisters, brothers and friends

To those who embarked on their journeys in search for knowledge

To the ones who are in their struggles for a brighter future

Acknowledgements

This thesis would not be possible without the helps and supports from everyone around me. Thus, I would like to specifically thanks following individuals given their significant roles throughout my study.

First of all I would like to convey my gratitude to my supervisor, Xavier, for giving me opportunity to advance my knowledge through this PhD study. You taught me a lot how to be patient during challenging times, and how to persist in each rough step I made during the study. All the independent times during this study and our discussions have renewed my life vision and reshaped my next life chapters for me to walk on. Thus, I sincerely appreciated all of your academic and non academic supports along with your commitment to this day. I also would like to thanks Federico for his helps during the research. The literacy and research skills that you exhibited and taught me are one of most important skills that I learned, and I appreciated during this study.

I would also like to mention my colleagues that I have met in CMTech : Adria, Dmytro, Oriol, Adriana, Mariceli, Quim and fellows from the other labs. Thank you for all of the discussions and interactions we made, it has been a fun and exciting time for me, and I will always remember you guys. Then, I would like to thanks the DTIC academic staffs for their supports throughout my study, and their patients for some headaches that I made in the past : Lydia, Lluis, Ruth, Jana and Montse. I also would like to thanks the students at UPF who I taught for being a source of my inspirations during the teaching. Lastly, my deepest and sincere thanks to my family and friends for their unconditional supports during difficult times.

Mengenai ini, saya ingin mengucapkan terima-kasih sedalam dalamnya kepada mama saya ibu Ipah, bapak saya pak Aspan, kakak Lia, adek Jaka dan adik Siti atas bantuan dan pengorbanannya selama ini. Saya bangga dan bersyukur bisa memiliki ibu, bapak, dan kakak adik yang selalu membantuku selama studi disini, di Spanyol, di Eropa. Semoga kalian juga saudara saudariku bisa lanjut studi yang tinggi sehingga kita bisa memiliki pemahaman yang lebih dalam. Hingga bisa membantu orang tua dan berkontribusi ke halayak ramai, ke negara Indonesia, dan ke dunia.

Abstract

Automatic Facial Analysis is one of the most important field of computer vision due to its significant impacts to the world we currently live in. Among many applications of Automatic Facial Analysis, Facial Alignment and Facial-Based Emotion Recognition are two most prominent tasks considering their roles in this field. That is, the former serves as intermediary steps enabling many higher facial analysis tasks, and the latter provides direct, real-world high level facial-based analysis and applications to the society. Together, they have significant impacts ranging from biometric recognition, facial recognition, health, and many others.

These facial analysis tasks are currently even more relevant given the emergence of big-data, that enables rapid development of machine learning based models advancing their current state of the arts accuracy. In regard to this, the uses of video-based data as the part of the development of current datasets have been more frequent. These sequence based data have been explicitly exploited in the other relevant machine learning fields through the use of inherent temporal information, that in contrast, it has not been the case for both of Facial Alignment and Facial-Based Emotion Recognition tasks. Furthermore, the in-the-wild characteristics of the data that exist on the current datasets present additional challenge for developing an accurate system to these tasks. In this context, the main purpose of this thesis is to evaluate the benefit of incorporating both temporal information and the in-the-wild data characteristics that are largely overlooked on both Facial Alignment and Facial-Based Emotion Recognition. We mainly focus in the use of deep learning based models given their capability and capacity to leverage on the current sheer size of input data. Also, we investigate the introduction of an internal noise modellings in order to assess their impacts to the proposed works.

Specifically, this thesis analyses the benefit of sequence modelling through progressive learning applied to facial tracking task, while it is also fully end to end trainable. This arrangement allows us to evaluate the optimum sequence length to increase the quality of our models estimation. Subsequently, we expand our investigations to the introduction of internal

noise modelling to benefit from the characteristics of each image degradation for single-image facial alignment, alongside the facial tracking task. Following this approach, we can study and quantify its direct impacts. We then combine both sequence based approach and internal noise modelling by proposing the unified systems that can simultaneously perform both of single-image facial alignment and facial tracking, with state of the art accuracy result.

Motivated by our findings from Facial Alignment task, we then expand these approaches to Facial-Based Emotion Recognition problem. We first explore the use of adversarial learning to enhance our image degradation modelling, and simultaneously increase the efficiency of our approaches through the formation of internal visual latent features. We then equip our base sequence modelling with soft attention modules to allow the proposed model to adjust their focus using the adaptive weighting scheme. Subsequently, we introduce a more effective fusion method for both facial features modality and visual representation of audio using gating mechanism. In this stage, we also analyse the impacts of our proposed gating mechanisms along with the attention enhanced sequence modelling. Finally, we found that these approaches improve our models estimation quality leading to the high level of accuracy, outperforming the results from other alternatives.

Resum

L'anàlisi facial es un dels camps importants en Visió per Ordinador degut a l'impacte que té en el mon on vivim. L'alineament facial i el reconeixement d'emocions basat en cares son dues tasques fonamentals en aquest camp. Mentre la primera tasca pot ser un pas intermedi per tasques d'anàlisi posterior, la segona aporta aplicacions directes, socialment útils. Les dues juntes tenen un impacte que va del reconeixement biomètric a captar l'estat emocional de la persona.

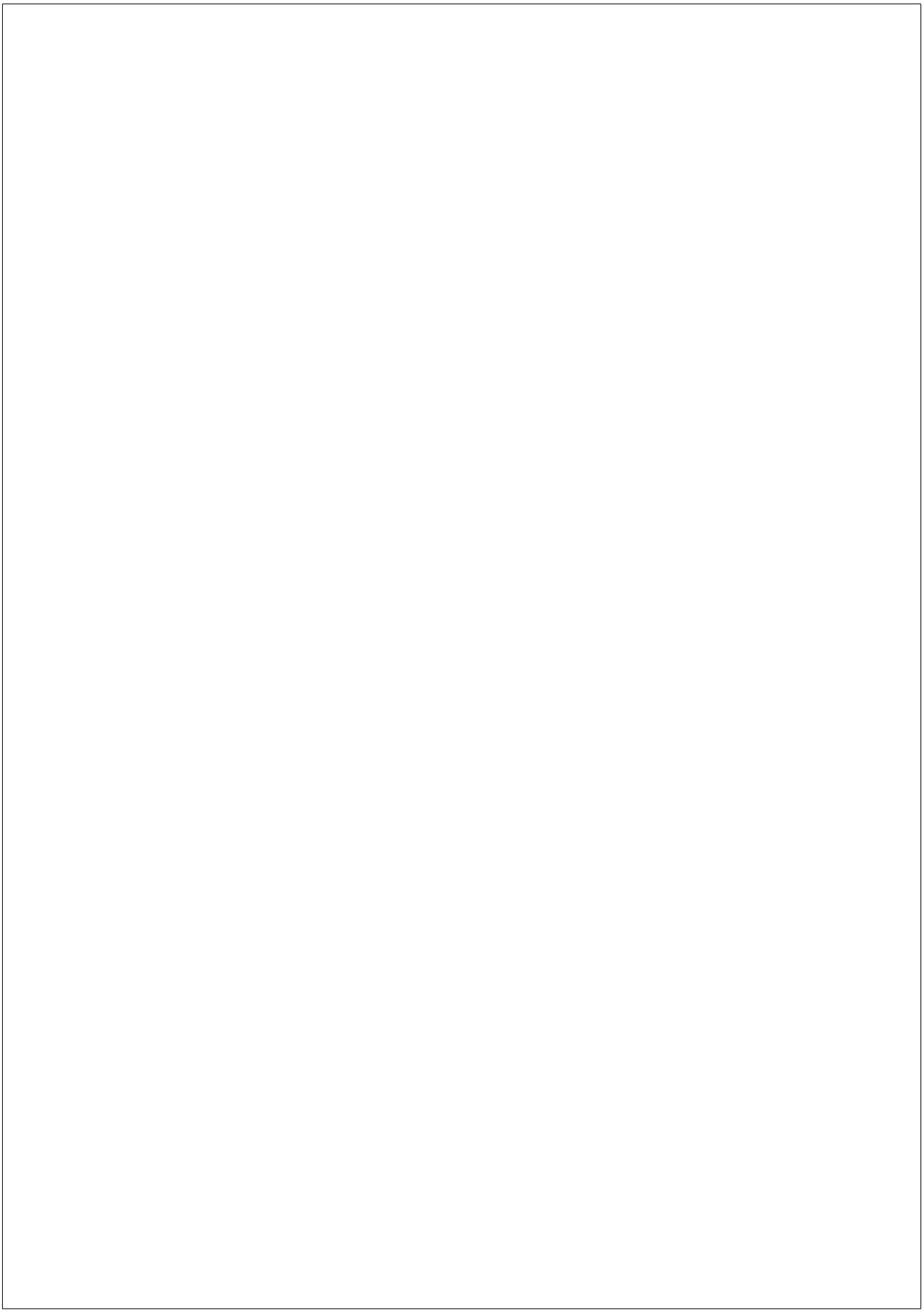
En l'era actual del Big Data, aquestes tasques d'anàlisi facial son encara més rellevants ja que es possible un progrés continuat de l'estat de l 'art. L'ús de grans bases de dades basades en vídeo ha permès l'ús de models temporals en l'aprenentatge automàtic i en Visió per Ordinador. Malgrat això, l'ús de models temporals es encara insuficient. A més a més, la presentació de les dades en forma natural -sense restriccions- afegeix nous desafiaments per desenvolupar sistemes precisos. En aquest context, el principal objectiu d'aquesta tesi consisteix en avaluar el benefici d'incorporar les dues coses, informació temporal i dades amb característiques naturals ja que aquests fets encara es tenen poc en compte tant en l'alineament facial com en el reconeixement d'emocions facials. Ens centrarem principalment en l'ús de models basats en l'aprenentatge profund, atesa la seva capacitat per aprofitar grans quantitats de dades, i també utilitzarem el modelatge del soroll en les dades per avaluar l'impacte sobre els algoritmes desenvolupats.

Concretament, en aquesta tesi s'analitza l'impacte de modelar les seqüències mitjançant aprenentatges progressius aplicades al seguiment facial i que es poden aprendre del principi al final. D'aquesta manera podem avaluar la longitud temporal òptima per evitar una precisió subòptima. Posteriorment, investiguem la incorporació de models de soroll interns per poder treure profit de les característiques de cada degradació visual i aconseguir l'alineació facial de cada imatge. D'aquesta manera, podem estudiar-ne els impactes i quantificar-ne els efectes directes. A continuació , combinant tant el modelatge basat en seqüències com el modelat de soroll intern, vam crear un sistema unificat que pot realitzar un seguiment de la

imatge i del rostre amb precisió.

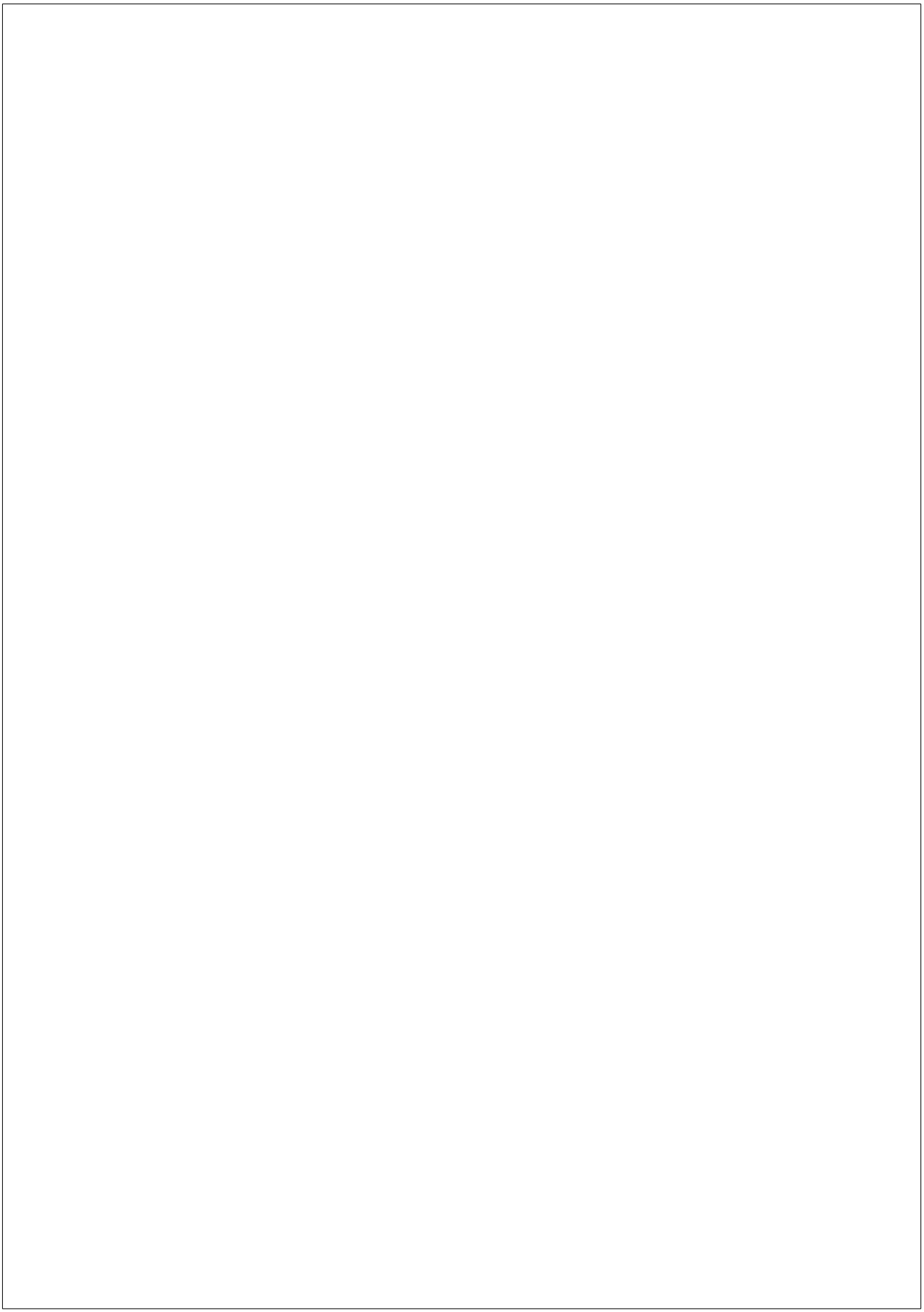
Aquest model de seguiment de l’alineació facial robust a imprevistos i a degradacions, l’ampliem a la computació afectiva, basada en el reconeixement d’emocions facials. Explorem primer l’ús de l’aprenentatge adversari per millorar tan el model de degradació de la imatge com el model de característiques latents. D’aquí resulta una millora de l’eficiència del sistema. A continuació, equipem el model amb mòduls d’atenció per deixar que el model processi la seqüència segons aquesta ponderació adaptativa. Finalment, introduïm un mètode de fusió més eficaç tant per model de trets facials com per a la representació visual d’àudio mitjançant un mecanisme de selecció (gated). A més, també analitzem els impactes d’aquests mecanismes de selecció i el modelatge de seqüències millorat per l’atenció. Hem trobat que aquests enfocaments milloren la qualitat de la nostra estimació i hem aconseguit la precisió actual de l’estat de l’art.

Contents



List of Figures

List of Tables



Chapter 1

INTRODUCTION

1.1 Automatic Facial Analysis

The emergence of *big-data*, especially in the form of video - based information has enabled rapid development of state of the art computer vision systems, especially to automatic facial analysis [Fasel and Luettin, 2003, Chow and Li, 1993] field. These can be seen on the rapid progress toward two essential tasks of this field: Facial Alignment and Automatic Facial Behavior Analysis. These two related, and inter-twined tasks, have provided a solid foundation for many of their applications to the society. Specifically, Facial Alignment serves as an essential task that enables more precise process for higher level facial analysis, such as facial recognition [Zhu et al., 2015b], head pose estimation [Wu et al., 2017] and including the affective analysis [Gopalan et al., 2018], which is part of Affective Computing Field. Furthermore, the Automatic Facial Behavior Analysis in other hand, provides higher level facial-based analysis that attempts to bridge the human affect with the advance of computer science that are importance to many fields, including health [Alhussein, 2016] and educations [Chen et al., 2013] among many others.

Both Facial Alignment and Facial Behaviour Analysis tasks have been directly influenced by the emergence of current rapid development of *big-data*, that can be seen by rapid development of respective datasets and

improvement state of the arts. However there are still several challenges to further advance current progress due to the nature of *big-data* itself resulting in the current lack of the respective analysis. One such characteristic is the sequential properties that are inherent to all varieties of video-based data, which are increasingly more common nowadays. These temporal information and their associated functions have been extensively investigated and exploited on the other computer science disciplines, such as natural language processing [Sha et al., 2016, Yao and Guan, 2018] for instance, with high level of success [Michael et al., 2019, Sutskever et al., 2014, Luong et al., 2015]. Another challenging property of the *big-data* is their *in-the-wild* characteristics which are native to all of the data that are currently taken in the unconstrained conditions. Lastly, their huge sizes also present other difficulties for any efficient modeling.

Given aforementioned shortcomings, this thesis tries to incorporate two fundamental aspects of spatial and temporal modelling into both of Facial Alignment and Facial Behavior Analysis tasks. We focus in creating an accurate models that can be utilised into these related tasks, and simultaneously introduce several mechanisms to enhance their predictions through effective learning mechanisms. We also concentrate on the utilization of fully end-to-end approaches to be able to fully benefit from the sheer size of the dataset used in these tasks. In the next section, we will provide relevant descriptions of both Facial Alignment and Facial Behavior Analysis along with their current challenges, that we address in this thesis.

1.1.1 Facial Alignment

Facial Alignment is a crucial task that serves as the primary processing pipeline to almost all spectrum of facial analysis, by providing more manageable and uniform data representations to be used for their specific applications, that includes face recognition [Zhu et al., 2015b], head pose estimation [Wu et al., 2017], facial reenactment [Thies et al., 2016] and emotion recognition [Kossaifi et al., 2019] (part of Facial Behaviour Analysis that we will explain on next Section). Figure 1.1 shows The

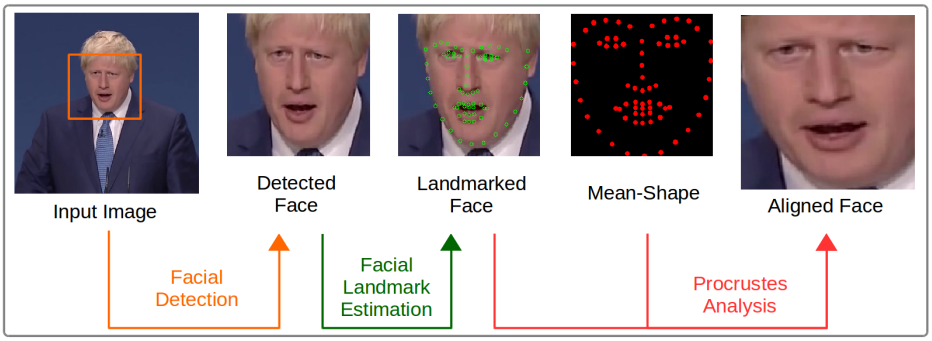


Figure 1.1: Standard pipeline of Facial Alignment: Facial Detection, Facial Landmark detection, and Procrustes Analysis.

standard Facial Alignment process that consists of three sub-processes: Facial Detection, Facial Landmark Estimations and Procrustes Analysis.

Facial Detection task is responsible to provide non-deformable, rough estimates of facial location, typically in the form of bounding box. Using this information, Facial Landmark Detection then estimates the predefined set of facial fiducial points, that will be used as reference for Procrustes Analysis to align the image with the associated mean-shape points of the face (typically calculated from the whole samples of dataset). In this thesis, we focus on the facial landmark estimation task given its central function for Facial Alignment. That is, the precise facial landmark estimations can subset or correct the rough estimated area from previous facial bounding box information, and it also provides fundamental information for Procrustes Analysis to function properly. The review of current state of the art facial detection can be found on [Kumar et al., 2019] while the description of the standardized Procrustes Analysis method can be found on [Gower, 1975].

Facial landmark estimations are generally conducted by predicting a predefined set of landmark positions which provide a structural representation of the facial geometry. Traditionally, these points have been estimated using either shape and appearance models [Yuille et al., 1992] or regres-

sion based models [Kazemi and Sullivan, 2011], with the latter offering some advantage due to their computational efficiency [Valstar et al., 2010]. Furthermore, due to the emergence of *big-data*, the use of deep learning approaches are also increased and improves the current state of the art accuracy significantly. However, there are still challenges for current facial landmark estimation models when targeting the images captured *in-the-wild*, that are increasingly frequents nowadays. These image degradations are quite varied, and can include large appearance variations due to heavy occlusion, severe pose or illuminations conditions, etc [Zhang et al., 2018a]. Some of current approaches to tackle this problems include the removal of outlier [Qu et al., 2015] and combining data and model driven estimators to enhance the robustness[Zhang et al., 2018a]. However, such approaches are largely built on assumption that that the model will be able to discard the noise and select only the meaningful features, effectively discarding the relevant information, that may beneficial when properly included in their learning process [Dong et al., 2018, Nada et al., 2018, Goswami et al., 2018].

Another aspect in facial landmark estimation tasks is related to the type of the data they are dealing with. When video data is used as the input, i.e in the form of multiple continuous frames, then this facial landmark estimations are translated into facial landmark tracking task [Shen et al., 2015, Chrysos et al., 2017]. The current most popular approach used in this task is Tracking-by-Detection method, that essentially works by converting existing facial landmark detection to perform estimations for each frames individually[Uricár et al., 2015, Zadeh et al., 2017]. However, these approaches have a fundamental limitation of not considering temporal information contained in the video sequences that have been shown to be beneficial on other machine learning field [Sha et al., 2016]. Thus, current approaches also have started their attempts to incorporate these information, even though they are still limited to include the adjacent frames [Yang et al., 2015, Chrysos et al., 2017].

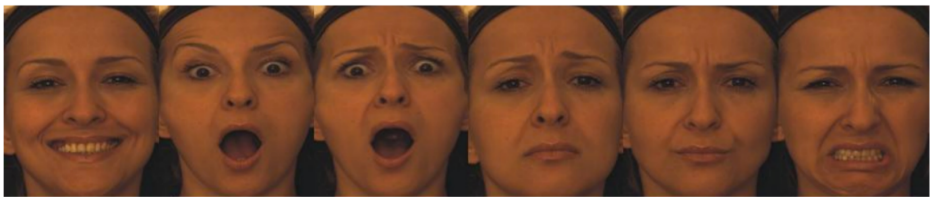


Figure 1.2: The six universal facial expressions. Samples are obtained from [Savran et al., 2008].

1.1.2 Facial Behavior Analysis

Facial Behavior Analysis is part of Affective Computing Field, an interdisciplinary research field that attempts to equip computers with the human-like capabilities to understand the human affects [Tao and Tan, 2005]. Given its crucial function, Affective Computing field has recently attracted the attention of the research community and applied them to diverse areas which include education [Duo and Song, 2010] or healthcare [Liu et al., 2008], among others. One of primary focus on Affective Computing is to build an automatic affect recognition that aims to estimate the personal emotional states given perceived sensory inputs. This automatic emotion recognition task has advanced quite rapidly due to the recently introduced big datasets allowing for more development of machine-learning based models with quite successful outcomes [Kossaifi et al., 2017, McKeown et al., 2010]. To this end, Facial Behaviour Analysis, that uses facial cues to infer respective personal emotional states is gaining more popularity compare to the other bio-signal [Correa et al., 2018] or speech based [El Ayadi et al., 2011] emotion recognition due to their intuitive nature [Tian et al., 2001], [Yeasin et al., 2006], [Comas et al., 2020].

One central task of Facial Behaviour Analysis, namely the facial-based emotion recognition, has developed quickly recently [Fasel and Luettin, 2003, Pantic and Rothkrantz, 2000] that can be seen

on the development of emotion dimensions type that are usually considered as perceived emotional labels. Originally, the user emotional dimension was characterises using six basic human emotions [Ekman and Friesen, 1971] such as happy, sad, angry, surprise, fear anger and disgust (the examples can be seen in Figure 1.2). Furthermore, it progressed toward the use of pre-defined 46 facial action based on facial muscle movements, that was developed in part of Facial Action Coding System (FACS) [Ekman et al., 1978]. In regard to this, their combinations can produce the previously mentioned emotional states as well [Savran et al., 2008] (the examples of several facial units can be seen in Figure 1.3). Recently, the use of Valence and Arousal dimension [Russell, 1980] to model the emotional states has become popular due its continuous nature, allowing it to represents wide array of emotional states (See Figure 1.4 for the respective visualization examples). This rapid progress, can also be reflected by growing number of datasets that primarily use this emotion representation for their experiments [Kossaifi et al., 2019, Kollias et al., 2020]. Because of this aspect, we chose to target this affect dimension in our facial-based emotion recognition task as a part of our focus in this thesis.

Automatic Valence/Arousal based emotion recognition task evolves quite considerably lately, especially with the availability of video based datasets [McKeown et al., 2010, Correa et al., 2018, Kossaifi et al., 2017]. This development further sparks the interest to use the inherent sequential information in their respective estimation recognition models [Kossaifi et al., 2017, Poria et al., 2019]. However, similar to the current state of facial alignment task, the progress to be able to fully utilise this information is still largely slow and overlooked, along with systematic analysis of their impacts that are also largely unexplored. One of the reason of these shortcomings may be attributed to the challenging characteristics of the *in-the-wild* data that usually consists of several image degradation types, and furthermore, their sheer sizes can present additional difficulties to create such accurate models [Kollias et al., 2020]. One way to improve the current facial based emotion recognition is to use additional modality in conjunction of facial features [Poria et al., 2019,

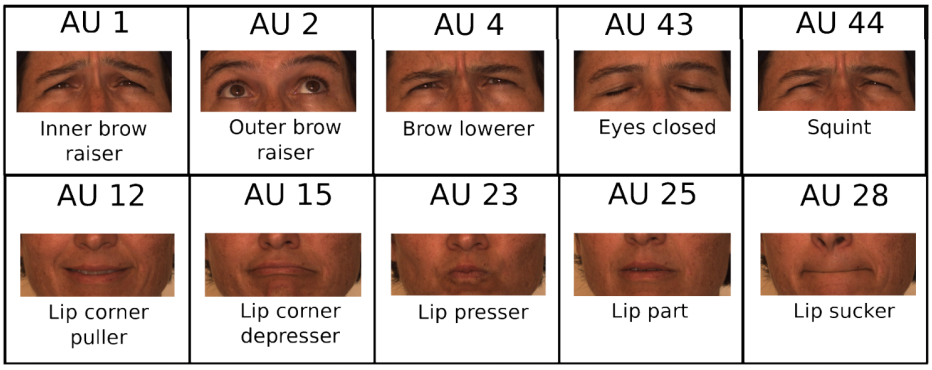


Figure 1.3: Examples of Action Units described in FACS. The samples are obtained from Bosphorus 3D facial expression database [Savran et al., 2008].

Comas et al., 2020, Tu et al., 2017]. However, typically the current approaches utilise a straight-forward approach through concatenation of these modalities that may resulted in the sub-optimal results [Zhang et al., 2018b], as it neglects relationships among these different modalities as shown in the other machine vision fields [Zhou et al., 2019].

1.2 Motivation and Contributions

Based on the current challenges faced on both Facial Landmark Estimations and Facial-Based Emotion Recognition tasks, we can ask this following question: *”Is it possible for us to include the temporal information, while simultaneously incorporate the inherent in-the-wild data characteristics to improve current state of the art in these related tasks?”*. In this respect, we defined the main objective of this thesis to address this particular question. Specifically, in this thesis, we focus on the investigations of temporal modelling to enhance the estimations of our proposed systems, that are relevant enough to be applied on these two related tasks. To achieve this, we use the well established Long Short

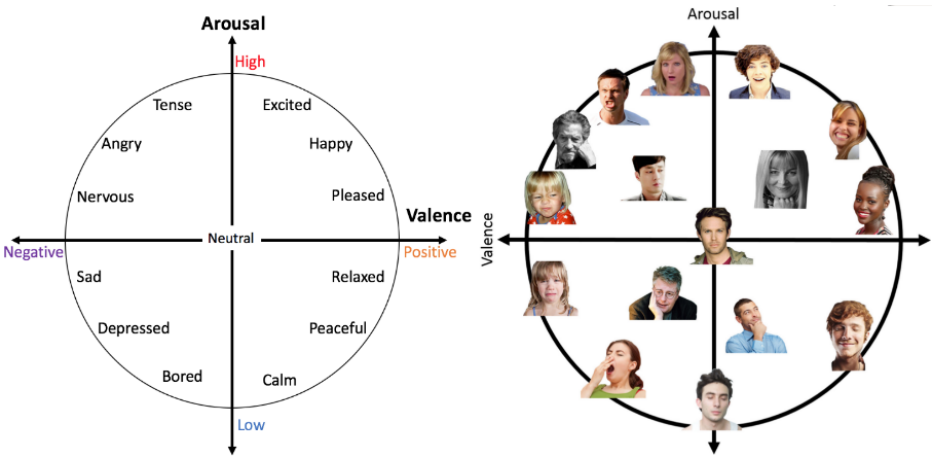


Figure 1.4: The left figure shows the example of Circumplex Model Diagram [Russell, 1980], the right figure provides the examples of facial expressions on their respective diagram positions. Image visualizations are obtained from [Chang et al., 2017], and [Mollahosseini et al., 2015].

Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] network as a principal sequential modeling as opposed to classical machine learning models such as Hidden Markov Model [Fine et al., 1998] and Bayesian Filtering [Chen et al., 2003], due to its end-to-end nature that enables our proposed system to scale up with the sheer size of currently available datasets. Thus it can potentially increase the attainable accuracy of our approach, as it has been shown on other related tasks, especially on natural language processing and text modeling [Luong et al., 2015, Michael et al., 2019, Sutskever et al., 2014]. Furthermore, we adopt end-to-end based approaches, with Convolutional Neural Network [LeCun et al., 1989] as main spatial modelling to allow for more representative features extractions to the large size datasets with their variety of image characteristics. Lastly, we further expand our analysis to include several key mechanisms that further enhance the effectiveness of our approaches including internal image degradation modellings and multiple

learning mechanisms. In the following sections, we will give the summary of our main contributions to both tasks, and their relationships to the progress made during the development of this thesis.

1.2.1 Facial Landmark Estimations

In Section 1.1.1 we described the essential functions of facial landmark estimation task for accurate Facial Alignment that enable current advance on higher level facial analysis. In this thesis, we focus on the estimation of facial landmarks using single image input, and in video-based facial tracking setting.

Our first contribution in this task is in our investigation of LSTM to take into account the sequence information from multiple frame inputs using progressive learning. This arrangement allows us to build, in a sense, a true facial landmark tracking. Furthermore, this approach also facilitates the analysis of the impacts of the number of sequence involved in the learning with respect to the final model estimates, that are currently lacking in the field. Here, we successfully build an accurate facial tracking models and reveal the respective optimum length to achieve their optimal accuracy. Next, we address the current challenges of Facial Alignment models when dealing with data taken in the wild. To do this, we propose auto-encoder based network to incorporate several types of image degradation information, and subsequently suppress them to improve our landmark estimates. In our extensive experiments involving number of current state of the arts models, we show the influence of each of the image degradations on their landmark estimations quality, and we demonstrate the importance of our internal image degradation modeling to improve the robustness of our approaches, thus maintaining high level of estimations accuracy. Then motivated by these findings, we develop a combined system from the previous two knowledge in a form of an unified model that can perform complete Facial Alignment for both still-image and video-based input. We do this by proposing several specialized sub-networks to focus on their tasks that are further chained together with our robust tracking algorithm. In addition to this, we also include a real-noise samples based on their known impacts

from previous experiments in their training. The comparisons involving more than 20 alternatives show our models superior accuracy, and our ablation study also confirms the positive impacts of each of our approach.

1.2.2 Facial-Based Emotion Recognition

Our findings regarding the significance of both sequential modelling and internal denoising strategy in Facial Alignment task motivated us to expand this approach into the Facial-Based Emotion Recognition task. This is mainly based on the fact that both tasks use similar input of facial appearances, and furthermore, this also allows us to investigate the robustness and relevance of our central idea of spatio-temporal modelling to both tasks. In addition to this, we also propose to explore the use of visual representations of audio modality in conjunction with facial appearance to benefit from their characteristics, that has been shown to be beneficial to improve the model estimates to each of valence and arousal dimension [Kossaifi et al., 2019, Kossaifi et al., 2017].

We first present our contribution in this task by adapting our generative noise suppression techniques and further improve it by the use of adversarial learning for more effective noise reduction, that currently is still largely unexplored. We also propose a structural arrangement that permits the formation of latent visual features, thus dramatically decreases the models training time. In addition to facial features, we also include the use of projected feature representations of audio modality to benefit from their characteristic to produce a more uniform affect estimations accuracy. In this initial investigation, we show that our models accuracy are able to outperform the baseline of respective emotional datasets, including the one used in the current biggest affect challenge [Kollias et al., 2020]. Furthermore, utilizing these latent based approaches, we further introduce sequential modelling by the use of internal LSTMs network. On top of this, we also propose the inclusion of adaptive weighting by means of sequential attention that allows us to inspect and evaluate the importance of certain sequence points. In this investigation, we found a correlative indication between the perceived changes in facial appearances with the

activation of the internal weights in our sequential attention module. This finding explains the ability of our approach to produce the state of the accuracy as shown in our comparisons. Lastly, we equip our sequence based modelling with latent features by introducing the gating mechanism for more effective multi-modal fusion, as opposed to standard concatenation approach that is heavily used in literature. We demonstrate the effectiveness of both our gating mechanisms and sequential modelling through our systematic ablation studies, that target both internal attention weight and gating coefficients. Lastly, we show our models state of the accuracy in the comparison experiments, and simultaneously highlight the benefit our combined approaches by comparing them with our previous results.

1.3 Thesis Outline

The thesis is organized into 8 chapters. Chapters 2 - 7 are self-contained that each of them contains a published or under review paper, while Chapter 8 provides the summary of this thesis and potential future works.

- In **Chapter 2**, we investigate the use of fully end to end temporal modelling to address current limitations on facial tracking task, that is central to video-based Facial Alignment. We build our model with the use of LSTM layers to capture both short and temporal modelling. We train our model using progressive learning to enable investigations of the impact of involved sequence length. We present the comparisons of our models results on rigid facial tracking and 3D/2D facial tracking setting using the most challenging 300 Videos in the Wild dataset. We first compare the performance of our tracking-by-detection variants, and found their comparable results with the state of the art trackers. However, upon activation of our tracking mechanisms, we observe improvement in our models estimates, reaching state of the art accuracy and far lower failure rates than competing approaches. Thus confirming the advantage of the introduction of temporal modelling in this task.

- In **Chapter 3**, we address the important challenges on current Facial Alignment tasks in improving their robustness when dealing with images taken under extreme conditions, such as severe occlusions or large variations in pose and illumination. As opposed to current approaches that neglect the importance of understanding the noise that characterizes unconstrained images, which has been shown to benefit computer vision models if used appropriately on the learning strategy. We instead incorporate an internal noise modelling which is capable to measure and identify the existing image degradation of the input image, and further reducing it to improve the overall facial landmark estimations quality. We build our model by combining the proposed noise detection and removal sub-network with our state of the art facial alignment modules, and show that such combinations lead to improved robustness both to synthesized noise and images taken under in-the-wild conditions. In our comparisons using both 300-W and Menpo datasets along with 300-VW tracking dataset, we found that our model are robust to both evaluated synthetic noises and to data taken in-the-wild, reaching state of the art accuracy.

  In **Chapter 4**, we make use of our previous findings on both single-based facial alignment and facial tracking to create an unified system that can address these two problems simultaneously. Our system is largely inspired by our previous knowledge of the benefits in incorporating the temporal modellings, and we further equip it with internal noise modeling to improve its general robustness, thus effectively uniting these two approaches. Specifically, we combine four different sub-networks that specialize in each of their associated tasks : facial detection, facial bounding-box tracking, facial region validator and facial alignment with internal denoising. We chained these sub-modules using our novel tracking algorithm resulting on both accurate, robust and resilient against drifting facial tracker. This is demonstrated by our state of the art accuracy from our comparison on each respective tasks, involving 20 other approaches.

- In **Chapter 5**, we expand the applicability of our generative noise

modelling to higher level of facial analysis of facial-based emotion recognition. We do so by adopting a recent advance in machine learning field of adversarial learning to improve model learning through artificially augmented samples. Furthermore, we also propose to include the use of latent feature extractions techniques to improve the efficiency of our training, which is critical considering huge size of dataset involved in this investigation. Specifically, our models operate by the Generator that cleans the noisy input images, while subsequently produces the intermediate latent visual features. These latent features together with concatenated high level sound features and input images are aggregated by the Discriminator to produce both true/fake identity of the generated cleaned images of generator and respective valence and arousal dimension. We show the effectiveness of our approaches in our comparison using the recently published SEWA and also in our participation to the first Affective Behavior Analysis in-the-Wild (ABAW) challenge.

- In **Chapter 6**, we adopt our previous sequential modellings from facial alignment tasks to facial-based affect Recognition to address a lack of incorporation and analysis of this sequence modelling in this task, while also simultaneously demonstrate the cross-task relevance of this approach. We do so by equipping the previous latent feature extractors with LSTM networks that are also enhanced with adaptive sequence weighting through attention modules. This approach involves three major networks of Generator, Discriminator and Combiner which are trained using combined adversarial and progressive learning. We show the effectiveness of our approaches in our comparison study against state of the art results on both AFEW-VA and SEWA dataset. In addition to this, we also notice that there is a degree of correlation between the existing facial movements with the temporal localisations, that explains the improvements made using our attention-enhanced sequence modeling. Finally, we also found that the medium length of sequences to be as optimum one, consistent with our previous findings from facial alignment task.

- In **Chapter 7**, we extend our previous sequential modeling for affect recognition to include gating mechanisms for more effective fusion of both facial feature and visual representation of audio modality. Specifically, we build three major networks that specialize in each of their own respective task: first is the latent-feature generators that extract efficient representations of both modalities while simultaneously clean the distorted version of each modality input; second is a multi-task discriminator that are trained in adversarial way with generator to predict both the real/fake status of input modality and first step quadrant emotion dimension; third is an attention-enhanced sequence modelling that uses extracted latent features as input, and fuses them together with gating mechanisms to estimate both valence/arousal intensity of current input. We show the significance of each of our approach through our comparisons using both SEMAINE and SEWA datasets with our models state of the art accuracy. Lastly, we also quantify and analyse the impacts of both attention and gating coefficients to our models accuracy in our ablation analysis.
- Finally in **Chapter 8**, we summarize the important findings and contributions that we made in the context of spatio-temporal modeling to both facial alignment and facial-based emotion recognition. We then conclude the thesis with some final remarks and potential future research lines.

1.4 Publications

The research developed during this thesis has resulted in the following list of publications:

Journals

1. **D. Aspandi**, F. Sukno, B. Schuller, and X. Binefa, "Audio-Visual Gated-Sequenced Neural Network for Affect Recognition" in *IEEE Transaction on Affective Computing, (Under Review)*.

2. **D. Aspandi**, O. Pujol, F. Sukno, and X. Binefa, "Composite recurrent network with internal denoising for Facial Alignment in still and video images in the wild" in *Image and Vision Computing, (Under Review)*.

3. R. Joshi, M. Rigau, M. Castro, D. Pineyro, S. Moran, V. Davalos, C. Fernandez-Tena, **D. Aspandi**, F. Sukno, X. Binefa, A. Valencia, and M. Esteller, "Look-Alike Humans Exhibit Closely Related Genomes, but Divergent Epigenomes and Microbiomes", in *Science Advances, (Under Review)*

4. N. Rodriguez-Diaz, **D. Aspandi**, F. Sukno, and X. Binefa, "Lie Detection based on Deep Learning applied to a collected and annotated Dataset", in *IEEE Transactions on Cognitive and Developmental System, (to be submitted)*

International Conferences

1. **D. Aspandi**, F. Sukno, B. Schuller, and X. Binefa, "An Enhanced Adversarial Network with Combined Latent Features for Spatio-Temporal Facial Affect Estimation in the Wild", in *16th International Conference on Computer Vision Theory and Applications (VISAPP), (To appear)*.

2. J. Comas, **D. Aspandi**, M. Ballester, L. Ballester, F. Carreras and X. Binefa, "Short-term Impact of Polarity Therapy on Physiological Signals in Chronic Anxiety Patients" in *8th International Conference on Bioinformatics and Computational Biology (ICBCB 2020)*. 16 May - 18 May 2020, Taiyuan, China.

3. *J. Comas, ***D. Aspandi**, and X. Binefa, "End-to-end facial and physiological model for Affective Computing and applications" in *15th IEEE International Conference on Facial and Gesture (FG) 2020*, pp. 1-8, 16-20 November 2020, Buenos Aries, Argentina.

4. **D. Aspandi**, O. Pujol, F. Sukno, and X. Binefa, ”Robust Facial Alignment with Internal Denoising Auto-Encoder” in *16th Conference on Computer Robot and Vision (CRV)*, pp. 143-150, 28 May - 31 May 2019, Kingston, Canada.

5. **D. Aspandi**, O. Pujol, F. Sukno, and X. Binefa, ”Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild,” in *14th IEEE International Conference on Facial and Gesture (FG) 2019*, 14 May - 18 May 2019, Lille, France.

International Workshops

1. **D. Aspandi**, A. Mullol-Ragorta, B. Schuller, and X. Binefa, ”Latent-Based Adversarial Neural Networks for Facial Affect Estimations” in *In Affective Behavior Analysis in-the-wild (ABAW) workshop in conjunction with International Conference on Facial and Gesture (FG) 2020*, pp. 348-352, 16-20 November 2020, Buenos Aries, Argentina.

2. **D. Aspandi**, O. Pujol and X. Binefa, ”Heatmap-Guided Balanced Deep Convolution Networks for Family Classification in the Wild”, in *Recognizing Families In the Wild (RFIW) workshop in conjunction with IEEE FG 2019*, 14 May - 18 May 2019, Lille, France.

* Both authors contributed equally to this publication

Chapter 2

FULLY END-TO-END COMPOSITE RECURRENT CONVOLUTION NETWORK FOR DEFORMABLE FACIAL TRACKING IN THE WILD

Adapted from: D. Aspandi, O. Pujol, F. Sukno, and X. Binefa, ”Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild,” in *14th IEEE International Conference on Facial and Gesture (FG) 2019*, 14 May - 18 May 2019, Lille, France.

Abstract

Human facial tracking is an important task in computer vision, which has recently lost pace compared to other facial analysis tasks. The majority of current available tracker possess two major limitations: their little use of temporal information and the widespread use of handcrafted features, without taking full advantage of the large annotated datasets that have recently become available. In this work we present a fully end-to-end facial tracking model based on current state of the art deep model architectures that can be effectively trained from the available annotated facial landmark datasets. We build our model from the recently introduced general object tracker Re^3, which allows modeling the short and long temporal dependency between frames by means of its internal Long Short Term Memory (LSTM) layers. Facial tracking experiments on the challenging 300-VW dataset show that our model can produce state of the art accuracy and far lower failure rates than competing approaches. We specifically compare the performance of our approach modified to work in tracking-by-detection mode and showed that, as such, it can produce results that are comparable to state of the art trackers. However, upon activation of our tracking mechanism, the results improve significantly, confirming the advantage of taking into account temporal dependencies.

2.1 Introduction

The human face is arguably one of the most important deformable objects for analysis, especially for tracking, with numerous real world applications, such as facial animation, human activity recognition and human - computer interaction [Wu and Ji, 2018]. The recent growth of facial datasets *in the wild* with annotated landmarks such as 300W [Sagonas et al., 2013] and 300 Videos in the Wild (300-VW) [Shen et al., 2015] has led to rapid development of facial analysis tools by introducing powerful deep learning models that are able to automatically extract more representative features from larger scale datasets. These new models have pushed forward the state of the art, outperforming the accuracy reported by earlier methods based on handcrafted features. We can find examples of such models targeting face detection [Triantafyllidou and Tefas, 2016, Zhang et al., 2016a], facial classification and verification [Parkhi et al., 2015, Taigman et al., 2014], and facial expression analysis [Li and Deng, 2018].

However, current progress in deformable facial tracking has been relatively slower when compared to other tasks and it has been less influenced by deep learning models [Chrysos et al., 2017]. Furthermore, currently available trackers make little use of temporal information. Indeed, most of them do not really take into account temporal information but process each frame independently and rely on doing so with sufficient precision to achieve *tracking-like* performance. In contrast, other trackers do some temporal modelings, but they are mostly limited to the adjacent frames [Xiao et al., 2015],[Rajamanoharan and Cootes, 2015]. This inhibits current facial trackers to take full advantage of the temporal information contained in video sequences [Xie et al., 2016].

In this chapter we present a fully end-to-end facial tracking model based on current state of the art deep model architectures that can be effectively trained from the available annotated facial landmark datasets. We build our model from the recently introduced general object tracker Re^3 [Gordon et al., 2018], which allows modeling the short and long temporal dependency between frames by means of its internal Long Short Term

Memory (LSTM) layers. While Re^3 is too generic to be directly used as facial tracker (its performance would be suboptimal), we introduce architectural modifications that lead to a robust facial tracker achieving state of the art performance. More specifically, the contributions in this work are:

1. We replaced the original Skip Convolution Networks from Re^3 by the more robust Inception Residual Networks [Szegedy et al., 2016] through transfer learning.

2. We embed our main tracker together with additional layers that validate the tracking results at every frame and trigger a re-initialization strategy if drifting is detected.

3. To the best of our knowledge, we are the first to successfully train an end-to-end network that can achieve state of the art face tracking on the 300-VW benchmark.

4. We investigate the impact of different temporal windows in the performance of face tracking.

2.2 Related Work

Currently, the most popular facial tracking technique is Tracking by Detection, which consists of performing facial detection and landmark localization at each frame. One example of this strategy is the work from Uricar et al. [Uricár et al., 2015] which uses tree-based Deformable Part Models (DPM) for facial landmark detection and localisation with Kalman Filter smoothing.

Other tracking methods perform face detection only in the first frame and then apply facial landmark localization using the fitting result from the previous frame as initialization. One such example is the work from Xiao et al. [Xiao et al., 2015] which adopts a multi-stage regression-based approach to initialize the shape of landmarks with high semantic meaning. Other examples include the work from Raja et al.

[Rajamanoharan and Cootes, 2015] which combines a global shape model with sets of response maps for different head angles indexed on the shape model parameters and the works from Wu et al [Wu and Ji, 2015] who apply shape augmented regression. There are also hybrid approaches which combine tracking by detection and initialization based on the latest fitting result. Among these, combinations of Coarse-To-Fine Shape Search (CFSS) [Zhu et al., 2015a] landmark localiser with multiple general-object trackers have shown to perform particularly well [Chrysos et al., 2017].

However, all methods derived from tracking by detection share the limitation of not considering the temporal information contained in video sequences. Furthermore, it is difficult to obtain consistent initializations from most face detectors, which tends to reduce the final landmark localisation accuracy [Lv et al., 2017a]. Some approaches try to mitigate this problem by including the information from the adjacent frames to capture short temporal dependencies. For example, Yang et al. [Yang et al., 2015] used time series regression on adjacent two frames, which led them to achieve the best result reported so far on the biggest deformable facial tracking dataset: 300 Videos in the Wild (300-VW) [Shen et al., 2015].

With the recent growth of facial landmark datasets, such as 300W [Sagonas et al., 2013], Menpo [Zafeiriou et al., 2018], 300-VW and LS3D-W [Bulat and Tzimiropoulos, 2017], current methodologies on facial analysis started to shift from systems based on handcrafted features towards incorporating deep learning architectures [Zafeiriou et al., 2017b, Greenspan et al., 2016]. Rapid progress can be seen on the development of various convolutional architectures as the main spatial feature extractor used on both facial detection [Zhang et al., 2016a] and landmark localisation models [Bulat and Tzimiropoulos, 2017, Zhu et al., 2017] and achieving state of the art accuracy. In spite of this, localization is still mainly performed on every single frame, without taking into the temporal information.

On the other hand, introduction of recurrent neural networks (RNN), especially Long Short Term Memory (LSTM) Network [Hochreiter and Schmidhuber, 1997], has allowed incorporating temporal information with great success in several applications [Greff et al., 2017].

This is the case of the recently introduced general object tracker Re^3 [Gordon et al., 2018], which is robust against image occlusions and can be trained on long sequences thanks to its internal LSTM networks. Nonetheless, RNN have received little attention in the context of facial tracking. The only exception so far has been the work by Jiang et al. [Gu et al., 2017], who proved that an end-to-end RNN is capable to work on multiple domains including facial landmark tracking. However, even though they obtained very low failure rates, their accuracy was still inferior to other state of the art facial trackers.

2.3 Fully-end-to end Recurrent Facial Tracker

Our tracking model is a composite network that receives raw frames as input and returns the localization of facial landmarks as the final output. It is composed by four sub-networks, arranged in a way that permits the end-to-end training of the whole network, without involving any hand-crafted features. Specifically, if $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ denote the current and previous frame, respectively, our Composite Recurrent Convolution Tracker (*CRCT*) will estimate the position of n facial landmarks in the current frame $\mathbf{l}_t$:

$$\mathbf{l}_t = \{(\hat{x}_1, \hat{y}_1)...(\hat{x}_n, \hat{y}_n)\} = CRCT_{\Phi}(X_t, X_{t-1}, \mathbf{b}_{t-1}) \tag{2.1}$$

where Φ are the parameters $\{\Phi^1, \Phi^2, \Phi^3\}$ of our composite networks *CRCT* and $\{\hat{x}_1...\hat{x}_n, \hat{y}_1...\hat{y}_n\} \in \mathbb{R}_{>0}$.

Our *CRCT* consists of four individual sub-networks: Multi-Task Cascaded Neural Network faces detector (*MTCNN*), facial bounding Box Tracker (*BT*), Facial Validator (*FV*) and Facial Landmark Localizer (*FLL*). Note that for face detection we relied on the state of the art *MTCNN* [Zhang et al., 2016a].

A schematic diagram of our tracker can be seen in Figure 2.1. We start by assuming a tracking scenario, where we have an existing estimate for the bounding box of the preceding frame.[2] This bounding box, together

[2]For initialization, this estimate can be obtained from the *MTCNN* detector or from an external input.

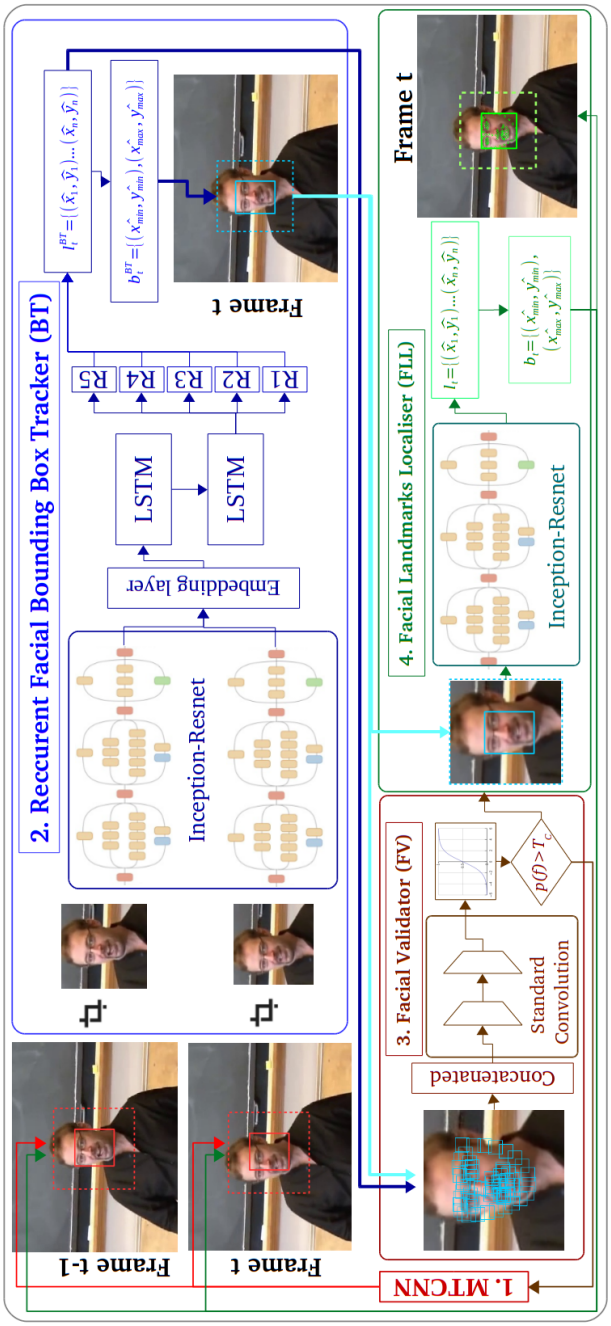


Figure 2.1: General overview of our tracker.

with the current and previous frames $\{\mathbf{X}_t, \mathbf{X}_{t-1}\}$ are fed to our BT network to produce a first estimate of the targeted landmarks ($\mathbf{l}_t^{BT}$) and bounding box ($\mathbf{b}_t^{BT}$), while at the same time updates its internal state.

Once we have our first landmarks estimate $\mathbf{l}_t^{BT}$, we use the FV network to validate the result obtained by the tracker. To do so, we train the FV network to estimate the probability that the objects tracked within $\mathbf{l}_t^{BT}$ is a face ($p(f)$). In case of obtaining a low probability, which would suggest that the BT network has lost track, we use the $MTCNN$ to perform face detection on the current frame and re-initialize the whole network for the next time step.

In contrast, if $\mathbf{l}_t^{BT}$ is successfully validated by the FV network, the current frame and its bounding box $\mathbf{b}_t^{BT}$ are fed to the FLL network, which produces the final estimates for the target landmarks, $\mathbf{l}_t^F$ and the corresponding bounding box, $\mathbf{b}_t^F$. Note that, while FLL and BT have similar convolutional layers, FLL works from an already detected and validated bounding box, which allows it to achieve a more accurate result.

2.3.1 The Recurrent Facial Bounding Box Tracker

We base our BT network on the structure of the Re^3 tracker [Gordon et al., 2018], which is a full end-to-end object tracker with LSTM networks to capture the temporal dependencies from video. Given input frames $\{\mathbf{X}_t, \mathbf{X}_{t-1}\}$ cropped as $\{\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}\}$ with the previous Bounding Box ($P_b = b_{t-1}$), the BT network estimates the landmark positions for the current frame $\mathbf{l}_t^{BT}$ and updates the internal state of the LSTM $\mathbf{h}_t$ as follows:

$$\begin{aligned} \mathbf{h}_t, \mathbf{l}_t^{BT} &= BT_{\Phi^1}(\mathbf{X}_t, \mathbf{X}_{t-1}, P_b, h_{t-1}) \\ &= BT_{\Phi^1}(\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}, h_{t-1}) \\ &= LSTM_{\Phi^1}(EL_{\Phi^1}(\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}), h_{t-1}) \odot W_{\Phi^1}^{BT} \end{aligned} \tag{2.2}$$

where $LSTM$ refers to the set of internal LSTM [Hochreiter and Schmidhuber, 1997] networks, EL stands for the Embedding Layer, W^{BT}, W^{EL} is the set of weight of each fully connected layers of BT and EL respectively and *res* is the Inception-Residual Network

[Szegedy et al., 2016] (Inception-Resnet). The Embedding Layer is a weighted concatenation of the residual network coefficients:

$$EL = [res_{\Phi^1}(\mathbf{X}_t^{Pb}); res_{\Phi^1}(\mathbf{X}_{t-1}^{Pb})] \odot W_{\Phi^1}^{EL} \tag{2.3}$$

We use Φ^1 to denote the parameters of all sub-networks contained in BT. Finally, we also generate an estimate of the bounding box for the current frame $\mathbf{b}_t^{BT}$ directly from the estimated landmarks:

$$\mathbf{b}_t^{BT} = \{(\hat{x}_{min}, \hat{y}_{min}), (\hat{x}_{max}, \hat{y}_{max}) | \hat{x}, \hat{y} \in \mathbf{l}_t^{BT}\} \tag{2.4}$$

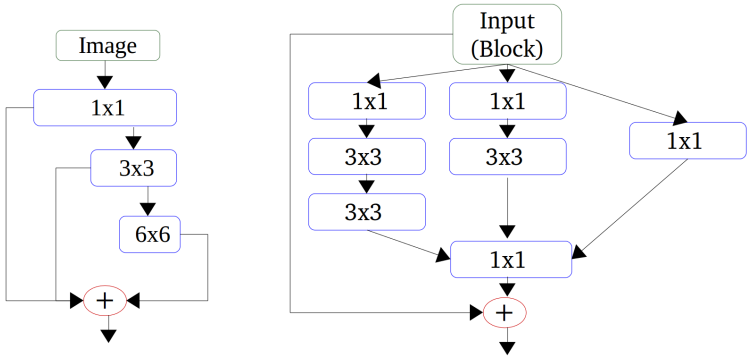
Note that, even though the architecture of BT is based on Re^3, we introduce several key modifications to adapt this recurrent tracker model into this new problem domain:

1. First we preconditioned the convolutional network of our BT to contain common facial features by replacing the internal Skip Convolution Networks (SkipNet) with the more sophisticated Inception-Resnet that has been pre-trained on the MS-Celeb [Guo et al., 2016] and CasiaWebFace [Yi et al., 2014] datasets[3] with triplet loss [Parkhi et al., 2015]. Figure 2.2 visualizes the differences between the original SkipNet on Re^3 versus the more complex structure of BT, which is inherited from the Inception-Resnet (Version 1). Each block of Inception-Resnet architecture can be expressed as below:

   $$\mathbf{r}_{i+1} = H(r_i) + F(r_i, W_i) \tag{2.5}$$

   Where r_i and r_{i+1} are the input and output of the i-th block, $H(b_i)$ is the identity matrix and F represents the combined effect of the various convolutional and ReLU layers. Notice that SkipNet does not have the advantage of residual connection as in the Inception-Resnet which eases the gradient flows in optimization [Szegedy et al., 2016].

[3]The trained inception resnet is publicly available on: https://github.com/davidsandberg/facenet



(a) Skip Connections (b) Inception-Resnet Block

Figure 2.2: Convolution architectures of Skip Network vs Inception-Residual Network block.

2. Second we use the BT network to produce a first estimate of landmark locations (L_t^b) following the work of [Gu et al., 2017], but we split the fully-connected layer that receives the output from the LSTMs into five independent fully-connected networks so that each of them is focused on a specifically facial region. Specifically, we divide the facial landmarks in the following regions: facial silhouette (our outer contour), eyebrows, eyes, nose and lips. Thus $W^{BT} = \{W^{R1}, W^{R2}, W^{R3}, W^{R4}, W^{R5}\}$.

3. Finally, we reduce by half the number of neurons from the original Re^3, which implies an input image size to 128x128. This enables us to train the network faster while still achieving state of the art accuracy.

2.3.2 The Facial Validator

After the initial estimates produced by the BT network we use the FV network to validate the results before further processing. The main reason for doing so is to avoid the drift problem, well known in the tracking literature [Wang et al., 2012a]. Specifically, the FV network can be understood as a

conditional function that determines whether to continue the processing pipeline based on the current estimates from BT or to reset the tracker and attempt to re-detect the facial region because the current estimates are not reliable enough.

We follow the methodology in [Chrysos et al., 2017] to build a strong classifier to estimate the probability $p(f)$ that the object currently being tracked by BT is a face. To this end, we use concatenated small patch regions from the estimated landmarks (l_t^{BT}) as follows:

$$\begin{aligned} \mathbf{p}(f|\mathbf{X}_t, \mathbf{l}_t^{BT}) &= FV_{\Phi^2}(\mathbf{X}_t, \mathbf{l}_t^{BT}) \\ &= \frac{1}{1+e^{-(W_{\Phi^2}^{FV} \odot cnn_{\Phi^2}(\mathbf{X}_t, \mathbf{l}_t^{BT}))}} \end{aligned} \tag{2.6}$$

Where cnn is the composite function of standard stacked convolution layers followed by a bottleneck layer with W^{FV} parameterized by Φ^2 and $0 < p(f) < 1$. We use the value of T_c as the threshold level.

2.3.3 The Facial Landmark Localiser

The FLL is built by reusing the same pretrained Inception-Resnet as in BT, with the assumption that the internally extracted facial feature should also be useful to estimate the locations of the facial landmarks. This landmark localization procedure can be expressed mathematically as below:

$$\mathbf{l}_t = FLL_{\Phi^3}(\mathbf{X}_t^{Pb}) = W_{\Phi^3}^{FLL} \odot res_{\Phi^3}(\mathbf{X}_t^{Pb}) \tag{2.7}$$

With FLL consisting of Inception-Resnet (res) and a regression layer of weight matrix W^{FLL} parameterized by Φ^3.

2.3.4 Recurrent Facial Tracking Algorithm

The operation of our Composite Recurrent Convolution Tracker, $CRCT$, is shown in Algorithm 1. When a suitable detection of the facial region is available, e.g. from initialization or the previous frame (lines 8 and 10), the BT network produces a first estimate of facial landmarks (line 13) and

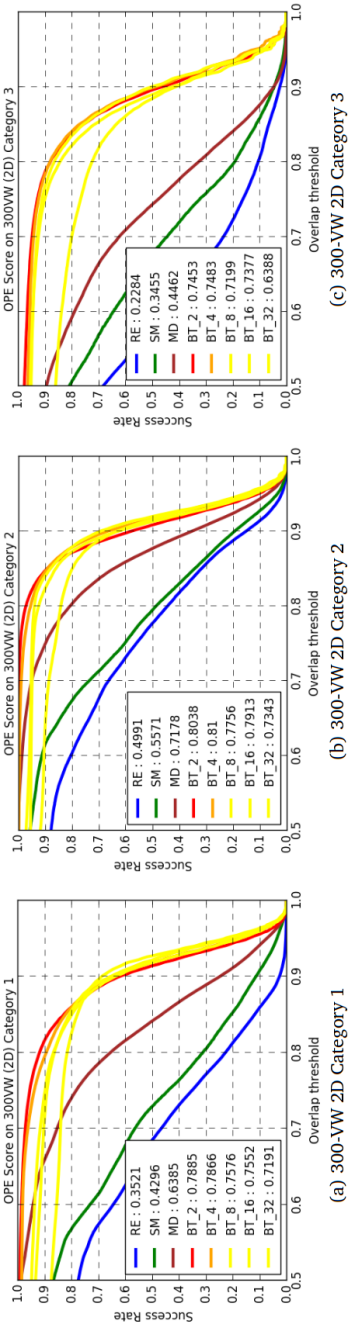
bounding box (line 14). Then, the FV network is used to estimate the probability $p(f)$ that the output from BT corresponds to a face. If $p(f)$ is sufficiently high (above threshold T_c), the initial estimate is refined by the FLL network to produce the final tracker estimate (lines 18 and 19). Otherwise, it is assumed that the BT has lost track and there is a need to re-initialize the tracker (line 16).

We perform reinitialization between lines 3 and 6. We start by detecting the face in the current frame by means of the $MTCNN$ network. This detector is likely to produce multiple detections, hence its outputs are validated with respect to the bounding box of the previous frame b_{t-1}. Specifically, we compare the Euclidean distance between each new detection and the center of the previous bounding vox d(b_{t-1}, b^{MT}) with respect to the magnitude of the previous bounding box, and keep the one that produces the minimum ratio:

$$P_b = \begin{cases} b, & \min\limits_{\forall b \in b^{MT}} \frac{d(b_{t-1},b)}{\|b_{t-1}\|} < T_B \\ b_0, & dim(b_{t-1}) < 0 \\ b_{t-1}, & otherwise \end{cases} \tag{2.8}$$

as long as there is at least one detection whose ratio is below threshold T_B. Otherwise, all new detections are too far from the previous tracking result and no re-initialization is performed. The latter is necessary to tackle the cases in which the face being tracked moves out of the visual field. In such cases, without threshold T_B the system might be incorrectly re-initialized to track another face. In contrast, by using T_B the tracker remains in its latest valid coordinates awaiting for the tracked object to *come back* to the field of view.

Finally, ($SeqBT$) controls the length of the temporal window that is considered by the tracker (in frame units), which is fixed at training time (see next section). If the tracker is re-initialized or if the sequence length ($SeqT$) exceeds the temporal window ($SeqBT$), then the internal state of the the BT network is reset (line 8).



(a) 300-VW 2D Category 1 (b) 300-VW 2D Category 2 (c) 300-VW 2D Category 3

Figure 2.3: OPE scores on the rigid facial bounding box tracking experiment.

Algorithm 1 Recurrent Facial Tracking Algorithms

Input : Frame of $\mathbf{X}_{0..N}$
Initial value of b_0 and h_0
Threshold value of T_B, T_C, and SeqBT
Network parameters of Φ^1, Φ^2 and Φ^3

Output : Facial Landmark of $\mathbf{l}_{1..N}$

redetect$\leftarrow$ FALSE, SeqT $\leftarrow 0$, $\mathbf{b}_t \leftarrow b_0$
for $t \leftarrow 1$ to N **do**
  if redetect **then**
    $b^{MT} \leftarrow MTCNN(\mathbf{X}_t)$
    if length(d$(b_{t-1}, b^{MT}) > T_B) > 0$ **then**
      $\mathrm{P}_b \leftarrow b^{MT}[\min(\mathrm{d}(b_{t-1}, b^{MT}))]$
  else
    $\mathrm{P}_b \leftarrow b_t$
  if dim$(\mathrm{P}_b) < 0$ **then**
    $\mathrm{P}_b \leftarrow b_0$
  if redetect OR SeqT $>$ SeqBT) **then**
    $h_t \leftarrow h_0$ and SeqT $\leftarrow 0$
  $\mathbf{h}_t, \mathbf{l}_t^{BT} \leftarrow BT_{\Phi^1}(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathrm{P}_{BB})$
  $\mathbf{b_t}^{BT} \leftarrow [max(\mathbf{l}_t^{BT}), min(\mathbf{l}_t^{BT})]$
  if $FV_{\Phi^2}(\mathbf{X}_t, \mathrm{l}_t^{BT}) < T_C$ **then**
    redetect $\leftarrow$ TRUE
  else
    $\mathbf{l}_t \leftarrow FLL_{\Phi^3}(\mathbf{X}_t, \mathbf{b}_t^{BT})$
    $\mathbf{b}_t \leftarrow [max(\mathbf{l}_t), min(\mathbf{l}_t)]$
    SeqT $\leftarrow$ SeqT + 1
    redetect $\leftarrow$ FALSE

2.3.5 Training procedure

We train BT,FLL and FV with ℓ_1,ℓ_2 and Cross Entropy Loss respectively. Specifically for BT, we follow the same curriculum learning as in Re^3 [Gordon et al., 2018] using sequence lengths between $SeqBT = 2$ to

$SeqBT = 32$ frames. We used multiple transfer learning to condition the pre-trained Inception-Resnet. To do so, we fine-tuned this network on the FLL network before its integration into BT.

We trained our BT network using 300-VW training dataset for 2D Landmark tracking and LS3D-W annotation [Bulat and Tzimiropoulos, 2017] for 3D-2DA landmark tracking. The FV and FLL networks were trained with the 300W [Sagonas et al., 2013] and Menpo datasets [Zafeiriou et al., 2018] for both 2D and 3DA-2D landmark localization. We performed data augmentations by means of horizontal flipping, -45° to 45° degree rotations and artificial strip boxes across the frames to simulate occlusions.

We trained our model using ADAM optimizer [Kingma and Ba, 2014] with scheduled weight learning decay every 10.000 iterations. Two NVIDIA tesla GPUs were used for training which took approximately two to three days to train a single BT for a defined sequence length, and around two days for both FLL and FV. Our pre-trained models and results are publicly available for additional reference [4].

2.4 Experiments

2.4.1 Experiment Settings

We conducted two main facial tracking experiments: rigid facial bounding box tracking and deformable 2D and 3DA-2D facial landmark tracking. We performed the experiments on the 300-VW dataset [Shen et al., 2015] comprising 55 videos divided into three categories according to the difficulty level. We used the original 2D facial landmarks directly as ground-truth for deformable 2D facial landmark tracking and their corresponding bounding box for rigid facial bounding box tracking.

We used the projected 3DA-2D dataset video dataset [Zafeiriou et al., 2018] for deformable 3DA-2D facial landmark tracking, which consists of a subset of the videos from 300-VW dataset. To facilitate

[4] https://github.com/deckyal/RT/tree/master

comparison to other works in all cases we report the projected result and follow the conventional 68 facial landmark locations. We set the thresholds $T_B = 1.0$ and $T_C = 0.5$ for all experiments.

2.4.2 Rigid - Facial bounding boxes tracking

In this experiment, we compare our bounding box tracker (BT) with three state of the art general object trackers:

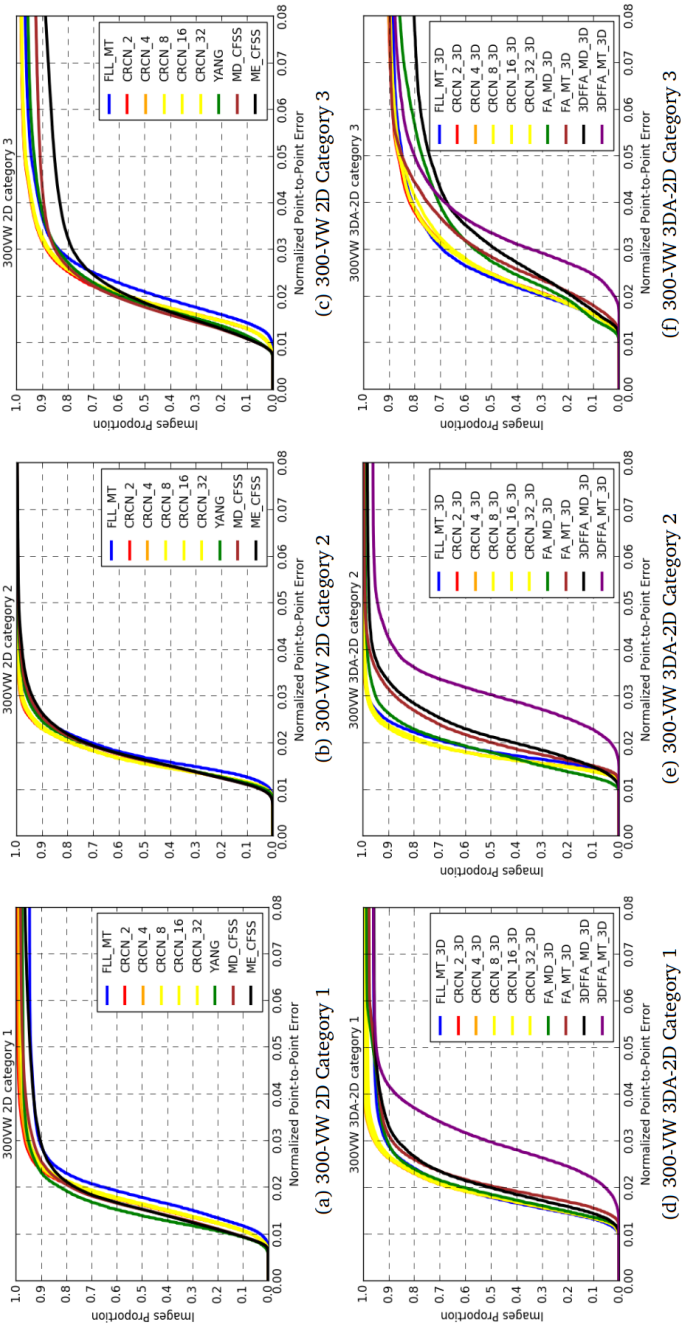
1. MDNET [Nam and Han, 2016] (abbreviated MD) which performs a series of convolutions and has a specialized regression layer on the single individual frames without taking any temporal information between frames

2. Siamese Net (abbreviated SM) [Bertinetto et al., 2016] which uses both the previous and the current frame to be fed to its Siamese Network based tracker. This can be seen as capturing a short temporal context of 2 frames.

3. Recurrent Tracker Re^3 (abbreviated RE), as provided in [Gordon et al., 2018], which is pre-trained on sequences of 32 frames.

In this test, all trackers are initialized with the same initial bounding box (the ground truth). Our system is tested without the FLL block, which means that $b_t = b_t^{BT}$, and we report results for different sequence lengths between $SeqBT = 2$ and $SeqBT = 32$ frames (BT_2, BT_4, BT_8, BT_16 and BT_32), to see the impact of longer temporal context in our BT.

Figure 2.3 shows the performance of each model, computed with One Pass Evaluation (OPE) [Wu et al., 2013] in terms of the success rate against the bounding box overlap ratios. We observe that our models, BT, achieve the best results in all three categories, outperforming all other trackers including the original Re^3. The main reason for these results is that, as opposed to our model, none of the compared trackers is specifically designed to track faces. Furthermore, with the exception of MDNET, other models lack any drift prevention mechanism, which



(a) 300-VW 2D Category 1 (b) 300-VW 2D Category 2 (c) 300-VW 2D Category 3

(d) 300-VW 3DA-2D Category 1 (e) 300-VW 3DA-2D Category 2 (f) 300-VW 3DA-2D Category 3

Figure 2.4: AUC graphs on 2D and 3DA-2D facial landmark tracking experiments.

Method	Temporal sequences					
	0	2	4	8	16	32
RE [Gordon et al., 2018]	-	-	-	-	-	0.363
SM [Bertinetto et al., 2016]	-	0.445	-	-	-	-
MD [Nam and Han, 2016]	0.616	-	-	-	-	-
BT	-	0.783	**0.784**	0.754	0.761	0.705

Table 2.1: AUC result of all categories of rigid facial tracking on 300-VW dataset.

explains the performance drop on category 3, where the extreme facial poses and illumination changes occur. As illustrated in Figure 2.5, our model demonstrated the ability to consistently track the facial bounding box on extreme pose and illumination conditions.

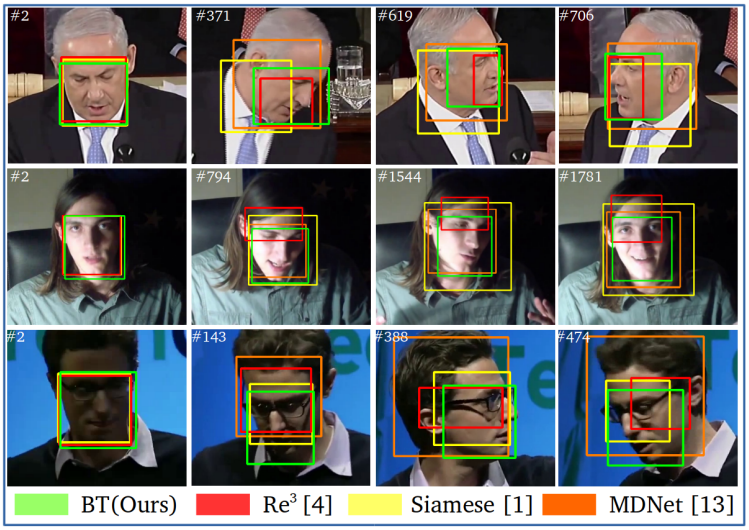


Figure 2.5: Some visual results of rigid facial tracking on 300-VW dataset.

Note that we show the results for our *BT* tracker under different train-

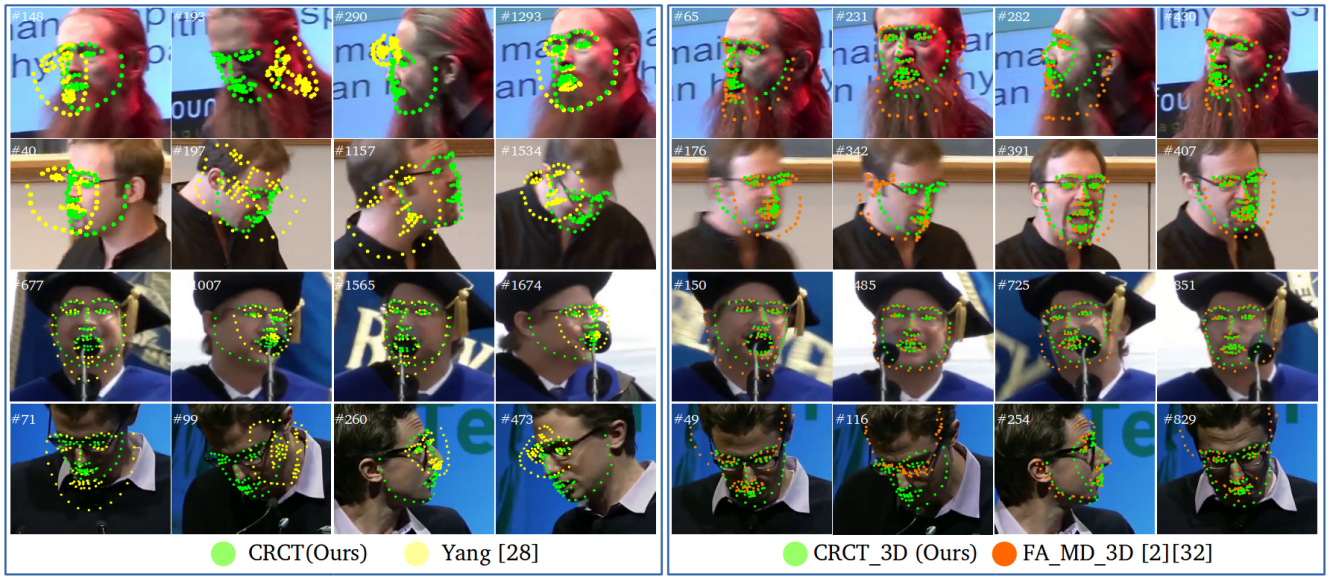
ing sequence lengths (2, 4, 8, 16 and 32). The highest scores were achieved for $SeqBT = 2$ and 4 frames in all categories, with very small differences between these two settings as shown in Table 2.1. Bigger *SeqBT* values generally resulted in lower scores. This suggests that a rather short temporal context is sufficient to optimize facial tracking. Nevertheless, these results must be read in relation to the test sequences, which show quite irregular (not necessarily natural) facial movements. This is especially noticeable in category 3, where rapid face movement with pose changes occur in relatively short sequences. In such cases, BT_2 and BT_4, trained to capture the temporal information from shorter sequences have an advantage since they are restarted more frequently.

2.4.3 2D and 3DA-2D facial landmark tracking

In this section we show the results of 2D and 3DA-2D facial landmark tracking. In the 2D setting, we compared our model with other 8 facial trackers: 1) two hybrid trackers, MEEM_CFSS and MD_CFSS, which showed the best performance in the recent facial tracking review from Chrysos et al. [Chrysos et al., 2017]; 2) the current state of the art tracker, from Yang et al. [Yang et al., 2015]; 3) the recent tracker from Gu et al. [Gu et al., 2017] based on Bayesian RNNs; 4) four other trackers from the original 300VW competition [Shen et al., 2015].

In 3DA-2D facial landmark tracking, we follow a similar procedure to [Chrysos et al., 2017] to build four hybrid trackers combining both MDNET and MTCNN with state of the art 3D facial localizers for comparison: 1) Facial Alignment Network [Bulat and Tzimiropoulos, 2017] resulting in FA_MD_3D and FA_MT_3D; 2) 3DFFA [Zhu et al., 2017] to create other two hybrid trackers: 3DFFA_MD_3D and 3DFFA_MT_3D.

Similarly to the previous section, we evaluate our full tracker, *CRCT*, under different operation conditions. First, analogously to the previous section, we build trackers with different lengths of training sequences, $SeqBT = 2$, 4, 8, 16 and 32. Then, we also report results for our model in tracking-by-detection mode (FLL_MT), where we use *MTCNN* for face detection in each frame and *FLL* for landmark localisation. This



(a) Visual results on 2D Facial Landmark Tracking

(b) Visual results on 3DA-2D Facial Landmark Tracking

Figure 2.6: Some visual results of landmark tracking on challenging case from 300-VW testset.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
CRCT_2	0.784	**0.50**	**0.790**	**0.05**	**0.729**	**1.75**
CRCT_4	0.778	1.01	**0.790**	0.07	0.725	1.80
CRCT_8	0.772	1.88	0.788	0.07	0.725	1.91
CRCT_16	0.773	1.64	0.787	0.07	0.725	2.02
CRCT_32	0.769	1.79	0.778	0.07	0.723	1.86
FLL_MT	0.729	5.38	0.769	2.60	0.691	3.39
MD_CFSS [Chrysos et al., 2017]	0.784	1.80	0.783	0.34	0.713	7.47
ME_CFSS [Chrysos et al., 2017]	0.758	3.56	0.772	0.38	0.659	11.3
Yang [Yang et al., 2015]	**0.791**	2.40	0.788	0.32	0.710	4.46
Jinwei [Gu et al., 2017]	0.718	1.20	0.703	0.20	0.617	4.83
Uricar [Uricár et al., 2015]	0.657	7.62	0.677	4.13	0.574	7.96
Xiao [Xiao et al., 2015]	0.760	5.90	0.782	3.84	0.695	7.38
Raja [Rajamanoharan and Cootes, 2015]	0.735	6.56	0.717	3.91	0.659	8.29
Wu [Wu and Ji, 2015]	0.674	13.9	0.732	5.60	0.602	13.1

Table 2.2: Results on the landmark 2D tracking dataset.

experiment is to assess the impact of the *BT* network on the performance of the full tracker.

Our results are summarized in Tables 2.2 and 2.3, while the curves for some of the trackers are also displayed in Fig. 2.4. In all cases, we use the Normalized Mean Error (NME) by Facial Bounding Box [Zafeiriou et al., 2018], and report the Area Under the Curve (AUC) and Failure Rate (FR) for NME scores up to 0.08 [Chrysos et al., 2017].

In Table 2.2 we see that our *CRCT* trackers trained with $SeqBT = 2$ and 4 frames achieves the highest AUC for Categories 2 and 3, while they rank within the top-3 trackers in the Category 1 dataset, slightly below [Yang et al., 2015]. Additionally, our models have far lower Failure Rates than all other compared trackers, Which for some applications is even more important than having smaller landmark localisarion errors [Gu et al., 2017]. We also see similar results in Table 2.3 for the 3DA-2D scenario, Where in overall terms our model outperforms other trackers across all categories with higher AUC and low Failure Rates.

Method	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
CRCT_2_3D	**0.760**	**0.09**	**0.772**	0.20	**0.605**	**9.97**
CRCT_4_3D	**0.760**	0.14	**0.772**	**0.19**	0.603	10.0
CRCT_8_3D	0.758	0.34	0.773	0.20	0.603	10.2
CRCT_16_3D	0.750	1.36	0.771	0.22	0.604	10.3
CRCT_32_3D	0.747	1.72	0.770	0.21	0.596	10.6
FLL_MT_3D	0.730	4.14	0.757	0.45	0.603	11.5
FA_MD_3D [Bulat and Tzimiropoulos, 2017],[Zhang et al., 2016a]	0.732	1.35	0.757	0.90	0.544	14.2
FA_MT_3D [Bulat and Tzimiropoulos, 2017],[Nam and Han, 2016]	0.706	2.41	0.722	0.57	0.566	10.3
3DFFA_MD_3D [Zhu et al., 2017],[Zhang et al., 2016a]	0.721	4.30	0.702	1.85	0.504	19.8
3DFFA_MT_3D [Zhu et al., 2017],[Nam and Han, 2016]	0.595	4.12	0.590	4.07	0.497	12.4

Table 2.3: Results on the landmark 3DA-2D tracking dataset.

Another observation is that our simpler tracking-by-detection model (FLL_MT) reaches fairly high AUC and low Failure Rates, with a performance comparable to other trackers. This demonstrates the maturity of tracking by detection models, as also reported in [Chrysos et al., 2017]. Nevertheless, these results are still inferior to those from our full *CRCT* models, which incorporates *BT* to benefit from the temporal dependency between frames. This proves that the *BT* network provides a more consistent facial bounding box which impacts the final landmark estimation from FLL. This effect has also been demonstrated in the recent work from Lv et al. [Lv et al., 2017a].

2.4.4 Visual Results Analysis

We provide several examples of 2D and 3DA-2D tracking in Figure 2.6, where we see that our tracker is able to accurately localize the facial landmarks in especially difficult cases. These include extreme head poses up to full profile, blurring (e.g. due to sudden movements of the face or the camera, see 2nd row of examples), partial occlusions (1st and 3rd rows) and strong illumination changes (4th row). Specifically, for 2D landmark tracking, our model performs well in cases in which the state of the art tracker from Yang et al. [Yang et al., 2015] often gives inaccurate landmark positions. Similarly, for 3DA-2D tracking, comparison of our

results to those from *FA_MD_3D*, highlights the robustness of our tracker to handle the difficulties mentioned above from this dataset.

2.5 Conclusions

In this chapter we present the first composite deformable facial tracker that, while being fully end-to-end, is able to achieve state-of-the-art results for *in the wild* benchmarks. Unlike other trackers, our model benefits from the temporal information captured by our internal recurrent tracker. Further, our model can be tuned to consider shorter or longer temporal contexts and analyze their impact on facial tracking performance.

Facial tracking experiments on the challenging 300-VW dataset show that our model can produce state of the art accuracy and far lower failure rates than competing approaches. We specifically compared the performance of our approach modified to work in tracking-by-detection mode and showed that, as such, it can produce results that are comparable to state of the art trackers. However, upon activation of our tracking mechanism, the results improve significantly, confirming the advantage of taking into account temporal dependencies.

Our results suggest that the optimal temporal context to consider for this dataset is between 2 and 4 frames ($\sim$ 70 to 160 ms). Nevertheless, these results must be read in relation to the test sequences, which show quite irregular (not necessarily natural) facial movements.

Chapter 3

ROBUST FACIAL ALIGNMENT WITH INTERNAL DENOISING AUTO-ENCODER

Adapted from: D. Aspandi, O. Pujol, F. Sukno, and X. Binefa, ”Robust Facial Alignment with Internal Denoising Auto-Encoder” in *16th Conference on Computer Robot and Vision (CRV)*, pp. 143-150, 28 May - 31 May 2019, Kingston, Canada.

Abstract

The development of facial alignment models is growing rapidly thanks to the availability of large facial landmarked datasets and powerful deep learning models. However, important challenges still remain for facial alignment models to work on images under extreme conditions, such as severe occlusions or large variations in pose and illumination. Current attempts to overcome this limitation have mainly focused on building robust feature extractors with the assumption that the model will be able to discard the noise and select only the meaningful features. However, such an assumption ignores the importance of understanding the noise that characterizes unconstrained images, which has been shown to benefit computer vision models if used appropriately on the learning strategy. Thus, in this work we investigate the introduction of specialized modules for noise detection and removal, in combination with our state-of-the-art facial alignment module and show that this leads to improved robustness both to synthesized noise and in-the-wild conditions. The proposed model is built by combining two major subnetworks: internal image denoiser (based on the Auto-Encoder architecture) and facial landmark localiser (based on the inception-resnet architecture). Our results on the 300-W and Menpo datasets show that our model can effectively handle different types of synthetic noise, which also leads to enhanced robustness in real-world unconstrained settings, reaching top state-of-the-art accuracy.

3.1 Introduction

Facial alignment aims to detect a set of facial landmark positions which can later be used for several facial analysis applications [Shen et al., 2015]. The development of facial alignment models is growing rapidly with the availability of large facial landmarked dataset such as 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2017b] dataset. This has made it possible the development of powerful deep learning models that have pushed forward the alignment accuracy and are considered the current state of the art [Bulat and Tzimiropoulos, 2017, Gu et al., 2017].

However the performance of current facial alignment models can severely deteriorate when dealing with images in highly unconstrained conditions, e.g. extreme pose or illumination changes, large occlusions [Zhang et al., 2018a] or, in general, whenever the test images can be considered to show less favorable conditions than those available for training. In other words, we may say that under challenging conditions, test images contain some form of distortion or noise that will impair the performance of facial alignment models. This limits the real-life applicability of such models to in-the-wild images which naturally contain such challenging conditions [Nada et al., 2018, Zhou et al., 2018]. While there have been attempts to improve the robustness of facial alignment models to target in-the-wild data [Qu et al., 2015, Zhang et al., 2018a], most of them have done so without modeling the effects of noise in their formulation. Nevertheless, understanding and incorporating such effects within the model training has proven beneficial to improve performance in other facial analysis tasks [Dong et al., 2018, Nada et al., 2018, Goswami et al., 2018].

To address the above shortcomings, in this chapter we investigate the introduction of specialized modules for noise detection and removal, in combination with our state-of-the-art facial alignment module and show that this leads to improved robustness both to synthesized noise and in-the-wild conditions. Our model is built by combining two major subnetworks: internal images denoiser and facial landmark localiser which work in parallel and can be trained end-to-end. We adopt the Auto-Encoder architecture [Chaitanya et al., 2017] with skip connection for the internal

denoising, and the inception-resnet architecture [Szegedy et al., 2016] for the facial landmark. We train our models with four of the most general image noise models: resolution down-sampling, gaussian blur, gaussian noise, and pixel intensities scaling. We show that the proposed architecture can successfully achieve robustness to these types of noise but also, and more importantly, we also find that such noise models are sufficient to train a network that achieves top state-of-the-art performance under real acquisition conditions, by testing on the 300-VW dataset, Specifically, our contributions in this chapter are:

1. To the best of our knowledge, we are the first to investigate the robustness of current state of the art facial alignment methods upon the introduction of synthetic image noise.

2. We propose a novel network for robust facial alignment that is capable to produce accurate facial landmark estimates on unconstrained images by means of its internal denoising strategy.

3. We show that our model reach the state of the art results on the most challenging category of 300-VW videos datasets, both on the single image alignment and facial landmark tracking.

3.2 Related Work

Face alignment has attracted considerable attention due to its importance for several applications, such as face recognition [Zhu et al., 2015b], head pose estimation [Wu et al., 2017], facial reenactment [Thies et al., 2016] and others. This task is generally conducted by estimating a predefined set of landmark positions which provide a structure representation of the facial geometry. Traditionally, facial landmarks have been estimated using either the shape and appearance models [Yuille et al., 1992], [Cootes et al., 1995] or regression based models [Kazemi and Sullivan, 2011], with the latter offering some advantages due to its computational efficiency [Valstar et al., 2010]. Recent examples of regression-based models include the work of Kazemi and Sullivan

[Kazemi and Sullivan, 2014], who use a cascade of regression functions to efficiently regress the landmark locations, and Zhu et al [Zhu et al., 2015a] who refined the regressors cascade with coarse to fine search.

The recent availability of large datasets such as 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2017b] allows for large scale data modeling and development of deep learning-based models that benefit from these load of data. For instance, the work of Bulat and Tzimiropoulos [Bulat and Tzimiropoulos, 2017] used multiple hourglass shaped convolutional networks with heatmap-guided layers to predict the final facial landmarks, while Zhang et al. [Zhang et al., 2016b] enforced the facial attributes as auxiliary features to their cascaded convolutional network to improve landmark estimates. These deep learning models currently hold the state of the art accuracy both on the single image facial alignment [Sagonas et al., 2013] and on the facial landmark tracking tasks [Shen et al., 2015].

Despite the maturity of current facial alignment models, there are still challenges for them to work on images with large appearance variations due to heavy occlusion, severe pose or illuminations conditions, etc [Zhang et al., 2018a]; especially on real world application targeting highly unconstrained settings [Nada et al., 2018]. Thus, there has been growing interest to improve the performance of facial models under such settings, including efforts such as iterative initialization of regression cascades to minimize the impact of outliers [Qu et al., 2015] or the combination of data- and model-driven estimators to enhance the robustness of landmark detection [Zhang et al., 2018a]. However, most efforts have focused on building robust feature extractors with the assumption that the model will be able to discard the noise and select only the meaningful features.

In contrast, we approach the problem differently by focusing on modelling image noise by means of our internal denoiser network and further minimizing it alongside of building an accurate facial landmark feature extractor. This strategy can be justified by recent findings such as those from Dong et al. [Dong et al., 2018], who aggregated different styles of images using Generative Adversarial Networks to improve their landmark estimates, revealing that even the slight color style variations between the same

images can impact the accuracy of landmark estimations. Similar findings have been reported in other facial analysis tasks, such as on the work from Zhou et al. [Zhou et al., 2018], who evaluated several facial detection models with synthetic noise, and Goswami et al. [Goswami et al., 2018] who investigated the robustness of facial classification to synthetically distorted images. The findings in these reports have led us to hypothesize that modeling certain types of noise may help to characterize the behavior of facial analysis systems with unconstrained images and incorporating such noise modeling into the training process could improve the robustness of their results.

3.3 Face Alignment with De-Noiser Network (FADeNN)

We argue that given a noisy input image, the final landmark estimation can be improved by first minimizing the existing noise and only then passing the image to any facial landmark estimator. To achieve this, we adopt a modular approach involving two major subnetworks to build our Face Alignment with De-Noiser Network (FADeNN) network: Internal Image Denoiser (*IID*) and Facial Landmark Localiser (*FLL*) networks. Given the input image I, which can be either a *clean* or contain some unknown amount of noise, the *IID* network will be trained to internally detect and model the existing and, if required, clean the input image to generate a normalized image that will be fed to the *FLL* network to estimate n landmark coordinates $\mathbf{l}_I$ $\{\hat{x}_1...\hat{x}_n, \hat{y}_1...\hat{y}_n\} \in \mathbb{R}_{>0}$.

$$l_I = FADeNN(I) = FLL(IID(I)) \tag{3.1}$$

3.3.1 Internal Image Denoiser (IID)

On this work, we explore two different models of *IID* : Direct Denoising AutoEncoder (Figure 3.1) and Classifier-Specialized Denoiser

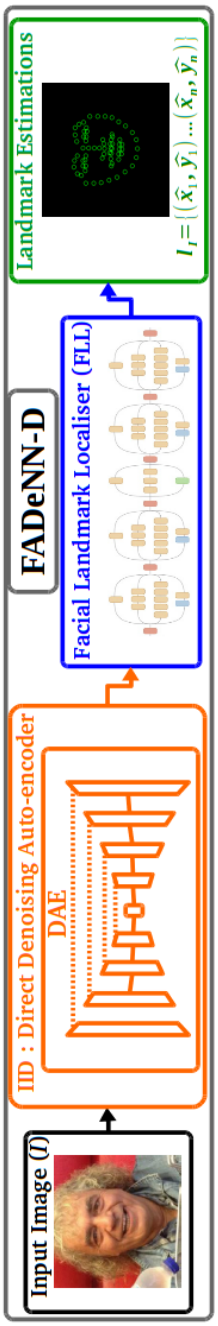


Figure 3.1: General overview of our robust facial alignment consisting two major subnetworks: Internal Image Denoiser and Facial Landmark Localiser.

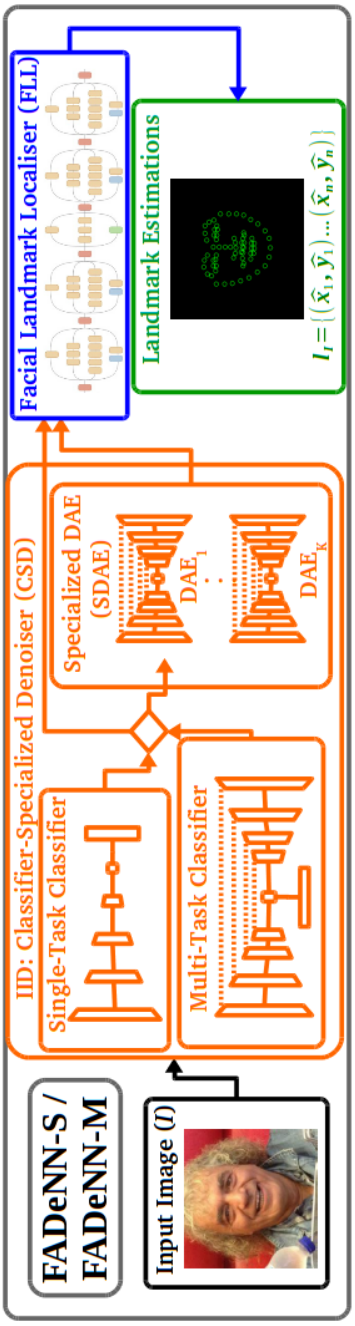


Figure 3.2: The structure of FADeNN-S and FADeNN-M which incorporates more deliberate Classifier-Specialized DAE.

(Figure 3.2). In both cases, the denoiser core is based on the Hourglass shaped Auto-Encoder Architecture (DAE) with skip connection [Mao et al., 2016]. The structure of our *DAE* is similar on the work of [Chaitanya et al., 2017] with a few additional mirror layers on both encoder and decoder parts. The code of the modified *DAE* is available at https://github.com/deckyal/FADeNN.

Direct Denoising AutoEncoder: we first explore the most straightforward approach, which consists of directly attempting to denoise any new input image with the same model. The basic operation of direct denoising can formally be expressed as follows :

$$IID(I) = DAE_{\Phi^1}(I) = enc_{\Phi^1}(dec_{\Phi^1}(I)) \tag{3.2}$$

where ϕ^1 are the parameters learned for subnetwork *DAE* consisting of coupled encoder $enc(x)$ and decoder $dec(x)$ layers. Due to the encoding and decoding operations, the original image will not be perfectly recovered after being processed by this network. In the ideal case, the output will be just a noise-free version of the input. However, in the cases when the input image is clean (i.e. without noise), the model may add itself a small amount of artificial noise due to this imperfect reconstruction. To tackle this issue, below we propose our second denoising model.

Classifier - Specialized Denoiser (CSD) is built with the introduction of a Noise Classifier (NC) network and Multiple Specialized DAE (SDAE) subnetworks which work in tandem to remedy the limitation of direct DAE. The main concept of CSD is to perform the denoising selectively based on the detected condition of the input image to trigger separate specialized denoiser sub-models. The formulation of this IID block is as follows:

$$IID(I) = SDAE_{\phi^3}(NC_{\phi^2}(I), I) \tag{3.3}$$

Given the number of K known noise class, the NC will estimate the probability that noise class c is present in the input image I :

$$NC_{\phi^2}(I) = argmax\{s_1, s_2..s_K\} \tag{3.4}$$

$$s_k = \left(\frac{e^{W^{NC}_{\phi^2_k} \odot conv^{NC}_{\phi^2_k}(I)}}{\sum_{k=1}^{K} e^{W^{NC}_{\phi^2_k} \odot conv^{NC}_{\phi^2_k}(I)}}\right) \tag{3.5}$$

Where W^{NC} are the multinomial bottleneck regression layers parameterized by Φ^2, s_k is the score for specific noise-class k estimated for the current input I, and *conv* is the set of convolutional layers for *NC*. If the classifier detects noise in the input image, then $c > 0$ and the image is denoised by one of the specialized denoisers $\{DAE_1, DAE_2 ... DAE_K\}$; each of these blocks can be trained to capture a different type of noise.

$$SDAE_{\phi^3}(c, I) = \begin{cases} DAE_{\phi^3_c}(I), \, if \, c > 0 \\ I, \, otherwise \end{cases} \tag{3.6}$$

Where Φ^3_c is the parameter learned for the specific DAE of class c. Otherwise, $c = 0$ and the denoising process is skipped to avoid unnecessarily distorting the input image. The idea behind our CSD network follows a similar spirit to the work in [Nam and Han, 2016], however we used distinct specialized subnetworks instead of specialized internal layers.

As observed in Fig. 3.2, we explored two architecture variants for our noise classifier: Single-Task Classifier (*STC*) and Multi-Task Classifier (*MTC*). The differences between *STC* and *MTC* are that *MTC* uses additional reconstruction loss as regularizer [Zhang and Yang, 2017, Bulat and Tzimiropoulos, 2017] alongside the classification loss, and *MTC* uses full encoder-decoder architecture similarly to *DAE* while *STC* uses only the encoder part.

3.3.2 Facial Landmark Localiser (FLL)

We build our *FLL* from our recently published Composite Recurrent Convolution Tracker *CRCT* [Aspandi et al., 2019b]. Specifically, we adopt the landmark localization subnetwork from *CRCT*, based on the Inception-Resnet [Szegedy et al., 2016] architecture, which we modified by replac-

ing the last layer with new bottleneck regression layers.[2]. The landmark localization procedure can be expressed mathematically as below:

$$\mathbf{l}_I = FLL_{\Phi^4}(X) = W^{INC}_{\Phi^4} \odot res_{\Phi^4}(X) \tag{3.7}$$

where FLL consists of the Inception-Resnet (res) and a regression layer of weight matrix W^{INC} parameterized by Φ^4. The input to FLL is the *clean* image X, which would be obtained as $X = IID(I)$ in the general case.

3.3.3 Overall Models and loss functions

To provide a comprehensive analysis of the denoising alternatives explored in this chapter, we provide results for three different models. We refer to these models based on their IDD subnetworks :

1. **FADeNN-D** which uses direct DAE.
2. **FADeNN-S** which uses the combination of STC and $SDAE$.
3. **FADeNN-M** which uses MTC and $SDAE$.

We use standard ℓ^2 loss to train FLL while the loss functions for each of the IID options are as follows :

$$FADeNN-D: \quad \mathbf{L_{IID}} = \min \ell^2(X, DAE(\hat{X})), \tag{3.8}$$

$$FADeNN-S: \quad \mathbf{L_{IID}} = \min(H(c, STC(I))), L_{SDAE}, \tag{3.9}$$

$$FADeNN-M: \quad \mathbf{L_{IID}} = \min \lambda_1(H(c, MTC(I)) + \lambda_2 \ell^2(I, MTC(I))), L_{SDA}. \tag{3.10}$$

[2]Our modified inception resnet network is publicly available on: https://github.com/deckyal/RT

where $\hat{X}$ is a noisy version of X (added synthetically), $H(x)$ is the cross entropy loss and λ is the regularizer parameter for each term. We use I to indicate both clean and noisy inputs (X and $\hat{X}$) and L_{SDA} is the loss of the specialized *SDA* networks:

$$\mathbf{L_{SDA}} = \min_{\forall D \in SDAE} \ell^2(X, D(\hat{X}))) \tag{3.11}$$

3.3.4 Training setup

We train our model progressively by first training each internal *IDD* and FLL component individually, and then jointly training them for further fine-tuning. We utilized the 300-W [Sagonas et al., 2013] training dataset for both *IDD* and FLL. We also introduce data augmentation procedures of -45° to 45° degree rotations and horizontal flipping, with additional artificial stripping to simulate occlusions for training FLL. Finally, we set our λ value to 0.5, and train our model using ADAM optimizer [Kingma and Ba, 2014] with initial learning rate of 0.0001 and progressive decaying every 1000 iterations.

3.4 Experiments

3.4.1 Image Degradation Models

To systematically evaluate the impact of image degradation on the final landmark estimations, we generate synthetically distorted images and evaluate their effect on the final landmark estimation of each model. We generate these distorted images by perturbing the original image with specific types of noises. We consider four types of noise: three of them following the recent work by Zhou et al. [Zhou et al., 2018] plus an additional noise class consisting of downsampling. Specifically, each noise class was obtained as follows:

1. **Down-sampling**. We down-sampled images successively by a factor of 2.0, resulting in images with half, one-quarter and one-eight of the original image resolution.

2. **Gaussian-Blurring**. We added Gaussian Blur by convolving the input image with two dimensional Gaussian filters with $\sigma_{gb} \in \{1, 3, 5\}$.

3. **Gaussian-Noise**. We added random pixel colors distributed according to normal distributions with zero-mean and standard deviation $\sigma^2_{gn} \in \{0.001, 0.005, 0.01\}$.

4. **Color Scaling**. We scaled the original pixel intensities linearly by a factor $s \in \{0.8, 0.5, 0.2\}$ from the original intensities.

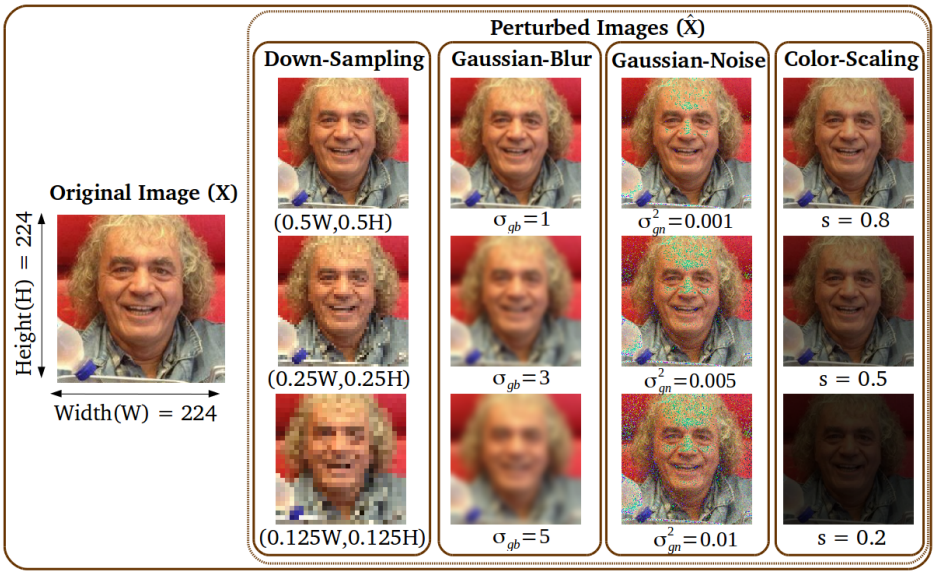


Figure 3.3: Example of a training image and its noise-distorted versions with specific types of noise.

Examples of the original and distorted images for each type of noise are depicted in Figure 3.3. Once trained, the robust models would be expected to produce small errors in the landmark estimates when given either normal or distorted input images.

3.4.2 Datasets and Experimental settings

To facilitate fair comparison, we used the standard benchmark of 300-W [Sagonas et al., 2013] and Menpo 2D test datasets [Zafeiriou et al., 2017b] to evaluate the robustness of the evaluated models. We cropped and added synthetic noise to the original images on both datasets following section 3.4.1 and named them as 300W-Test-Nx and Menpo-Test-Nx respectively, with *x* standing for the type of noise that was added. We followed the standard testing procedure as in [Zafeiriou et al., 2017b, Chrysos et al., 2017] consisting of initializing all models with the bounding box obtained from the ground-truth points and quantitatively evaluating their final landmark estimations with Area Under the Curve (AUC) and Failure Rate (FR) for Normalized Mean Error (NME) by Facial Bounding Box scores up to 0.08.

To further evaluate the robustness of our models in real-world conditions, we also tested our model trained with the noisy datasets on the Category 3 from 300-VW dataset (the most challenging one). This subset contains low quality images with several illuminations, blurry images and challenging poses which are ideal to evaluate the robustness of the models [Gu et al., 2017, Shen et al., 2015]. We refer to this dataset as 300-VW-3 and produce also a modified version with cropped images according to the facial bounding boxes 300-VW-3-C.

3.4.3 Impact of Synthetically Degraded Images on Landmark Estimations

To evaluate the robustness to degraded images of our model we report our results under 3 different configurations: *FADeNN-D*, *FADeNN-S* and *FADeNN-M*, as detailed in Section 3.3.3. Additionally, we also introduce stand-alone *FLL* to establish a baseline for the performance of our model with no denoising block at all. We compare our results with respect to six other facial alignment models. The first four consists of state of the art deep-learning based models: Robust Estimation-Correction-Tuning (ECT) Network [Zhang et al., 2018a], Style Aggregated Neural Network (SAN) [Dong et al., 2018], Facial Alignment Network (FAN)

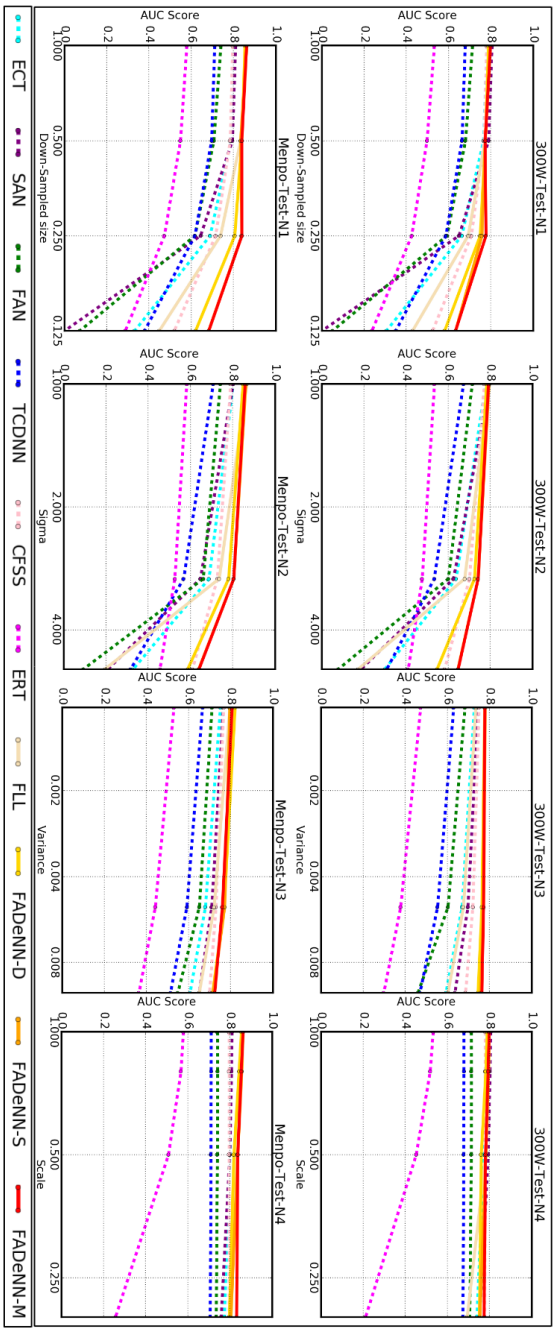


Figure 3.4: The AUC value degradation of the evaluated models on perturbed images with different noise levels.

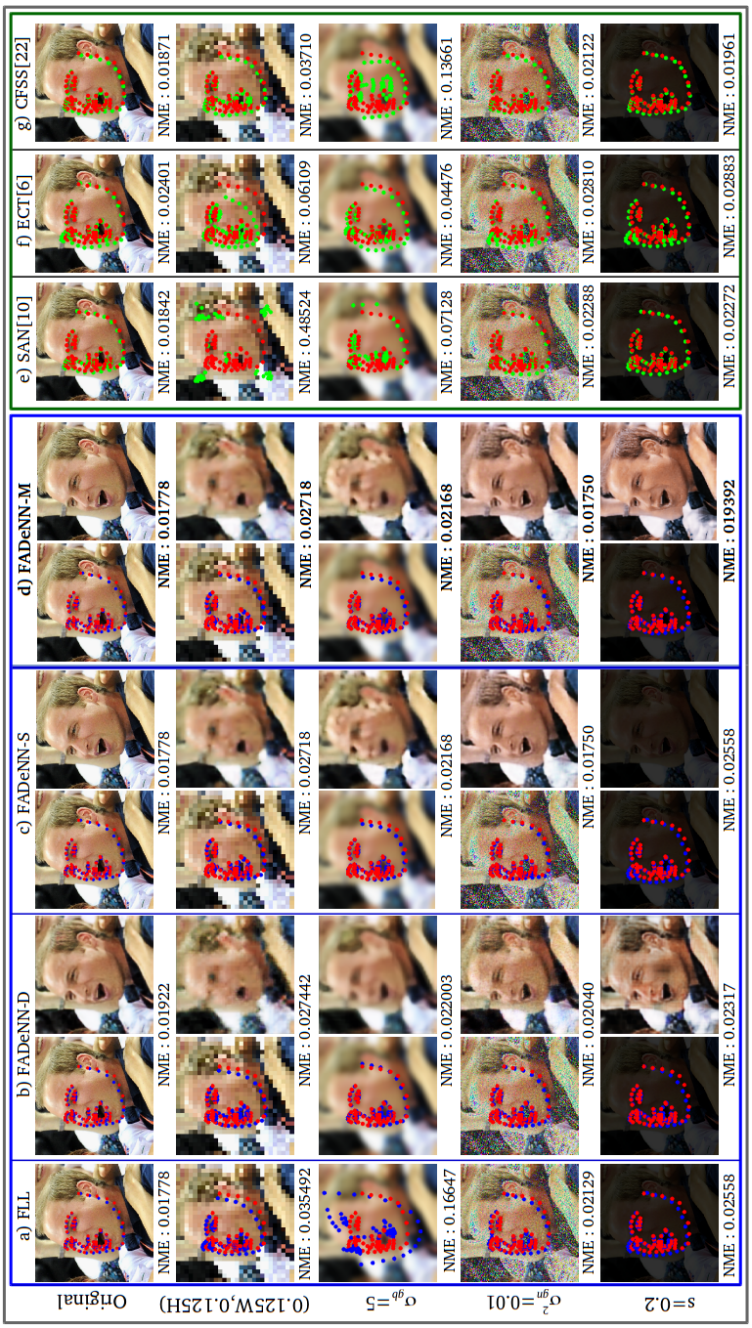


Figure 3.5: Example of facial landmark estimations of our models compared to other alternatives: a) FLL, b) FADeNN-D, c) FADeNN-S, d) FADeNN-M, e) SAN, f) ECT, g) CFSS.

model [Bulat and Tzimiropoulos, 2017] and Task Constrained Convolution Deconvolution Network (TCDCN) [Zhang et al., 2016b]. While the other two are traditional hand-crafted based models of facial alignment: Ensemble Regression Trees (ERT) [Kazemi and Sullivan, 2014] and Coarse to Fine Shape Searching (CFSS) [Zhu et al., 2015a].

On figure 3.4 we show the impact of each type of noise in the performance of the different models (full AUC and FR values of each models on each graph can be seen in the Appendix B). The deep learning models, including our baseline *FLL* model are quite sensitive to both image down-sampling and Gaussian blurring. This can be seen on the rapid AUC score drop especially at the highest levels of distortion (one eighth of image resolution and gaussian blur with σ_{gb}=5). The [Zhang et al., 2018a] method on the other hand performs slightly better compared with other models given that it is designed to be robust to noise, albeit its performance is still in par with our base model *FLL*. We also see that the hand-crafted methods do not experience so significant performance drop and, indeed, CFSS [Zhu et al., 2015a] shows quite competitive results. For the gaussian and color scale noise, most models showed only moderate degradation with the exception of ERT, which performed slightly worse than the rest. This color and gaussian noise tolerance may be attributed to the internal image normalization steps used in the pre-processing pipelines of most methods. These results are consistent with findings from [Zhou et al., 2018], suggesting universal weakness of current models against blur and resolution changes which may be due to lacks of blurry features in the designed or learned filter banks.

Further analysis of our own models performance, we find that our base *FLL* model performs comparatively well, on par with other state-of-the art models in spite of its simplicity, probably as a result of our rigorous transfer learning setup. In the heavily blurred and down-sampled images however, it still fails to locate landmark accurately (Figure 3.5 which is avoided by our *FADeNN* models thanks to their internal denoiser blocks (Figure 3.5). However, on the reasonably clean image (first row of Figure 3.5.b), we observe a slight accuracy drop from *FLL* to *FADeNN-D*, which is explained by the use of the direct approach in which the

IID block attempts to denoise all input images, even when they may not require it. This is further solved by the selective denoising approach of *FADeNN-S* and *FADeNN-M* with even cleaner denoised image. Finally we find that *FADeNN-M* works slightly better than *FADeNN-S* because its classification scheme is more accurate in correctly identifying the type of denoising required for each input image.

3.4.4 Comparison on 3rd category of 300-VW Test dataset

To compare the performance of our model under real-world conditions, in which the noise of the input images cannot be explicitly synthesized for training, we performed further experiments using the 300-VW dataset. We considered two different settings: facial landmark localization on single images (using the cropped dataset 300-VW-3-C) and facial landmark tracking (on the original 300-VW-3 dataset, without cropping). For the latter, we convert our model to perform model-free tracking [Chrysos et al., 2017] by using the result from the previous frame as initialization for the current one. In case of fitting failure, the external facial detector [Zhang et al., 2016a] will be called for re-initialization. Fitting failure is automatically determined by means of an independent facial logistic regressor over the patches around landmark estimates as in [Chrysos et al., 2017]. Table 3.1 summarizes the results obtained by our method as well as by state-of-the-art alternatives (corresponding AUC curves can be seen in the Appendix B). For the single-image setting we compare our results to the same methods used for comparison in the previous section, while for the tracking setting we compare our models with to the competitors of the original challenge in [Shen et al., 2015] and also to MD_CFSS and ME_CFSS, which were the highest-performing methods in the recent evaluation from [Chrysos et al., 2017].

We can see that our model achieves state-of-the-art accuracy, with a slight improvement with respect to the best-performing alternatives, and at the same time produces far lower failure rates than all other compared

methods. This behavior is consistent across both the single-image and the tracking settings. As in the experiments reported under synthetic noise in the previous section, we find that our base model FLL produces a comparable accuracy to other state-of-the-art models. Furthermore, with the introduction of the IID block, the result is progressively improved. The best performance is achieved by the FADeNN-M model.

We can also validate our result by visual inspection on Figure 3.6, where our models consistently provide the most accurate facial landmark estimates under very challenging conditions: strong illumination differences on the first row, blurry images on the second row, and also partially occluded facial parts on the third row. We can also see that the use of the internally normalized image on $FADeNN-M$, even though not perfect, helps to correct the position of several landmark points, explaining its effectiveness.

3.5 Conclusions

In this chapter, we evaluate the impact of synthetically degraded images on the performance of current facial alignment models and further use this understanding to build a robust facial alignment modes. We do this by building a deep neural network composed of a state-of-the-art Facial Landmark Localisation network in combination with an internal image denoising network.

Our systematic experiments with synthetically added noise reveal that the addition of synthetic noise degrades the performance of all the compared facial alignment models. In particular, the accuracy of deep learning models is severely degraded in the case of blurring and down-sampling but seems more resistant to gaussian noise and color scaling. In contrast, our proposed *FADeNN* models show enhanced robustness to all types of tested noise while achieving top-level accuracy. Remarkably, we also find that models trained with simple types of synthesized noise maintain their robustness to challenging input images from real-World datasets, such as 300-VW (category 3). In such *in-the-wild* settings, our model inter-

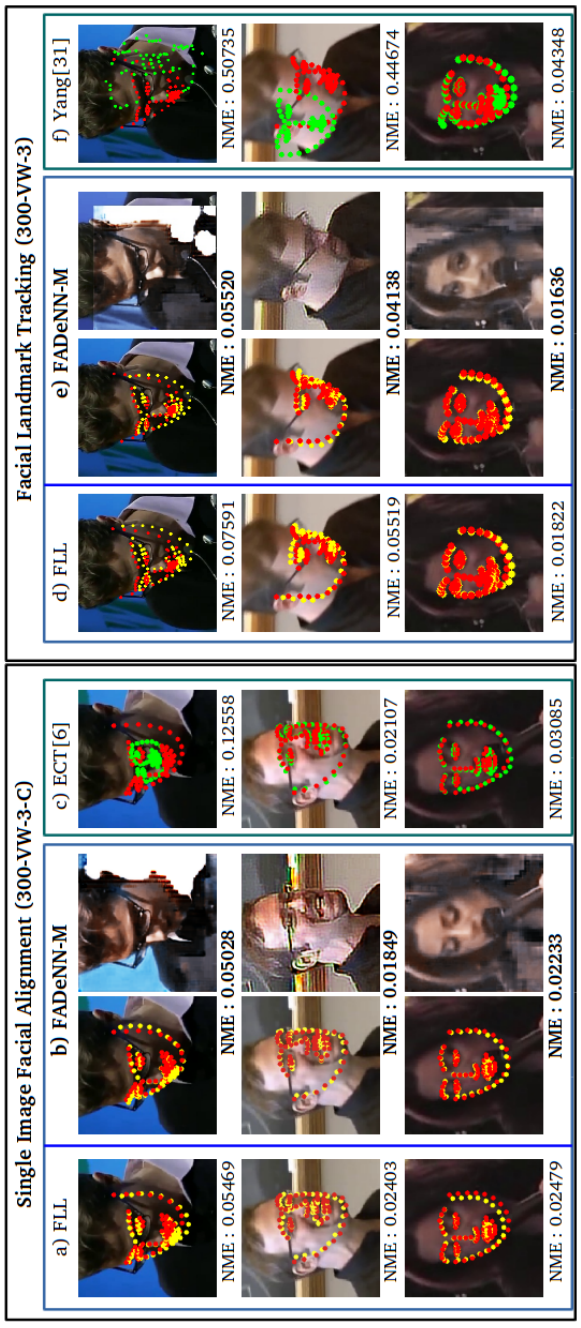


Figure 3.6: Visualization of single image facial landmark alignments : a) FLL, b) FADeNN-M, c)ECT. Facial landmark tracking : d) FLL-T, e) FADeNN-M-T, f) Yang.

Method	Single-Image		Tracking	
	AUC	FR	AUC	FR
ECT [Zhang et al., 2018a]	0.763	0.037	-	-
SAN [Dong et al., 2018]	0.747	0.038	-	-
FAN [Bulat and Tzimiropoulos, 2017]	0.733	0.015	-	-
TCDNN [Zhang et al., 2016b]	0.728	0.034	-	-
CFSS [Zhu et al., 2015a]	0.754	0.016	-	-
ERT [Kazemi and Sullivan, 2014]	0.680	0.120	-	-
MD_CFSS [Chrysos et al., 2017]	-	-	0.726	0.075
ME_CFSS [Chrysos et al., 2017]	-	-	0.659	0.114
Yang [Yang et al., 2015]	-	-	0.710	0.045
Jinwei [Gu et al., 2017]	-	-	0.617	0.048
Uricar [Uricár et al., 2015]	-	-	0.574	0.080
Xiao[Xiao et al., 2015]	-	-	0.695	0.074
Raja [Rajamanoharan and Cootes, 2015]	-	-	0.659	0.083
Wu [Wu and Ji, 2015]	-	-	0.602	0.131
FLL	0.752	0.017	0.711	0.022
FADeNN-D	0.756	0.015	0.718	0.020
FADeNN-S	0.760	0.013	0.726	0.020
FADeNN-M	**0.764**	**0.012**	**0.729**	**0.019**

Table 3.1: AUC and FR score on the 300-VW dataset, category 3rd.

nally estimates a cleaned-up version of the input image before performing landmark localisation, which leads to significantly lower failure rates than competing approaches while maintaining top state-of-the-art accuracy.

Chapter 4

COMPOSITE RECURRENT NETWORK WITH INTERNAL DENOISING FOR FACIAL ALIGNMENT IN STILL AND VIDEO IMAGES IN THE WILD

Adapted from: D. Aspandi, O. Pujol, F. Sukno, and X. Binefa, ”Composite recurrent network with internal denoising for facial alignment in still and video images in the wild” in *Image and Vision Computing, (Under Review)*.

Abstract

Facial alignment is an essential task for many higher level facial analysis applications, such as animation, human activity recognition and human - computer interaction. Although the recent availability of big datasets and powerful deep-learning approaches have enabled major improvements on the state of the art accuracy, the performance of current approaches can severely deteriorate when dealing with images in highly unconstrained conditions, which limits the real-life applicability of such models. In this work, we propose a composite recurrent tracker with internal denoising that jointly address both single image facial alignment and deformable facial tracking in the wild. Specifically, we incorporate multilayer LSTMs to model temporal dependencies with variable length and introduce an internal denoiser which selectively enhances the input images to improve the robustness of our overall model. We achieve this by combining 4 different sub-networks that specialize in each of the key tasks that are required, namely face detection, bounding-box tracking, facial region validation and facial alignment with internal denoising. These blocks are endowed with novel algorithms resulting in a facial tracker that is both accurate, robust to in-the-wild settings and resilient against drifting. We demonstrate this by testing our model on 300-W and Menpo datasets for single image facial alignment, and 300-VW dataset for deformable facial tracking. Comparison against 20 other state of the art methods demonstrates the excellent performance of the proposed approach.

4.1 Introduction

The human face is arguably one of the most important deformable objects for analysis in real world applications, such as facial animation, human activity recognition and human - computer interaction [Wu and Ji, 2018]. Facial alignment, which aims to detect a set of facial landmark positions, is essential for the high level analysis required in those applications to function well [Shen et al., 2015]. Currently, the development of facial alignment models is growing rapidly with the availability of large facial landmarked datasets such as 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2017b]. This has made it possible the development of powerful deep learning models that have pushed forward the alignment accuracy and are considered the current state of the art [Bulat and Tzimiropoulos, 2017, Gu et al., 2017].

However the performance of current facial alignment models can severely deteriorate when dealing with images in highly unconstrained conditions, e.g. extreme pose or illumination changes, large occlusions [Zhang et al., 2018a] or, in general, whenever the test images can be considered to show less favorable conditions than those available for training. In other words, we may say that under challenging conditions, test images contain some form of distortion or noise that will impair the performance of facial alignment models. This limits the real-life applicability of such models to *in-the-wild* images which naturally contain such challenging conditions [Nada et al., 2018, Zhou et al., 2018]. While there have been attempts to improve the robustness of facial alignment models to target in-the-wild data [Qu et al., 2015, Zhang et al., 2018a], most of them have done so without modeling the effects of noise in their formulation. Nevertheless, understanding and incorporating such effects within the model training has proven beneficial to improve performance in other facial analysis tasks [Dong et al., 2018, Nada et al., 2018, Goswami et al., 2018].

A very important aspect of facial alignment algorithms is their ability to operate reliably in the presence of image sequences (i.e. video data). In such cases, the problem becomes more challenging as there is the need for persistent stability to keep track of the facial features throughput the

whole sequence [Wang et al., 2012a]. Progress on facial tracking has been relatively slower when compared to single image facial alignment, and it has been less influenced by deep learning models [Chrysos et al., 2017]. Furthermore, the majority of currently available trackers make little use of temporal information: most of them process each frame independently and rely on doing so with sufficient precision to achieve *tracking-like* performance [Chrysos et al., 2015, Zheng et al., 2013], or convert single-image facial alignment to perform tracking by using the results of the previous frame as initialization [Yang et al., 2015, Asthana et al., 2014, Aspandi et al., 2019c]. However, the latter approach makes the models vulnerable to drifting since they are not designed specifically to track [Chrysos et al., 2017]. In contrast, other trackers incorporate some temporal modeling, but they are mostly limited to the adjacent frames [Xiao et al., 2015],[Rajamanoharan and Cootes, 2015]. All of the above suggests that current facial trackers may not take full advantage of the temporal information contained in video sequences [Xie et al., 2016].

In this chapter we present a composite end to end facial tracking model which jointly addresses both single image and video sequence alignment. We introduce a robust facial landmark estimator equipped with an internal denoiser autoencoder to selectively enhance the images on a case-by-case basis to boost the accuracy of single-image landmark estimation. Temporal information is handled by means of internal Long Short Term Memory (LSTM) layers that allows modeling of short and long temporal dependencies between frames when video is available. Combination of these two main modules results in an algorithm that produces very accurate facial alignment and is also resilient against drifting, leading to a robust facial tracker.

The contributions of this chapter are as follow:

1. We present a unified approach for both single image alignment and facial tracking with our joint robust facial alignment and tracker.

2. Our model employs an internal image denoising network to obtain an enhanced intermediate facial image that helps to improve the accuracy of landmark localization.

3. We incorporate temporal modelling between frames using internal multilayer LSTMs, which unlike other approaches, allow to consider time dependencies at various time scales.

4. We achieve state of the art results for both targeted tasks: Single Image Alignment (on 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2017b] datasets), and Deformable Facial Tracking on the wild (using 300-VW dataset [Shen et al., 2015]).

5. We investigate the impact of the considered temporal scale for tracking as well as the impact of the internal denoiser in the overall accuracy of our model.

Preliminary results of landmark detection by denoising autoencoder networks and composite recurrent tracking were presented in [Aspandi et al., 2019b, Aspandi et al., 2019c]. The rest of the chapter is organized as follows: Section 4.2 describes the related work in context of single image facial alignment and facial tracking. In Section 4.3, we explain our Denoised Composite Recurrent Tracker, which consists of multiple sub-networks operating in tandem and merged using our novel tracking algorithm. Section 4.4 reports our experiments on both Single Image Facial Alignment and Deformable Facial Tracking. Finally in Section 4.5 we derive our conclusions.

4.2 Related Work

In this section, we describe the prior works related to the two fundamental problems addressed by our approach, namely Single Image Facial Alignment and Facial Tracking.

4.2.1 Single Image Facial Alignment

Face alignment has attracted considerable attention due to its importance for several applications, such as face recognition [Zhu et al., 2015b], head

pose estimation [Wu et al., 2017], facial reenactment [Thies et al., 2016] and others. This task is generally conducted by estimating a predefined set of landmark positions which provide a structure representation of the facial geometry. Traditionally, facial landmarks have been estimated using either shape and appearance models [Yuille et al., 1992], [Cootes et al., 1995] or regression based models [Kazemi and Sullivan, 2011], [Xiong and la Torre, 2015], with the latter offering some advantage due to their computational efficiency [Valstar et al., 2010]. Recent examples of regression-based models include the work of Kazemi and Sullivan [Kazemi and Sullivan, 2014], who use a cascade of regression functions to efficiently regress the landmark locations, and Zhu et al [Zhu et al., 2015a] who refined the regressors cascade with coarse to fine search.

The recent availability of large datasets such as 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2017b] allows for large scale data modeling and development of deep learning-based models that benefit from these load of data. For instance, the work of Bulat and Tzimiropoulos [Bulat and Tzimiropoulos, 2017] used multiple hourglass shaped convolutional networks with heatmap-guided layers to predict the final facial landmarks, while the Zadeh et al [Zadeh et al., 2017] introduce multiple patch experts to adjust each of their final landmark estimations, which has been incorporated in the Openface framework [Baltrusaitis et al., 2018]. Zhang et al. [Zhang et al., 2016b] in addition, enforced the facial attributes as auxiliary features to their cascaded convolutional network to improve landmark estimates. These deep learning models currently hold the state of the art accuracy both on the single image facial alignment [Sagonas et al., 2013] and on the facial landmark tracking tasks [Shen et al., 2015].

Despite the maturity of current facial alignment models, there are still challenges for them when targeting images with large appearance variations due to heavy occlusion, severe pose or illuminations conditions, etc [Zhang et al., 2018a]. This is especially relevant in real world application deployed in highly unconstrained settings [Nada et al., 2018]. Thus, there has been growing interest to improve the performance of facial mod-

els under such settings, including efforts such as iterative initialization of regression cascades to minimize the impact of outliers [Qu et al., 2015] or the combination of data- and model-driven estimators to enhance the robustness of landmark detection [Zhang et al., 2018a]. However, most efforts have focused on building robust feature extractors with the assumption that the model will be able to discard the noise and select only the meaningful features.

In contrast, we address this problem from a different perspective: we focus on modelling image noise by means of our internal denoiser network, which is conceived as an auxiliary block that aims to automatically enhance the quality of the input image to improve the accuracy of landmark localization. This strategy can be justified by recent findings such as those from Dong et al. [Dong et al., 2018], who aggregated different styles of images using Generative Adversarial Networks to improve their landmark estimates, revealing that even a slight color style variations in a given facial images (which is otherwise kept fixed) can impact the accuracy of landmark estimations. Similar findings have been reported in other facial analysis tasks, such as those from Zhou et al. [Zhou et al., 2018], who evaluated several facial detection models with synthetic noise, and Goswami et al. [Goswami et al., 2018] who investigated the robustness of facial classification to synthetically distorted images. The findings in these reports have led us to hypothesize that modeling certain types of noise may help to characterize the behavior of facial analysis systems with unconstrained images and that incorporating such noise modeling into the training process could improve the robustness of their results.

4.2.2 Facial Tracking

Currently, the most popular facial tracking technique is Tracking by Detection, which consists of performing facial detection and landmark localization at each frame. Some examples of this strategy include the work from Uricar et al. [Uricár et al., 2015] and the OpenFace tracker. Uricar et al. use tree-based Deformable Part Models (DPM) for facial landmark detection and localization with Kalman Filter smoothing, while the OpenFace

tracker [Baltrusaitis et al., 2018] uses the multiple convolutional experts method described in the previous section [Zadeh et al., 2017] initialized with the bounding box from a Multi-Task Cascaded Neural Network face detector [Zhang et al., 2016a].

Other tracking methods perform face detection only in the first frame and then apply facial landmark localization using the fitting result from the previous frame as initialization. One such example is the work from Xiao et al. [Xiao et al., 2015] which adopts a multi-stage regression-based approach to initialize the shape of landmarks with high semantic meaning. Other examples include the work from Raja et al. [Rajamanoharan and Cootes, 2015] which combines a global shape model with sets of response maps for different head angles indexed on the shape model parameters and the work from Wu et al [Wu and Ji, 2015] who apply shape augmented regression. There are also hybrid approaches that combine tracking by detection and initialization based on the latest fitting result. Among these, combinations of Coarse-To-Fine Shape Search (CFSS) [Zhu et al., 2015a] landmark localizer with multiple general-object trackers have shown to perform particularly well [Chrysos et al., 2017].

However, all methods derived from tracking by detection share the limitation of not considering the temporal information contained in video sequences. Furthermore, it is difficult to obtain consistent initializations from most face detectors, which tends to reduce the final landmark localization accuracy [Lv et al., 2017a]. Some approaches try to mitigate this problem by including the information from the adjacent frames to capture short temporal dependencies. For example, Yang et al. [Yang et al., 2015] used time series regression on pairs of adjacent frames, which led them to achieve the best result reported so far on the original challenge in the 300 Videos in the Wild dataset (300-VW) [Shen et al., 2015], which is the largest deformable facial tracking benchmark to date.

With the recent growth of facial landmark datasets, such as 300-W [Sagonas et al., 2013], Menpo [Zafeiriou et al., 2018], 300-VW and LS3D-W [Bulat and Tzimiropoulos, 2017], current methodologies on facial analysis started to shift from systems based on handcrafted features towards in-

corporating deep learning architectures [Zafeiriou et al., 2017b], [Greenspan et al., 2016]. Rapid progress can be seen on the development of various convolutional architectures as the main spatial feature extractor used on both facial detection [Zhang et al., 2016a] and landmark localization models [Bulat and Tzimiropoulos, 2017, Zhu et al., 2017], achieving state of the art accuracy. In spite of this, localization is still mainly performed on every single frame, without taking into account the temporal information.

On the other hand, introduction of recurrent neural networks (RNN), especially Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997], has allowed incorporating temporal information with great success in several applications [Greff et al., 2017]. This is the case of the recently introduced general object tracker Re^3 [Gordon et al., 2018], which is robust against image occlusions and can be trained on long sequences thanks to its internal LSTM networks. Nonetheless, RNN have received little attention in the context of facial tracking. The only exceptions so far are the methods by Jiang et al. [Gu et al., 2017] and Peng et al. [Peng et al., 2018]. Jiang et al. [Gu et al., 2017] proved that an end to end RNN is capable to work on multiple domains including facial landmark tracking, with very low failure rates but without reaching state of the art accuracy. Peng et al. [Peng et al., 2018] used an internal recurrent encoder-decoder network with heatmap guidance, thus allowing temporal modeling, but at the expense of high model complexity.

In summary, even though there exist models that allow both the single image alignment and facial tracking tasks, such as OpenFace [Baltrusaitis et al., 2018], they consist of separate sub-models designed independently for each task, which is likely to produce sub-optimal results [Zhang et al., 2016a]. To the best of our knowledge, we are the first to present a facial tracker that incorporates a joint model for single-image denoising, facial landmark localization and temporal modeling. As it will be shown in the experiment section 4.4, this strategy improves the single image alignment accuracy and also enhances the stability of the tracking operation.

4.3 End to end Denoised Composite Recurrent Facial Tracker

Our tracking model is a composite network that receives raw frames as input and returns the localization of facial landmarks as the final output. It is composed by four sub-networks, arranged in a way that permits end-to-end training for each of them without involving any hand-crafted features. Specifically, let $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ denote the current and previous frame; our Denoised Composite Recurrent Tracker (*DCRT*) will estimate the position of n facial landmarks in the current frame $\mathbf{l}_t$:

$$\mathbf{l}_t = \{(\hat{x}_1, \hat{y}_1)...(\hat{x}_n, \hat{y}_n)\} = DCRT_{\Phi}(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{b}_{t-1}) \tag{4.1}$$

$$\mathbf{l}_t = DCRT_{\Phi}(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{b}_{t-1}) \tag{4.2}$$

where Φ are the parameters $\{\Phi^1, \Phi^2, ..\Phi^5\}$ of our composite networks *DCRT* and $\{\hat{x}_1...\hat{x}_n, \hat{y}_1...\hat{y}_n\} \in \mathbb{R}_{>0}$.

Our *DCRT* consists of four individual sub-networks: Multi-Task Cascaded Neural Network face detector (*MTCNN*), facial bounding Box Tracker (*BT*), Facial Validator (*FV*) and Denoised Facial Alignment (*DFA*). Note that for face detection we relied on the state of the art *MTCNN*[Zhang et al., 2016a]. This task-specific arrangement allows us to optimise each network to each of their task characteristics, and simultaneously to inspect their behaviour and contributions to our final landmark estimates when chained together using our tracking algorithm (cf. Section 4.3.4).

A schematic diagram of our tracker can be seen in Figure 4.1. We start by assuming a tracking scenario, where we have an existing estimate for the bounding box of the preceding frame.[2] This bounding box, together with the current and previous frames $(\mathbf{X}_t, \mathbf{X}_{t-1})$ are fed to our *BT* network to produce a first estimate of the targeted landmarks ($\mathbf{l}_t^{BT}$) and bounding box ($\mathbf{b}_t^{BT}$), while at the same time its internal state is updated.

[2]For initialization, this estimate can be obtained from the *MTCNN* detector or from an external input.

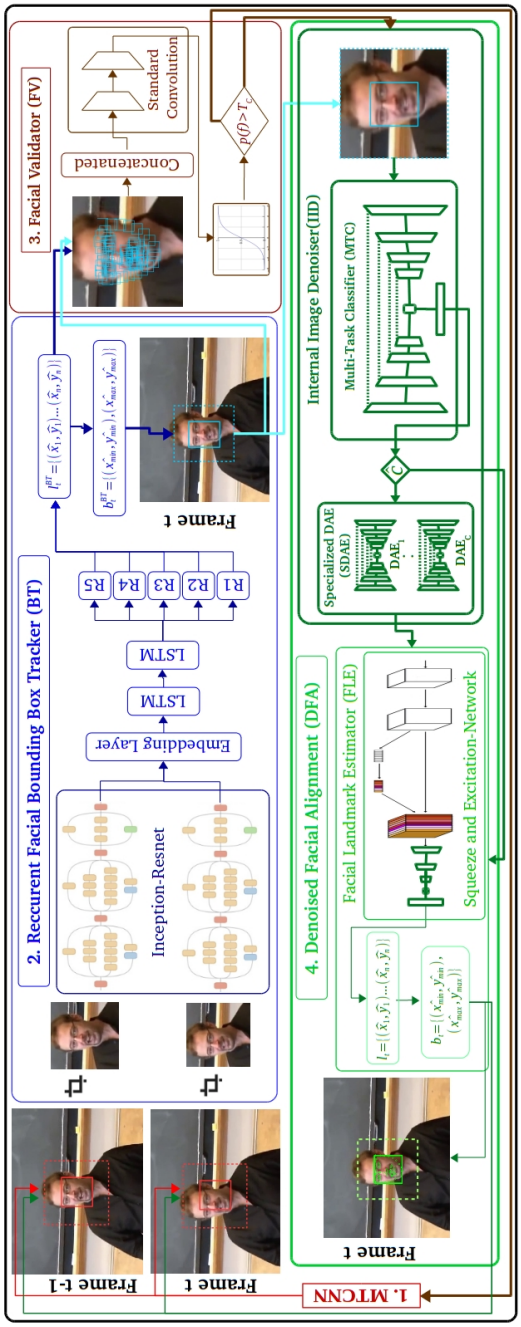


Figure 4.1: Overview of our Denoised Composite Recurrent Tracker architecture which consists of : 1) External Facial detector using MTCNN, 2) Recurrent Facial Bounding Box Tracker (BT) with internal multi layer LSTMs, 3) Facial Validator (FV), 4) Denoised Facial Alignment (DFA) which consists of Internal Image Denoiser (IID) and Facial Landmak Estimator (FLE).

Once we have our first landmarks estimate $\mathbf{l}_t^{BT}$, we use the *FV* network to validate the result obtained by the tracker. To do so, we train the *FV* network to estimate the probability that the object tracked within $\mathbf{l}_t^{BT}$ is a face ($p(f)$). In case of obtaining a low probability, which would suggest that the *BT* network has lost track, we use the *MTCNN* to perform face detection on the current frame and re-initialize the whole network for the next time step.

In contrast, if $\mathbf{l}_t^{BT}$ is successfully validated by the *FV* network, the current frame and its bounding box $\mathbf{b}_t^{BT}$ are fed to the *DFA* network, which produces the final estimates for the target landmarks $\mathbf{l}_t^F$ and the corresponding bounding box, $\mathbf{b}_t^F$. *DFA* works by first evaluating the image conditions and, if necessary, applying denoising to the image; subsequently, landmark coordinates are estimated from a *cleaned* version of the input image. Note that, *DFA* will operate from an already detected and validated bounding box, which allows it to achieve a more accurate result. Further, the *DFA* network would be able to operate by itself, independently of the rest of the network (e.g. in the single image facial alignment task), but in such case it would lose the temporal modeling given by the previous sub-networks.

4.3.1 Bounding Box tracker

We base our *BT* network on the structure of the Re^3 tracker [Gordon et al., 2018], which is a full end to end object tracker with internal LSTM networks to capture the temporal dependencies from video. Given input frames $\{\mathbf{X}_t, \mathbf{X}_{t-1}\}$ cropped as $\{\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}\}$ with the previous Bounding Box ($P_b = b_{t-1}$), the *BT* network estimates the landmark positions for the current frame $\mathbf{l}_t^{BT}$ and updates the internal state of the LSTM $\mathbf{h}_t$ as follows:

$$\begin{aligned}\mathbf{h}_t, \mathbf{l}_t^{BT} &= BT_{\Phi^1}(\mathbf{X}_t, \mathbf{X}_{t-1}, P_b, h_{t-1}) \\ &= BT_{\Phi^1}(\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}, h_{t-1}) \\ &= LSTM_{\Phi^1}(EL_{\Phi^1}(\mathbf{X}_t^{Pb}, \mathbf{X}_{t-1}^{Pb}), h_{t-1}) \odot W_{\Phi^1}^{BT}\end{aligned} \tag{4.3}$$

where $LSTM$ refers to the set of internal LSTM [Hochreiter and Schmidhuber, 1997] networks, EL stands for the Embedding Layer, W^{BT}, W^{EL} is the set of weights of each fully connected layers of BT and EL respectively and res is the Inception-Residual Network [Szegedy et al., 2016] (Inception-Resnet). The Embedding Layer is a weighted concatenation of the residual network coefficients:

$$EL = [res_{\Phi^1}(\mathbf{X}_t^{Pb}); res_{\Phi^1}(\mathbf{X}_{t-1}^{Pb})] \odot W_{\Phi^1}^{EL} \tag{4.4}$$

We use Φ^1 to denote the parameters of all sub-networks contained in BT. Finally, we also generate an estimate of the bounding box for the current frame $\mathbf{b}_t^{BT}$ directly from the estimated landmarks:

$$\mathbf{b}_t^{BT} = \{(\hat{x}_{min}, \hat{y}_{min}), (\hat{x}_{max}, \hat{y}_{max}) | \hat{x}, \hat{y} \in \mathbf{l}_t^{BT}\} \tag{4.5}$$

Note that, even though the architecture of BT is based on Re^3, we introduce several key modifications to adapt this recurrent tracker model into this new problem domain:

1. First we preconditioned the convolutional network of our BT to contain common facial features by replacing the internal Skip Convolution Networks (SkipNet) with the more sophisticated Inception-Resnet that has been pre-trained on the MS-Celeb [Guo et al., 2016] and CasiaWebFace [Yi et al., 2014] datasets[3] with triplet loss [Parkhi et al., 2015]. Figure 4.2 visualizes the differences between the original SkipNet on Re^3 versus the more complex structure of BT, which is inherited from the Inception-Resnet (Version 1). Each block of Inception-Resnet architecture can be expressed as below:

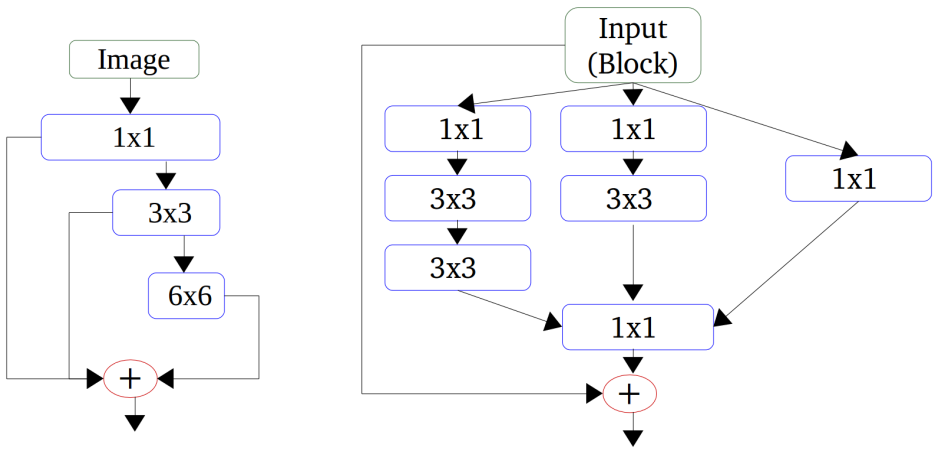
   $$\mathbf{r}_{i+1} = H(r_i) + F(r_i, W_i) \tag{4.6}$$

   Where r_i and r_{i+1} are the input and output of the i-th block, $H(b_i)$ is the identity matrix and F represents the combined effect of the various convolutional and ReLU layers. Notice that SkipNet does not have the advantage of residual connection as in the Inception-Resnet which eases the gradient flow in optimization [Szegedy et al., 2016].

[3]The trained inception resnet is publicly available on: https://github.com/davidsandberg/facenet

2. Second we use the BT network to produce a first estimate of landmark locations ($\mathbf{l}_t^{BT}$) following the work of [Gu et al., 2017], but we split the fully-connected layer that receives the output from the LSTMs into five independent fully-connected networks so that each of them is focused on a specific facial region. To this end, we divide the facial landmarks in the following regions: facial silhouette (our outer contour), eyebrows, eyes, nose and lips. Thus $W^{BT} = \{W^{R1}, W^{R2}, W^{R3}, W^{R4}, W^{R5}\}$.

3. Finally, we reduce the dimensionality of the input image size to 128x128 to accelerate the training process. This in turn also reduces the number of the neurons of the original Re^3 by half and helps to avoid over-fitting [Biau et al., 2016] while still achieving state of the art accuracy.



(a) Skip Connections (b) Inception-Resnet Block

Figure 4.2: Convolution architectures of Skip Network vs Inception-Residual Network block.

4.3.2 The Facial Validator

After the initial estimates produced by the *BT* network we use the *FV* network to validate the results before further processing. The main reason for doing so is to avoid the drift problem, well known in the tracking literature [Wang et al., 2012a]. Specifically, the *FV* network can be understood as a conditional function that determines whether to continue the processing pipeline based on the current estimates from *BT* or to reset the tracker and attempt to re-detect the facial region because the current estimates are not reliable enough.

We follow the methodology in [Chrysos et al., 2017] to build a strong classifier to estimate the probability $p(f)$ that the object currently being tracked by *BT* is a face. To this end, we use concatenated small patch regions from the estimated landmarks ($\mathbf{l}_t^{BT}$) as follows:

$$\begin{aligned}\mathbf{p}(f \,|\, \mathbf{X}_t, \mathbf{l}_t^{BT}) &= FV_{\Phi^2}(\mathbf{X}_t, \mathbf{l}_t^{BT}) \\ &= \frac{1}{1+\exp\big(-(W_{\Phi^2}^{FV} \odot CNN_{\Phi^2}(\mathbf{X}_t, \mathbf{l}_t^{BT}))\big)}\end{aligned} \tag{4.7}$$

Where *CNN* is the composite function of standard stacked convolution layers followed by a bottleneck layer with W^{FV} parameterized by Φ^2 and $0 < p(f) < 1$. We set a threshold T_c to determine the lowest probability that is acceptable from *FV* to validate the current estimate.

4.3.3 Denoised Facial Alignment

We argue that when given a noisy input image, final landmark estimation can be improved by first minimizing the existing noise and only then passing the image to any facial landmark estimator. To achieve this, we adopt a joint approach involving two major sub-networks to build our Denoised Facial Alignment network (DFA): Internal Image Denoiser (*IID*) and Facial Landmark Estimator (*FLE*) networks. We use the current cropped facial area $\mathbf{X}_t^{Pb}$ as input image, which we assume to be potentially contaminated with unknown noise. In other words, $\mathbf{X}_t^{Pb}$ is assumed to be an estimate of a noiseless image $\mathbf{Y}$ to be estimated by the *IID* network. We

do so in a two-step approach in which we first detect if the noise in $\mathbf{X}_t^{Pb}$ complies with any of our noise models and, if so, clean the input image to generate an enhanced estimate of $\mathbf{Y}$. namely $\hat{\mathbf{Y}} = IID(\mathbf{X}_t^{Pb})$. Otherwise, if $\mathbf{X}_t^{Pb}$ does not match any of the internal noise models, then $\hat{\mathbf{Y}} = \mathbf{X}_t^{Pb}$. The resulting image $\hat{\mathbf{Y}}$ will be fed to the FLE network to estimate n landmark coordinates $\mathbf{l}_I$

$$\mathbf{l_t} = DFA_{\{\Phi^3..\Phi^5\}}(\mathbf{X}_t^{Pb}) = FLE_{\Phi^5}(IID_{\{\Phi^3,\Phi^4\}}(\mathbf{X}_t^{Pb})) \tag{4.8}$$

Internal Image Denoiser

We adopt a selective denoising approach [Aspandi et al., 2019c], [Nam and Han, 2016] by combining Multi-Task noise Classifier(MTC) network and multiple Specialized Denoise Auto Encoders (SDAE) sub-networks to work in tandem. This joint model selectively performs denoising based on the detected condition of the input image to trigger separate specialized denoiser sub-models. This has been shown to perform better than directly denoising all input images, as the latter may actually distort an already clean input image [Aspandi et al., 2019c]. The formulation of the IID block is as follows:

$$\hat{\mathbf{Y}} = IID_{\{\Phi^3,\Phi^4\}}(\mathbf{X}_t^{Pb}) = SDAE_{\phi^4}(MTC_{\phi^3}(\mathbf{X}_t^{Pb}), \mathbf{X}_t^{Pb}) \tag{4.9}$$

Given C known image degradation types, the MTC will estimate the probability that degradation type c is present in the input image $\mathbf{X}_t^{Pb}$) :

$$s_c = \frac{\exp\left(W_{\phi_c^3}^{MTC} \odot CNN_{\phi_c^3}^{MTC}(\mathbf{X}_t^{Pb})\right)}{\sum_{c=1}^{C} \exp\left(W_{\phi_c^3}^{MTC} \odot CNN_{\phi_c^3}^{MTC}(\mathbf{X}_t^{Pb})\right)} \tag{4.10}$$

$$\hat{c} = MTC_{\phi^3}(\mathbf{X}_t^{Pb})) = argmax_c\{s_c\} \quad \{s_c\} = \{s_0, s_1, \ldots s_C\} \tag{4.11}$$

where W^{MTC} are the multinomial bottleneck regression layers parameterized by Φ^3, CNN is the set of convolutional layers for MTC, and s_c is the estimated probability that the current input $\mathbf{X}_t^{Pb}$) is contaminated with

specific noise-class c, with s_0 denoting the noiseless case. If the classifier detects noise in the input image, then $\hat{c} > 0$ and the image is denoised by one of the specialized denoisers $\{DAE_1, DAE_2 ... DAE_C\}$. We build our Internal Image Denoiser based on the Hourglass shaped Auto-Encoder Architecture with skip connection [Mao et al., 2016]. The structure of our *DAE* is similar to the work of [Chaitanya et al., 2017] with a few additional mirror layers on both encoder and decoder parts. Each of these blocks can be trained to capture a different type of noise:

$$SDAE_{\phi^4}(\hat{c}, \mathbf{X}_t^{Pb})) = DAE_{\hat{c},\phi_c^4}(\mathbf{X}_t^{Pb})) = enc_{\hat{c},\Phi^4}(dec_{\hat{c},\Phi^4}(\mathbf{X}_t^{Pb}))) \quad (4.12)$$

Where Φ^4 contains the parameters learned for all DAE blocks. In case $\hat{c} = 0$, the denoising process is skipped to avoid unnecessarily distorting the input image, i.e. $SDAE_{\phi^4}(0, \mathbf{X}_t^{Pb})) = \mathbf{X}_t^{Pb}$.

Facial landmark estimator

We build our final pipeline of *FLE* using the state-of-the-art Squeeze and Excitation Network (SENET) [Hu et al., 2018] which has been pre-trained on the recently published VGGFace2 [Cao et al., 2018] facial dataset[4]. This landmark localization procedure can be expressed mathematically as below:

$$FLE_{\phi^5}(\hat{\mathbf{Y}}) = W_{\Phi^5}^{FLE} \odot SEN_{\Phi^5}(\hat{\mathbf{Y}}) \quad (4.13)$$

4.3.4 Recurrent Denoised Facial Tracking Algorithm

The operation of our Denoised Composite Recurrent Tracker, *DCRT*, which combines all the blocks explained previously in this section, is shown in Algorithm 2. When a suitable detection of the facial region is available, e.g. from initialization or the previous frame (lines 8 and 10), the *BT* network produces a first estimate of facial landmarks (line 13) and bounding box (line 14). Then, the *FV* network is used to estimate the

[4] https://github.com/ox-vgg/vgg_face2

probability $p(f)$ that the output from *BT* corresponds to a face. If $p(f)$ is sufficiently high (above threshold T_c), the initial estimate is refined by the *DFA* network to produce the final tracker estimate (lines 18 and 19) with its selective internal denoisers. Otherwise, it is assumed that the *BT* has lost track and there is a need to re-initialize the tracker (line 16).

We perform re-initialization between lines 3 and 6. We start by detecting the face in the current frame by means of the *MTCNN* network. This detector is likely to produce multiple detections, hence its outputs are validated with respect to the bounding box of the previous frame b_{t-1}. Specifically, we compare the Euclidean distance between each new detection and the center of the previous bounding box $\mathrm{d}(b_{t-1}, b^{MT})$ with respect to the magnitude of the previous bounding box, and keep the one that produces the minimum ratio:

$$P_b = \begin{cases} b, & \min\limits_{\forall b \in b^{MT}} \frac{d(b_{t-1},b)}{\|b_{t-1}\|} < T_B \\ b_0, & dim(b_{t-1}) < 0 \\ b_{t-1}, & otherwise \end{cases} \tag{4.14}$$

as long as there is at least one detection whose ratio is below threshold T_B. Otherwise, all new detections are too far from the previous tracking result and no re-initialization is performed. The latter is necessary to tackle the cases in which the face being tracked moves out of the visual field. In such cases, without threshold T_B the system might be incorrectly re-initialized to track another face. In contrast, by using T_B the tracker remains in its latest valid coordinates awaiting for the tracked object to *come back* to the field of view.

Finally, (*SeqBT*) controls the length of the temporal window that is considered by the tracker (in frame units), which is fixed at training time (see next section). If the tracker is re-initialized or if the sequence length (*SeqT*) exceeds the temporal window (*SeqBT*), then the internal state of the *BT* network is reset to 0 (line 10). This allows the network to refresh the facial appearances encoding to adapt to the surrounding characteristic of the current view

Algorithm 2 Recurrent Denoised Facial Tracking Algorithms

Input : Frame of $\mathbf{X}_{0..N}$
Initial value of b_0 and h_0
Threshold value of T_B, T_C, and SeqBT
Network parameters of $\Phi^1 ... \Phi^5$

Output : Facial Landmark of $\mathbf{l}_{1..N}$

redetect$\leftarrow$ FALSE, SeqT $\leftarrow 0$, $\mathbf{b}_t \leftarrow b_0$
for $t \leftarrow 1$ to N **do**
  if redetect **then**
    $b^{MT} \leftarrow MTCNN(\mathbf{X}_t)$
    if length(d$(b_{t-1}, b^{MT}) > T_B) > 0$ **then**
      $\mathrm{P}_b \leftarrow b^{MT}[\min(\mathrm{d}(b_{t-1}, b^{MT}))]$
  else
    $\mathrm{P}_b \leftarrow b_t$
  if dim$(\mathrm{P}_b) < 0$ **then**
    $\mathrm{P}_b \leftarrow b_0$
  if (redetect OR SeqT $>$ SeqBT) **then**
    $h_t \leftarrow h_0$ and SeqT $\leftarrow 0$
  $\mathbf{h}_t, \mathbf{l}_t^{BT} \leftarrow BT_{\Phi^1}(\mathbf{X}_t, \mathbf{X}_{t-1}, \mathrm{P}_{BB})$
  $\mathbf{b_t}^{BT} \leftarrow [max(\mathbf{l}_t^{BT}), min(\mathbf{l}_t^{BT})]$
  if $FV_{\Phi^2}(\mathbf{X}_t, \mathrm{l}_t^{BT}) < T_C$ **then**
    redetect $\leftarrow$ TRUE
  else
    $\mathbf{l}_t \leftarrow DFA_{\{\Phi^3..\Phi^5\}}(\mathbf{X}_t, \mathbf{b}_t^{BT})$
    $\mathbf{b}_t \leftarrow [max(\mathbf{l}_t), min(\mathbf{l}_t)]$
    SeqT $\leftarrow$ SeqT + 1
    redetect $\leftarrow$ FALSE

4.3.5 Overall Loss and Model Training

To train the *BT* sub-network, we performed curriculum learning with sequence lengths between $SeqBT = 2$ to $SeqBT = 32$ frames [Gordon et al., 2018]. Multiple stages of transfer learning

[Christodoulidis et al., 2017] were used to condition the pre-trained Inception-Resnet. To do so, we fine-tuned this network to perform single image landmark estimation by an adding auxiliary Fully Connected layer and training the resulting network with ℓ_2 loss using the 300-W [Sagonas et al., 2013] and Menpo [Zafeiriou et al., 2018] datasets. After convergence, we integrated the internal convolutional layers into our *BT* network, which was then trained end-to-end for landmark tracking using the 300-VW training dataset with ℓ_1 loss.

We used the same facial landmark alignment datasets to train both *FV* and *DFA* sub-networks. Standard cross entropy loss was used for *FV*, while *DFA* required joint losses from both *FLE* and *IID* to allow them to be optimized in parallel. We argue that synchronous optimization of *IID* and *FLE* allows *IID* to produce *cleaner* intermediate images $\hat{\mathbf{Y}}$ since that favors their mutual aim to achieve more accurate landmark estimates. Hence the complete loss for *DFA* training is as follows:

$$\mathbf{L_{DFA}} = \min \lambda_1 \ell^2(FLE(\hat{\mathbf{Y}}), \mathbf{l}) + \lambda_2 L_{IID} \tag{4.15}$$

$$\mathbf{L_{IID}} = \min \lambda_3 (H(\hat{c}, MTC(\mathbf{X})) + \lambda_4 \ell^2(\hat{\mathbf{X}}, \mathbf{X}) + \mathbf{L_{SDA}} \tag{4.16}$$

$$\mathbf{L_{SDA}} = \min_{\forall D \in SDAE} \ell^2(\hat{\mathbf{Y}}, D(\mathbf{X}))) \tag{4.17}$$

where $\mathbf{X}$ is the input image, $\hat{\mathbf{Y}}$ is the input image after denoising (i.e. an estimate of the unknown *clean* image $\mathbf{Y}$), $\mathbf{l}$ are the ground-truth landmarks, $\hat{c}$ is the estimated noise class, $H()$ is the cross-entropy and $\hat{\mathbf{X}}$ is the autoencoder reconstruction of $\mathbf{X}$. The λ coefficients act as regularization parameters for each term [Zhang and Yang, 2017], [Bulat and Tzimiropoulos, 2017]. In the initial training phase, we set a higher value for λ_2 compared to λ_1, for instance λ_2 of 0.75 and λ_1 of 0.25, to accelerate the training of the denoising part (λ_3 and λ_4 are set with equal value of 0.5), since we found that this part took more time to converge than the former. As training progresses, we gradually increase the value of λ_1 by a step value of 0.1, and reduce the value of λ_2 with a similar value to better balance the training priority, reaching a final value of 0.5 for both coefficients.

Figure 4.3 illustrates the impact of our proposed joint training when the input images present some sort of noises or distortions. We can see that

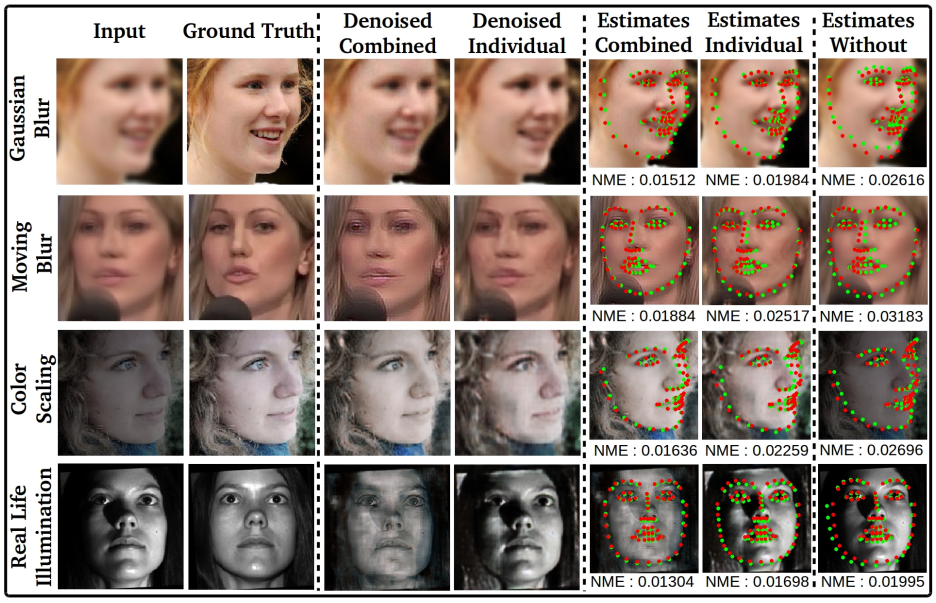


Figure 4.3: Visual examples of different results of *DFA* raining with respect to the use of our proposed join training of *IID* and FLE modules. The first two columns shows the degraded input images and their respective cleaned (ground truth) versions. Columns 3 and 4 consist of the denoised output of *IID* with and without joint training of FLE respectively. Columns 5 and 6 provide the predicted landmark estimates of $FLE + IID$, with and without joint training. Column 7 shows the FLE results without the use of *IDD*.

using this joint training scheme, *DFA* is able to improve the accuracy of the estimated landmarks (5th column), compared to the results using separate loss (6th column). We can also see that the internally denoised images are also less distorted (4th and 5th column), suggesting that these cleaner image charactheristics are beneficial for FLE to produce more accurate landmark estimates [Chrysos et al., 2019]. Finally, as we expected, the use of FLE (7th column) by itself, i.e without *IIE*, yields the worst performance given its lack of any internal denoising modeling (more details can also be seen on the next Section 4.4).

We consider two types of image degradations to train the *IID* subnetwork: Image Blur and Illumination Differences. This is motivated by our previous findings about the common weaknesses of current facial alignment models against such degradation models[Aspandi et al., 2019c]. Specifically, each degradation model was obtained as follows:

1. **Blur Model**. We added Gaussian Blur by convolving the input image with two dimensional Gaussian filters with $\sigma \in \{1,3,5\}$ [Zhou et al., 2018, Aspandi et al., 2019c]. In addition, we also included motion blur [Chrysos et al., 2019] by averaging L frames, with $L \in \{10,15,25\}$ to further simulate real world blur distortion.

2. **Illumination Model**. On this degradation models, we incorporated both synthetic and real life illumination degradation. First we synthetically created images with low pixel intensities by scaling the original pixel values linearly by a factor $s \in \{0.8,0.5,0.2\}$ from the original intensities. Second, we used the Yale [Georghiades et al., 2001, Belhumeur et al., 1997] and SOF datasets [Afifi and Abdelhamed, 2019] which contain facial images captured under large illumination variations. Specifically, Yale dataset provides some facial examples taken under different poses, while SOF dataset offers the coloured facial images with different emotional expressions, both are under different illumination settings. We used two versions of Yale dataset: original Yale dataset [Belhumeur et al., 1997] that consists of 15 subjects taken under different lighting (left, right and center lighting); the extended Yale Dataset B [Georghiades et al., 2001] consists of 28 subjects with different light source directions and poses. For the original version, we obtain the ground-truth (considered well-lit images) by using the centered-lit faces. Whereas for the extended Yale, we use the face taken with neutral position lighting, i.e indicated by azimuth and elevation degree with the value of 0. As for the SOF dataset, we use its second version that includes videos of 12 subjects with four different emotions (normal, happy, sad, angry, disgusted, and surprises), and filmed under extreme lighting conditions (using arbitrarily lo-

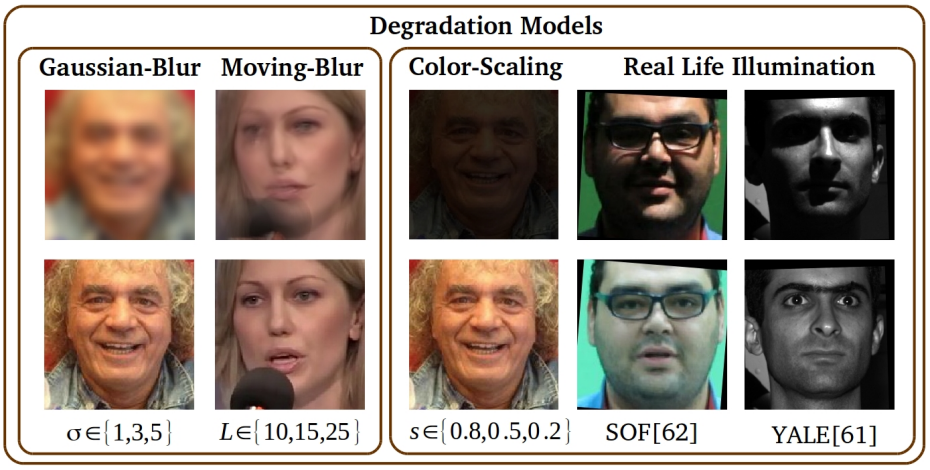


Figure 4.4: Example of noise-distorted and cleaned versions of input images for each degradation model. From left to right: images with Gaussian Blur, Moving Blur, synthetic illumination degradation (color scaling) and real life illumination changes from SOF[Afifi and Abdelhamed, 2019] and YALE[Lee et al., 2005] datasets.

cated wheel-whirled lamp). We used the faces taken on the ideal lighting, indicated through the provided metadata of 'illuminationQuality' label of valued 0 for each associated emotion expression as the ground truth for this dataset. Finally, we crop-centered the faces from SOF dataset using the provided 17 facial landmarks, while our previous tracker [Aspandi et al., 2019b] and facial alignment [Aspandi et al., 2019c] were used to obtain the landmarks for the Yale data given that they are not available for this dataset.

Examples of some training images together with their degraded versions are depicted in Figure 4.4.

In all the training process, we performed data augmentations by means of horizontal flipping, -45° to 45° degree rotations and artificial strip boxes across the images to simulate occlusions. We trained our model using both Stochastic Gradient Descent (SGD) and ADAM optimizer

[Kingma and Ba, 2014] with scheduled weight learning decay every 10.000 iterations. We first initialized training with SGD to speed up convergence, and then used ADAM for fine-tuning at final training stages. Five NVIDIA Titan GPUs were used for training which took approximately one to two days to train a single BT for a defined sequence length, and around two days for both DFA and FV.

4.4 Experiments

We present our results on two major experiments: Single Image Face Alignment and Deformable Facial Tracking. For fair comparisons, we use the widespread 68 facial landmark points from [Sagonas et al., 2013, Zafeiriou et al., 2017b, Shen et al., 2015] and compare our model against other state of the art methods by either gathering their reported results or reproducing them from their publicly available code.

4.4.1 Facial Alignment Experiments

In this section, we compare our internal Denoised Facial Alignment (DFA) presented in section 4.3.3 for single image facial alignment. In addition, we also report results from our FLE component operating without IID to analyze the impact of internal denoising in our final landmark estimates.

Experiment setup

We compare the performance of our DFA model against 8 alternative state of the art models using 300-W [Sagonas et al., 2013] and Menpo Challenge [Zafeiriou et al., 2017b] test datasets. We follow the standard procedure as in [Zafeiriou et al., 2017b] by first initializing each models with corresponding bounding box obtained from the ground-truth point for each images. Subsequently we calculate the Normalized Mean Square Error (NMSE), defined as the average error for all landmarks divided by the bounding box size [Bulat and Tzimiropoulos, 2017, Yang et al., 2015]. Finally, we also include the Area Under the Curve (AUC) and Failure Rate

(FR) using the standard NMSE threshold of 0.08 [Bulat and Tzimiropoulos, 2017] as additional metric.

We compare our results against both traditional and deep learning models, including also the results from our previous FADeNN model [Aspandi et al., 2019c], to highlight the relative improvement gain provided by DFA. Specifically, we compare the following models:

1. **ECT**: Estimation-Correction-Tuning [Zhang et al., 2018a].
2. **SAN**: Style Aggregated Neural Network [Dong et al., 2018].
3. **FAN**: Heatmap based Facial Alignment Network [Bulat and Tzimiropoulos, 2017].
4. **CFSS**: Coarse to Fine Shape Searching [Zhu et al., 2015a].
5. **TCDNN**: Task Constrained Convolution Deconvolution Network [Zhang et al., 2016b].
6. **ERT**: Ensemble Regression Trees [Kazemi and Sullivan, 2014].
7. **OpenFace**: Widely used facial alignment model of OpenFace [Baltrusaitis et al., 2018] which internally uses multi patch experts [Zadeh et al., 2017].
8. **FADeNN**: Our previous models of Facial Alignment with Internal Denoising Auto-Encoder [Aspandi et al., 2019c].

Results on 300-W and Menpo datasets

Table 4.1 shows the results on both 300-W and Menpo test datasets, while the corresponding AUC graph is shown on Figure 4.5. Based on these results we can see the improvement obtained by incorporating the internal denoiser (IID) in our facial alignment model. Firstly, DFA obtains higher performance than FLE (namely, the same model without denoising) in all the considered metrics. Secondly, DFA produces state of the art results for both datasets, consistently achieving the highest AUC and the lowest FR and

NMSE among all compared methods. Finally, We also observe a noticeable improvement against our previous FaDeNN [Aspandi et al., 2019c] model, which highlights the effectiveness of the newly proposed approach.

We also observe that generally, models based on deep learning perform particularly well, with SAN [Dong et al., 2018] and ECT [Zhang et al., 2018a] achieving the second and third best performance in terms of AUC, followed by FAN [Bulat and Tzimiropoulos, 2017], which seems to produce more stable estimates judging from its low FR values. Overall, these models perform better than those using hand-crafted features, such as ERT [Kazemi and Sullivan, 2014] and CFSS [Zhu et al., 2015a], which reflects the maturity of deep learning approaches and their advantage compared to the traditional approaches.

Method	300-W-Test			Menpo-Test		
	NMSE	AUC	FR	NMSE	AUC	FR
ECT	0.027	0.781	0.008	0.022	0.802	0.011
SAN	0.020	0.804	0.005	0.021	0.809	0.015
FAN	0.027	0.712	**0.000**	0.026	0.739	0.001
TCDNN	0.030	0.678	0.012	0.029	0.710	0.016
CFSS	0.022	0.779	0.001	0.029	0.795	0.019
ERT	0.051	0.532	0.196	0.051	0.578	0.200
OpenFace	0.049	0.555	0.178	0.058	0.524	0.249
FADeNN	0.021	0.798	**0.000**	0.017	0.844	**0.000**
FLE (Ours)	0.016	0.864	0.002	**0.013**	0.889	**0.000**
DFA (Ours)	**0.014**	**0.883**	**0.000**	**0.012**	**0.908**	**0.000**

Table 4.1: MSE, AUC and FR values for single image facial alignment on 300-W and Menpo test datasets.

Visual Results on 300-W and Menpo

Figure 4.6 visualizes examples of estimated facial landmarks for our model without (FLE) and with (DFA) denoising, against the other three

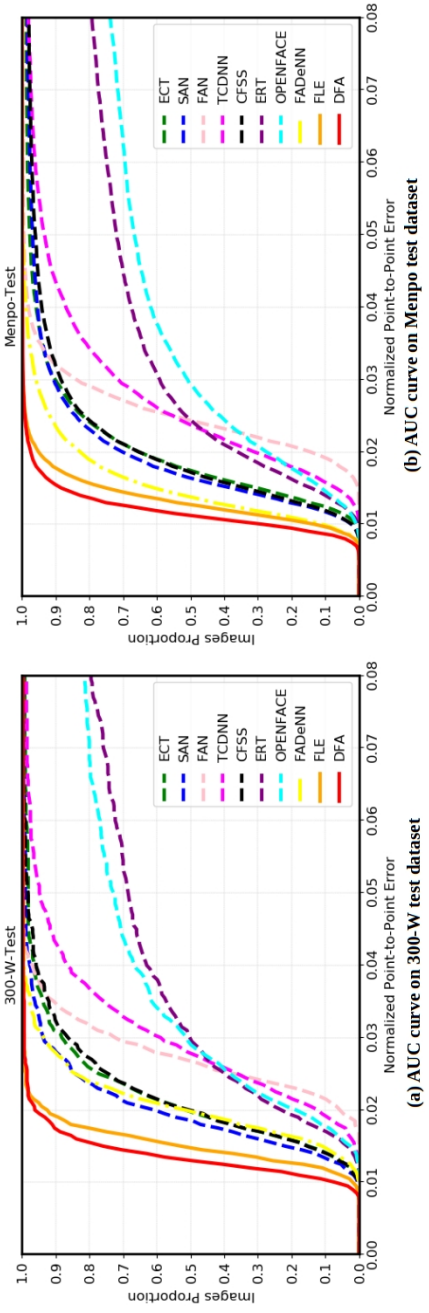


Figure 4.5: AUC Graph for the experiments on single image facial alignment: 300-W (left), and Menpo (right) test datasets.

best performing approaches in Table 4.1: SAN [Dong et al., 2018], ECT [Zhang et al., 2018a], and FAN [Bulat and Tzimiropoulos, 2017]. For each dataset, the different columns show examples of input images under different imaging conditions. The first column of each dataset shows what could be considered an optimal (clean) image (columns 1 and 6). The second and third columns show blurring artifacts, and the fourth and fifth columns show sub-optimal illumination.

By visually investigating these examples, we first notice that in case of relatively cleaned image input, our models produce more accurate predictions than the compared alternatives. This might be due to the inclusion of the denoising block during training, which removes part of the artifacts present in the training data and thus enhances the quality of the training set. On other hand, in case of degraded images, i.e with noisy input is introduced, we can see that the internal denoiser enhances the image, thus improving the quality of the final landmark estimates as shown on both of blurry and low illumination examples. This finding of the correlation between cleaned input image and accurate landmark estimates, i.e cleaner the images input will yield more accurate landmark predictions, also conform with a recent related study [Chrysos et al., 2019] suggesting the benefit of our denoising approach. Finally, we also notice that the internally enhanced images are visually better than the original ones, which further shows the effectiveness of each of our internal denoiser modules for removing their specific noise class.

4.4.2 Facial Tracking Experiments

In this section, we compare our full DCRT model described in section 4.3 for deformable facial tracking in the wild. We empirically set the threshold values of $T_B = 1.0$ and $T_C = 0.5$ for our model to perform tracking. As will be shown later, our model produces the best results when the sequence length is fixed to 2 frames and the internal denoiser module is activated (see Section 4.4.2).

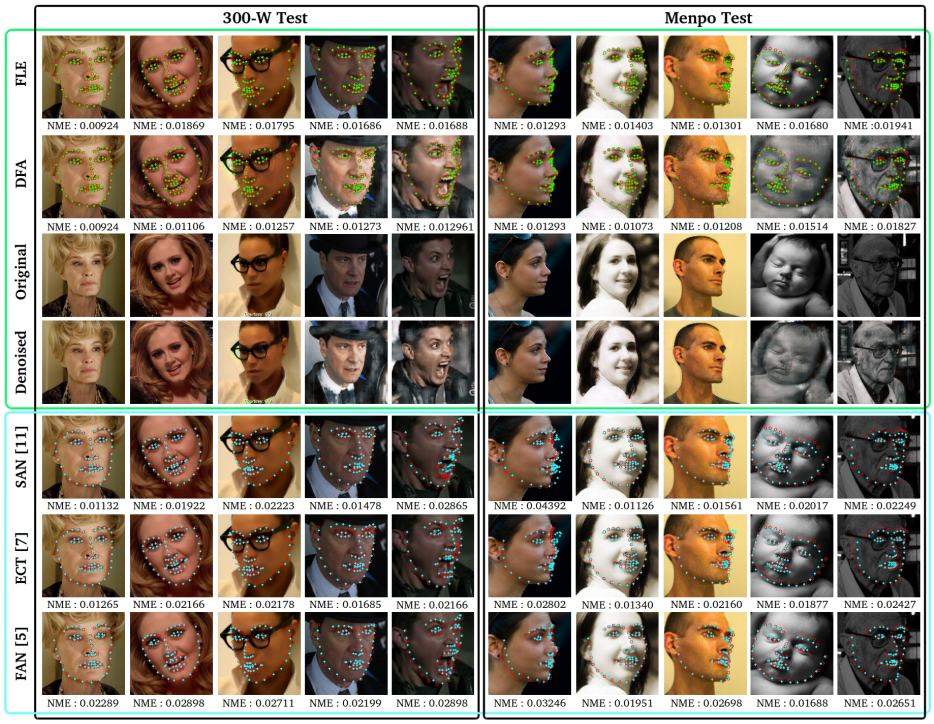


Figure 4.6: Visual examples of estimated landmarks on the 300-W Test dataset (left) and Menpo Challenge (right). Each column on each dataset provides examples of input images under different conditions: relatively clean images (column 1), blurry input (columns 2 and 3) and low illumination (columns 4 and 5). The first two rows are the results of our models of FLE and DFA respectively, followed with the original and internally enhanced images on the third and four rows. Finally, rows fifth to seven show the results from SAN [Dong et al., 2018], ECT [Zhang et al., 2018a] and FAN [Bulat and Tzimiropoulos, 2017], respectively.

Experiment setup

We compared our model against other 13 available state of the art facial tracking models, including our previous results using the CRCT model [Aspandi et al., 2019b] and FADeNN [Aspandi et al., 2019c]. We also provide results form the SAN model [Dong et al., 2018], which obtained the second-best performance in the facial alignment experiments from the previous section, by converting it to a facial tracker. We do so by adopting: 1) a tracking by detection approach, which we refer to as (SAN_DT); and 2) by using the result from the previous frame as initialization for the current one (SAN_PREV). This is interesting to asses the ability of current facial alignment models when converted to minimum-effort trackers (i.e. using none or very little temporal information). The list of compared trackers is as follows:

1. **Yang**: The original winner of 300-VW deformable facial tracking challenge from Yang et al [Yang et al., 2015].
2. **MEEM_CFSS** and **MD_CFSS**: Two hybrid trackers that showed the best performance in the recent facial tracking review from Chrysos et al. [Chrysos et al., 2017].
3. **Gu**: The recent tracker from Gu et al. [Gu et al., 2017] based on Bayesian RNNs
4. **Uricar, Xiao, Raja** and **Wu**: Four other trackers from the original 300VW competition [Shen et al., 2015] consisting of models from Uricar et al. [Uricár et al., 2015], Xiao et al. [Xiao et al., 2015], Raja et al. [Rajamanoharan and Cootes, 2015], and Wu et al [Wu and Ji, 2015].
5. **OpenFace**: OpenFace tracker [Baltrusaitis et al., 2018] which operates with tracking by detection approach.
6. **SAN_DT** and **SAN_PREV**: The SAN facial image alignment model [Dong et al., 2018] converted into a tracker by using the MTCNN face detector (SAN_DT) or the previous frame result (SAN_PREV) as initialization.

7. **CRCT**: Our previous tracking model [Aspandi et al., 2019b], which in contrast to the currently proposed model, does not include internal denoising.

8. **FADeNN**: Our previous facial alignment model [Aspandi et al., 2019c] which uses internal denoising, converted into a tracker by using the result from the previous frame as initialization.

We performed our experiments using the 300-VW dataset [Shen et al., 2015], which is the largest available deformable facial landmark tracking dataset. It comprises 55 videos divided into three categories according to the difficulty levels:

1. The first category contains people recorded in well-lit conditions, which is intended to evaluate facial tracking with images acquired in nearly ideal conditions.

2. The second category consists of people recorded in unconstrained conditions (variable illumination, dark rooms, etc.) with arbitrary poses. This setting evaluates facial tracking models in real-world human-computer applications.

3. The Third category contains videos of people recorded in fully unconstrained conditions which include cases of ambient illumination differences, large occlusions, expressions, etc. This scenario aims to asses the models under arbitrary recording conditions.

We use the original 2D facial landmarks directly as ground-truth and first-frame initialization for all models. We follow [Shen et al., 2015, Chrysos et al., 2017] to evaluate each tracked landmark with Area Under the Curve (AUC) and Failure Rate (FR) for Normalized Mean Error (NME), as previously explained in Section 4.4.1.

Results on 300-VW test dataset

Table 4.2 shows the results of all compared models for each category, while the respective AUC curves for several trackers are shown in Fig-

ure 4.7. We can observe that our model achieves top state of the art performance, consistently outperforming all compared alternatives. We also notice a large improvement with respect to our previous models, namely CRCT [Aspandi et al., 2019b] and FADeNN [Aspandi et al., 2019c], which were partial implementations of the short/long-range tracking and denoising, respectively; this highlights the need for a unified approach as the one presented here.

Regarding the other compared methods, we found the results of the OpenFace library (one of the most popular tools for nowadays) among the less accurate ones, which contrasts with the relatively good results obtained by this library in the single image alignment experiments. This occurs because, as stated previously, OpenFace is not actually designed for tracking and is indicative of the accuracy loss that can be experienced by not taking into account temporal dependencies. This is also supported by the results obtained by SAN_PV, which showed comparable or better performance than several other trackers and was clearly superior than its detection counterpart (SAN_DT).

Visual Results on 300-VW

Figure 4.8 showcases results of our DCRT model against the state of the art Yang et al. [Yan et al.,] and the widely used OpenFace [Baltrusaitis et al., 2018]. We can point out several observations based on these visual examples. Firstly, our model performs better on the majority of clean inputs, as shown in the first two rows, consistently producing accurate estimations even in cases of extreme and profile poses.

Secondly, we see that our model is also robust against severe image degradation, e.g. blur, occlusion and illumination. This is especially clear when processing the extremely dark frames from the last two rows, where the faces sometimes are hardly visible to a human observer. In these examples, our model still manages to produce consistent landmark estimates thanks to the denoising block, while other models struggle. When we visually inspect such difficult examples, we find that the intermediate images generated by the internal denoiser are considerably changed with

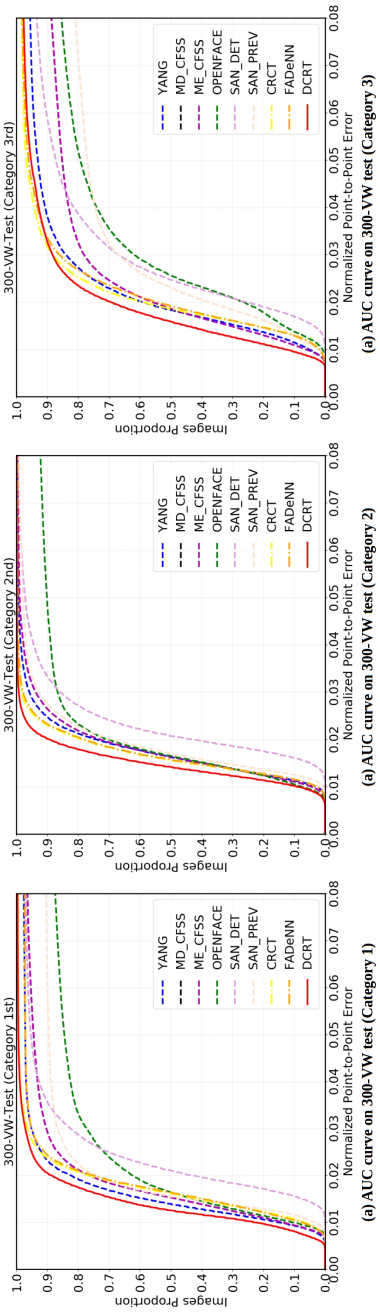


Figure 4.7: AUC Graph for the experiments on 300-VW Test dataset: a) results from the images in the first category, b) results from the second category and c) results from the third category.

Methods	Category 1		Category 2		Category 3	
	AUC	FR	AUC	FR	AUC	FR
Yang	0.791	2.4	0.788	0.32	0.710	4.46
MD_CFSS	0.784	1.8	0.783	0.34	0.713	7.47
ME_CFSS	0.758	3.56	0.772	0.38	0.659	11.3
Gu	0.718	1.2	0.703	0.2	0.617	4.83
Uricar	0.657	7.62	0.677	4.13	0.574	7.96
Xiao	0.760	5.9	0.782	3.84	0.695	7.38
Raja	0.735	6.56	0.717	3.91	0.659	8.29
Wu	0.674	13.9	0.732	5.6	0.602	13.1
OpenFace	0.668	12.58	0.715	7.78	0.566	14.62
SAN_DT	0.690	3.0	0.705	0.76	0.611	6.57
SAN_PV	0.713	9.61	0.756	0.83	0.571	19.06
CRCT	0.784	0.5	0.790	0.05	0.729	**1.75**
FADeNN	0.768	2.73	0.787	0.26	0.729	1.9
DCRT (Ours)	**0.822**	**0.4**	**0.813**	**0.03**	**0.752**	2.32

Table 4.2: AUC and FR values for deformable facial tracking using 300-VW test dataset, split by category.

respect to the original ones. This is especially noticeable in cases of poor illumination (e.g. last 2 rows of Fig. 4.8), where the denoised images not always look visually "cleaner" from a perception point of view, but we can still see that the illumination has been considerably enhanced, especially in the facial area. Thirdly in some cases, we also notice that our tracker produced facial landmarks that are visually more convincing than the actual ground-truth locations, as can be seen on frames 346, 356, 358 and 706 of the first row. These occurrences, though small, occur due to the need to use semi-automatic annotation methods to generate ground truths for very large datasets [Bulat and Tzimiropoulos, 2017].

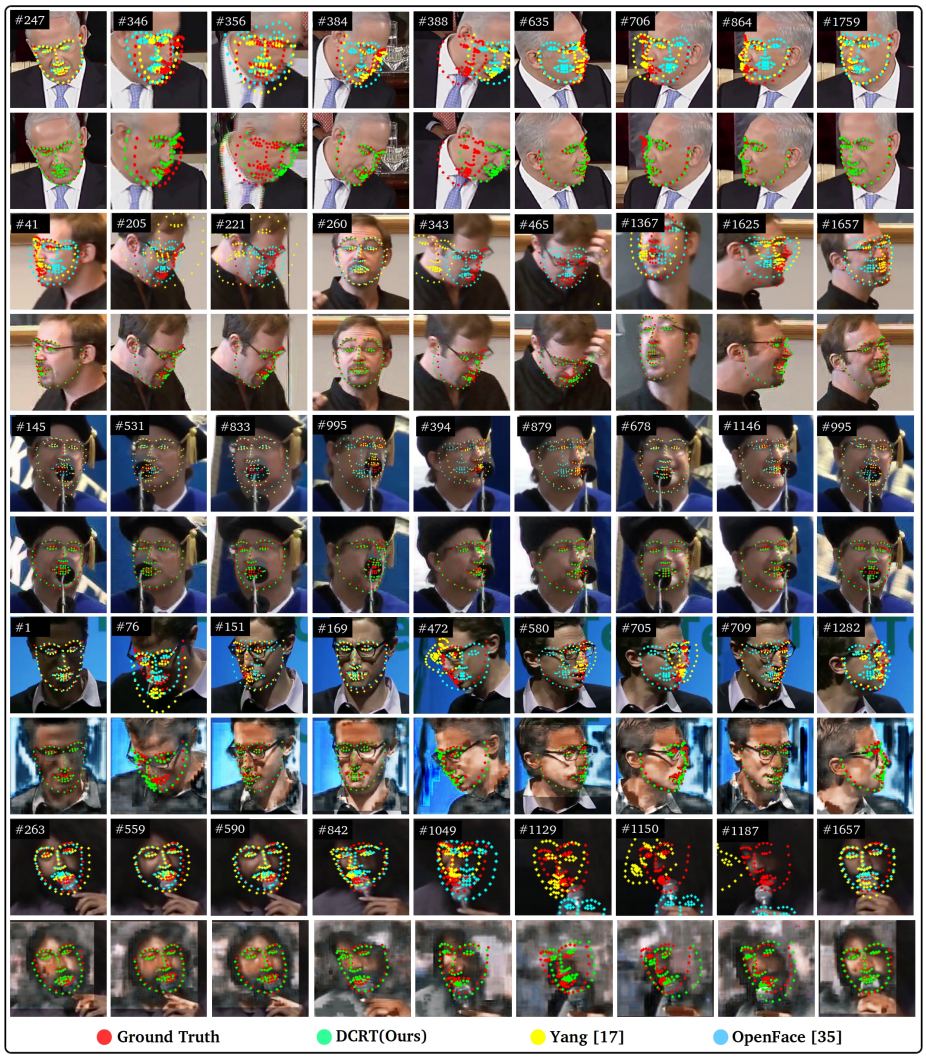


Figure 4.8: Visual examples of tracked facial landmarks on the 300-VW dataset. The odd rows show the results from Yang and OpenFace overlaid on the original image against the ground truth. Even row shows our results, displayed on the internally denoised images. From top to bottom, we show examples of clean input images (rows 1 and 2), blurred inputs (rows 3 to 6) and low ambient illumination (rows 7 to 10).

Ablations Study

In this section, we perform a detailed analysis of the behaviour of our tracker under different operation conditions. Firstly, we analyze the impact of the temporal context in our full tracker, DCRT, by varying length of the training sequences i.e SeqBT = 2, 4, 8, 16 which we refer to as DCRT-2, DRCT-4, ..., DCRT-32. Next, we also report the results obtained by a partial implementation of our tracker that does not include denoising, namely CRT-2, CRT-2, ..., CRT-32. Finally, we also report results under minimum-effort tracking:

1. Tracking by detection without denoising (FLE-Det)

2. Tracking by detection with denoising (DFA-Det)

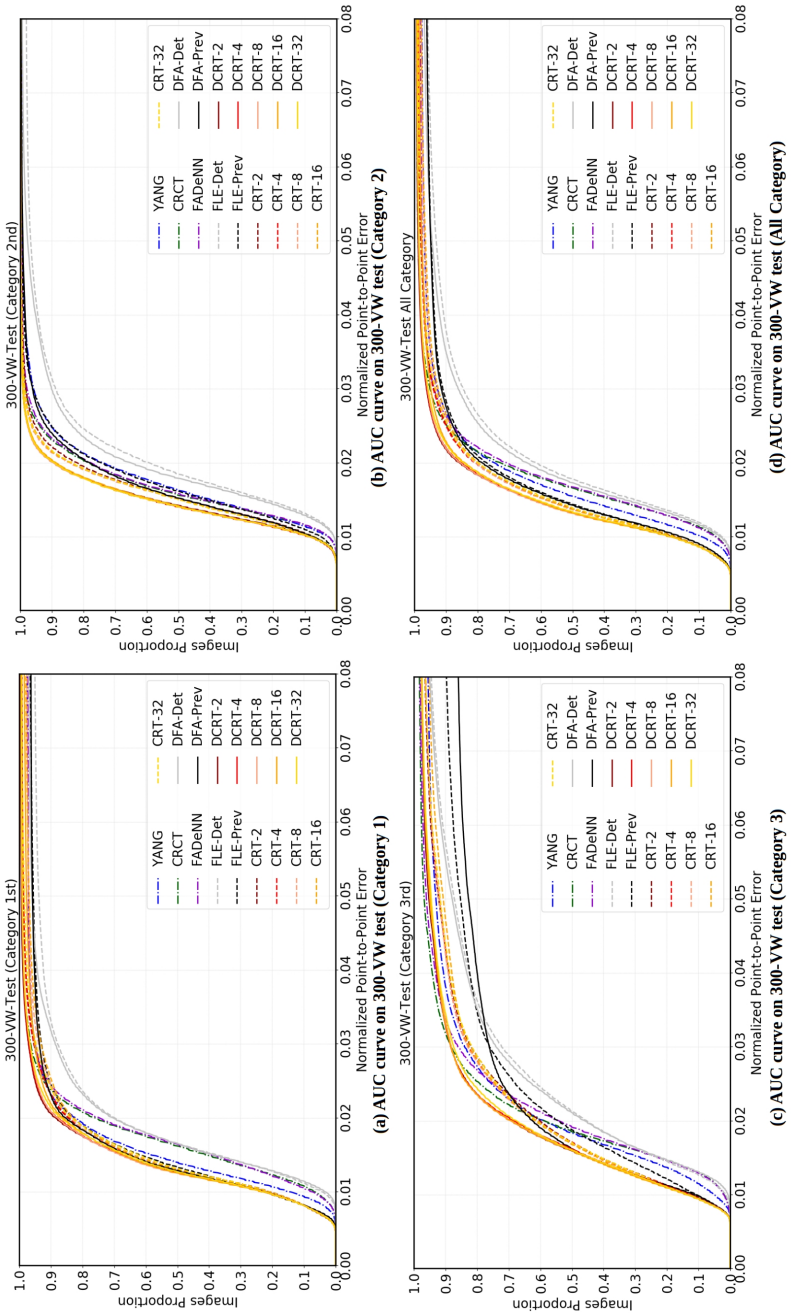
3. Initialization from the previous frame, without denoising (FLE-Prev)

4. Initialization from the previous frame, with denoising (DFA-Prev)

Table 4.3 shows the results of all the above variants of our tracker separately for each category of the test dataset, and jointly considering the images in all categories to sumarize the overall performance of each model. The respective AUC curve are displayed in Figure4.9. We also include the result of Yang etl. [Yang et al., 2015] as the baseline, as well as our previous results for both CRCT [Aspandi et al., 2019b] and FaDeNN [Aspandi et al., 2019c].

We can summarize our findings based on these results as follows:

- The minimum-effort tracking, when using initialization from the previous frame, yields quite comparable accuracy to the one provided by the baseline [Yang et al., 2015]. However, these results are still inferior against our models with internal temporal modeling (DCRT), in all cases with quite large margin. Further, all of our full models (DRCT-x, with internal temporal modeling), achieve better results than all tracking by detection approaches.



(a) AUC curve on 300-VW test (Category 1)
(b) AUC curve on 300-VW test (Category 2)
(c) AUC curve on 300-VW test (Category 3)
(d) AUC curve on 300-VW test (All Category)

Figure 4.9: AUC graph of our models on 300-VW using different settings: a) results on images from the first category, b) second category, c) third category, and d) overall results including the images in all categories.

Models	Category 1		Category 2		Category 3		All Category	
	AUC	FR	AUC	FR	AUC	FR	AUC	FR
YANG	0.7910	2.4	0.7880	0.32	0.7100	4.46	0.7690	2.31
CRCT	0.7835	0.5	0.7901	0.05	0.7293	**1.75**	0.7731	0.65
FADeNN	0.7682	2.73	0.7869	0.26	0.7290	1.9	0.7621	1.9
FLE-Det	0.7332	4.92	0.7307	1.9	0.6616	5.5	0.7142	4.25
FLE-Prev	0.7872	3.45	0.7811	0.33	0.6616	10.41	0.7574	4.15
CRT-2	0.8103	0.56	0.7991	0.03	0.7185	3.50	0.7947	0.90
CRT-4	0.8091	0.79	0.8023	0.03	0.7200	3.67	0.7946	1.08
CRT-8	0.7983	2.39	0.8031	0.03	0.7217	3.62	0.7878	2.15
CRT-16	0.7956	2.05	0.8022	0.04	0.7171	4.77	0.7847	2.10
CRT-32	0.8052	0.80	0.8022	0.05	0.7174	3.66	0.7916	1.09
DFA-Det	0.7483	3.01	0.7459	1.66	0.6603	5.19	0.7279	3.14
DFA-Prev	0.7965	3.04	0.7971	0.09	0.6755	10.27	0.7695	3.91
DCRT-2	**0.8219**	**0.40**	0.8130	**0.03**	**0.7520**	2.32	**0.8099**	**0.62**
DCRT-4	0.8199	0.65	**0.8145**	0.04	0.7497	2.40	0.8085	0.80
DCRT-8	0.8204	0.54	**0.8145**	**0.03**	0.7511	2.36	0.8091	0.72
DCRT-16	0.8087	1.55	0.8129	0.05	0.7513	2.52	0.8009	1.43
DCRT-32	0.8173	0.60	0.8140	0.04	0.7466	2.44	0.8062	0.77

Table 4.3: AUC and FR on 300-VW test dataset for the ablation study, split by category and combined over the images in all categories.

- The difference between results of models trained on different sequence lengths is minimal. In the majority of cases, shorter lengths were slightly better than longer ones. Nevertheless, these results must be read in relation to the test sequences, which show quite irregular (not necessarily natural) facial movements [Aspandi et al., 2019b].

- We observe a general improvement when introducing the Internal Denoiser on all metrics. This correlates well with the results from previous experiments on single image alignment, demonstrating the effectiveness of the internal denoiser in improving the final landmark estimations.

- We also notice lower Failure Rates (FR) when the internal denoiser is incorporated suggesting that, apart from increasing accuracy, it also improves the stability of the tracker.

4.5 Conclusions

In this chapter, we present an end to end composite recurrent tracker with internal denoising that is capable of performing joint single facial alignment and deformable facial tracking. We incorporate multilayer LSTMs to model temporal dependencies with variable length and introduce an internal denoiser which selectively enhances the input images to improve the robustness of our overall model. We achieve this by combining 4 different sub-networks that specialize in each of the key tasks that are required, namely face detection, bounding-box tracking, facial region validation and facial alignment with internal denoising.

We tested out model against state of the art alternatives in the most popular datasets for facial alignment and tracking. We started by comparing our results on single image facial alignment to other 8 facial alignment models in the 300-W and Menpo datasets. In this experiment, we found that our model consistently outperformed all compared alternatives, producing higher AUC and lower NME and FR values, respectively. Qualitative assessment highlighted the usefulness of our internal denoiser, which successfully enhanced the input images to produce intermediate representations in which facial details were easier to visualize, thus facilitating the task of the landmark localization algorithm.

Facial tracking experiments were performed using the 300-VW test dataset, in which we compared our model against 13 other facial trackers. We found that our unified approach outperformed all other compared models consistently over all categories of the test set. Further, our models showed to be considerably more robust than the compared alternatives in cases of input image with severe degradation, again due to our internal denoiser. We also noticed that, overall, tracking-by-detection approaches produced comparatively lower accuracy than models that utilized tempo-

ral modelling, which supports the advantage of incorporating temporal dependencies.

Finally, to investigate the impact of the different building blocks in our composite model, we presented a set of ablation experiments using 300-VW, in which one or more of the building blocks were removed, or some of its parameters were modified. These tests allowed to quantitatively confirm: 1) that inclusion of an internal denoising block outperformed the results without internal denoiser, even in images from category 1 that are expected to show nearly ideal conditions; 2) that the denoising block also contributed to the stability of the tracker, by reducing the failure rates, 3) that inclusiong of temporal modelling produced more accurate landmark estimates than either tracking-by-detection or tracking by initialization with the results from the previous frame. We also found that the improvements introduced by the temporal modeling were optimized for a temporal context between $SeqBT = 2$ and $SeqBT = 8$ frames, although differences were usually small between models as long as $SeqBT \geq 2$. Nevertheless, the latter must be read in relation to the test sequences that were used, which show quite irregular (not necessarily natural) facial movements.

Chapter 5

LATENT-BASED ADVERSARIAL NEURAL NETWORKS FOR FACIAL AFFECT ESTIMATIONS

Adapted from D. Aspandi, A. Mullol-Ragorta, B. Schuller, and X. Binefa, ”Latent-Based Adversarial Neural Networks for Facial Affect Estimations” in *In Affective Behavior Analysis in-the-wild (ABAW) workshop in conjunction with International Conference on Facial and Gesture (FG) 2020*, pp. 348-352, 16-20 November 2020, Buenos Aries, Argentina.

Abstract

There is a growing interest in affective computing research nowadays given its crucial role in bridging humans with computers. This progress has recently been accelerated due to the emergence of bigger dataset. One recent advance in this field is the use of adversarial learning to improve model learning through augmented samples. However, the use of latent features, which is feasible through adversarial learning, is not largely explored, yet. This technique may also improve the performance of affective models, as analogously demonstrated in related fields, such as computer vision. To expand this analysis, in this work, we explore the use of latent features through our proposed adversarial-based networks for valence and arousal recognition in the wild. Specifically, our models operate by aggregating several modalities to our discriminator, which is further conditioned to the extracted latent features by the generator. Our experiments on the recently released SEWA dataset suggest the progressive improvements of our results. Finally, we show our competitive results on the Affective Behavior Analysis in-the-Wild (ABAW) challenge dataset.

5.1 Introduction

Affective computing has recently attracted the attention of the research community, due to its applications in multiple and diverse areas, including education [Duo and Song, 2010] or healthcare [Liu et al., 2008], among others. Furthermore, the growing availability of affect-related datasets, such as SEWA [Kossaifi et al., 2019] and the recently introduced Aff-Wild2 [Kollias and Zafeiriou, 2019], enable the rapid development of deep learning-based techniques, which currently hold the state of the art [Kossaifi et al., 2017, Kossaifi et al., 2019, Kollias et al., 2019].

In computer vision tasks, such as natural image generation [Radford et al., 2015] and image classification [Odena, 2016], adversarial learning techniques from the family of generative models have been extensively investigated [Radford et al., 2015], [Odena, 2016], [Choi et al., 2018]. This learning technique enables rapid progress, not only to create additional data, but also to improve the performance of predictive models. Nevertheless, in the context of affective computing-related applications, this technique is still young and confined to its usage for data augmentation purposes [Han et al., 2019].

To expand the investigation of generative models in the field of affective computing, we investigate the use of latent features that are extracted in adversarial manners to improve the predictive capabilities of our model estimations. Specifically, we extract the visual latent features of the generator, which are then used to condition the discriminator on its estimations. Furthermore, we also aggregate the audio modality during training. We later show in our experiments on the SEWA [Kossaifi et al., 2019] and Aff-Wild2 [Kollias and Zafeiriou, 2019] datasets the benefits of our proposed approach with our competitive results. Specifically, the contributions of this work are:

1. We are the first to introduce the utilisation of latent features arranged in an adversarial way to improve affect-related model estimates.
2. We show the progressive improvements on our proposed works on the SEWA and Aff-Wild2 datasets and achieve competitive results on both datasets.

5.2 Related Work

Early approaches on automatic affect estimations involved the use of classical machine learning techniques with some degree of success. Several techniques explored include linear and partial least square regression [Povolny et al., 2016], and support vector machines [Nicolaou et al., 2011]. Furthermore, given the number of available modalities (e. g. video, audio, and bio-signals), several fusion techniques were also introduced to improve affect-related estimates. Different examples of these methods include early, late, model, and output-associative fusion [Zeng et al., 2009]. Diverse affective information has been progressed, starting from Action Units detection, emotion detection, to more recently continuous valence and arousal estimation [Kossaifi et al., 2017, Kossaifi et al., 2019].

Current progress relates to the emergence of big data that creates the opportunity to introduce large scale datasets in many fields, including affective computing. Examples of these datasets are SEMAINE [McKeown et al., 2010], AVEC [Ringeval et al., 2019], AFEW [Kossaifi et al., 2017], RECOLA [Ringeval et al., 2013], SEWA [Kossaifi et al., 2019], and the recently introduced Aff-Wild2 dataset [Kollias and Zafeiriou, 2019]. These datasets enable the development of powerful deep learning models that improve the accuracy of current state of the art [Kollias and Zafeiriou, 2018],[Kollias et al., 2019]. The investigations of deep learning-based techniques onto affective computing include the introduction of Convolutional Neural Networks (CNN) [Cardinal et al., 2015], incorporation of Recurrent Neural Network (RNN) [Kossaifi et al., 2019], and recently the fusion with Tensor-based methods [Mitenkova et al., 2019].

Adversarial learning [Radford et al., 2015] as a generative approach has been intensively studied in other machine learning research, especially in computer vision [Choi et al., 2018]. Given its potential, this method has also been explored in the field of affective computing, usually to augment the training data available for training [Han et al., 2019]. However, there is another aspect of generative models that is largely unexplored in this

field, which is the use of latent features to improve the models estimations, as shown in previous works from other fields, such as computer vision [Trumble et al., 2018], machine learning [Garciarena et al., 2018], and bio-signal analysis [Comas et al., 2020]. This inspired us to investigate the use of adversarial learning to improve our proposed models' performance through extracted latent features.

5.3 Latent-Based Adversarial Networks

We build our model based on the Star-GAN network[Choi et al., 2018], with architectural modifications to allow the extraction of latent features and use of the audio features. Figure 5.1 shows the overview of our proposed network. Our model operates by aggregating two main modalities: facial and audio features. There are two main sub-networks involved in our overall networks as already outlined above: the Auto-Encoder-based Generator(AEG), and the Conditional Discriminator-based affect estimator (CD) [Kumar et al., 2017]. The main role of the AEG is to produce cleaned images from noisy images to fool the discriminator, while simultaneously extracting robust latent features. On the other hand, the CD tries to recognise the fake images created by the AEG, and, at the same time, estimates the actual valence and arousal values. We train the AEG and CD in an adversarial way as below:

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E}_x[\log CD_{adv}(x)] + \\ & \mathbb{E}_x[\log(1 - CD_{adv}(AEG(\hat{x})))], \end{aligned} \tag{5.1}$$

where x corresponds to the noisy image and $\hat{x}$ is the cleaned input image approximated by the AEG. We use similar noise introduction methods as in [Aspandi et al., 2019c], which consist of four different types of artifacts: Gaussian blurring, Gaussian noise, image downsampling, and colour scaling.

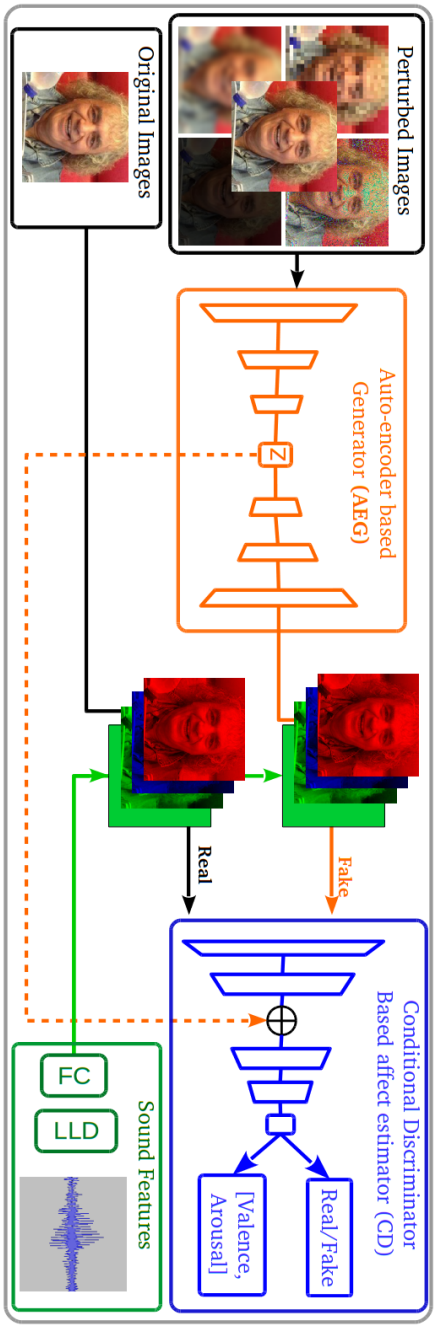


Figure 5.1: Complete architecture of our proposed models which incorporate two main networks: first is an Auto-Encoder-based Generator (AEG) which denoises the image and creates robust latent features. Second is a Conditional Discriminator-based affect estimator (CD) that aggregates both sounds and image input which is conditioned by latent features from the CD to estimate both real/fake and valence/arousal values.

5.3.1 Auto-Encoder-based Generator

Given the noisy input image, x, the AEG will approximate the cleaned version of the input image, $\hat{x}$. This is done by utilising coupled mirrored convolutions and deconvolutions with intermediate 2D bottleneck latent kernels; i. e. without skip connections. This scheme enforces the AEG to create latent robust features in order to effectively clean the input image. To improve the denoising and reconstruction process, we use the cycle loss [Choi et al., 2018, Kim et al., 2017] defined below :

$$\mathcal{L}_{rec} = \mathbb{E}_x[||x - AEG(AEG(\hat{x}))||]. \tag{5.2}$$

5.3.2 Conditional Discriminator-Based Affect Estimator

The CD employs both facial and audio features to identify the real/fake status of the current input and the corresponding valence and arousal values of $\hat{\theta}$. The facial features correspond to the cleaned image (denoised or reconstructed from the generator and the corresponding latent features of z. From the audio modality, we use the low-level descriptors (LLDs) of the eGeMAPS feature set [Eyben et al., 2016] (cf. Section 5.3.4). Both latent and audio features are combined through late fusion [Gunes and Piccardi, 2005, Snoek et al., 2005, Zeng et al., 2009] alongside the main RGB input images. Specifically, the audio features are merged by feeding them into a 1D fully connected layer to enlarge its dimension and converting it to a single 2D kernel, which is then concatenated with the denoised image. The latent features are combined in middle pipelines of the CD by concatenating them with intermediate kernels.

To detect both real and fake status and estimate valence and arousal values, we add another classifier [Odena, 2016] on top of the main classifier, which consists of a 2x2 pixels layer [Isola et al., 2017]. In the adversarial training, the CD will be optimised using real (r) and fake (f) images to minimise the affect loss ($\mathcal{L}_{afc}$) that judges the accuracy of the estimated valence and arousal values (cf. Equation 5.5). The corresponding loss of training the CD for both real ($\mathcal{L}_{va}^{r}$) and false examples ($\mathcal{L}_{va}^{f}$) can be seen

below :

$$\mathcal{L}_{va}^{r} = \mathbb{E}_{x,\theta}[-\mathcal{L}_{afc}(\theta'|x)], \tag{5.3}$$

$$\mathcal{L}_{va}^{f} = \mathbb{E}_{x,\theta}[-\mathcal{L}_{afc}(D(\theta|G(x)))], \tag{5.4}$$

where $\hat{\theta}$ is the ground truth valence/arousal value, and the affect loss, $\mathcal{L}_{afc}$, corresponds to the amalgamations of multiple affect metrics: Mean Square Error(MSE) (Eq. 5.6), Correlation(COR) (Eq. 5.7), and Concordance Correlation Coefficients (CCC) (Eq. 5.8) [Kossaifi et al., 2017], [Kollias and Zafeiriou, 2019] :

$$\mathcal{L}_{afc} = \sum_{i=1}^{N} \frac{n_i}{N}(\mathcal{L}_{MSE} + \mathcal{L}_{COR} + \mathcal{L}_{CCC}) \tag{5.5}$$

$$\mathcal{L}_{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i, \theta_i)}, \tag{5.6}$$

$$\mathcal{L}_{COR} = \frac{\mathbb{E}[(\hat{\theta} - \mu_\theta) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}\sigma_\theta} \tag{5.7}$$

$$\mathcal{L}_{CCC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \mu_\theta) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}, \tag{5.8}$$

where n_i is the total number of instances of discrete valence/arousal class i, and N is the normalisation factor [Aspandi et al., 2019a] for the total valence/arousal class. This normalisation factor is crucial given considerably unbalanced class instance on the Aff-Wild2 dataset [Kollias et al., 2020].

5.3.3 Overall Objective

Finally, the overall objective functions to train both AEG and CD are expressed as follows:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{afc}\,\mathcal{L}_{afc}^{r}, \tag{5.9}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{afc}\,\mathcal{L}_{afc}^{f} + \lambda_{rec}\,\mathcal{L}_{rec}, \tag{5.10}$$

in which λ_{afc} and λ_{rec} are the regulariser parameters for affect estimations and reconstruction loss.

5.3.4 Audio Feature Extraction

One of the first challenges when combining audio and video signals is the difference in term of sampling rates between both modalities. To overcome this issue, we first generate audio frames from the original audio signal by selecting the portions of the audio signal corresponding to one frame of video. We then enlarge the audio frame with the samples corresponding to the previous and future video frame to ensure information overlap between consecutive audio frames. We finally extract the LLDs of the EGEMAPS [Eyben et al., 2016] feature set using OPENSMILE [Eyben et al., 2010], and concatenate the first two sets of LLDs for further analysis. These LLDs are extracted from windows of 0.060 seconds with a step size of 0.010 seconds. Selecting the first two sets of LLDs only, we ensure the same dimensionality of the audio features in spite of videos recorded at different sampling rates.

5.3.5 Model Training

To train our model, we use the respective training subset of each dataset. On the Sentiment Analysis in the Wild dataset (SEWA) [Kossaifi et al., 2019], we followed original person-independence protocols, and apply the feature extraction techniques described on previous sections. Moreover, we also use the external tracker of [Aspandi et al., 2019b] to refine the given bounding box. We also includes the experiments on Aff-Wild2 dataset as part of the Affective Behavior Analysis in-the wild (ABAW) 2020 Competition to provide more actual analysis of our models performance. In this dataset, we only utilise the training subset to obtain our validation results. We then use the full available data (training and validation) to train our final models to produce our test results. Specifically, we used the crop-aligned samples provided by the organisers as facial features, in addition to the audio signals from the

available videos.

For both datasets, we trained our model progressively to allow us to analyse the impact of each proposed step. We first train both of generator and discriminator together, and proceed by adding the extracted latent z features alongside the audio features. Our models were trained using an NVIDIA Titan X GPU and it took approximately two days to converge. The source code of our models is available at our github page[2]

5.4 Experiments

5.4.1 Datasets and Experiment Settings

In this section, we describe our results on SEWA [Kossaifi et al., 2019] and recently introduced Aff-Wild2 [Kollias et al., 2020] datasets to confirm the advantages of each of our proposed approaches.

- The SEWA dataset [Kossaifi et al., 2019] is a recently published affect dataset which consists of video and audio recording involving 398 subjects from multiple cultures. It is split into 538 sequences with various meta-data (e.g. subject id, culture etc) are available alongside the actual affect ground truth of valence/arousal and liking/disliking.

- The Aff-Wild2 challenge dataset is being published as part of the first ABAW 2020 competitions [Kollias et al., 2020] which consists of three main challenges : valence-arousal, basic expression and eight action units. Aff-wild2 is considered to be the current, largest affect in the wild dataset with more than 558 videos and 458 total number of subjects. Specific on the valence and arousal challenge, there are 545 annotated videos with 2.786.201 frames which is split into three subsets : 346 videos of training, 68 videos of validation and 131 videos of test.

[2]https://github.com/deckyal/ALN

In each experiment, we provide the results from the variant of our models to highlight the important of each approach. First is the method Disc which corresponds to our results utilizing only plain Discriminator (CD) trained using standard ℓ^2 loss. Second is method AEG-CD that constitute to our model which uses adversarial training for both of AEG and CD. Lastly, the AEG-CD-ZS shows the results of our previous model trained with the inclusion of both latent features z from AEG and the mapped audio features.

We use MSE, COR and CCC metrics to evaluate the quality of each affect estimations [Kossaifi et al., 2019, Ringeval et al., 2019]. That on the Aff-Wild2 dataset, we compared our results on the validation stage against the baseline provided by the organizers [Kollias et al., 2020]. While for the SEWA dataset, we report our results from original five cross validation settings [Kossaifi et al., 2019] and compared them with the respective baseline [Kossaifi et al., 2019] and recent state of the art of [Mitenkova et al., 2019]

5.4.2 Experiment Results

Table 5.1 provides our results on the SEWA dataset, where we can see that our models able to produce quite competitive results, with a quite high accuracy on the arousal dimension. Specifically, we observe a relatively high accuracy obtained by our discriminator (Disc) that is enough to outperform the current baseline, albeit still lower than the results of [Mitenkova et al., 2019]. Using the adversarial training further improve the accuracy which conforms the previous findings of the benefit in using the adversarial learning upon standard ℓ^2 loss [Choi et al., 2018, Odena, 2016, Gan et al., 2015]. Another potential explanation of this improvement can be attributed to the generated images that may reduce the available noises on the input images (cf. Section 5.4.3). Finally, incorporating both of the latent and audio features improves the overall accuracy of our results, surpassing the current state of the art on this dataset on the arousal dimension, highlighting the benefit of incorporating such features.

Methods	MSE		COR		CCC	
	Val	Aro	Val	Aro	Val	Aro
Baseline [Kossaifi et al., 2019]	-	-	0.322	0.4	0.195	0.427
Tensor [Mitenkova et al., 2019]	0.334	0.380	**0.503**	0.439	**0.469**	0.392
Disc	0.336	0.399	0.395	0.457	0.349	0.379
AEG-CD	0.329	0.394	0.429	0.467	0.380	0.429
AEG-CD-SZ	**0.323**	**0.350**	0.442	**0.478**	0.405	**0.430**

Table 5.1: Experiment results on SEWA dataset.

Methods	MSE		COR		CCC	
	Val	Aro	Val	Aro	Val	Aro
Baseline [Kollias et al., 2020]	-	-	-	-	**0.14** (0.11)	0.24 (**0.27**)
Disc	0.44	0.30	0.07	0.19	0.07	0.20
AEG-CD	0.42	0.28	0.10	0.22	0.08	0.22
AEG-CD-SZ	0.42	0.28	0.11	0.29	0.10 (**0.17**)	**0.26** (0.16)

Table 5.2: Experiment results on Aff-Wild2, ABAW Challenge dataset. The values in parentheses denote the testing results.

We also found similar findings in our results on the Aff-Wild2 dataset as shown in Table 5.2. In this dataset, we observe identical improvements toward our results on the validation stage, with the lowest accuracy produced by our Disc model and progressively increased to the best accuracy of AEG-CD-SZ, attaining superior accuracy on arousal domain compared to the baseline. In the test set however, we found that AEG-CD-SZ produces a quite balanced accuracy for both valence and arousal, with higher accuracy against the baseline on the valence domain. This may be a result from the incorporation of the validation split in our training, that further altered the distribution of valence and arousal instances.

5.4.3 Latent Feature and Visual Analysis

In this section, we further visualize the learned latent kernel features to explain the observed progressive improvement of our models. Figure 5.2

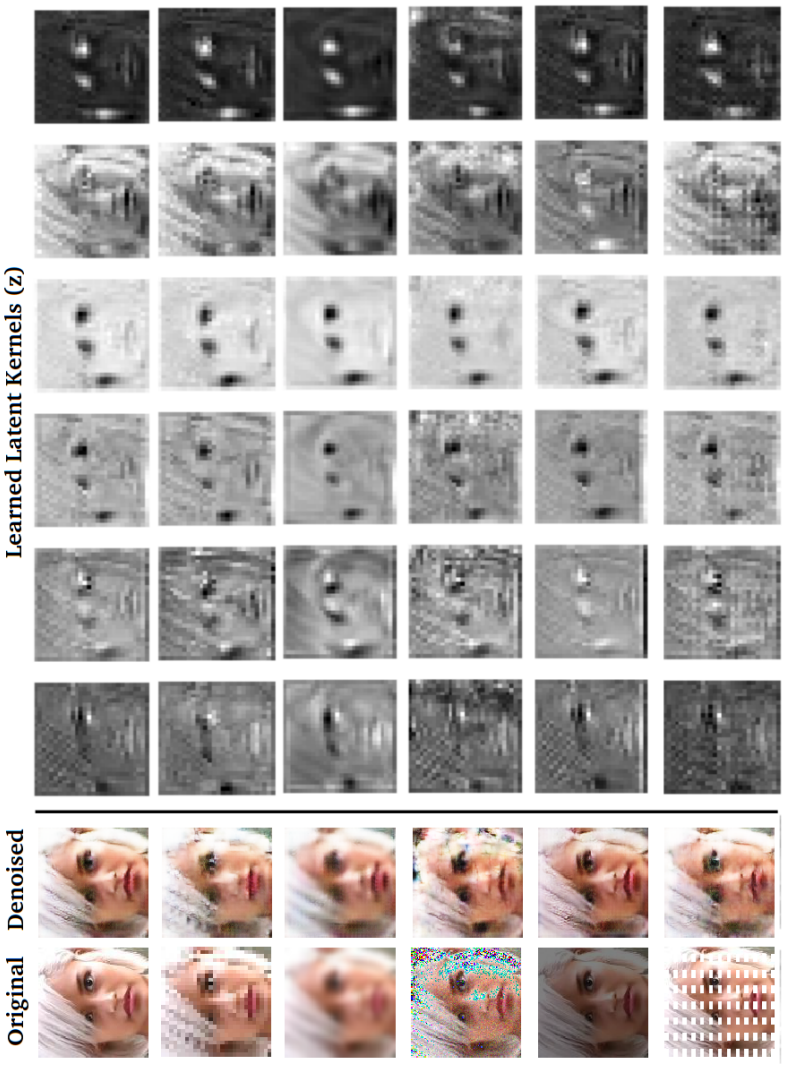


Figure 5.2: Example of denoised images and the corresponding latent kernels (selected randomly). As we can see, the denoised images are quite cleaned, and the latent kernels are also consistent across different input conditions.

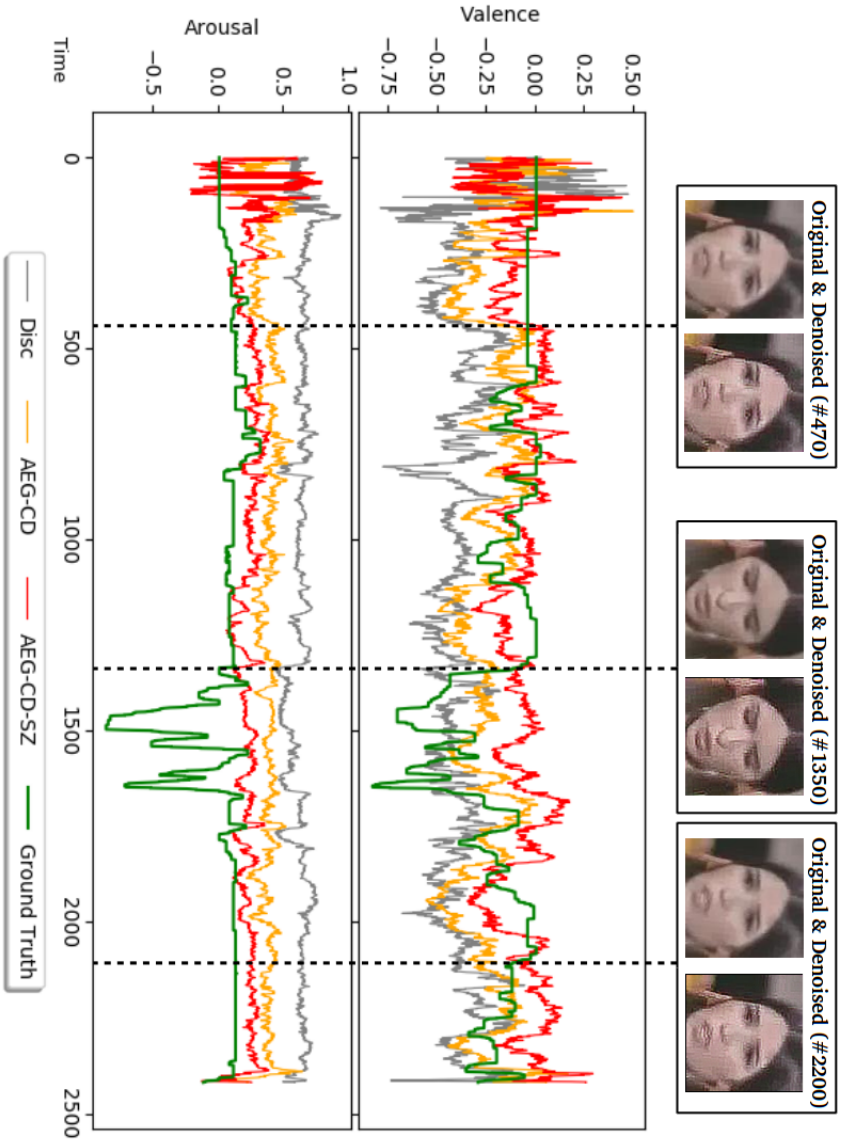


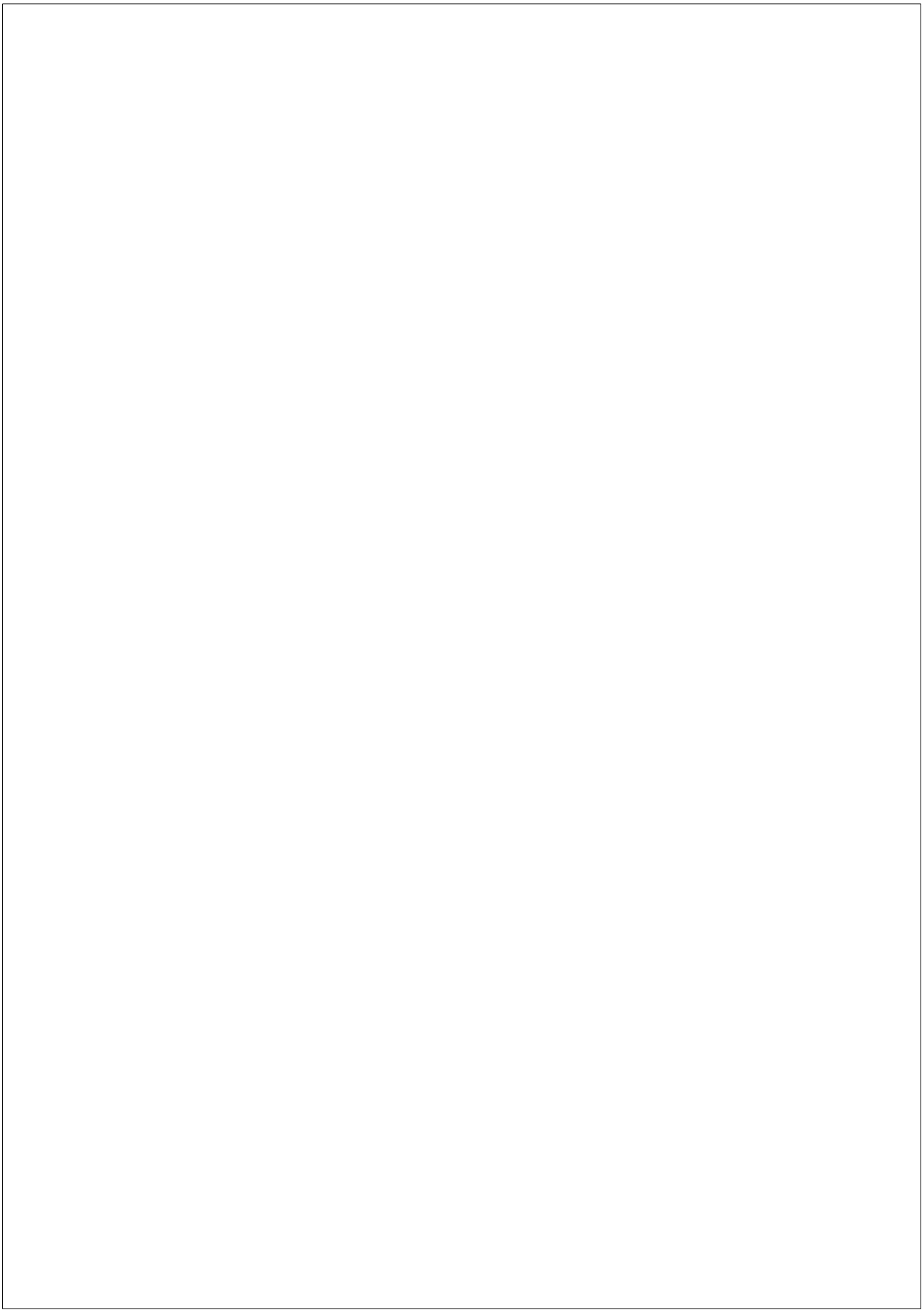
Figure 5.3: Visualization of the results from of our model variants. Notice that the results from AEG-CD-ZS, both in Valence and Arousal domain are the most closely resemble to the ground truth compared to the others. Furthermore, the denoised images appear to be clearer compared to the original input.

shows the examples of original input images and their denoised (cleaned) versions, followed by randomly selected six latent kernels. In regards to the denoising quality, we found that our model manages to clean the underlying noise on the input image considerably well. Furthermore, we notice the consistency of the learned latent features, i.e their spatial structures are not drastically altered, regardless the input conditions. These robust representations help the Discriminator in its inference as complementary features [Comas et al., 2020] resulting in overall higher accuracy, as shown on the previous section.

Furthermore in Figure 5.3, we can also see the continous results of our model variants. By observing this, we could concur that AEG-CD-SZ produces the most accurate predictions compared to the rest indicated by their resemblances to the ground truth. Also notice that, the denoised input images are also clearer and sharper compared to the original input, which demonstrates the real-life applicability of denoising functions of our model. Finally, these cleaned inputs, may also further aids the discriminator training in conjunction with learned latent features, hence improved its overall results.

5.5 Conclusions

In this chapter, we presented the first investigation using latent features extracted through adversarial learning in Affective Computing domain. Specifically, we performed progressive training on our generator to extract robust features given noisy inputs paired with a discriminator through adversarial learning. Then, we employed a conditional discriminator to aggregate several modality's inputs to achieve our affect estimations. We tested the performance of our models on two datasets: SEWA and Aff-wild2. In our experiments, we observed progressive improvements made by each of our approaches ultimately leading to competitive results on both datasets. In the future, we seek to incorporate temporal modelling to further increase the accuracy of the proposed models.



Chapter 6

AN ENHANCED ADVERSARIAL NETWORK WITH COMBINED LATENT FEATURES FOR SPATIO-TEMPORAL FACIAL AFFECT ESTIMATION IN THE WILD

Adapted from: D. Aspandi, F. Sukno, B. Schuller, and X. Binefa, ”An Enhanced Adversarial Network with Combined Latent Features for Spatio-Temporal Facial Affect Estimation in the Wild”, in *16th International Conference on Computer Vision Theory and Applications (VISAPP), (To Appear)*

Abstract

Affective Computing has recently attracted the attention of the research community, due to its numerous applications in diverse areas. The emergence of video-based data allows to enrich the widely used spatial features with the inclusion of temporal information. However, such spatio-temporal modeling often results in very high-dimensional feature spaces and large volumes of data, making training difficult and time consuming. This work addresses both the lack of incorporation and analysis of temporal modeling on affective estimates. We propose a novel model that efficiently extracts both spatial and temporal features of the data by means of its enhanced temporal modeling based on latent features. We do so by incorporating three major networks, coined Generator, Discriminator, and Combiner, which are trained in an adversarial setting combined with curriculum learning to enable our adaptive attention modules. In our experiments, we show the effectiveness of our approach by reporting our competitive results on both AFEW-VA and SEWA dataset, suggesting that temporal modeling improves the affect estimates both in qualitative and quantitative terms. Furthermore, we found that the inclusion of attention mechanisms lead to the highest accuracy improvements, as its weights seem to correlate well with the appearance of facial movements, both in terms of temporal localisation and intensity. Finally, we conclude that a sequence length of around 160 ms to be the optimum one for temporal modeling, which is consistent with other relevant findings utilising similar lengths.

6.1 Introduction

Affective Computing has recently attracted the attention of the research community, due to its numerous applications in diverse areas which include education [Duo and Song, 2010] or healthcare [Liu et al., 2008], among others. The growing availability of affect-related datasets, such as AFEW-VA [Kossaifi et al., 2017] and the recently introduced SEWA [Kossaifi et al., 2019] enable the rapid development of deep learning-based techniques, which currently hold the state of the art.

Further, the emergence of video-based data allows to enrich the widely used spatial features with the inclusion of temporal information. To this end, several authors have explored the use of long-short term memory (LSTM) recurrent neural networks (RNNs) [Tellamekala and Valstar, 2019, Ma et al., 2019], endowed also with attention mechanisms [Luong et al., 2015, Li et al., 2020, Xiaohua et al., 2019]. However, such spatio-temporal modelling often results in very high-dimensional feature spaces and large volumes of data, making training difficult and time consuming. Moreover, it has been shown that the sequence length considered during training can be a decisive factor for successful temporal modelling [Kossaifi et al., 2017, Xia et al., 2020, Gordon et al., 2018, Aspandi et al., 2019b], and yet a detailed study of this aspect is lacking in the field.

This chapter addresses both the lack of incorporation and analysis of temporal modelling on affective analysis. We propose a novel model which can be efficiently used to extract both spatial and temporal features of the data by means of its enhanced temporal modelling based on latent features. We do so by incorporating three major networks, coined Generator, Discriminator, and Combiner, which are trained in an adversarial setting to estimate the affect domains of Valence (V) and Arousal (A). Furthermore, we capitalise on these latent features to enable temporal modelling using LSTM RNNs, which we train progressively using curriculum learning enhanced with adaptive attention. Specifically, the contributions of this chapter are as follows:

1. We upgrade the standard adversarial setting, consisting of a Gen-

erator and a Discriminator, with a third network that combines the features from these networks, which are modified accordingly. This yields features that combine the latent space from the autoencoder-based Generator and a V-A Quadrant estimate produced by the modified Discriminator, resulting in a compact but meaningful representation that helps reduce the training complexity.

2. We propose the use of curriculum learning to enable analysis and optimisation of the temporal modelling length.

3. We incorporate dynamic attention to further enhance our model estimates and show its effectiveness by reporting state of the art accuracy on both AFEW-VA and SEWA datasets.

6.2 Related Work

Affective Computing started by exploiting the use of classical machine learning techniques to enable automatic affect estimation. Examples of early approaches include partial least squares regression [Povolny et al., 2016], and support vector machines [Nicolaou et al., 2011]. Subsequently, to further progress the investigations in this field, the development of larger and bigger datasets was addressed. Several datasets have been introduced so far, starting with SEMAINE [McKeown et al., 2010], AFEW-VA [Kossaifi et al., 2017], RECOLA [Ringeval et al., 2013], OMG [Barros et al., 2018], AffectNet [Mollahosseini et al., 2015] and more recently SEWA [Kossaifi et al., 2019], aff-wild [Kollias et al., 2019], [Zafeiriou et al., 2017a] and aff-wild2 [Kollias and Zafeiriou, 2019], [Kollias et al., 2020]. Furthermore, the V-A labels have become the standard emotional dimensions over time, as opposed to hard emotion labels, given their continuous nature [Kossaifi et al., 2017, Kossaifi et al., 2019].

Throughout the last few years, models based on Deep Learning have emerged and currently hold the state of the art in the context of affective analysis, given their ability to learn from large scale data. A recent example along this line is the work from Mitenkova et al. [Mitenkova et al., 2019],

who introduce tensor modelling for affect estimations by using spatial features. In their work, they use tucker tensor regression optimised by means of deep gradient methods, thus allowing to preserve the structure of the data and reduce the number of parameters. Other works, such as [Handrich et al., 2020], adopt the multi-task approach to simultaneously address face detection and affective states prediction. Specifically, they use YOLO-based CNN models [Huang et al., 2018] to estimate the facial locations alongside V-A values through their proposed combined losses. As such, their models are able to incorporate the characteristics of facial attributes and estimate their relevance to affect inferences.

The recent growth of video-based datasets has encouraged the inclusion of temporal modelling, which has shown to improve models training [Xie et al., 2016, Cootes et al., 1998]. Relevant examples in Affective Computing include the works of Tellamekala et al. [Tellamekala and Valstar, 2019] and Ma et al. [Ma et al., 2019]. In their work, Tellamekala et al. [Tellamekala and Valstar, 2019] enforce temporal coherency and smoothness aspects on their feature representation by constraining the differences between adjacent frames, while Ma et al. resort to the utilisation of LSTM RNNs with residual connections applied to multi-modal data. Furthermore, the use of attention has also been recently explored by Xiaohua et al. [Xiaohua et al., 2019] and Li et al. [Li et al., 2020]. Xiahoua et al. adopt multi-stage attention, which involves both spatial and temporal attention, on their facial based affect estimations. Meanwhile, using spectrogram data as input, Li et al. propose a deep network that utilises an attention mechanism [Luong et al., 2015] on top of their LSTM networks to predict the affective states.

Unfortunately, to our knowledge, all previous works involving temporal modelling on affective computing miss one important aspect of the analysis: the involved sequence length in their training. While the specified length of temporal modelling has been shown to affect the final results on other related facial analysis tasks [Kossaifi et al., 2017, Xia et al., 2020, Gordon et al., 2018, Aspandi et al., 2019b], the computational cost required to train large spatio-temporal models hampers one to address such analysis. However, these problems could be mitigated

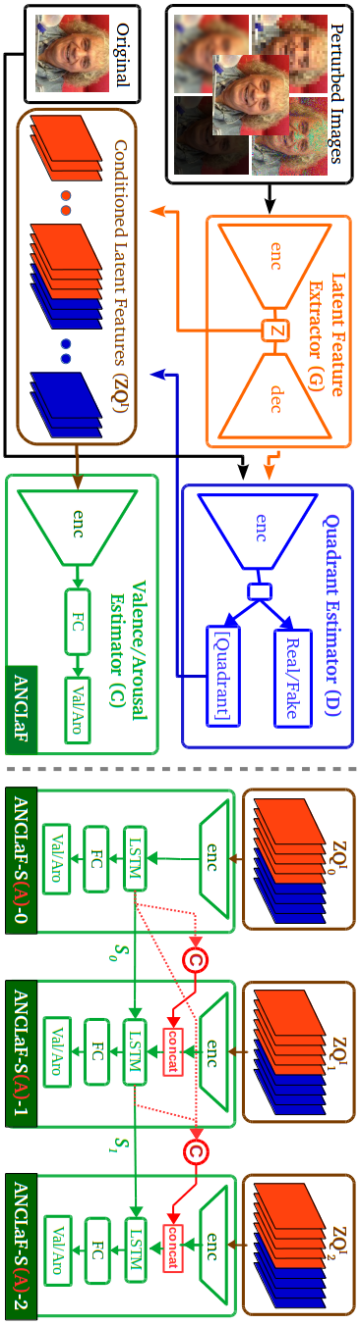


Figure 6.1: Schematic representation of our Full ANCLaF Networks. Left is our base model, which consists of three networks jointly trained in an adversarial setting: Latent Feature Extractor (G), Quadrant Estimator (D), and Valence Arousal Estimator (C). On the right, we see our network endowed with sequence modelling (ANCLaF-S) and attention mechanism (ANCLaF-SA).

by: 1) the use of progressive sequence learning to permit step-wise observations of various sequence lengths; this approach has been shown in the recent work of [Aspandi et al., 2019b] on facial landmark estimations, which uses curriculum learning enabling more robust training analysis and tuning of the temporal length; 2) the use of reduced feature sizes, enabling more efficient training process [Comas et al., 2020]; this has been explored in the affective computing field by the recent works such as [Aspandi et al., 2020a], which uses generative modelling to extract a latent space of representative features. These two aspects have inspired us to propose the combined models presented in this work, as explained in the next section.

6.3 Methodology

Figure 6.1 shows the overview of our proposed models, which consist of three networks: a Latent Feature Extractor (acting as Generator, G), a Quadrant Estimator (or Discriminator, D), and a Valence/Arousal Estimator (or Combiner, C). Given input image I which contains the facial area, both G and D will be responsible to learn low dimensional features that the combiner will use to estimate the associated Valence (V) and Arousal (A) state θ. The architecture of both the G and D networks follows the recent work from [Aspandi et al., 2020a], and we propose to use LSTM enhanced with attention to create our C network. We proposed two main architecture variants: the **ANCLaF** network (left), which uses single images as input and estimates V and A values independently for each frame, and **ANCLaF-S** and **ANCLaF-SA** (right) that uses sequences of latent features extracted from n frames as input, and utilises LSTMs for the inference (**-S**), optionally combined with internal attention layers (**-SA**).

6.3.1 Adversarial Network with Combined Latent Features (ANCLaF)

The pipeline of our base model **ANCLaF** starts with the G network. It receives either the original input image I, or a distorted version of it, $\tilde{I}$, as detailed in [Aspandi et al., 2019c, Aspandi et al., 2019a]. It simultaneously produces the cleaned reconstruction of the input image $\hat{I}$ and a 2D latent representation that will be used as features ($\mathbb{Z}$):

$$G(I)_{\Phi^G} = dec_{\Phi^G}(enc_{\Phi^G}(I)) \, with \, \mathbb{Z}^I \approx enc_{\Phi^G}(I), \tag{6.1}$$

where Φ are the parameters of the respective networks, *enc* and *dec* are the encoder and decoder, respectively. Subsequently, the D network receives $\hat{I}$ and tries to estimate whether it was obtained from a true or fake example (namely, original or distorted input image), as well as a rough estimate of the affective state. In contrast with the formulation in [Aspandi et al., 2020a], in which D targets directly the intensity of V and A, we propose to base the estimated affect on the circumplex quadrant ($\mathbb{Q}$) [Russell, 1980] which discretize emotions along valence and arousal dimensions (four quadrants). This in turn reduces the training complexity. Thus, letting FC stand for fully connected layer:

$$D(I)_{\Phi^D} = FC_{\Phi^D}(enc_{\Phi^D}(I)) = \mathbb{Q}^I \; and \; \{0,1\}. \tag{6.2}$$

Then, $\mathbb{Q}$ is used to condition the extracted latent features $\mathbb{Z}$ through layer-wise concatenation, which we call $\mathbb{ZQ}$ [Dai et al., 2017, Ye et al., 2018]. Given these conditioned latent features, the C network performs the final stage of affect estimation, producing refined affect predictions of both V-A intensity [Lv et al., 2017b],[Triantafyllidou and Tefas, 2016], [Aspandi et al., 2019b]. Thus, if $\hat{\theta}$ denote the estimated V and A:

$$\begin{aligned} ANCLaF(I) &= C_{\Phi^C}([G_{\Phi^G}(I); D_{\Phi^D}(G_{\Phi^G}(I))]) \\ &= FC_{\Phi^C}(enc_{\Phi^C}([G_{\Phi^G}(I); D_{\Phi^D}(G_{\Phi^G}(I))])) \\ &= FC_{\Phi^C}(enc_{\Phi^C}([\mathbb{Z}^I; \mathbb{Q}^I])) \Rightarrow \hat{\theta}^I_{ANCLaF}. \end{aligned} \tag{6.3}$$

6.3.2 Attention Enhanced Sequence Latent Affect Networks

We propose two sequence-based variants of our models: ANCLaF-S and -SA. Both of them use the combined features extracted by the G and D networks $\mathbb{ZQ}$, which are fed to LSTM networks to allow for temporal modelling [Hochreiter and Schmidhuber, 1997] and complemented with an FC layer to produce the final estimates. These networks are trained using Curriculum Learning [Bengio et al., 2009], [Gordon et al., 2018], [Aspandi et al., 2019b], in which the number of frames is progressively increased, allowing more throughout analysis of the training progress. Moreover, the training outcome for a given length facilitates the subsequent training of larger sequences [Gordon et al., 2018]. In this work, we considered a series of 2,4,8,16 and 32 successive frames ($\mathrm{N} = \{2,4,8,6,32\}$) for both training and inference stages. Depending on the number of frames to take into account (n), we use ANCLaF-S-n and ANCLaF-SA-n to name the respective variants of both ANCLaF-S and ANCLaF-SA networks. Lastly, the main difference between the two sequence models is that ANCLaF-SA also includes internal attentional modelling using the current and previous internal states from the LSTM layers. Thus, V-A predictions of ANCLaF-S-n are:

$$\begin{aligned} \forall n \in \mathrm{N}, ANCLaF\text{-}S\text{-}n(I_n), h_n = FC_{\Phi^C}(LSTM_{\Phi^C}([Z_n^I, Q_n^I], h_{n-1})) \\ \Rightarrow FC_{\Phi^C}(LSTM_{\Phi^C}(\mathbb{ZQ}_n^I, h_{n-1})). \end{aligned} \tag{6.4}$$

where LSTM is the LSTM network [Hochreiter and Schmidhuber, 1997] and h_n is LSTM states (h) after n successive frames. Built upon *ANCLaF-SA*, we further use attention modelling [Luong et al., 2015] to enable adaptive weights on model inferences by calculating the context vectors ($\mathbb{C}$) that summarise the importance of each previous state h. Differently from the original method, however, here, we also propose to include both the LSTM inner state (c) and outgoing states (h) [Kim et al., 2018] to provide the full previous information, and also to adapt these techniques to only consider n previous states following our curriculum learning approach.

Hence, given the combined LSTM states at frame t, denoted ($S_t = [c_t, h_t]$), and n previous states ($\bar{S}$), the alignment score is calculated as:

$$\begin{aligned} a_n(t) &= \text{align}(S_t, \bar{S}_t)\,, with S_x = [h_x; c_x] \\ &= \frac{\exp\left(W_a[S_t^\top; \bar{S}_n]\right)}{\sum_{N'} \exp\left(W_a[S_t^\top; \bar{S}_{n'}]\right)}. \end{aligned} \tag{6.5}$$

Then, the location-based function computes the alignment scores from the previous states ($\bar{S}$):

$$a_n = \text{softmax}(W_a \bar{S}). \tag{6.6}$$

Given the alignment vector, it is used to compute the context vector $\mathbb{C}_t$ as the weighted average over the considered n previous hidden states:

$$\mathbb{C}_t = \frac{\sum_n a_n \odot S_n}{n} \tag{6.7}$$

Finally, the context vector is concatenated with the current $\mathbb{ZQ}$ to be used as input to the C network pipeline:

$$\begin{aligned} \forall n \in \text{N}\,, ANCLaF\text{-}SA\text{-}n(I_n), h_n = \\ FC_{\Phi^C}(LSTM_{\Phi^C}([\mathbb{C}_n; \mathbb{ZQ}^I_n], h_{n-1})). \end{aligned} \tag{6.8}$$

6.3.3 Training Losses

We use the modified adversarial training from [Aspandi et al., 2020a] to train both the G and D networks, and incorporate the training of the C network by providing the latter with the features extracted from both the G and D nets on the fly. With this setup, we allow C to benefit from the improved quality of the features extracted by G and D as their training progresses. The equations for the modified adversarial training of these three networks are:

$$\begin{aligned} \mathcal{L}_{adv} = &\mathbb{E}_I\left[\log D(I)\right] + \\ &\mathbb{E}_I[\log\left(1 - D(G(\tilde{I}))\right) + \mathbb{E}_{afc}[C(I), \theta_I]. \end{aligned} \tag{6.9}$$

We used similar L_{afc} losses as in [Aspandi et al., 2020a], which incorporates multiple affect metrics: Rooted Mean Square Error (RMSE) (Eq. 6.11), Correlation(COR) (Eq. 6.12), Concordance Correlation Coefficients (CCC) (Eq. 6.13) [Kossaifi et al., 2017] with the addition of Intra-class Correlation Coefficient (ICC) [Kossaifi et al., 2019]. Thus, with $\{\hat{\theta},\theta\}$ as the predicted and the ground truth V-A values, the $\mathcal{L}_{afc}$ is defined as follows:

$$\mathbb{E}_{afc} = \sum_{i=1}^{F} \frac{f_i}{F}(\mathcal{L}_{RMSE} + \mathcal{L}_{COR} + \mathcal{L}_{CCC} + \mathcal{L}_{ICC}) \tag{6.10}$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}, \tag{6.11}$$

$$\mathcal{L}_{COR} = \frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}\sigma_\theta} \tag{6.12}$$

$$\mathcal{L}_{CCC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2 + (\mu_{\hat{\theta}} - \mu_\theta)^2} \tag{6.13}$$

$$\mathcal{L}_{ICC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}, \tag{6.14}$$

where f_i is the total number of instances of discrete V-A classes i, and F is a normalisation factor [Aspandi et al., 2019a] for the total V-A classes (discretised by a value of 10). This normalisation factor is crucial in cases of large imbalance in the number of instances per class, like in the AFEW-VA dataset (see Section 6.4.1).

6.3.4 Model Training

We use both the AFEW [Kossaifi et al., 2017] and SEWA [Kossaifi et al., 2019] datasets to train all our model variants, by following their original subject-independent protocol (5-fold cross validation). We conducted two training stages for each of our proposed models. Firstly, we trained the G, D, and C networks simultaneously using adversarial loss

as indicated in Equation 6.9. This training stage produced our baseline results without any sequential modelling, and conditional latent features $\mathbb{ZQ}$ to be used for the next stages of ANCLaF-S(A) Training.

In the second stage, The training of both ANCLaF-S and ANCLaF-SA was performed using the combined latent and quadrant features, under the previously defined curriculum learning scheme. We progressively train our ANCLaF-S models from 2, 4, 8, 16 to 32 steps of temporal modelling with multi-stage transfer learning [Christodoulidis et al., 2016]. Subsequently, we added our proposed attention mechanism to the pre-trained ANCLaF-S models, thus obtaining our ANCLaF-SA models. In both cases, we optimised the affect loss defined in Equation 6.10 with the same experimental settings used to train ANCLaF. We need to note that this combined training setup translates to more than 100 experiments in total. Hence, the use of latent features (known as a good choice to achieve reduced dimensionality representations) is critical to speed up our training process and make our experiments feasible. We observed a saving up to 1 : 4 of the original times during training each of our models by using the extracted latent features, with respect to using the original image (around 12 hours versus 2 days) on a single NVIDIA Titan X GPU. Full definitions of our models are available at Appendix A, with the respective source code is also available online .

6.4 Experiments and Results

6.4.1 Datasets and Experiment Settings

To quantify the impact of our temporal modelling, we opted to use two of the most popular and accessible video datasets available: Acted Facial Expressions in the Wild (AFEW-VA) [Kossaifi et al., 2017] and Automatic Sentiment Analysis in the Wild (SEWA) [Kossaifi et al., 2019]. On the one hand, AFEW-VA has more individual examples (600 versus 538) than SEWA, however, the latter has more frame examples, more contextual

https://github.com/deckyal/Seq-Att-Affect

information (such as subject, id of the associated culture) and is more balanced in terms of V-A labels [Mitenkova et al., 2019]. Furthermore, both datasets contain *in the wild* situations, enabling real time model evaluations. Finally, the labels provided are in the form of continuous V-A values, together with additional facial landmark locations, that we refined further using other external models [Aspandi et al., 2019b] to obtain more stable detection of the facial area.

In each experiment, we provide the results from all variants of our models to highlight the contribution of each module: first, we evaluate the ANCLaF model, which operates by exclusively using the latent features extracted on each frame ($\mathbb{ZQ}$) without any temporal modelling. Then, we provide results from both ANCLaF-S and ANCLaF-SA, which incorporate temporal modelling (and attention in the case of -SA). We report both RMSE and COR results, on both datasets, adding also ICC and CCC metrics for the AFEW-VA and SEWA datasets, respectively, to facilitate quantitative analysis to other results reported in the literature. Finally, for fair comparisons, we compare our models against external results which followed similar experimental protocols, i.e using exclusively this dataset in their training stage.

6.4.2 Comparative Results

Table 6.1 and table 6.2 provide the full comparisons of our proposed models against other reported results for both the AFEW-VA and SEWA datasets, respectively (specific results for each fold can be seen in the Appendix B). We can identify several findings based on these results: Firstly, that our base ANCLaF model, relying on a single image at a time, can produce quite competitive accuracy compared to other results from the literature. Furthermore, its accuracy is also higher than the results from the original AEG-CD-SZ models in which it is based upon [Aspandi et al., 2020a], as shown by its higher accuracy on the SEWA datasets, especially for Valence. This may indicate its better processing capabilities of the visual features, considering that AEG-CD-SZ also incorporates audio features, which in a way also explains its higher accuracy on the prediction of Arousal.

Model	RMSE ↓			COR ↑			ICC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Baseline [Kossaifi et al., 2017]	2.680	2.275	2.478	**0.407**	0.450	**0.429**	0.290	0.356	0.323
Coherent [Tellamekala and Valstar, 2019]	-	-	-	0.293	0.426	0.360	-	-	
Simul [Handrich et al., 2020]	2.600	2.500	2.550	0.390	0.290	0.340	**0.320**	0.210	0.265
ANCLaF	2.682	2.344	2.513	0.306	0.399	0.353	0.219	0.309	0.264
ANCLaF-S-2	2.675	2.295	2.485	0.314	0.410	0.362	0.236	0.296	0.266
ANCLaF-S-4	2.654	2.279	2.467	0.303	0.420	0.361	0.224	0.307	0.266
ANCLaF-S-8	2.595	2.202	2.398	0.328	0.425	0.377	0.272	0.344	0.308
ANCLaF-S-16	2.617	2.292	2.454	0.302	0.401	0.351	0.224	0.299	0.261
ANCLaF-S-32	2.568	2.328	2.448	0.288	0.405	0.346	0.214	0.304	0.259
ANCLaF-S-AVG	2.622	2.279	2.450	0.307	0.412	0.360	0.234	0.310	0.272
ANCLaF-SA-2	2.540	2.241	2.390	0.373	0.454	0.413	0.291	0.353	0.322
ANCLaF-SA-4	2.586	2.260	2.423	0.386	0.445	0.415	0.302	0.342	0.322
ANCLaF-SA-8	**2.481**	2.239	**2.360**	0.371	**0.467**	0.419	0.294	**0.367**	**0.331**
ANCLaF-SA-16	2.601	**2.225**	2.413	0.377	0.467	0.422	0.294	0.363	0.328
ANCLaF-SA-32	2.581	2.256	2.419	0.361	0.436	0.399	0.270	0.332	0.301
ANCLaF-SA-AVG	2.558	2.244	2.401	0.373	0.454	0.414	0.290	0.352	0.321

Table 6.1: Quantitative comparisons on the AFEW-VA dataset.

Model	RMSE ↓			COR ↑			CCC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Baseline [Kossaifi et al., 2019]	-	-	-	0.350	0.350	0.350	0.350	0.290	0.320
Tensor [Mitenkova et al., 2019]	0.334	0.380	0.357	0.503	0.439	0.471	0.469	0.392	0.431
AEG-CD-SZ [Aspandi et al., 2020a]	**0.323**	0.350	0.337	0.442	**0.478**	0.460	0.405	**0.430**	0.418
ANCLaF	0.354	0.347	0.351	0.530	0.395	0.462	0.492	0.364	0.428
ANCLaF-S-2	0.349	0.345	0.347	0.533	0.396	0.464	0.503	0.368	0.436
ANCLaF-S-4	0.344	0.336	0.340	0.536	0.403	0.469	0.510	0.382	0.446
ANCLaF-S-8	0.341	0.339	0.340	0.538	0.404	0.471	0.514	0.381	0.448
ANCLaF-S-16	0.354	0.344	0.349	0.527	0.395	0.461	0.490	0.369	0.429
ANCLaF-S-32	0.353	0.346	0.349	0.527	0.396	0.461	0.494	0.368	0.431
ANCLaF-S-AVG	0.348	0.342	0.345	0.532	0.399	0.465	0.502	0.374	0.438
ANCLaF-SA-2	0.343	0.333	0.338	0.545	0.420	0.482	0.509	0.390	0.449
ANCLaF-SA-4	0.336	**0.328**	**0.332**	0.550	0.429	0.490	0.526	0.399	0.463
ANCLaF-SA-8	0.336	0.332	0.334	**0.558**	0.424	**0.491**	**0.529**	0.405	**0.467**
ANCLaF-SA-16	0.334	0.331	**0.332**	0.556	0.421	0.488	0.528	0.393	0.461
ANCLaF-SA-32	0.336	0.362	0.349	0.550	0.418	0.484	0.513	0.389	0.451
ANCLaF-SA-AVG	0.337	0.337	0.337	0.552	0.422	0.488	0.521	0.395	0.458

Table 6.2: Quantitative comparisons on the SEWA dataset.

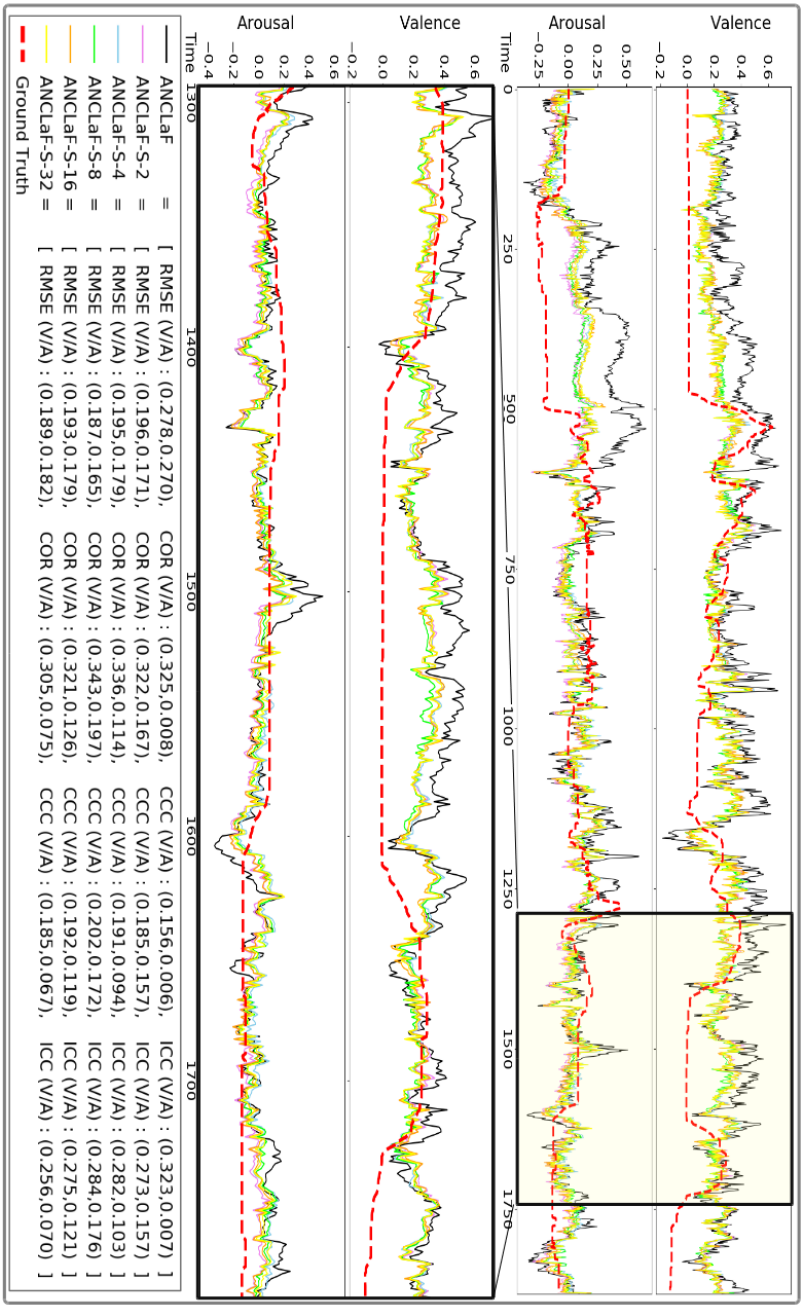


Figure 6.2: Analysis of prediction results from a single frame (ANCLaF) and from multiple frames with temporal modelling (ANCLaF-S-n). Top: the overview of the overall results; Bottom:, a closer look at the prediction results.

Secondly, we notice a slight accuracy improvement when our models incorporate sequence modelling (ANCLaF-S), especially in terms of Correlations, namely, CCC, and ICC. This finding demonstrates the benefit of the temporal modelling, yielding more stable results than those achieved by ANCLaF (cf. Section 6.4.3). However, even though the overall accuracy of ANCLaF-S is better than that of ANCLaF (and quite comparable to other state of the art models), the improvement can be considered modest, especially if we compare it with the improvement achieved when we include attention in our models. Indeed, we can see that our ANCLaF-SA outperforms almost all compared models across the different affect metrics. These findings suggests that the plain utilisation of LSTMs may not be enough to attain a considerable and substantial increase of accuracy [Schmitt et al., 2019], justifying the inclusion of the attention mechanism in our approach.

Thirdly, we further observe a noticeable trend of steady increase in accuracy from the predictions of both ANCLaF-S and ANCLaF-SA as the number of considered frames grows from 2 to 8, and then it plateaus (or even worsens a bit) as n continues to increase. This trend suggests that generally, a medium sequence length (between 4 to 16 frames) is optimal to produce more accurate predictions and that too short and too long sequences degrade temporal modelling. This finding is quite consistent with those from [Aspandi et al., 2019b], indicating the importance of progressive learning, which allows us to analyse and choose the optimal sequence length during training. Lastly, this sequence length selection may also impact the context vector along with its weights learnt in our attentional module, which explains why a similar trend was observed in the results from these models (see Section 6.4.4 for more details).

6.4.3 Analysis of the Impact of Sequence Modelling

Figure 6.2 shows an example of V-A predictions for ANCLaF and ANCLaF-S-n, together with the ground-truth annotations. Specifically, in the top part, we can see the predicted affect states from our models that, in general, are quite related to the ground truth values. However, we notice that the

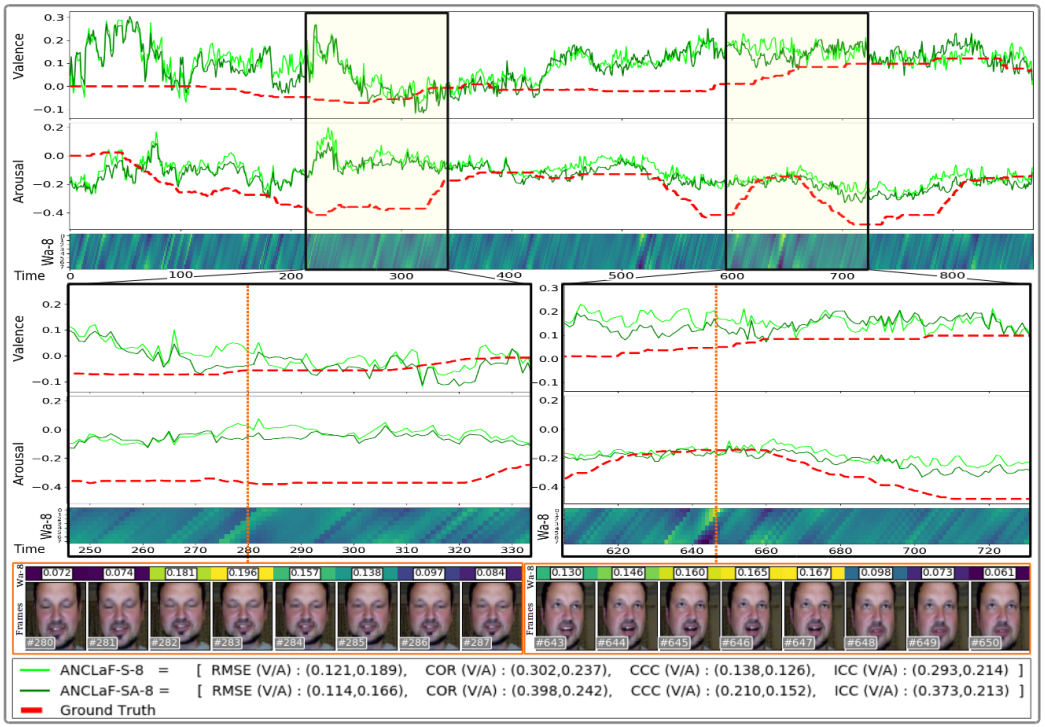


Figure 6.3: Analysis of the attention impact on the prediction results of our sequence modelling (results from ANCLaF-S-8 and ANCLaF-SA-8, which correspond to the best ANCLaF-S and ANCLaF-SA models, respectively). Top: overview of the overall results; Bottom: two examples of a closer view on the prediction graph. The column Wa-8 shows the attention weights learnt for the eight considered frames.

results of our sequence based models are more accurate than their non-sequential counterparts. We can also see that the the predicted values from ANCLaF are quite sparse, thus, quite unstable compared to ANCLAF-S, which explains its lower COR, CCC, and ICC values. Our sequence modelling, on the other hand, is able to create smooth predictions with higher overall accuracy.

On the bottom part of the figure, showing a magnified portion of the same example, we further notice that the results for all ANCLaF-S-n are quite similar, with those from ANCLaF-S-8 showing the highest resemblance to the ground-truth. Thus, inclusion of too short or too long sequences yields sub-optimal results due to the complexity of the facial movements included between frames (see the next section for further details).

6.4.4 Analysis of the Role of the Learnt Attentions Weights

To analyze the impact of the attention mechanism on our sequence modelling, we first show in Figure 6.3 a comparison of our baseline sequence modelling (ANCLaF-S) against ANCLaF-SA with attention activated. In the top part, we can see the predictions from the best performing models with and without attention (ANCLaF-S-8 and ANCLaF-SA-8). Comparing the predictions from both models, we find that the results are quite similar, though in some cases ANCLaF-SA seems to be more accurate and closer to the ground truth. The quantitative accuracy results indicated on the respective legends confirm this observation.

The attention weights learnt by ANCLaF-SA, involving the previous eight frames, are also displayed at the bottom of the prediction plots. We can see that the weights calculated with respect to the associated frames seem to be higher in the presence of changes. Indeed, we observe that the attention weights are usually activated prior to subsequent facial movements. Interestingly, the intensity of the activations also appears to highlight the level of these facial movements, or the changes between frames. For instance, from frames 280-287, we can see that the different

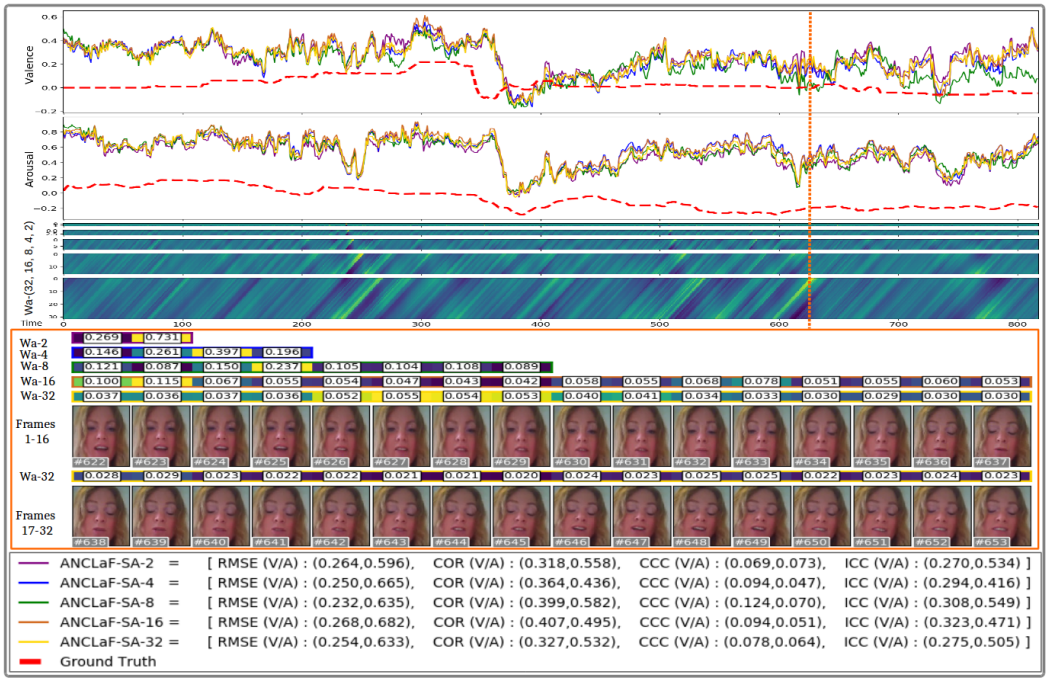


Figure 6.4: Analysis of the relationship between the selection of sequence length (n) and the learnt weights of our attentional approach. Top: overview of the prediction results of all variants of our models with attention mechanism (ANCLaF-SA-n) alongside their learnt weights. Middle: details for frames 622 to 653 with their associated weights for each model. Bottom: legend containing the quantitative comparisons.

level of the weight intensity seems to be small, which also correlates to the subtle changes observed in those frames (e. g., closing of the eyes). In contrast, in frames 643-650, we see high levels of activation on the first few frames that correspond to the more discernible facial movements on the respective frames, such as the changes observed in the mouth area. These correlations illustrate how our models are capable of learning temporal changes.

Figure 6.4 provides further details on the attention mechanism for different temporal modelling lengths. We can see that all the displayed models show quite smooth results, thanks to the temporal modelling, but not all of them achieve the same accuracy on the predictions. The bottom part of the figure, highlighting the input sequence from frames 622 to 653, can help to provide an intuition about the optimal temporal modelling length, which was found to be about 8 frames. To this end, let us start by looking at the whole set of 32 frames: we can see that such sequence of frames comprises multiple facial changes, and considering all of them together makes the training task harder to optimise. On the other hand, if we consider groups of very few frames (e. g., 2 or 4 frames), the system is likely to capture only part of a given facial action, which may impede it to properly interpret it. Therefore, we see that the optimal sequence length is the one that contains enough frames to interpret facial changes without extending too much the temporal context, which may unnecessarily increase training complexity and reduce accuracy.

Finally, it is important to emphasise that the optimal sequence length needs to take into account the frame rate and the specific facial movements that are present in each dataset. In the considered dataset, with an overall frame rates of 50 fps, this length corresponds to 160 ms.

6.5 Conclusions

In this work, we have successfully built a sequence-attention based neural network for affect estimations in the wild. We did so by incorporating three major sub-networks: the Generator, which is responsible to extract

latent features on each frame; the Discriminator, which is used to supply the first step of affect estimates of emotional quadrant, and the Combiner, which merges latent features and quadrant information to produce the final refined affect estimates of Valence and Arousal on a frame by frame basis. We then added an LSTM layer to allow temporal modelling, which we further enhanced by using step-wise attention modelling. We trained these three major sub-networks in an adversarial setting, and used curriculum learning on the sequential training stages.

We show the effectiveness of our approach by reporting top state of the art results on two of the most widely used video datasets for affect analysis, namely AFEW-VA and SEWA. Specifically, our baseline models, which operate without any sequence modelling, yield quite competitive results with other models reported in the literature. On the other hand, our more advanced models, which are sequence-based, clearly helped to improve the affect estimates both in qualitative and quantitative terms. Qualitatively, the temporal modelling helped to produce more stable results, with visibly smoother transitions between affect predictions. Quantitatively, our models produced the overall best accuracy results reported so far on both tested datasets.

Within sequence-based models, we observed the highest accuracy improvements when the attention mechanism was included. Detailed analysis of the attention weights highlighted their correlation with the appearance of facial movements, both in terms of (temporal) localisation and intensity. Finally, we found a sequence length of around 160 ms to be the optimum one for temporal modelling, which is consistent with other relevant findings utilising similar lengths. Lastly, future work will need to explore further optimisation of the considered adversarial topologies and attention mechanisms as well as their transferability across databases, cultures, and domains.

Chapter 7

AUDIO-VISUAL BASED GATED-SEQUENCED NEURAL NETWORK FOR AFFECT RECOGNITION

Adapted from: D. Aspandi, F. Sukno, B. Schuller, and X. Binefa, ”Audio-Visual Gated-Sequenced Neural Network for Affect Recognition” in *IEEE Transactions on Affective Computing, (Under Review)*.

Abstract

The interest in automatic emotion recognition and the larger field of Affective Computing has recently gained momentum. The current emergence of large, video-based affect datasets offering rich multi-modal inputs facilitates the development of deep learning-based models for automatic affect analysis that currently holds the state of the art. However, recent approaches to process these modalities cannot fully exploit them due to the use of oversimplified fusion schemes. Furthermore, the efficient use of temporal information inherent to these huge data are also largely unexplored hindering their potential progress. In this work, we propose a multi-modal, sequence-based neural network with gating mechanisms for affect recognition. Our model consists of three major networks: Firstly, a latent-feature generator that extracts compact representations from both modalities that have been artificially degraded to add robustness. Secondly, a multi-task discriminator that estimates both input identity and a first step emotion quadrant estimation. Thirdly, a sequence-based combiner with attention and gating mechanisms that effectively merges both modalities and uses this information through sequence modelling. In our experiments on the SEMAINE and SEWA datasets, we observe the impact of both proposed methods with progressive increase in accuracy. We further show in our ablation studies how the internal attention weight and gating coefficient impact our models’ estimates quality. Finally, we demonstrate state of the art accuracy through comparisons with current alternatives on both datasets.

7.1 Introduction

Emotions of humans made accessible and 'readable' to computing devices keeps trending, as we near a robust automatic recognition which actually opens up real world usage such as in education [Duo and Song, 2010] or healthcare [Liu et al., 2008], among others. The growing availability of affect-related datasets, such as SEMAINE [McKeown et al., 2010] as was used in the first Audio/Visual Emotion Challenge (AVEC) and the recently introduced SEWA [Kossaifi et al., 2019] database enable the rapid development of automatic visual-based emotion recognition. While the field started from handcrafted methods, it currently heavily relies on deep learning-based approaches [Kossaifi et al., 2017, Kossaifi et al., 2020]. The use of other modalities such as sound and text has also improved the current systems in other emotional aspects that the visual modality lacks or situations, where it is not accessible or disturbed. This in turn also encourages the combination of these modalities, typically by direct concatenation approaches [Chen and Jin, 2015, Yan et al., 2018, Wang et al., 2012b]. However, such a straightforward approach may produce sub-optimal results given the difference characteristics of each modality [Arevalo et al., 2020].

Another aspect to consider is the need to deal with *bigger-data*, given the emergence of video-based datasets that enrich the widely used modality features with the inclusion of temporal information. To this end, several authors have explored the use of deep-learning based sequence modelling of long-short term memory (LSTM) recurrent neural networks (RNNs) [Tellamekala and Valstar, 2019, Ma et al., 2019], endowed also with attention mechanisms [Luong et al., 2015, Xiaohua et al., 2019] to exploit these sequence based data inputs. However, such spatio-temporal modelling often results in very high-dimensional feature spaces and large volumes of data, making training difficult and time consuming.

This work addresses current lack of efficient temporal modellings and effective multi-fusion approaches to affect analysis, by proposing the use of latent sequence networks combined with gating mechanisms to effectively fuse multi-modal inputs. We do so by incorporating three ma-

jor networks, coined Generator, Discriminator, and Combiner, which are trained in an adversarial setting to estimate the affect domains of Valence (V) and Arousal (A). Furthermore, we capitalise on these latent features to enable temporal modelling using internal LSTMs, that are trained progressively using curriculum learning enhanced with adaptive attention. Finally, we combine the input modalities through gating mechanisms for more effective modality fusion, leading to our state of the art accuracy. Specifically, the contributions of this chapter are as follows:

1. We upgrade the standard adversarial setting with a third network that fuses features from the Generator and Discriminator. This produces features that combine the latent space from the autoencoder-based Generator and a V-A Quadrant estimate produced by the modified Discriminator, resulting in a compact, but meaningful representation that helps reduce the training complexity.

2. We propose the use of sequential modelling with attention to enhance our model estimates, and also quantify the relative impact of these adaptive attention mechanism by calculating the respective internal weight activation differences.

3. We extend our temporal modelling with gating networks for more effective fusion of both, audio and visual modalities. We further evaluate its effectiveness in our ablation study using thresholding analysis.

4. We report state of the art accuracy of our models on both the SEMAINE and SEWA datasets and compare our results to other alternatives.

Preliminary results of our modified adversarial training with latent features, and the respective sequence modelling with attention, can be found in [Aspandi et al., 2020a, Aspandi et al., 2020b]. The rest of this chapter is organised as follows: Section 7.2 describes the related work in the context of facial-based emotion recognition and the use of different modalities and other relevant temporal modelling; in Section 7.3, we explain our

Audio-Visual Gated-Sequenced Neural Networks consisting of three major networks combined using our gated-sequence modelling. In Section 7.4, we report our results on both, the SEMAINE and SEWA datasets in relation to each of our methods, and further compare our results with current state of the arts models. Finally, Section 7.5 provides the conclusions.

7.2 Related Work

Multi-modal emotion recognition started by the use of classical machine learning techniques, applied to visual features to enable automatic affect estimation. Examples of early approaches include partial least squares regression [Povolny et al., 2016], and support vector machines [Nicolaou et al., 2011]. Subsequently, to further progress the investigations in this field, the development of larger and bigger datasets was initiated, with the SEMAINE [McKeown et al., 2010] dataset as one popular instance. This audio-visual dataset facilitates direct analysis for human and agent interactions, and has been used by many authors. One of the early works is Gunes et al. [Gunes and Pantic, 2010], who used global head motions consisting of nod and shake to be fed to individual Hidden Markov Models (HMM) to construct the baseline features, which are then utilised by Support Vector Regression (SVR) to estimate the final affect dimension. Using a person independent scheme in their experiment, they proved that automatic affect recognition is indeed possible to a certain degree. Progressing ahead, Kossaifi et al. [Kossaifi et al., 2017] introduced a hybrid system that used deep learning alongside classical geometrical and texture features for affect recognition. Specifically, they proposed to include the use of features extracted from Scale-Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), and facial Landmarks combined with several classifiers, such as Bag of Words (BOW) and conditional Random Field (RF). In addition, they performed transfer learning to several Convolutional Network-based models to investigate the effectiveness of these deep learning models. Using the SEMAINE database, alongside the other related affect dataset of AFEW-VA, they found that deep learning-based

methods constantly outperformed other classical approaches and provides new baselines for each dataset.

More recently, the SEWA dataset [Kossaifi et al., 2019] was published to allow more extensive deep learning-based modelling under unconstrained settings (in-the-wild) and offer multiple languages and cultures at the same time. Such deep learning-based approaches can be seen in the recent works of [Mitenkova et al., 2019] and [Kossaifi et al., 2020]. Mitenkova et al. [Mitenkova et al., 2019] introduced tensor modelling for affect estimations by using spatial features. In their work, they used tucker tensor regression optimised by means of deep gradient methods, thus allowing their model to preserve the structure of the data and reduce the number of parameters. Similarly, Kosaifi et al. [Kossaifi et al., 2020] introduced the use of tensor decomposition to enable their multi-dimensional convolutional approach for visual-based emotion recognition. Specifically, they applied a generalised factorised higher-order framework to several convolutional models, such as ResNet, Inception, and Mobile net. Furthermore, they proposed to perform a more efficient tensor decomposition on Convolutional Operations by the introduction of weight vector coefficients with non-linearities affecting the magnitude of the decomposed factors. Then, they also added higher-order transduction and automatic rank selections in their pipelines to further improve the calculation efficiency. Using this approach, they arrived at state of the art results.

Visual-based approaches have been considerably gaining attention lately, since facial expressions are considered one of the dominant channels to display affective information. However, facial expressions not always provide the full emotional information [Kossaifi et al., 2019]. Indeed, it has been shown that modalities such as Electrocardiogram (ECG) and audio can complement and enhance the performance obtained from visual-features [Correa et al., 2018]. Specifically, the audio modality has been highlighted for its accurate arousal estimates[Kossaifi et al., 2019, Poria et al., 2019]. One example of audio-based emotion recognition is the work of Yang et al. [Yang and Hirschberg, 2018] that exploited both the sound-wave and its spectrogram derivatives as main features for affect recognition. In their work, they used a 3D Convolutional Neural Network

(3DCNN) to extract the individual waveform and spectral features. Then, these features were combined using basic concatenation approaches and passed to a Bidirectonal Long Short-term Memory (BLSTM) network to estimate final valence and arousal values. Another recent approach is the Dialogue-RNN [Poria et al., 2019] that tries to incorporate the notion of dialogue in a multi-modal approach. In their work, the authors also used the text modality alongside visual and sound information as main feature. They chose LSTM and Gated Recurrent Units (GRU) to explicitly model the interaction between the user (global, speaker, listener) through sequential learning, thus benefiting from this additional knowledge. They found that adding attention modules improved the accuracy of their models, suggesting the importance of modelling the interaction between modalities.

The recent growth of video-based datasets has encouraged the inclusion of temporal modelling, which has shown to improve models' training [Xie et al., 2016, Cootes et al., 1998]. Relevant examples in Affective Computing include the works of Tellamekala et al. [Tellamekala and Valstar, 2019] and Ma et al. [Ma et al., 2019]. In their work, Tellamekala et al. [Tellamekala and Valstar, 2019] enforced temporal coherency and smoothness aspects on their feature representation by constraining the differences between adjacent frames, while Ma et al. resort to the utilisation of LSTMs with residual connections applied to multi-modal data. Furthermore, the use of attention has also been recently explored by Xiaohua et al. [Xiaohua et al., 2019] and Li et al. [Li et al., 2020]. Xiahoua et al. adopted multi-stage attention, which involved both, spatial and temporal attention for their facial-based affect estimation pipeline. Meanwhile, using spectrogram data as input, Li et al. proposed a deep network that utilised an attention mechanism [Luong et al., 2015] on top of their LSTM networks to predict the affective states.

In summary, recent developments of large scale datasets such as SEMAINE [McKeown et al., 2010] and SEWA [Kossaifi et al., 2019] have facilitated the development of automatic affect recognition. The starting point was the use of handcrafted features and classifiers applied to visual features, typically the facial area. The field then progressed toward the use

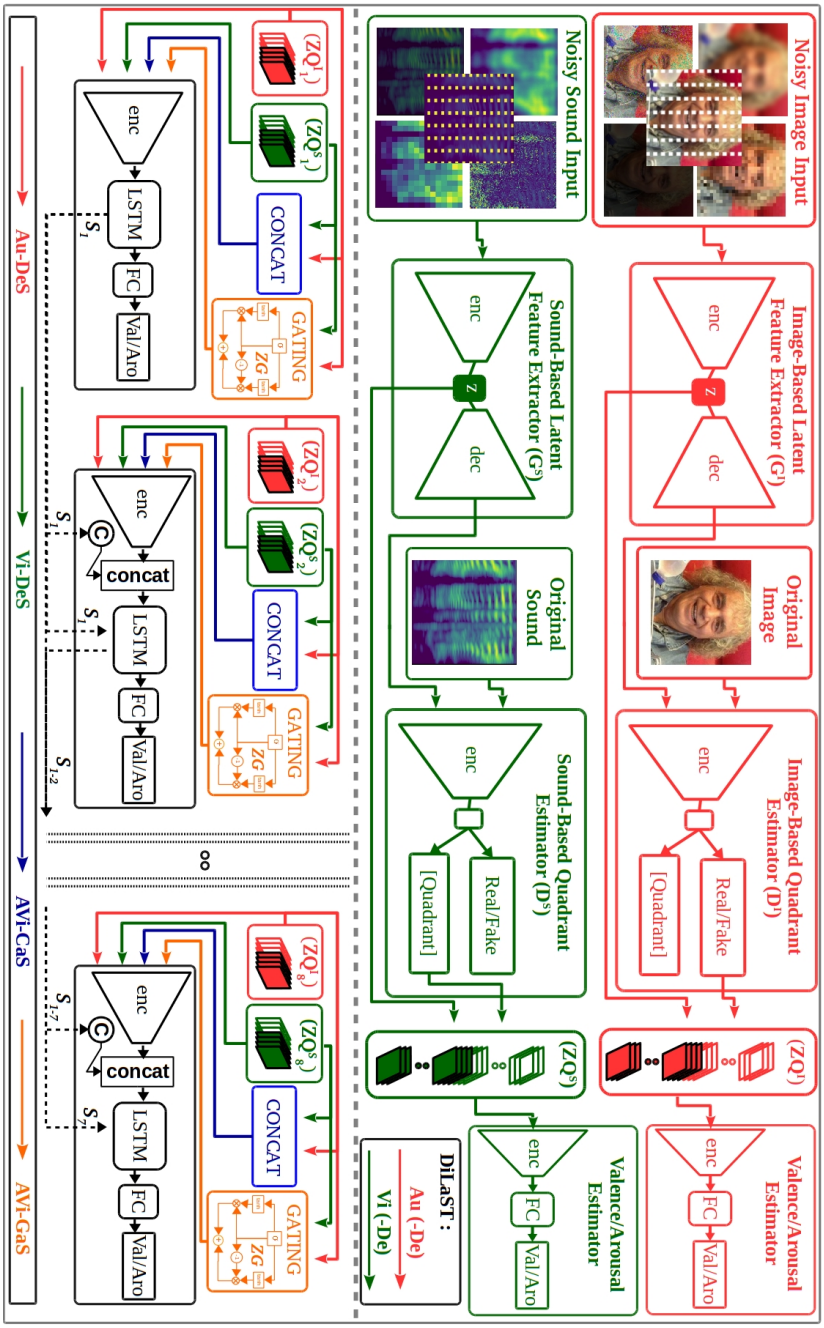


Figure 7.1: Schematic representation of our Full AViGaS Networks (AviGaS-NET). First top part shows the pipelines of individual modality version of our AViGaS networks: AU-De Net and VI-De Net. The bottom part visualizes the process of our sequence based models of AU-DeS Net, Vi-DeS Net, AVi-CaS Net and our complete model of AVi-GaS Net (AViGaS).

of Deep Learning approaches. Furthermore, the use of other modalities has emerged due to the limitation of the visual features in regards to the accuracy obtained, and some works also tried to combine modalities by the use of simple concatenation. However, these straight-forward approaches have the limitation of their tendency to give equal weights to the different modalities [Zhang et al., 2018b, Zhou et al., 2019]. This can be problematic, in the situation where the importance of one modality may be considered higher than that of the others [Kossaifi et al., 2019]. This problem could be mitigated by the use of Gating Mechanisms [Arevalo et al., 2020] that permit the adaptive weighting for the considered modalities. Furthermore, we also propose to combine it with our temporal modelling including attention, to allow for more accurate results as recently shown in other related studies [Xie et al., 2016, Cootes et al., 1998], including our preliminary results in [Aspandi et al., 2020a, Aspandi et al., 2020b]. To the best of our knowledge, we are the first to explore the adaptive combinations of multiple modalities with attention for temporal modelling within affective computing approaches.

7.3 Audio-Visual Gated-Sequenced Neural Network (AViGaS-NET)

Figure 7.1 shows the overview of our proposed approach that consists of two major sub-operations: a direct latent-based Valence/Arousal (V/A) estimations (top), and combining mechanisms with sequence modelling (bottom). The direct latent-based V/A estimator (DiLaST) consists of a coupled Generator and Discriminators that are trained under adversarial settings to denoise input images, and subsequently create latent features. These latent features are then combined with the first step V/A quadrants and later used for final V/A estimates using the V/A estimator parts (C). In this work, we consider two modality inputs: visual and audio features. The second part of our models pipeline consists of the inclusion of two combining mechanisms that are later used as input for a sequence-based V/A estimator with attention [Luong et al., 2015].

Several combinations of the two modalities, temporal modelling and combination strategies lead to possible sub-models, as listed below, which we will include in our experiments to motivate the need for our full system:

1. Au-De: direct latent-based V/A estimator using only the audio stream (without temporal modelling).
2. Vi-De: direct latent-based V/A estimator using only the video stream (without temporal modelling).
3. Au-DeS: sequence-based V/A estimator using only the audio stream.
4. Vi-DeS: sequence-based V/A estimator using only the video stream.
5. AVi-CaS: sequence-based V/A estimator using concatenated audio and video streams.
6. AVi-GaS: sequence-based V/A estimator using the gating mechanism to fuse the audio and video streams.

7.3.1 The Direct Latent based V/A Estimator (DiLaST)

The first part of of our approach uses latent features extracted through adversarial learning combined with first step V/A quadrant estimates given the noisy image inputs. This approach is similar to our previous model [Aspandi et al., 2020a], however, we further consider to use two different modalities separately to asses their respective impact. We use RGB images as visual input, while a spectrogram of corresponding sounds is used for the audio modality. The details of the data pre-processing of both modalities can be seen in Section 7.3.4. The pipeline of **DiLaST** models starts with the G network that receives either the original inputs F, consisting either on the images I or sounds S, and also a distorted version of it, $\tilde{F} \in \{\tilde{I}, \tilde{S}\}$, as detailed in [Aspandi et al., 2019c, Aspandi et al., 2019a]. Given the noisy versions of these modalities, G simultaneously produces an estimate of the cleaned reconstruction of both the input modality $\hat{F}$ and a 2D latent representation that will be used as subsequent features ($\mathbb{Z}$):

$$G(F)_{\Phi^G} = dec_{\Phi^G}(enc_{\Phi^G}(F)) \, with \, \mathbb{Z}^F \approx enc_{\Phi^G}(F), \tag{7.1}$$

where Φ are the parameters of the respective networks, *enc* and *dec* are the encoder and decoder, and G consists of both G^I and G^S respectively. Subsequently, the D network receives $\hat{F}$ and tries to estimate whether the sample was obtained from a true or fake examples (namely, original or distorted input image), as well as a rough estimate of the affective state in the form of Circumplex Quadrant ($\mathbb{Q}$) [Russell, 1980] which discretises emotions along the valence and arousal dimensions (four quadrants) [Aspandi et al., 2020b]. This in turn reduces the training complexity. Thus, letting FC stand for fully connected layer:

$$D(F)_{\Phi^D} = FC_{\Phi^D}(enc_{\Phi^D}(F)) = (\mathbb{Q}^F, \Upsilon^F). \tag{7.2}$$

where Υ^F is a binary variable indicating whether the sample is classified as real (1) or fake (0). Then, $\mathbb{Q}$ is used to condition the extracted latent features $\mathbb{Z}$ through layer-wise concatenation, which we call $\mathbb{ZQ}$ [Dai et al., 2017, Ye et al., 2018]. Given these conditioned latent features, the C network performs the final stage of affect estimation to produce refined affect predictions of both, the V and A intensity [Lv et al., 2017b, Triantafyllidou and Tefas, 2016, Aspandi et al., 2019b]. Thus, if $\hat{\theta}$ denotes the estimated V and A values:

$$\begin{aligned} DiLaST(F) &= C_{\Phi^C}([G_{\Phi^G}(F); D_{\Phi^D}(G_{\Phi^G}(F))]) \\ &= FC_{\Phi^C}(enc_{\Phi^C}([G_{\Phi^G}(F); D_{\Phi^D}(G_{\Phi^G}(F))])) \\ &= FC_{\Phi^C}(enc_{\Phi^C}([\mathbb{Z}^F; \mathbb{Q}^F])) = \hat{\theta}^F_{DiLaST}. \end{aligned} \tag{7.3}$$

Depending on the modality inputs, we call the DiLaST as Vi-De and Au-De Net when it uses Visual and Audio input, respectively.

7.3.2 Multi-Modal Fusion with Attention Enhanced Sequence Modelling (DiLaST-SA)

The compact size of ZQ extracted from the previous pipeline allows us to perform more complex processing to reach a higher accuracy. Motivated by our previous findings in [Aspandi et al., 2020b] about the importance of sequence modelling, we propose to use such approaches on both available latent features. This is reached by employing LSTM combined with attention mechanisms [Luong et al., 2015] and training with Curriculum Learning [Bengio et al., 2009, Gordon et al., 2018, Aspandi et al., 2019b]. Our sequence modelling (DiLaST-S) uses the extracted ZQ as the primary input for processing. Furthermore, alongside the use of individual ZQ to the sequence pipelines, we also propose to investigate the impact of two different fusion strategies to merge the sound ZQ^S and image ZQ^I features:

1. By direct concatenation, which has been the most popular in the field, and consists of simply concatenating both inputs ZQ as new features. Thus
$$ZQ^C = [ZQ^I; ZQ^S] \tag{7.4}$$

2. By gating mechanisms, where we use gated multi unit approach [Arevalo et al., 2020] that relates these two distinct modalities. It is calculated as follows:
$$\begin{aligned} GMU(ZQ^I, ZQ^S) &= ZQ^G \\ &= ZG(ZQ^I, ZQ^S) \odot h_v + \\ &\quad (1 - ZG(ZQ^I, ZQ^S)) \odot h_s) \end{aligned} \tag{7.5}$$
with ZG, h_v and h_s are calculated as:
$$ZG(ZQ^I, ZQ^S) = \sigma(W_{ZG}[ZQ^I; ZQ^S]) \tag{7.6}$$
$$h_v = tanh(W_v \odot (ZQ^I)^T) \tag{7.7}$$
$$h_s = tanh(W_s \odot (ZQ^S)^T) \tag{7.8}$$
thus the ZG coefficient control the importance of each modality input.

Subsequently, these features (ZQ, ZQ_C and ZQ_G) will be individually fed to our LSTM modeling that is based the C network but with attention modules. Thus, given the ZQ as example of the input, the final results of our models are:

$$\forall n \in \mathrm{N}\,, DiLaST-SA, h_n = FC_{\Phi^C}(LSTM_{\Phi^C}([\mathbb{C}_n; \mathbb{ZQ}_n], h_{n-1})). \quad (7.9)$$

where LSTM is the Long Short Term Memory network [Hochreiter and Schmidhuber, 1997] and h_n is the set of LSTM states (h) after n successive frames. The context vector [Luong et al., 2015] of ($\mathbb{C}$) enables adaptive weights on model inferences that summarise the importance of each previous state h. ($\mathbb{C}$) consists of both the LSTM inner state (c) and outgoing states (h) [Kim et al., 2018] to provide the full previous information, and also to adapt these techniques to only consider 8 sequence of previous states (n= 8) following our curriculum learning approach [Aspandi et al., 2020b]. Hence, using combined LSTM states at frame t, denoted ($S_t = [c_t, h_t]$), and n previous states ($\bar{S}$), the alignment score is computed as:

$$\begin{aligned} a_n(t) &= \mathrm{align}(S_t, \bar{S}_t)\,, with\, S_x = [h_x; c_x] \\ &= \frac{\exp\left(W_a[S_t^\top; \bar{S}_n]\right)}{\sum_{N'} \exp\left(W_a[S_t^\top; \bar{S}_{n'}]\right)}. \end{aligned} \quad (7.10)$$

Then, the location-based function calculates the alignment scores from the previous states ($\bar{S}$):

$$a_n = \mathrm{softmax}(W_a \bar{S}). \quad (7.11)$$

The alignment vector then is used to quantify the context vector $\mathbb{C}_t$ as the weighted average over the considered n previous hidden states:

$$\mathbb{C}_t = \frac{\sum_n a_n \odot S_n}{n} \quad (7.12)$$

Depending on the configurations, the above will yield three different models: *i*) a sequence-based single modality affect estimator (Au/Vi-Des Net), when the direct ZQ^I and ZQ^S are used as input; *ii*) a concatenated-based affect estimator (AVi-CaS Net), when the concatenated latent features from both the visual and sound modality (ZQ^C) are used as input; and *iii*) our full model of a gated affect estimator (AVi-GaS Net), when the gating mechanism is used to fuse both modalities (ZQ^G).

7.3.3 Training Losses

To train our models, we adopt the modified adversarial training from [Aspandi et al., 2020a, Aspandi et al., 2020b] to train both the G and D networks, and incorporate the training of the C network by providing the latter with the features extracted from both the G and D nets on the fly. This arrangement allows the C to benefit from the improved quality of the features extracted by G and D as their training progresses. Thus the equations for the modified adversarial training of these networks are:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_F\left[\log D(F)\right] + \\ &\mathbb{E}_F[\log\left(1 - D(G(\tilde{F}))\right) + \mathbb{E}_{afc}[C(F), \theta_F].\end{aligned} \tag{7.13}$$

We apply similar L_{afc} losses as in [Aspandi et al., 2020a], that incorporate multiple affect metrics: Rooted Mean Square Error (RMSE) (Eq. 7.15), Correlation(COR) (Eq. 7.16), Concordance Correlation Coefficients (CCC) (Eq. 7.17) [Kossaifi et al., 2017] along with the Intra-class Correlation Coefficient (ICC) [Kossaifi et al., 2019]. Hence, letting $\{\hat{\theta},\theta\}$ as the predicted and the ground truth V-A values, the $\mathcal{L}_{afc}$ is defined as follow:

$$\mathbb{E}_{afc} = \sum_{i=1}^{K} \frac{k_i}{K}(\mathcal{L}_{RMSE} + \mathcal{L}_{COR} + \mathcal{L}_{CCC} + \mathcal{L}_{ICC}) \tag{7.14}$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}, \tag{7.15}$$

$$\mathcal{L}_{COR} = \frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}\sigma_\theta} \tag{7.16}$$

$$\mathcal{L}_{CCC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2 + (\mu_{\hat{\theta}} - \mu_\theta)^2} \tag{7.17}$$

$$\mathcal{L}_{ICC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}, \tag{7.18}$$

where K is a normalisation factor [Aspandi et al., 2019a] for the total V-A classes (discretised by a value of 10) and k_i is the total number of instances of discrete V-A classes i. The normalisation factor is crucial in cases of large imbalance in the number of instances per class in the dataset.

7.3.4 Data Pre-Processing and Model Training

We use both the SEMAINE [McKeown et al., 2010] and SEWA [Kossaifi et al., 2019] datasets to train all variants of our models by following their original subject-independent protocol (5-fold cross validation). Using these datasets, we obtain the facial area by running a state of the art facial tracker [Aspandi et al., 2019b]. To extract the sound features, we first calculate the whole Mel-spectrogram of the respective sound files of the video inputs. We extract the Mel-scaled spectrogram using Librosa library [McFee et al., 2020] with the parameters of a fast Fourier transform at a sample rate of 22 kHz, and with number of Mel dimensions of 128 to match the input image dimension of 128 x 128. Subsequently, we convert the obtained power spectrogram to decibel units using its maximum amplitude. We then crop the parts of spectrogram centred with the time-stamp of the input frames, with the left and right pad of half of the input image size, i. e., 64. Finally, we replicate these spectrograms into 3 channels, to allow them to be processed with a similar network structure as the one used for the visual input.

We start the training process with the DiLaST network, which involves the respective G, D, and C networks simultaneously using an adversarial

loss as indicated in Equation 7.13. To quantify the impact of the denoising, we also choose to train the standard DiLaST without any noisy image inputs. This stage produces our baseline consisting of individual results for each modality with and without noise modelling (Au/Vi and Au-De/Vi-De Nets, respectively), and the conditional latent features $\mathbb{ZQ}$ of each modality to be used on the sequence modelling (DiLaST-SA). This is done by the use of multi-stage transfer learning from 2, 4, and 8 [Aspandi et al., 2019b] with attention enabled. We use the individual *ZQ* directly to the sequence modelling pipelines to produce the sequence variants of the original DiLaST of both modalities, i. e., Au-DeS net and Vi-DeS nets. Furthermore, we combine both *ZQ* altogether to be used as the input to our sequence modelling to produce the $AVi-CaS$ network results. Lastly, we also use the gating mechanism to perform selective merging as the input to create our final $AVi-GaS$ networks. In all cases, we train all of our model variants using the affect loss defined in Equation 7.14.

We need to note that this combined training requires considerable computation power. Hence, the uses of latent features from both modalities (known as a good choice for reduced dimensionality representations) are critical to speed up our training process, making our experiments feasible. We observe a reduction of up to a quarter of the original times required for the training each of our models by using the extracted latent features, with respect to using the original inputs size (around 12 hours versus the original 2 days) on a single NVIDIA Titan X GPU.

7.4 Experiment Results

In this section, we first describe the datasets used in our experiments with the respective metrics to quantify the performance of each model in (Section 7.4.1). Secondly, we perform an ablation study to highlight the importance of each element of our model: we first analyse the results produced by each of our modality approaches and its correlations with the sequence modelling with attention (Section 7.4.2). Then, we focus

Model	RMSE ↓		COR ↑		CCC ↑		ICC ↑	
	VAL	ARO	VAL	ARO	VAL	ARO	VAL	ARO
Vi	0.268	0.315	0.364	0.238	0.350	0.233	0.368	0.235
Vi-De	0.247	0.297	0.391	0.246	0.373	0.234	0.399	0.238
Vi-DeS	**0.232**	0.289	**0.441**	0.250	**0.412**	0.234	**0.455**	0.238
Au	0.302	0.228	0.261	0.495	0.238	0.476	0.249	0.484
Au-De	0.290	0.217	0.266	0.506	0.240	0.486	0.249	0.489
Au-DeS	0.306	**0.226**	0.272	**0.509**	0.250	**0.495**	0.262	**0.500**

Table 7.1: Quantitative comparisons of our models utilising each modality (DiLaST) on the SEMAINE dataset.

Model	RMSE ↓		COR ↑		CCC ↑		ICC ↑	
	VAL	ARO	VAL	ARO	VAL	ARO	VAL	ARO
Vi	0.340	0.345	0.444	0.375	0.434	0.375	0.466	0.381
Vi-De	0.335	0.343	0.463	0.383	0.447	0.377	0.484	0.388
Vi-DeS	**0.328**	0.333	**0.501**	0.400	**0.476**	0.398	**0.520**	0.405
Au	0.363	0.341	0.391	0.483	0.379	0.467	0.386	0.480
Au-De	0.343	0.332	0.411	0.530	0.397	0.512	0.405	0.523
Au-DeS	0.342	**0.327**	0.430	**0.534**	0.420	**0.520**	0.424	**0.528**

Table 7.2: Quantitative comparisons of our models utilising each modality (DiLaST) on the SEWA dataset.

on the importance of our gating mechanisms to aggregate both modalities to reach our best results (Section 7.4.3). Lastly, Section 7.4.4 compares our best results with other alternatives on both, the SEMAINE and SEWA datasets.

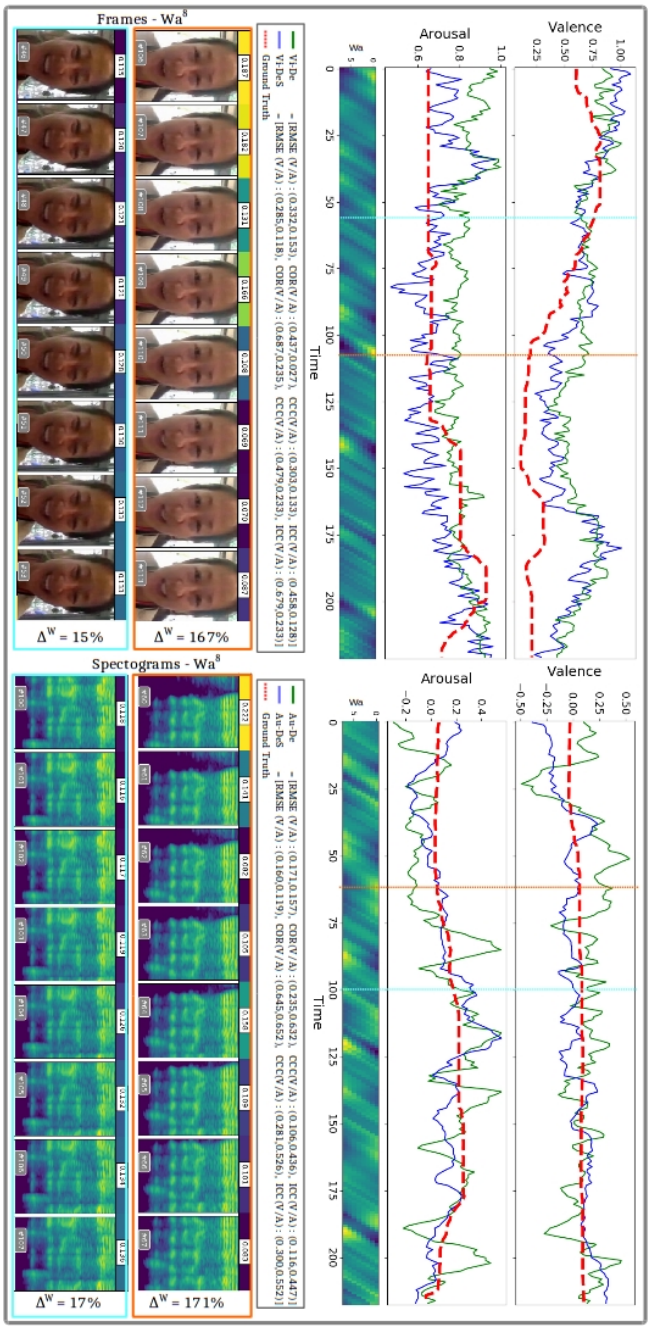


Figure 7.2: The impact of attention on both image and sound modalities as input to our model. The left part shows examples of sequence modelling with attention improving our model estimates in regards to the change captured on the visual input. The right part shows the changes captured with our attention modelling using sound inputs.

7.4.1 Dataset and Experiment Settings

We use two relevant affective datasets to provide comprehensive analysis and comparison of our models' results : SEMAINE [McKeown et al., 2010] and SEWA [Kossaifi et al., 2019] datasets.

- The SEMAINE dataset [McKeown et al., 2010] is a large audio-visual database built from the interactions between an agent and users from stimulated settings. It consists of recordings from 150 participants with a total of 959 conversations. Alongside the emotion labels, It also includes other annotations race, gender and fully transcribed character to allow rich data analysis.

- The SEWA dataset [Kossaifi et al., 2019] is a recently published affect dataset which consists of video and audio recording involving 398 subjects from multiple cultures. It is split into 538 sequences and meta-data (e.g. subject id, culture etc) are available alongside the actual affect ground truth of valence/arousal and liking/disliking.

In each experiment, we provide the results from the variants of our models to highlight the importance of each approach. All results are reported by following the original subject-independent protocol (5-fold cross validation) for both datasets. The RMSE and COR metrics[Kossaifi et al., 2019, Ringeval et al., 2019] are calculated for both datasets, adding also the ICC and CCC metrics for the SEMAINE and SEWA datasets, respectively, to facilitate the quantitative comparison to other results reported in the literature.

7.4.2 Single-Modality and Sequence Modelling Analysis

Tables 7.1 and 7.2 provide the comparisons of each of our single-modality approaches with denoising and sequential modelling. In these tables, we can see that the result of the Vi-DeS network that utilises visual input, produces higher accuracy in the Valence domain, while the Au-DeS network attains higher accuracy in the Arousal domain. These results confirm the previously reported studies [Kossaifi et al., 2019, Kossaifi et al., 2017]

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)
Vi-De	(0.47,0.48)	—	(0.31,0.15)	—	(0.42,0.49)	—	(0.35,0.24)	—	(0.55,0.59)	—	(0.12,0.09)	—
Vi-DeS	(0.45,0.46)	$\Downarrow$ (3.54%, 2.63%)	(0.32,0.16)	$\Uparrow$ (2.80%, 6.49%)	(0.38,0.45)	$\Downarrow$ (8.33%, 8.52%)	(0.38,0.28)	$\Uparrow$ (7.71%, 16.7%)	(0.49,0.52)	$\Downarrow$ (10.5%, 11.9%)	(0.18,0.15)	$\Uparrow$ (62.2%, 46.1%)
Au-De	(0.40,0.32)	—	(0.19,0.45)	—	(0.35,0.25)	—	(0.15,0.21)	—	(0.29,0.29)	—	(0.07,0.11)	—
Au-DeS	(**0.39**,**0.30**)	$\Downarrow$ (3.6%, 7.1%)	(**0.2**,**0.53**)	$\Uparrow$ (10.8%, 16.9%)	(**0.30**,**0.20**)	$\Downarrow$ (15.2%, 18.9%)	(**0.20**,**0.35**)	$\Uparrow$ (34.5%, 61.3%)	(**0.14**,**0.14**)	$\Downarrow$ (51.2%, 49.4%)	(**0.12**,**0.25**)	$\Uparrow$ (65.6%, 119%)

Table 7.3: The relative impacts of attentions on their level of relative differences (Δ^{Wa}) on the involved sequences on SEMAINE Dataset.

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)	$\downarrow$ *RMSE* (*V*,*A*)	GAIN (V,A)	$\uparrow$ *COR* (*V*,*A*)	GAIN (V,A)
Vi-De	(0.36,0.37)	—	(0.46,0.36)	—	(0.35,0.36)	—	(0.43,0.32)	—	(0.44,0.40)	—	(0.22,0.18)	—
Vi-DeS	(0.34,0.37)	$\Downarrow$ (0.68%, 4.54%)	(0.47,0.38)	$\Uparrow$ (0.95%, 3.44%)	(0.31,0.34)	$\Downarrow$ (6.83%, 10.9%)	(0.45,0.36)	$\Uparrow$ (6.08%, 11.1%)	(0.38,0.33)	$\Downarrow$ (13.5%, 18.4%)	(0.40,0.32)	$\Uparrow$ (78.8%, 81.4%)
Au-De	(0.46,0.45)	—	(0.19,0.29)	—	(0.40,0.47)	—	(0.25,0.34)	—	(0.55,0.59)	—	(0.09,0.12)	—
Au-DeS	(**0.45**,**0.44**)	$\Downarrow$ (2.73%, 3.84%)	(**0.20**,**0.31**)	$\Uparrow$ (7.24%, 5.35%)	(**0.37**,**0.44**)	$\Downarrow$ (6.19%, 8.32%)	(**0.29**,**0.40**)	$\Uparrow$ (18.2%, 15.5%)	(**0.47**,**0.50**)	$\Downarrow$ (15.7%, 15.6%)	(**0.13**,**0.16**)	$\Uparrow$ (40.7%, 30.6%)

Table 7.4: The relative impacts of attentions on their level of relative differences (Δ^{Wa}) on the involved sequences on SEWA Dataset.

that these modalities are more relevant to each of these domains due to the very nature of each modality.

In these results, we further notice an increase in accuracy for both of our baseline Vi and Au networks when we add the Denoiser operations. This finding is in agreement with our previous work [Aspandi et al., 2020a], where we found that the inclusion of the denoiser improves the robustness of the learnt latent features, leading to higher accuracy. In regard to this, examples of the denoising results for both image and audio input of our models can be seen in Figure 7.3. Notice that our models can clean both input modalities quite well, which is remarkable considering the different characteristics of these modalities.

Another finding is that the activation of temporal modelling provides further improvement of the accuracy, which confirms the benefit of including such sequential inputs [Aspandi et al., 2020b]. Examples of the learnt attention, in regards to sequence modelling can be seen in Figure 7.2. There, we can see examples for the results of models with attention enabled (Vi-DeS and Au-DeS) and disabled (Vi-De and Au-De). We also introduced the coefficient of Δ^{Wa} that is calculated from the percentage of the disparity between the minimum and maximum w values for each sequence. This Δ^{Wa} thus provides the level of the relative activation weight in the sequence. Based on this figure, we see that the overall results of our sequence-based modelling are more accurate for both modalities. We also notice that the attention mechanism is able to capture relevant changes for both modalities as well, since the attention intensity correlates well with changes observed in the input, as shown in this figure. For instance, in the first row, we see that both, the facial input and the spectrogram changed slightly compared to the second rows, and this is reflected in the respective attention intensities on top of each figure. This capability of our sequence modelling to capture such changes, explains the observed gain of the accuracy [Aspandi et al., 2020b].

To quantitatively confirm the above analysis, we consider three levels of differences for Δ^{Wa}: 12.5%, 25%, and 50%, and evaluate the portion of the datasets with Δ^{Wa} activation weights above each level. Visual examples, as well as the portion size and accuracy computed for each level Δ^{Wa} can

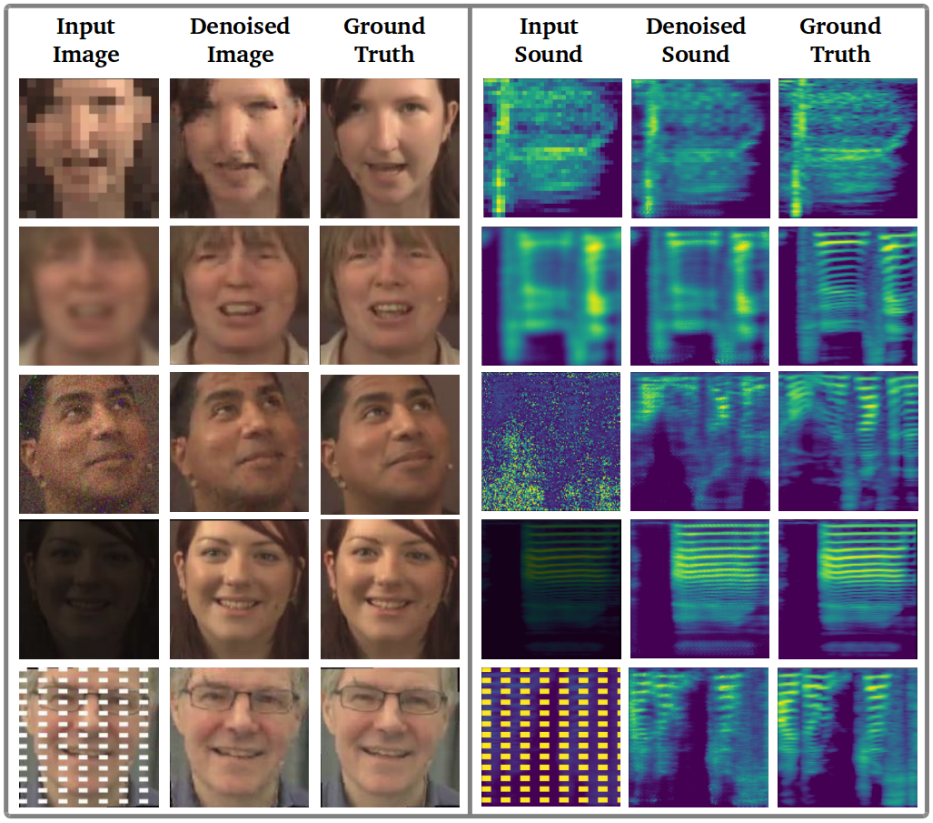


Figure 7.3: Visual examples of our denoised input of both modalities. Columns 1 and 4 show the noisy inputs. Next, columns 2 and 5 show the corresponding denoised examples of our models. Finally, columns 3 and 6 show the ground-truth, e.g the clean versions.

be seen in Figure 7.4. We see that in general, there is a decrease in the size of the frames included as the Δ^{Wa} accompanied by a raise in accuracy.

Tables 7.3 and 7.4 further show the accuracy of our single-based models without (Au-De and Vi-De) and with the sequences (Au-DeS and Vi-DeS) for different Δ^{Wa} values. Based on these results, we observe that the accuracy gain (in terms of RMSE and COR) increases, as we raise Δ^{Wa} for both datasets. For instance, the highest accuracy gain at $\Delta^{W} = 12.5$ is 16%

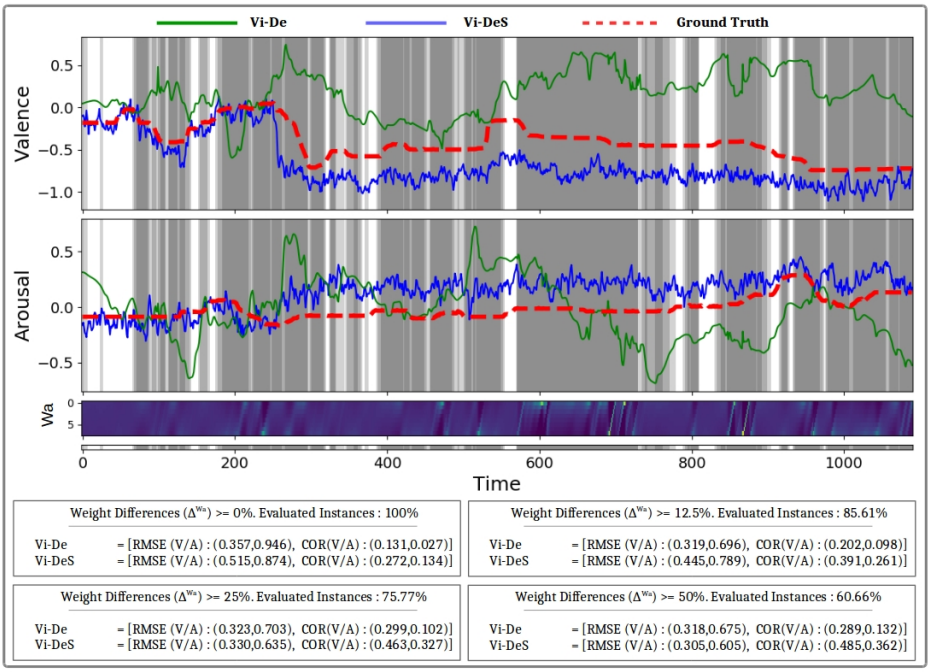


Figure 7.4: The examples of the results from Vi-DeS using several W value differences (Δ^{Wa}).

compared to more than 119% when the threshold is higher ($\Delta^{W} = 50\%$). This indicates that attention impacts the results more, when its weights are (relatively) high. Furthermore, we also see that the gain is pretty balanced across modalities, suggesting its compatibility to both types of input.

7.4.3 The Impact of Multi-Modality Approach with Concatenation and Internal Gating Mechanism

Table 7.5 and Table 7.6 present the results of our multi-modal approaches by means of concatenation (AVi-CaS) and Gating (AVi-GaS) together with the comparison with the best preforming results from previous sections of each modality. In this comparison, we can see that the combination of these

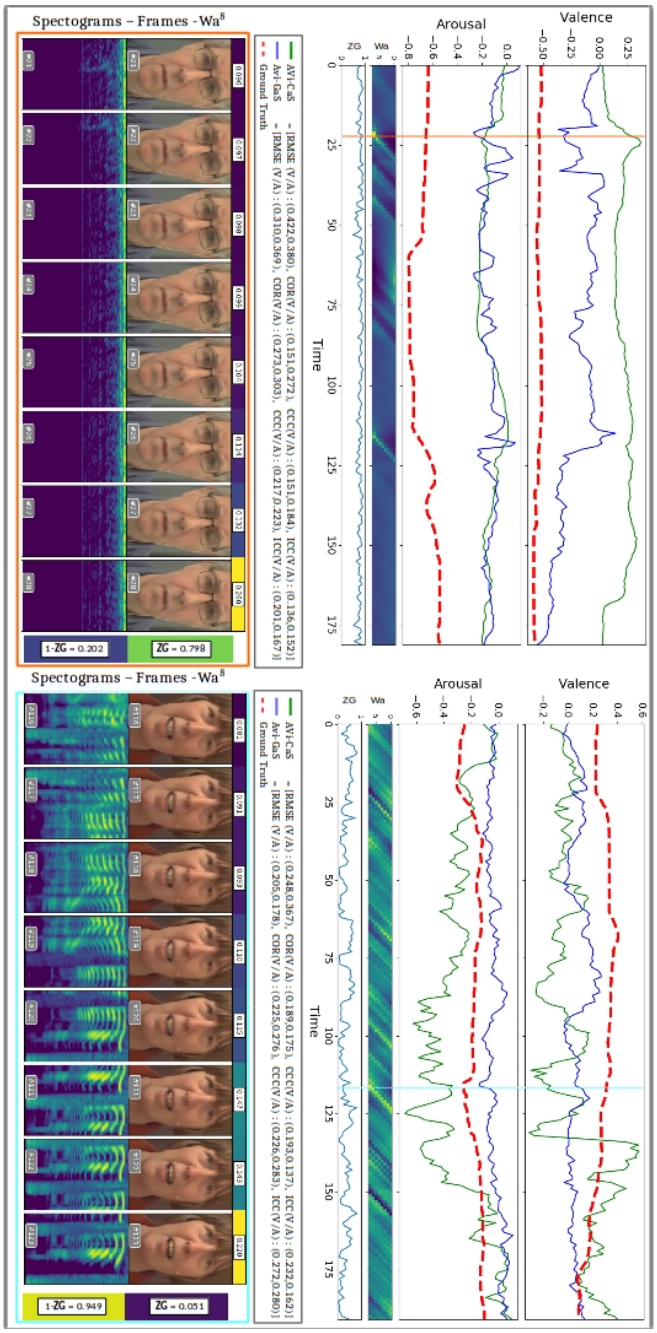


Figure 7.5: Example of the results of our model with concatenation (AViCaS) and gating (AviGaS) approaches. The bottom left and right show examples of how the internal *ZG* of AViGaS network detected the change happened on the visual and audio input respectively.

Model	RMSE ↓			COR ↑			CCC ↑			ICC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Vi-DeS	0.232	0.289	0.260	0.441	0.250	0.346	0.412	0.234	0.323	0.455	0.238	0.347
Au-DeS	0.306	0.226	0.266	0.272	0.509	0.390	0.250	0.495	0.372	0.262	0.500	0.381
AVi-CaS	0.239	0.219	0.229	0.459	0.516	0.487	0.405	0.507	0.456	0.466	0.512	0.489
AVi-GaS	**0.224**	**0.180**	**0.202**	**0.618**	**0.656**	**0.637**	**0.587**	**0.642**	**0.615**	**0.600**	**0.650**	**0.625**

Table 7.5: The results of our multi-modality approach of concatenation and gating mechanisms compared to the single-modality based approaches on the SEMAINE dataset.

Model	RMSE ↓			COR ↑			CCC ↑			ICC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Vi-DeS	0.328	0.333	0.331	0.501	0.400	0.450	0.476	0.398	0.437	0.520	0.405	0.462
Au-DeS	0.342	0.327	0.334	0.430	0.534	0.482	0.420	0.520	0.470	0.424	0.528	0.476
AVi-CaS	0.321	0.312	0.317	0.541	0.510	0.525	0.525	0.502	0.513	0.535	0.506	0.521
AVi-GaS	**0.282**	**0.282**	**0.282**	**0.697**	**0.604**	**0.651**	**0.686**	**0.583**	**0.634**	**0.693**	**0.589**	**0.641**

Table 7.6: The results of our multi-modality approach of concatenation and gating mechanisms compared to the single-modality based approaches on the SEWA dataset.

modalities yields an increase in accuracy for both approaches with more balanced results in both domains. This suggests the benefit of aggregating these modalities. However, comparing the results of AVi-GaS with AVi-CaS, we find that in general, the results from our gating mechanisms are better than the basic concatenation approach. This supports the need of more sophisticated approaches to combine these modalities.

Examples of the effectiveness of our gating approach compared to the standard concatenation counterparts are visualised in Figure 7.5, where we can see more accurate predictions of our gating mechanisms over the other compared models. The respective bottom sections provide two different examples of learnt *ZG* coefficients that are able to 'control' the importance of each modality. That is, in the the first column, we can see examples

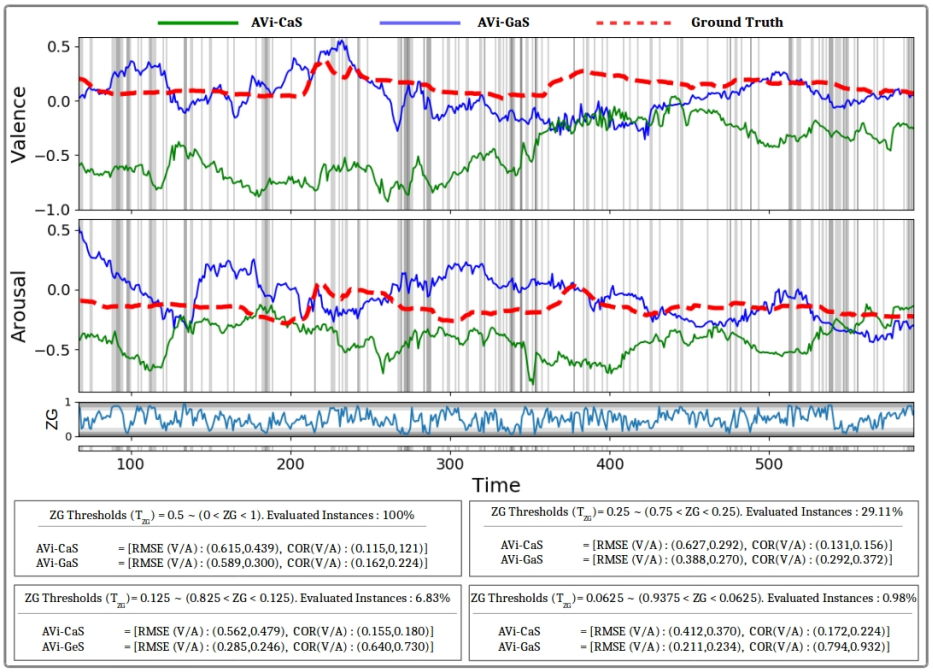


Figure 7.6: The visualization of the results on AVi-GaS using several threshold (T_{ZG}) values of 0.5,0.25, 0.125 and 0.0625. The first row shows the example of the area of each respective T_{ZG} values. The second row provides associated accuracy.

where the higher values of ZG indicate changes detected in the visual features. This is also synchronously detected by the sequence attention. We also see the other instances, where ZG is able to detect changes in the sound features, thus giving a higher priority to this modality. We also see, again, that these coefficients correlate well with the sequence activation, in line with the perceived activation of ZG. All of this explains the quite substantial improvement on accuracy of the AVi-GaS network compared to the marginal improvement achieved by the AVi-CaS network with respect to the single-modality variants.

Analogously to the analysis of high level of attention weight activations in the previous section, we evaluate the importance of ZG using different

thresholds T_{ZG}. Specifically, we chose three different T_{ZG} thresholds (0.25, 0.125, and 0.0625) that affect the range of T_{ZG}. However, because we are now evaluating a gating block, we are interested in deviations around the central value (i. e., 0.5); for instance, with $T_{ZG} = 0.125$, we evaluate $ZG < 0.125$ and $ZG > 0.875$. Examples of the considered segments using these different thresholds with their respective accuracy can be seen in Figure 7.6. Similar to the analysis in the previous section, we can see that increasing the threshold reduces the amount of data that is considered, but also raises the associated accuracy.

Tables 7.9 and 7.10 show the quantitative results of our $AVi-GaS$ with respect to the different threshold T_{ZG} for both, SEMAINE and SEWA, while the results of different Δ^{Wa} can be seen in the Tables 7.7 and 7.8, respectively. Notice that the results of the AVi-GaS networks are consistently better than those of the other models, including the concatenation-based approach AVi-CaS. An analysis of the ZG values reveals that the accuracy improvement grows as the threshold values decrease (which implies using only the most activated ZG values), showing the benefit of the gating approach in successfully controlling each modality to boost performance. For example, the highest gains with the threshold at 0.25 are 13.4% and 17.5%, but they grow to 35% and 20% for the SEMAINE and SEWA datasets using the narrowest threshold value of 0.0625.

7.4.4 Comparison to the State of the Art

In this section, we present the comparison of our best performing models from the previous ablation analysis (AVi-GaS network), including the results of our single-modality-based models (the Au-DeS, and Vi-DeS networks) against other alternative models on both, the SEMAINE and SEWA datasets. Specifically, we compare our model to the following ones:

1. FT-DCNN [Kossaifi et al., 2017], a hybrid deep learning based system that uses handcrafted visual and geometrical facial features.

2. BLSTM-WS [Yang and Hirschberg, 2018], a sequence based neural

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)
Vi-DeS	(0.42,0.45)	⇓ (7.50%,21.9%)	(0.17,0.13)	⇑ (147%,136%)	(0.40,0.44)	⇓ (17.9%,21.4%)	(0.33,0.24)	⇑ (70.1%,56.2%)	(0.41,0.41)	⇓ (22.6%,19.8%)	(0.41,0.33)	⇑ (25.6%,44.6%)
Au-DeS	(0.49,0.42)	⇓ (24.8%,13.1%)	(0.18,0.19)	⇑ (65.6%,91.4%)	(0.43,0.40)	⇓ (24.2%,11.1%)	(0.26,0.32)	⇑ (80.1%,28.2%)	(0.40,0.41)	⇓ (19.2%,20.9%)	(0.36,0.40)	⇑ (62.0%,27.6%)
AVi-CaS	(0.45,0.42)	⇓ (20.0%,17.3%)	(0.25,0.21)	⇑ (27.7%,44.9%)	(0.40,0.41)	⇓ (22.2%,17.0%)	(0.36,0.34)	⇑ (35.1%,17.8%)	(0.39,0.39)	⇓ (21.7%,17.6%)	(0.44,0.39)	⇑ (30.1%,24.1%)
AVi-GaS	(**0.37**,**0.36**)	—	(**0.35**,**0.39**)	—	(**0.33**,**0.35**)	—	(**0.55**,**0.42**)	—	(**0.32**,**0.33**)	—	(**0.63**,**0.51**)	—

Table 7.7: The results of our models variant compared to our full model of AVi-GaS on their level of relative differences (Δ^{Wa}) on SEWA dataset.

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)
Vi-DeS	(0.43,0.46)	⇓ (5.87%,14.6%)	(0.37,0.34)	⇑ (2.70%,9.17%)	(0.42,0.43)	⇓ (21.52%,20.6%)	(0.40,0.35)	⇑ (16.01%,26.2%)	(0.30,0.36)	⇓ (7.25%,27.82%)	(0.41,0.30)	⇑ (22.5%,82.7%)
Au-DeS	(0.42,0.40)	⇓ (4.14%,2.49%)	(0.29,0.30)	⇑ (31.5%,21.9%)	(0.46,0.43)	⇓ (28.52%,20.9%)	(0.33,0.34)	⇑ (41.92%,28.7%)	(0.38,0.30)	⇓ (26.7%,13.96%)	(0.32,0.35)	⇑ (57.1%,54.8%)
AVi-CaS	(0.43,0.40)	⇓ (4.50%,0.89%)	(0.33,0.36)	⇑ (16.8%,2.37%)	(0.41,0.36)	⇓ (19.60%,6.56%)	(0.40,0.42)	⇑ (15.69%,4.43%)	(0.33,0.29)	⇓ (15.0%,11.49%)	(0.48,0.42)	⇑ (4.24%,29.6%)
AVi-GaS	(**0.41**,**0.39**)	—	(**0.38**,**0.37**)	—	(**0.33**,**0.34**)	—	(**0.46**,**0.44**)	—	(**0.28**,**0.26**)	—	(**0.50**,**0.54**)	—

Table 7.8: The results of our models variant compared to our full model of AVi-GaS on their level of relative differences (Δ^{Wa}) on SEMAINE dataset.

Model	T_{ZG} = 0.25,($ZG < 0.25, ZG > 0.75$)				T_{ZG} = 0.125, ($ZG < 0.125, ZG > 0.875$)				T_{ZG} = 0.0625, ($ZG < 0.0625, ZG > 0.9375$)			
	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)
Au-DeS	(0.32,0.29)	⇓ (6.70%,6.29%)	(0.43,0.46)	⇑ (4.71%,1.39%)	(0.34,0.31)	⇓ (11.9%,14.9%)	(0.44,0.48)	⇑ (14.5%,9.42%)	(0.31,0.30)	⇓ (4.79%,15.3%)	(0.46,0.50)	⇑ (22.1%,16.4%)
Vi-DeS	(0.30,0.31)	⇓ (1.90%,12.1%)	(0.44,0.41)	⇑ (2.81%,13.4%)	(0.30,0.33)	⇓ (2.03%,19.9%)	(0.46,0.41)	⇑ (9.90%,27.4%)	(0.31,0.34)	⇓ (4.61%,24.3%)	(0.45,0.43)	⇑ (24.8%,35.8%)
AVi-CaS	(0.31,0.28)	⇓ (4.68%,1.88%)	(0.44,0.45)	⇑ (3.39%,3.34%)	(0.38,0.37)	⇓ (21.3%,28.6%)	(0.44,0.47)	⇑ (15.4%,12.1%)	(0.33,0.31)	⇓ (9.45%,18.3%)	(0.47,0.49)	⇑ (21.63%,17.8%)
AVi-GaS	(**0.30**,**0.27**)	—	(**0.45**,**0.46**)	—	(**0.30**,**0.27**)	—	(**0.50**,**0.53**)	—	(**0.30**,**0.26**)	—	(**0.57**,**0.58**)	—

Table 7.9: The results of our models variant with respect to different threshold T_{ZG} of learned ZG on SEMAINE dataset.

Model	T_{ZG} = 0.25,($ZG < 0.25, ZG > 0.75$)				T_{ZG} = 0.125, ($ZG < 0.125, ZG > 0.875$)				T_{ZG} = 0.0625, ($ZG < 0.0625, ZG > 0.9375$)			
	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)	$\downarrow RMSE\ (V,A)$	GAIN (V,A)	$\uparrow COR\ (V,A)$	GAIN (V,A)
Au-DeS	(0.41,0.40)	⇓ (10.6%,7.01%)	(0.35,0.40)	⇑ (17.3%,6.57%)	(0.43,0.41)	⇓ (13.3%,15.1%)	(0.43,0.45)	⇑ (4.48%,1.62%)	(0.43,0.40)	⇓ (18.4%,6.27%)	(0.39,0.41)	⇑ (22.1%,4.11%)
Vi-DeS	(0.39,0.38)	⇓ (5.75%,2.03%)	(0.41,0.38)	⇑ (0.50%,12.5%)	(0.38,0.41)	⇓ (1.86%,13.7%)	(0.42,0.36)	⇑ (7.32%,25.3%)	(0.38,0.41)	⇓ (8.15%,8.89%)	(0.41,0.36)	⇑ (19.3%,20.9%)
AVi-CaS	(0.38,0.38)	⇓ (2.55%,3.87%)	(0.41,0.40)	⇑ (2.44%,7.97%)	(0.40,0.38)	⇓ (6.52%,8.36%)	(0.41,0.41)	⇑ (7.79%,11.6%)	(0.40,0.39)	⇓ (13.8%,5.68%)	(0.40,0.39)	⇑ (20.9%,10.4%)
AVi-GaS	(**0.37**,**0.37**)	—	(**0.42**,**0.43**)	—	(**0.38**,**0.35**)	—	(**0.45**, **0.45**)	—	(**0.35**,**0.37**)	—	(**0.48**,**0.43**)	

Table 7.10: The results of our models variant with respect to different threshold T_{ZG} of learned ZG on SEWA dataset.

network that utilises both direct sound wave input and its spectrogram derivatives as input to LSTM networks.

3. DialogueRNN [Poria et al., 2019], a deep learning model that uses multiple modalities such as visual, sound and text features that are aggregated using GRU networks modelling. Additionally, they also include the interaction properties on their modelling.

4. ResNet-18 [Kossaifi et al., 2019] is a CNN based model (with residual connection) that operates directly on the video frames to produce emotion estimates.

5. Tensor [Mitenkova et al., 2019], a tensor based neural network that processes visual input and is optimised using the tucker tensor regression.

6. Factorized [Kossaifi et al., 2020], a deep learning model that uses a similar approach to [Mitenkova et al., 2019] but uses generalised factorizations to allow more efficient decomposition.

7. AEG-CD-SZ [Aspandi et al., 2020a], our previous latent based approach that uses also adversarial training, using both visual and sound modalities.

8. ANCLaF-SA [Aspandi et al., 2020b] the precursor of our sequential modelling with attention, relying only on the visual input.

Table 7.11 and Table 7.12 provide the comparisons of the evaluated models on the SEMAINE and SEWA dataset, respectively. In general, we can see that the results of our full model (AVi-GaS) compare favourably with respect to other state of the art approaches on both datasets. Indeed, our model produces the highest accuracy in terms of both, the CCC and ICC metrics on SEMAINE. While on the SEWA datasets, we attain the best accuracy on Arousal, and rank only slightly lower than the Factorised model [Kossaifi et al., 2020] on Valence. In this respect, we need to note that in their according work [Kossaifi et al., 2020], the authors include

Model	Modalities	RMSE ↓			COR ↑			ICC ↑		
		VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
FT-DCNN	VIS	**0.160**	0.208	**0.165**	0.283	0.296	0.290	0.331	0.277	0.304
BLSTM-WS	VIS	-	-	-	**0.692**	0.423	0.558	0.320	0.210	0.265
DialogueRNN	VIS+AUD+TXT	0.165	**0.165**	**0.165**	0.350	0.590	0.470	-	-	
AEG-CD-SZ	VIS+AUD	0.303	0.262	0.283	0.175	0.301	0.238	0.173	0.291	0.232
ANCLaF-SA	VIS	0.258	0.297	0.278	0.410	0.279	0.345	0.423	0.298	0.360
Vi-DeS	VIS	0.232	0.289	0.260	0.441	0.250	0.346	0.455	0.238	0.347
Au-DeS	AUD	0.306	0.226	0.266	0.272	0.509	0.390	0.262	0.500	0.381
Avi-GaS	VIS+AUD	0.224	0.180	0.202	0.618	**0.656**	**0.637**	**0.600**	**0.650**	**0.625**

Table 7.11: Quantitative comparisons on the SEMAINE dataset.

far larger CNN models than we do (they evaluated three different sub-networks) to process the visual input, which may explain their slightly higher accuracy for the valence emotion domain.

Further observation shows that our single modality-based models (Vi-DeS and Au-DeS) also perform quite well, and even outperform some of the other alternatives. Their results are also slightly better when compared with our previous approaches, with comparable results for the choice Vi-DeS with ANCLaF-SA due to the similar designs. Lastly, we also notice that our models produce a quite balanced accuracy on both emotional dimensions (Valence / Arousal) when compared to other alternatives. That shows the effectiveness of our model to aggregate these modalities in support of each other.

Figure 7.7 visualises examples of the predictions of our AVi-GaS networks compared to our previous approaches (AEG-CD-SZ and ANCLaF-SA). Notice that our current results are more accurate than our previous results for both datasets. This highlights the importance of our proposed methods to increase the effectiveness of our models. Lastly, the results also show more balanced accuracy on both dimensions, compared to ANCLaF-SA that is comparatively less accurate for Arousal due to the lack of the audio modality.

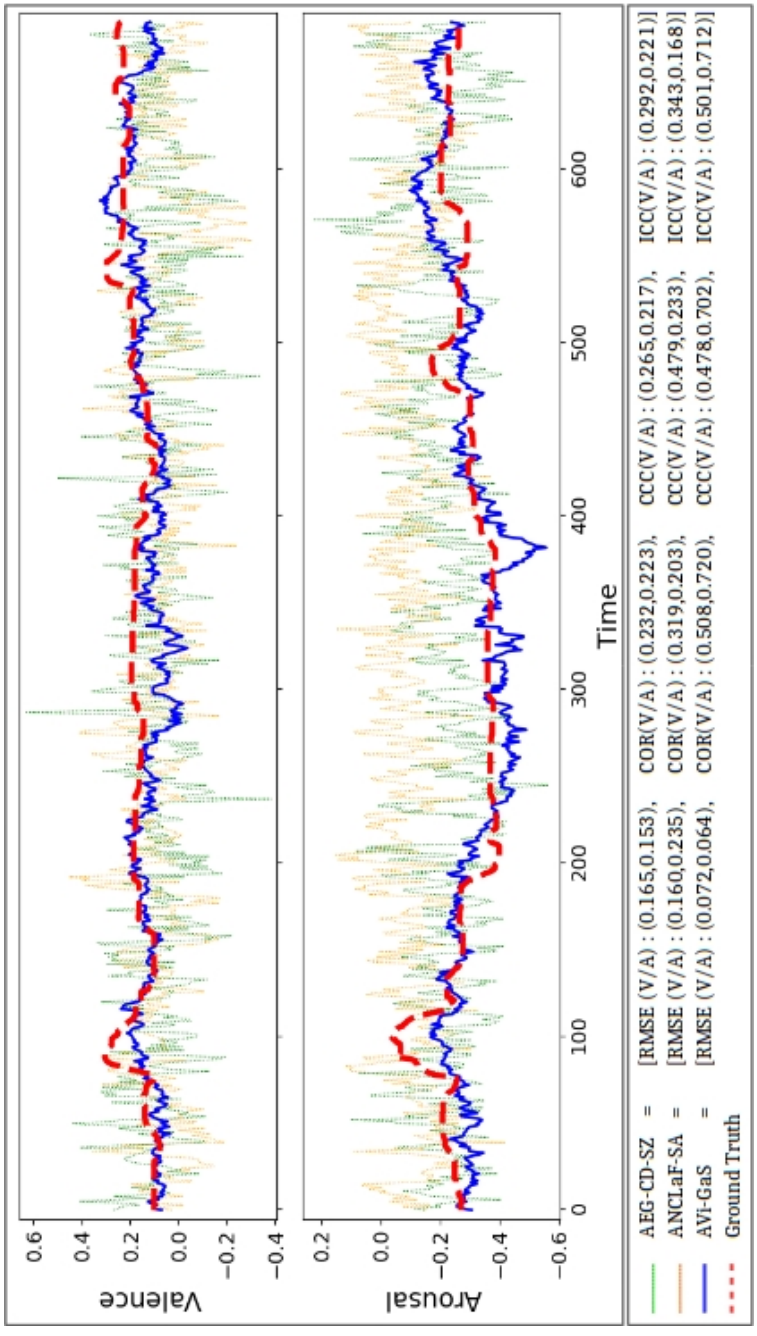


Figure 7.7: The comparison of our gated multi modality based sequence model approach (AviGaS) versus our previous approaches (AEG-CD-SZ and ANCLaF-SA). Notice that current proposed approach is able to produce accurate predictions on both emotion dimensions and outperforms our previous results.

Model	Modalities	RMSE ↓			COR ↑			CCC ↑		
		VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
ResNet-18	VIS	-	-	-	0.350	0.350	0.350	0.350	0.290	0.320
Tensor	VIS	0.334	0.380	0.357	0.503	0.439	0.471	0.469	0.392	0.431
Factorized	VIS	**0.240**	0.320	**0.280**	**0.840**	0.600	**0.720**	**0.750**	0.520	**0.635**
AEG-CD-SZ	VIS+AUD	0.323	0.350	0.337	0.442	0.478	0.460	0.405	0.430	0.418
ANCLaF-SA	VIS	0.336	0.328	0.332	0.558	0.332	0.445	0.405	0.529	0.467
Vi-DeS	VIS	0.328	0.333	0.331	0.501	0.400	0.450	0.476	0.398	0.437
Au-DeS	AUD	0.342	0.327	0.334	0.430	0.534	0.482	0.420	0.520	0.470
Avi-GaS	VIS+AUD	0.282	**0.285**	0.284	0.697	**0.604**	0.651	0.686	**0.583**	0.634

Table 7.12: Quantitative comparisons on the SEWA dataset.

7.5 Conclusions

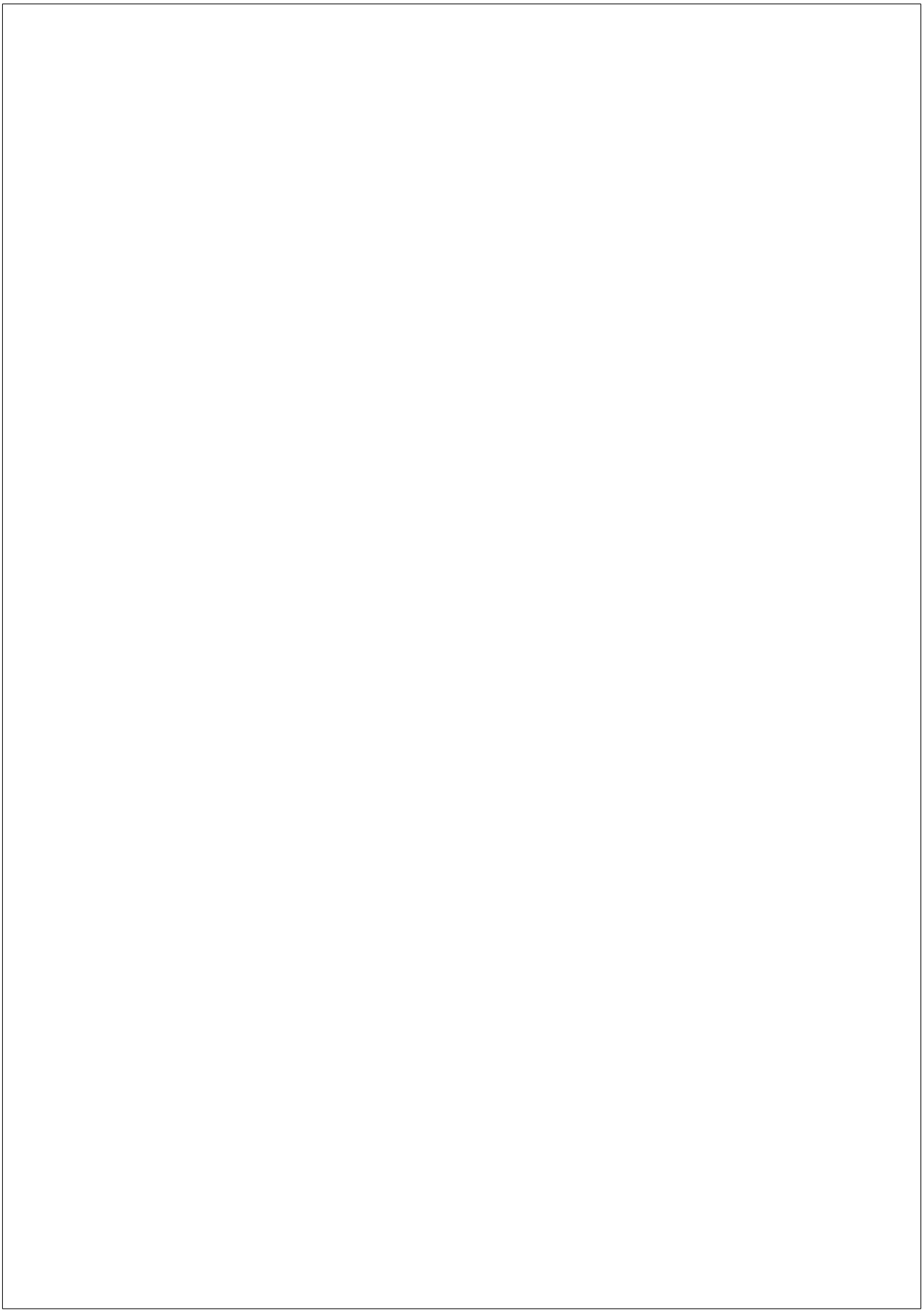
In this chapter, we presented multi-modal affect networks that are capable to efficiently process bi-modal inputs using our combined latent-based representations with sequence modelling and attention mechanisms. We then equipped our networks with gating to allow for more effective multi-modal fusion. We trained our models using adversarial learning to extract more representative latent features given the noisy inputs of both visual and sound modality. We then used these latent features from both modalities fused together with our gating mechanisms, and fed the result to our sequential modelling, which was trained through curriculum learning to allow for progressive training.

We demonstrated the effectiveness of our approach as a whole as well as for each of its components, on the two most widely used and accessible affective datasets: SEMAINE and SEWA. In our ablation studies, we firstly showed the impact of denoising to improve the base results of our single-modal input models, and we also observed the consistently cleaner results from noisy modality inputs; secondly, we find that sequence modelling with attention further improved the results in terms of accuracy and provided a detailed quantitative analysis to support this conclusion by thresholding on the relative learnt attention differences; thirdly, we

observed a noticeable gain in accuracy when both modalities were merged, either by concatenation or by our gating mechanism, the latter producing the highest improvement.

Finally, we compare our best performing models against current state of the art alternatives on both datasets, including our two previous approaches. In the comparison, we show that our model is able to consistently produce high accuracy, comparable to the top results from the state of the art, with an outstanding balance in the performance obtained for Valence and Arousal emotion dimension estimates with respect to alternative approaches.

Future efforts may consider other types of attention and an extension to other modalities such as physiology or textual cues. It will be exciting to see the benefits of the proposed architecture in a self-learning potentially cross-modal context.



Chapter 8

CONCLUSION AND FUTURE RESEARCH

In **Chapter 1**, we provide the descriptions of two fundamental tasks of Automatic Facial Analysis: Facial Alignment and Facial Behavior Analysis. Subsequently, we present existing challenges in these tasks that become the main focus of this thesis. These challenges mainly consist of the lack of analysis to the impact of incorporating the temporal modeling, that have been heavily utilised on the related machine learning fields. Furthermore, the in-the-wild characteristics of the big data alongside the size of the datasets present another layer of difficulty to improve current state of the arts in these task. We then describe our contributions to address these shortcomings by proposing our spatio-temporal modelling specifically applied to facial landmark estimations and facial-based emotion recognition tasks. This is considering their central roles to both Facial Alignment and Facial Behavior Analysis respectively. Lastly, we provide the summaries of each of chapter in this thesis to highlight the associated key research findings and contributions.

In **Chapter 2**, we introduce the use of temporal modelling to facial landmark estimations under facial tracking setting, which is still largely unexplored in the field. We do so by proposing a fully end-to-end composite facial tracker that uses internal temporal modelling by means of LSTM net-

work. We use progressive learning to train our model, that simultaneously allows us to examine the impacts of shorter and longer temporal context to the accuracy of landmark estimates. We present state of the art results of our models in our comparisons using the biggest facial tracking dataset: the 300-VW dataset. In this comparison, we mainly found that our full models, which utilize temporal modelling are performing better compared to our tracking by detection counterparts, suggesting the importance of such approach. We also further notice that sequence length between 2 to 4 frames as optimum one to consider in our approach. Nonetheless, we need to also note that these results are also dependent on the test sets of respective dataset.

Following our findings on facial tracking task, in **Chapter 3**, we evaluate the impact of synthetically degraded images to the performance of current facial alignment models, and further use this understanding to build a robust facial alignment model. This is done by combining final pipeline part of our facial tracker with internal specialized noise denoiser network. We conducted experiments to test the robustness of our model and other facial alignments against number of synthetic noises using 300-W and Menpo dataset. This systematic comparison reveals that the majority of deep learning models are severely deteriorated in case of blurring and down-sampling, but seems more resistant to gaussian noise and colour scaling. In contrast, our proposed models are less affected and are able to maintain their high accuracy results. Lastly, when we tested our model using 300-VW dataset under single image facial alignment and tracking setting, we also found that our proposed models are able to enhance the intermediate image taken in the wild. As such, it improves the quality of our models landmark predictions maintaining top state-of-the-art accuracy.

In **Chapter 4** we combine our previous sequence modellings with internal noise suppression technique resulted in an unifying approach to solve both single and multiple frames based facial alignment. We do so by proposing four sub-networks that are focused on their specialized task: facial detection, facial bounding-box tracking, facial region validator and facial alignment with internal denoising. One key aspect of our approach is to train our model with real-life noise models to increase their

robustness. Furthermore, we also propose to incorporate joint learning between sub-modules of facial alignment part given their mutual aims to achieve accurate landmark estimations. We compare the results of our proposed models on single-image facial alignment task using 300-W and Menpo datasets, and video-based facial tracking task using 300-VW dataset. In these comparisons, we found that our models consistently outperform other compared alternatives, with higher AUC value and lower NME and FR value respectively. Specific to facial tracking experiments, we also found that, in overall, tracking-by-detection approaches produce comparatively lower accuracy than the other models with some forms of temporal modeling, that supports the advantage of incorporating this temporal information. Additionally, we also perform the ablation study to confirm the benefit of each of our sequential and denoising approach in maintaining our state of the art results. Finally, we found sequence length of between 2 to 8 as optimum one, that are related to the test sequence used in the experiment.

In **Chapter 5** we expand our image degradation modelling to affect recognition in the form of latent features that are extracted through adversarial learning. We do so by utilizing paired generator and conditional discriminator that are trained using adversarial setting. Specifically, the generator is trained using progressive learning to create the cleaned versions of distorted image inputs, and simultaneously extract the compact visual kernels features. The discriminator then takes the concatenated input images with the projected audio features, to be internally aggregated by latent visual features to estimate both real/fake identity of current input and final valence and arousal estimates. We tested the capability of our models using two recently published datasets of SEWA, and Aff-Wild2 dataset that is part of ABAW challenge. In this experiment, we see the progressive improvements made from each of our approaches that ultimately lead to our competitive results on both datasets, confirming the benefit of these approaches.

In **Chapter 6** we explore the use of our previous sequential modelling from facial alignment to affect recognition by adapting it to our adversarial based visual latent feature extractions, that further enhanced by sequential

modelling with attention. This is done by introducing a Combiner network to the existing paired Generator and Discriminator setup. The pipeline of this combined model is started by generator receiving distorted images and produces both latent visual features and its cleaned versions. Then, the discriminator produces both real/fake identity of input images, and associated first step emotion quadrant. Subsequently, the combiner network receives latent visual features that are conditioned by emotion quadrant information to predict the respective valence/arousal value. In regard to this, the base version of our combiner directly infers the valence/arousal value, while the respective complete version also incorporates temporal modeling using LSTM, that is enhanced by the use of internal sequence weighting of attentions to process the conditioned visual features. We show the effectiveness of our approach by reporting top state of the art results on two of the most widely used video datasets for affect analysis, namely AFEW-VA and SEWA. In this comparison, we found that our base results are quite comparable to the state of the art. In other hand, our more advanced model, that uses internal temporal modeling yields higher accuracy in terms of qualitative and quantitative comparison. Within our sequence modelling, we observed that the highest accuracy improvements occurred when attention mechanisms are activated. Further detailed inspection also reveals the correlation between the attention intensity and observed facial movements. Finally, we found that the sequence length around 160 ms to be as optimum one, consistent with our previous findings.

Finally in **Chapter 7**, we investigate the benefit of the use of gating mechanisms to allow more effective multi-modal fusion onto our latent-based sequence modelling with attention for affect recognition. We build our model by combining multiple sub-networks, namely Generator, Discriminator and Combiner with attention-enhanced sequence modeling enabled. The pipeline of our approach started by the formation of latent visual features from both distorted images and Spectrogram input by Generator. These latent features are then combined with quadrant information produced by Discriminator during adversarial learning with Generator. The conditioned features are further passed into our Combiner network with gating mechanism to control the importance of both modality inputs.

Lastly, these fused information are used by our LSTM modeling with attention enabled to be trained using our progressive learning to estimate the final affect dimension. To demonstrate the benefit of each of our approach, we first present the ablation analysis the results of our models variant using both SEMAINE and SEWA dataset. In this analysis, we found the consistent findings with our previous investigations, that both denoising and sequence modeling with attention help to improve the accuracy of our estimates for both modality inputs. Furthermore, we show the effectiveness of our internal gating mechanisms to fuse both modalities, as observed by their higher accuracy compared to our models variants that use standard concatenation approach. We also notice the constant increases of our models performance in our throughout analysis on both attention weight and gating coefficients, demonstrating their positive influences to our models performance. Lastly, the comparisons of our best performing models against other alternatives, including our previous approach, show the superiority of our approach demonstrated by their state of the art accuracy.

In **conclusion**, research presented on this thesis have demonstrated the evidences of the importance and benefit of spatio-temporal modelling for both facial alignment and facial emotion recognition tasks, that are fundamental parts of facial analysis field. We have learned and showed several key aspects that serve the important roles in enabling our proposed models to achieve high accuracy results on their targeted tasks. First respective aspect is the throughout analysis of the length of sequence to be incorporated, that we found to be crucial to allow for optimal results. Another aspect is the incorporation of internal noise modelling that facilitates the reduction of the inherent input image degradations, which are important in dealing with data that are taken in the wild. The use and analysis of attention modellings are also significant in allowing the sequence modelling to adapt their focus on the changes perceived through the sequences, while simultaneously allow the inspections of our models inner work. Next, the utilization of several modalities that are aggregated by gating mechanisms for more effective feature fusions are also relevant to fully benefit from each data stream characteristics. Lastly, the use of

fully end-to-end approaches and adoptions of several learning mechanisms, such as curriculum learning and adversarial learning to enable more comprehensive model training, and to scale well in regards to the sheer size of currently available datasets.

In our opinion, the **future work** in this research directions will gain significant traction, thus attentions, given its ability to make use of inherent characteristics from big-data that are becoming more accessible. There are several potential venues that the presented approaches may also feasibly adapted to benefit from their capacity, this can include Economic (e.g stock market analysis [Vỳrost et al., 2015]), Health (e.g pandemic evolution monitoring [Tang et al., 2020]), Agriculture (e.g crop progress analysis [Dong et al., 2017]) and so on. Another future research line to explore, based on the presented research can be in terms of the other learning schemes to be used in the investigation, such as the use of Fully Unsupervised Learning [Bengio et al., 2012] and Reinforcement Learning [Mnih et al., 2016] to allow for more flexible training and to increase current learning capacity for potentially larger scale of dataset size. Lastly, the use of recent approaches as the core of sequential modelling including Neural Turing Machine [Gulcehre et al., 2018] and Transformer Network [Vaswani et al., 2017] may potentially be explored to introduce another aspect of the way the sequential modelling is designed. Although we also need to consider their stability and adaptability to current systems that have been relying on the current approaches. As a **final remark**, we hope that research presented in this thesis can encourage other investigations in this research lines of spatio-temporal modelling, that in our opinion, may ultimately help us to advance current state of the arts in the field.

Appendix A

MODEL DEFINITIONS

A.1 Models Definitions of ANCLaF

Layer	Kernel size	Activation	Filters No.	Stride	Input
Conv1	4x4	ReLU+instanceNorm2D	64	2	Input
Conv2	4x4	ReLU+instanceNorm2D	128	2	Conv1
Conv31	3x3	ReLU+instanceNorm2D	128	1	Conv3
Conv32	3x3	ReLU+instanceNorm2D	128	1	Conv31
Conv33	3x3	ReLU+instanceNorm2D	128	1	Conv32
Layer-wise addition [R1]	-	ReLU+instanceNorm2D	Conv3 + Conv33	-	-
Conv41	3x3	ReLU+instanceNorm2D	128	1	R1
Conv42	3x3	ReLU+instanceNorm2D	128	1	Conv41
Conv43	3x3	ReLU+instanceNorm2D	128	1	Conv42
Layer-wise addition [Z]	-	ReLU+instanceNorm2D	R1 + Conv43	-	-
DConv1	4x4	ReLU+instanceNorm2D	128	2	Z
DConv2	4x4	ReLU+instanceNorm2D	64	2	DConv1
Conv5	3x3	TanH	3	1	DConv2

Table A.1: Architecture of the Generator Network (G).

Layer	Kernel size	Activation	Filters No.	Stride	Input
Conv1	4x4	LReLU	64	2	Input
Conv2	4x4	LReLU	128	2	Conv1
Conv3	4x4	LReLU	256	2	Conv2
Conv4	4x4	LReLU	512	2	Conv3
Conv5	4x4	LReLU	1024	2	Conv4
Conv61 [Real/Fake]	3x3	LReLU	512	1	Conv5
Conv62 [Q])	3x3	LReLU	512	1	Conv5

Table A.2: Architecture of the Discriminator Network (D).

Layer	Kernel size	Activation	Filters No.	Stride	Input
Conv1	4x4	LReLU	64	2	Input
Conv2	4x4	LReLU	128	2	Conv1
Conv3	4x4	LReLU	256	2	Conv2
Conv4	4x4	LReLU	512	2	Conv3
Conv5	4x4	LReLU	1024	2	Conv4
W_a	512	LReLU	-	-	Conv5, h_lstm
LSTM	512	LReLU	-	-	[Conv5, W_a*Conv5]
FC [V/A]	512	LReLU	-	-	o_LSTM

Table A.3: The architecture of the sequence based combiner with attention (C).

Appendix B

ADDITIONAL RESULTS

B.1 Impact of Synthetic noise on facial landmark estimation experiments

Table B.1,B.2,B.3 and B.4 show the AUC and FR value of each models on the noisy version of 300-W and Menpo dataset.

Method	300-W-Test-N1						Menpo-Test-N1					
	0.5		0.25		0.125		0.5		0.25		0.125	
	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR
ECT [Zhang et al., 2018a]	0.765	0.008	0.662	0.018	0.306	0.228	0.787	0.015	0.689	0.029	0.332	0.238
SAN [Dong et al., 2018]	**0.789**	0.005	0.649	0.020	0.001	0.996	0.794	0.016	0.643	0.061	0.001	0.997
FAN [Bulat and Tzimiropoulos, 2017]	0.679	**0.000**	0.592	**0.001**	0.051	0.853	0.706	0.003	0.617	0.009	0.073	0.800
TCDNN [Zhang et al., 2016b]	0.667	0.013	0.587	0.016	0.349	0.100	0.695	0.018	0.616	0.031	0.378	0.1074
CFSS [Zhu et al., 2015a]	0.767	0.003	0.707	0.010	0.523	**0.025**	0.784	0.021	0.714	0.032	0.519	0.0594
ERT [Kazemi and Sullivan, 2014]	0.497	0.216	0.425	0.243	0.237	0.401	0.549	0.206	0.473	0.232	0.288	0.352
FLL	0.773	0.001	0.692	0.003	0.429	0.058	0.833	**0.000**	0.738	**0.001**	0.448	0.057
FADeNN-D	0.773	0.002	0.749	0.003	0.581	0.056	0.832	**0.000**	0.805	**0.001**	0.619	0.048
FADeNN-S	0.773	0.001	0.757	0.006	0.632	0.035	**0.837**	**0.000**	**0.838**	0.002	0.681	0.025
FADeNN-M	0.773	0.001	**0.776**	0.003	**0.633**	0.035	**0.839**	**0.000**	**0.838**	**0.001**	**0.683**	**0.024**

Table B.1: AUC and FR for different level of down-sampling on 300-W-Test-N1 and Menpo-Test-N1.

Method	300-W-Test-N2						Menpo-Test-N2					
	$\sigma_{gb} = 1$		$\sigma_{gb} = 2$		$\sigma_{gb} = 5$		$\sigma_{gb} = 1$		$\sigma_{gb} = 2$		$\sigma_{gb} = 5$	
	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR
ECT [Zhang et al., 2018a]	0.778	0.007	0.638	0.042	0.299	0.365	0.802	0.010	0.687	0.036	0.328	0.356
SAN [Dong et al., 2018]	0.796	0.005	0.625	0.025	0.186	0.515	0.801	0.015	0.647	0.052	0.208	0.513
FAN [Bulat and Tzimiropoulos, 2017]	0.714	**0.000**	0.601	0.092	0.078	0.845	0.741	0.002	0.660	0.049	0.092	0.834
TCDNN [Zhang et al., 2016b]	0.673	0.012	0.534	0.055	0.308	0.213	0.707	0.019	0.567	0.068	0.316	0.259
CFSS [Zhu et al., 2015a]	0.771	0.003	0.701	0.013	0.582	0.050	0.789	0.019	0.726	0.032	0.608	0.072
ERT [Kazemi and Sullivan, 2014]	0.535	0.208	0.477	0.243	0.410	0.275	0.582	0.199	0.525	0.2372	0.455	0.275
FLL	**0.796**	0.003	0.678	0.032	0.172	0.638	**0.865**	**0.000**	0.753	0.019	0.194	0.621
FADeNN-D	0.782	0.003	0.725	0.008	0.547	0.121	0.848	0.001	0.779	0.004	0.586	0.110
FADeNN-S	0.793	**0.000**	**0.742**	0.007	0.646	0.033	0.861	0.001	**0.804**	0.002	0.638	0.066
FADeNN-M	0.791	**0.000**	**0.742**	**0.006**	**0.647**	**0.032**	0.856	0.001	**0.804**	**0.001**	**0.640**	**0.064**

Table B.2: AUC and FR for different level of σ_{gb} on gaussian blurring of 300-W-Test-N2 and Menpo-Test-N2.

Method	300-W-Test-N3						Menpo-Test-N3					
	σ^2_{gn} = 0.001		σ^2_{gn}= 0.005		σ^2_{gn} = 0.01		σ^2_{gn} = 0.001		σ^2_{gn} = 0.005		σ^2_{gn} = 0.01	
	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR
ECT [Zhang et al., 2018a]	0.727	0.038	0.669	0.067	0.593	0.110	0.751	0.040	0.680	0.084	0.606	0.138
SAN [Dong et al., 2018]	0.746	0.025	0.697	0.038	0.635	0.006	0.765	0.029	0.710	0.056	0.650	0.087
FAN [Bulat and Tzimiropoulos, 2017]	0.683	0.010	0.603	0.081	0.460	0.261	0.711	0.018	0.651	0.069	0.549	0.182
TCDNN [Zhang et al., 2016b]	0.629	0.017	0.553	0.042	0.469	0.097	0.666	0.032	0.592	0.071	0.514	0.117
CFSS [Zhu et al., 2015a]	0.750	0.012	0.720	0.023	0.686	0.028	0.766	0.032	0.780	0.049	0.699	0.007
ERT [Kazemi and Sullivan, 2014]	0.473	0.247	0.378	0.340	0.297	0.403	0.530	0.228	0.442	0.305	0.364	0.372
FLL	0.734	0.020	0.675	0.037	0.602	0.062	0.795	0.008	0.721	0.023	0.650	0.052
FADeNN-D	0.777	0.007	0.763	0.008	0.745	0.010	0.821	0.004	0.768	0.016	0.714	0.041
FADeNN-S	**0.779**	**0.003**	0.770	**0.005**	0.757	**0.007**	0.795	0.008	**0.771**	0.015	0.716	0.034
FADeNN-M	**0.779**	**0.003**	**0.771**	**0.005**	**0.763**	**0.007**	**0.806**	**0.003**	0.760	**0.010**	**0.724**	**0.023**

Table B.3: AUC and FR for different level of σ^2_{gn} on gaussian noises of 300-W-Test-N3 and Menpo-Test-N3.

Method	300-W-Test-N4						Menpo-Test-N4					
	s = 0.8		s = 0.5		s = 0.2		s = 0.8		s = 0.5		s = 0.2	
	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR	AUC	FR
SAN [Zhang et al., 2018a]	0.780	0.008	0.773	0.005	0.739	0.117	0.802	0.012	0.799	0.012	0.773	0.015
SAN [Dong et al., 2018]	0.801	0.005	0.793	0.005	0.753	0.007	0.806	0.016	0.799	0.017	0.758	0.022
FAN [Bulat and Tzimiropoulos, 2017]	0.713	**0.000**	0.712	**0.000**	0.706	0.002	0.740	0.001	0.739	0.002	0.734	0.005
TCDNN [Zhang et al., 2016b]	0.678	0.001	0.677	0.001	0.673	**0.001**	0.710	0.016	0.709	0.016	0.706	0.017
CFSS [Zhu et al., 2015a]	0.780	0.001	0.778	0.003	0.776	0.003	0.796	0.018	0.795	0.019	0.795	0.018
ERT [Kazemi and Sullivan, 2014]	0.518	0.202	0.452	0.273	0.212	0.458	0.567	0.203	0.508	0.248	0.257	0.440
FLL	0.795	0.002	**0.782**	0.005	0.694	0.033	0.855	**0.000**	0.830	**0.001**	0.800	0.002
FADeNN-D	0.782	0.002	0.760	0.005	0.752	0.005	0.843	0.001	0.818	**0.001**	0.797	0.003
FADeNN-S	0.795	0.002	**0.782**	0.003	0.754	0.003	0.853	0.004	0.834	**0.001**	0.804	0.002
FADeNN-M	**0.796**	0.002	0.780	0.005	**0.773**	0.003	**0.856**	**0.000**	**0.836**	**0.001**	**0.831**	**0.001**

Table B.4: AUC and FR for different level of *s* of color scaling noise on 300-W-Test-N4 and Menpo-Test-N4.

Table B.5 shows the results of each evaluated models on original 300-W and Menpo test dataset with their corresponding AUC on the Figure B.1 for Single Image Facial Landmark Localisations.

Method	300-W		Menpo	
	AUC	FR	AUC	FR
ECT [Zhang et al., 2018a]	0.781	0.008	0.802	0.011
SAN [Dong et al., 2018]	**0.804**	0.005	0.809	0.015
FAN [Bulat and Tzimiropoulos, 2017]	0.712	**0.000**	0.739	0.001
TCDNN [Zhang et al., 2016b]	0.678	0.012	0.710	0.016
CFSS [Zhu et al., 2015a]	0.779	0.001	0.795	0.019
ERT [Kazemi and Sullivan, 2014]	0.532	0.196	0.578	0.200
FLL	0.797	0.001	**0.861**	**0.000**
FADeNN-D	0.788	0.001	0.853	**0.000**
FADeNN-S	0.797	0.001	**0.861**	**0.000**
FADeNN-M	0.798	**0.000**	**0.861**	**0.000**

Table B.5: Results on the original 300-W and Menpo test dataset.

B.2 Comparison on the 300-VW Test dataset, category 3rd

Figure B.2 shows the AUC curve of Single Image Facial Landmark Localisations and Facial Landmark Tracking on original (i.e. without additional synthetic noise) 300-VW test dataset category 3rd.

B.3 Results on ANCLaf on both AFEW and SEWA dataset

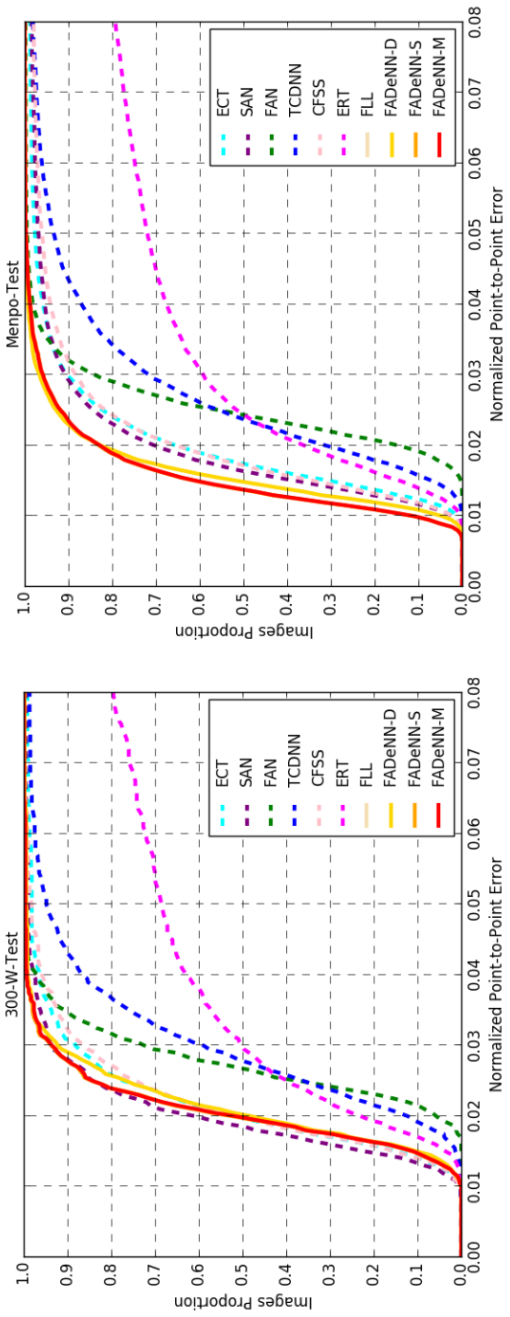


Figure B.1: Left : AUC curve of 300-W test dataset, Right : AUC curve of Menpo challenge test dataset.

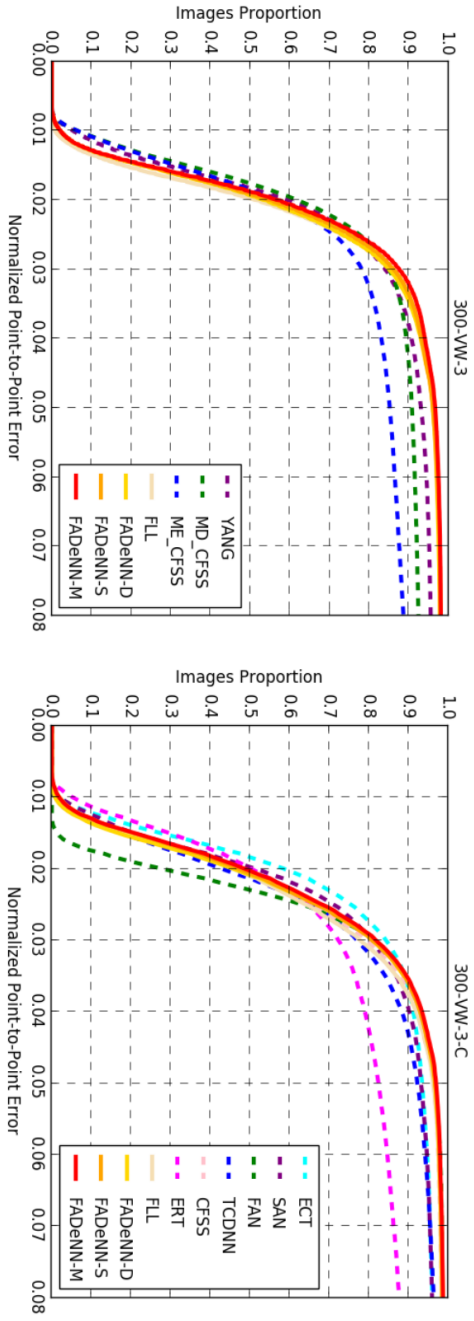


Figure B.2: Left : single image facial landmark localisation (300-VW-3), right : facial landmark tracking (300-VW-3-C).

Method	Fold-1		Fold-2		Fold-3		Fold-4		Fold-5		**Average**	
ANCLaF	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	3.1300	1.9632	2.2829	1.9235	2.3615	2.3052	2.4996	2.7035	3.1348	2.8253	**2.6818**	**2.3441**
COR	0.1837	0.4650	0.4194	0.3130	0.3373	0.2970	0.2843	0.4730	0.3063	0.4490	**0.3062**	**0.3994**
ICC	0.1181	0.3151	0.3016	0.3136	0.2528	0.2713	0.2071	0.3458	0.2157	0.2975	**0.2191**	**0.3087**
ANCLaF-S-2	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.8696	1.8404	2.3541	1.9668	2.4454	2.2805	2.3473	2.6124	3.3577	2.7727	**2.6748**	**2.2946**
COR	0.3368	0.4640	0.3670	0.4090	0.2036	0.3710	0.4399	0.3230	0.2239	0.4810	**0.3142**	**0.4096**
ICC	0.2563	0.3783	0.2732	0.2210	0.1303	0.2042	0.3693	0.3830	0.1511	0.2956	**0.2360**	**0.2964**
ANCLaF-S-4	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.8279	1.8493	2.4249	1.8921	2.4279	2.1921	2.5805	2.7077	3.0083	2.7560	**2.6539**	**2.2794**
COR	0.3553	0.4010	0.3517	0.4110	0.2469	0.3750	0.2029	0.4370	0.3595	0.4740	**0.3033**	**0.4196**
ICC	0.2645	0.3716	0.2733	0.3204	0.1754	0.2680	0.1394	0.2459	0.2696	0.3300	**0.2244**	**0.3072**
ANCLaF-S-8	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.9766	1.7686	2.2579	1.8630	2.2923	2.0819	2.4411	2.7379	3.0069	2.5565	**2.5950**	**2.2016**
COR	0.2534	0.4882	0.3866	0.3742	0.3316	0.4348	0.3214	0.2956	0.3491	0.5342	**0.3284**	**0.4254**
ICC	0.2040	0.4379	0.3134	0.2890	0.2764	0.3448	0.2801	0.2196	0.2861	0.4266	**0.2720**	**0.3436**
ANCLaF-S-16	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.8260	1.8924	2.3803	1.8966	2.3774	2.1887	2.5636	2.7905	2.9352	2.6936	**2.6165**	**2.2923**
COR	0.3378	0.4213	0.3500	0.4017	0.2555	0.3724	0.1845	0.3070	0.3819	0.5022	**0.3020**	**0.4009**
ICC	0.2568	0.3397	0.2717	0.3107	0.1783	0.2705	0.1254	0.2322	0.2874	0.3415	**0.2239**	**0.2989**
ANCLaF-S-32	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.7454	1.8543	2.3867	1.9507	2.3382	2.2519	2.4474	2.8264	2.9214	2.7546	**2.5678**	**2.3276**
COR	0.3324	0.4589	0.3501	0.4227	0.2278	0.3501	0.1564	0.2967	0.3713	0.4946	**0.2876**	**0.4046**
ICC	0.2527	0.3689	0.2719	0.3286	0.1574	0.2604	0.1029	0.2283	0.2847	0.3341	**0.2139**	**0.3041**
ANCLaF-SA-2	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.7318	1.8853	2.3437	1.8666	2.3252	2.1337	2.3400	2.6460	2.9570	2.6744	**2.5395**	**2.2412**
COR	0.3934	0.4208	0.3833	0.4313	0.3118	0.4292	0.3886	0.4705	0.3860	0.5181	**0.3726**	**0.4540**
ICC	0.3070	0.3368	0.3066	0.3415	0.2280	0.3301	0.3186	0.3950	0.2938	0.3623	**0.2908**	**0.3531**
ANCLaF-SA-4	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.8320	1.9702	2.3610	1.8697	2.3381	2.1170	2.4068	2.6471	2.9923	2.6936	**2.5860**	**2.2595**
COR	0.3984	0.3866	0.3938	0.4296	0.3354	0.4482	0.4022	0.4553	0.4006	0.5038	**0.3861**	**0.4447**
ICC	0.3132	0.3106	0.3109	0.3338	0.2497	0.3349	0.3330	0.3812	0.3037	0.3516	**0.3021**	**0.3424**
ANCLaF-SA-8	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.6957	1.8168	2.3419	1.9170	2.2703	2.2004	2.1885	2.5519	2.9076	2.7087	**2.4808**	**2.2390**
COR	0.3872	0.4829	0.3850	0.4486	0.3070	0.4029	0.4054	0.4846	0.3701	0.5164	**0.3709**	**0.4671**
ICC	0.3063	0.3922	0.3071	0.3540	0.2326	0.3201	0.3419	0.4127	0.2823	0.3574	**0.2940**	**0.3673**
ANCLaF-SA-16	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.8467	1.8231	2.3511	1.8587	2.3429	2.1184	2.4515	2.6190	3.0122	2.7074	**2.6009**	**2.2253**
COR	0.3818	0.4706	0.4065	0.4340	0.3305	0.4474	0.3835	0.4713	0.3828	0.5105	**0.3770**	**0.4668**
ICC	0.2944	0.3795	0.3203	0.3383	0.2484	0.3446	0.3176	0.3981	0.2898	0.3529	**0.2941**	**0.3627**
ANCLaF-SA-32	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	2.6919	1.9060	2.5055	1.9637	2.3340	2.1085	2.3709	2.6211	3.0042	2.6796	**2.5813**	**2.2558**
COR	0.3772	0.3871	0.3194	0.3636	0.3249	0.4533	0.3915	0.4644	0.3907	0.5133	**0.3607**	**0.4363**
ICC	0.2431	0.2776	0.2503	0.2846	0.2448	0.3527	0.3170	0.3854	0.2962	0.3612	**0.2703**	**0.3323**

Table B.6: The results of all variants of our proposed models on AFEW dataset for all folds.

Method	Fold-1		Fold-2		Fold-3		Fold-4		Fold-5		**Average**	
ANCLaF	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3337	0.3523	0.3400	0.3846	0.3946	0.3076	0.3943	0.3515	0.3057	0.3415	**0.3537**	**0.3475**
COR	0.5358	0.4837	0.6654	0.2887	0.4103	0.4348	0.4079	0.3645	0.6281	0.4033	**0.5295**	**0.3950**
CCC	0.5068	0.4405	0.6364	0.2550	0.3752	0.4207	0.3309	0.3224	0.6116	0.3799	**0.4922**	**0.3637**
ANCLaF-S-2	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3345	0.3509	0.3380	0.3771	0.3921	0.3105	0.3859	0.3444	0.2942	0.3415	**0.3489**	**0.3449**
COR	0.5304	0.4862	0.6584	0.2955	0.4203	0.4336	0.4135	0.3566	0.6407	0.4090	**0.5327**	**0.3962**
CCC	0.5037	0.4461	0.6412	0.2660	0.3916	0.4217	0.3447	0.3233	0.6352	0.3826	**0.5033**	**0.3679**
ANCLaF-S-4	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3333	0.3502	0.3273	0.3555	0.3913	0.3064	0.3761	0.3430	0.2901	0.3271	**0.3436**	**0.3364**
COR	0.5338	0.4844	0.6662	0.3152	0.4149	0.4372	0.4242	0.3624	0.6401	0.4134	**0.5358**	**0.4025**
CCC	0.5065	0.4443	0.6547	0.3011	0.3855	0.4233	0.3638	0.3358	0.6387	0.4077	**0.5098**	**0.3825**
ANCLaF-S-8	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3312	0.3507	0.3291	0.3707	0.3915	0.3056	0.3606	0.3436	0.2901	0.3256	**0.3405**	**0.3392**
COR	0.5353	0.4851	0.6732	0.3028	0.4219	0.4394	0.4209	0.3635	0.6405	0.4281	**0.5384**	**0.4038**
CCC	0.5045	0.4435	0.6570	0.2809	0.3872	0.4266	0.3838	0.3350	0.6387	0.4192	**0.5143**	**0.3810**
ANCLaF-S-16	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3345	0.3545	0.3415	0.3789	0.3927	0.3088	0.3990	0.3397	0.3039	0.3357	**0.3543**	**0.3435**
COR	0.5277	0.4803	0.6650	0.2929	0.4110	0.4306	0.4048	0.3589	0.6266	0.4114	**0.5270**	**0.3948**
CCC	0.4978	0.4370	0.6319	0.2624	0.3780	0.4144	0.3273	0.3342	0.6136	0.3949	**0.4897**	**0.3686**
ANCLaF-S-32	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3354	0.3562	0.3447	0.3820	0.4063	0.3125	0.3832	0.3490	0.2954	0.3285	**0.3530**	**0.3456**
COR	0.5200	0.4856	0.6564	0.2953	0.4096	0.4311	0.4130	0.3503	0.6345	0.4134	**0.5267**	**0.3951**
CCC	0.4855	0.4355	0.6362	0.2645	0.3678	0.4221	0.3479	0.3110	0.6331	0.4079	**0.4941**	**0.3682**
ANCLaF-SA-2	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3255	0.3411	0.3148	0.3586	0.3844	0.2998	0.4034	0.3400	0.2860	0.3263	**0.3428**	**0.3332**
COR	0.5419	0.5068	0.6801	0.3300	0.4346	0.4592	0.4165	0.3651	0.6495	0.4397	**0.5445**	**0.4201**
CCC	0.5078	0.4638	0.6708	0.3012	0.4024	0.4440	0.3158	0.3195	0.6491	0.4204	**0.5092**	**0.3898**
ANCLaF-SA-4	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3219	0.3418	0.3212	0.3499	0.3863	0.3007	0.3679	0.3294	0.2852	0.3178	**0.3365**	**0.3279**
COR	0.5483	0.5050	0.6894	0.3387	0.4311	0.4573	0.4396	0.3904	0.6428	0.4535	**0.5502**	**0.4290**
CCC	0.5283	0.4790	0.6740	0.3060	0.4060	0.4400	0.3732	0.3260	0.6482	0.4430	**0.5259**	**0.3988**
ANCLaF-SA-8	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3218	0.3370	0.3305	0.3548	0.3774	0.2994	0.3581	0.3484	0.2923	0.3195	**0.3360**	**0.3318**
COR	0.5565	0.5204	0.6864	0.3288	0.4442	0.4512	0.4472	0.3717	0.6571	0.4499	**0.5583**	**0.4244**
CCC	0.5265	0.4620	0.6579	0.3190	0.4162	0.4460	0.3910	0.3600	0.6513	0.4370	**0.5286**	**0.4048**
ANCLaF-SA16	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3239	0.3391	0.3177	0.3588	0.3794	0.2954	0.3607	0.3324	0.2892	0.3274	**0.3342**	**0.3306**
COR	0.5519	0.5003	0.6846	0.3246	0.4361	0.4559	0.4457	0.3783	0.6615	0.4435	**0.5560**	**0.4205**
CCC	0.5254	0.4617	0.6714	0.2977	0.4053	0.4397	0.3892	0.3498	0.6504	0.4184	**0.5284**	**0.3935**
ANCLaF-SA-32	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	**Valence**	**Arousal**
RMSE	0.3233	0.3434	0.3280	0.5185	0.3788	0.2975	0.3641	0.3285	0.2842	0.3219	**0.3357**	**0.3620**
COR	0.5489	0.5229	0.6391	0.1989	0.4397	0.4783	0.4547	0.4168	0.6655	0.4737	**0.5496**	**0.4181**
CCC	0.5185	0.4763	0.6044	0.1593	0.4076	0.4644	0.3793	0.3850	0.6575	0.4583	**0.5135**	**0.3887**

Table B.7: The results of all variants of our proposed models on SEWA dataset for all folds.

Bibliography

[Afifi and Abdelhamed, 2019] Afifi, M. and Abdelhamed, A. (2019). Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*.

[Alhussein, 2016] Alhussein, M. (2016). Automatic facial emotion recognition using weber local descriptor for e-healthcare system. *Cluster Computing*, 19(1):99–108.

[Arevalo et al., 2020] Arevalo, J., Solorio, T., Montes-y Gomez, M., and González, F. A. (2020). Gated multimodal networks. *Neural Computing and Applications*, pages 1–20.

[Aspandi et al., 2020a] Aspandi, D., Mallol-Ragolta, A., Schuller, B., and Binefa, X. (2020a). Latent-based adversarial neural networks for facial affect estimations. In *2020 15th IEEE FG*, pages 348–352, Los Alamitos, CA, USA. IEEE Computer Society.

[Aspandi et al., 2019a] Aspandi, D., Martinez, O., and Binefa, X. (2019a). Heatmap-guided balanced deep convolution networks for family classification in the wild. In *2019 14th IEEE FG 2019*, pages 1–5.

[Aspandi et al., 2019b] Aspandi, D., Martinez, O., Sukno, F., and Binefa, X. (2019b). Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild. In *2019 14th IEEE FG*, pages 1–8.

[Aspandi et al., 2019c] Aspandi, D., Martinez, O., Sukno, F., and Binefa, X. (2019c). Robust facial alignment with internal denoising auto-encoder. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 143–150.

[Aspandi et al., 2020b] Aspandi, D., Sukno, F., Schuller, B., and Binefa, X. (2020b). An enhanced adversarial network with combined latent features for spatio-temporal facial affect estimation in the wild. In *2020 16th International Conference on Computer Vision Theory and Applications (VISAPP)*. In Press.

[Asthana et al., 2014] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2014). Incremental face alignment in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

[Barros et al., 2018] Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., and Wermter, S. (2018). The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

[Belhumeur et al., 1997] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720.

[Bengio et al., 2012] Bengio, Y., Courville, A. C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1:2012.

[Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual*

International Conference on Machine Learning, ICML '09, pages 41–48, New York, NY, USA. Association for Computing Machinery.

[Bertinetto et al., 2016] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. S. (2016). Fully-convolutional siamese networks for object tracking. In *ECCV 2016 Workshops*, pages 850–865.

[Biau et al., 2016] Biau, G., Scornet, E., and Welbl, J. (2016). Neural random forests. *Sankhya A*, pages 1–40.

[Bulat and Tzimiropoulos, 2017] Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030.

[Cao et al., 2018] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*.

[Cardinal et al., 2015] Cardinal, P., Dehak, N., Koerich, A. L., Alam, J., and Boucher, P. (2015). Ets system for avec 2015 challenge. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pages 17–23, New York, NY, USA. ACM.

[Chaitanya et al., 2017] Chaitanya, C. R. A., Kaplanyan, A. S., Schied, C., Salvi, M., Lefohn, A., Nowrouzezahrai, D., and Aila, T. (2017). Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Trans. Graph.*, 36(4):98:1–98:12.

[Chang et al., 2017] Chang, W.-Y., Hsu, S.-H., and Chien, J.-H. (2017). Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation.

[Chen et al., 2013] Chen, M.-Y., Lughofer, E. D., Lin, K. C., Huang, T.-C., Hung, J. C., Yen, N. Y., and Chen, S. J. (2013). Facial emotion

recognition towards affective computing-based learning. *Library Hi Tech*.

[Chen and Jin, 2015] Chen, S. and Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56.

[Chen et al., 2003] Chen, Z. et al. (2003). Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182(1):1–69.

[Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE CVPR*, pages 8789–8797.

[Chow and Li, 1993] Chow, G. and Li, X. (1993). Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755.

[Christodoulidis et al., 2016] Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE journal of biomedical and health informatics*, 21(1):76–84.

[Christodoulidis et al., 2017] Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., and Mougiakakou, S. (2017). Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE Journal of Biomedical and Health Informatics*, 21(1):76–84.

[Chrysos et al., 2017] Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., and Zafeiriou, S. (2017). A Comprehensive Performance Evaluation of Deformable Face Tracking In-the-Wild. *International Journal of Computer Vision*, pages 1–35.

[Chrysos et al., 2015] Chrysos, G. G., Antonakos, E., Zafeiriou, S., and Snape, P. (2015). Offline deformable face tracking in arbitrary videos. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

[Chrysos et al., 2019] Chrysos, G. G., Favaro, P., and Zafeiriou, S. (2019). Motion deblurring of faces. *International Journal of Computer Vision*, 127(6):801–823.

[Comas et al., 2020] Comas, J., Aspandi, D., and Binefa, X. (2020). End-to-end facial and physiological model for affective computing and applications. In *15th IEEE FG*.

[Cootes et al., 1998] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In Burkhardt, H. and Neumann, B., editors, *Computer Vision — ECCV'98*, pages 484–498, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

[Correa et al., 2018] Correa, J. A. M., Abadi, M. K., Sebe, N., and Patras, I. (2018). Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*.

[Dai et al., 2017] Dai, B., Fidler, S., Urtasun, R., and Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In *The IEEE International Conference on Computer Vision (ICCV)*.

[Dong et al., 2017] Dong, J., Burnham, J. G., Boots, B., Rains, G., and Dellaert, F. (2017). 4d crop monitoring: Spatio-temporal reconstruction for agriculture. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3878–3885. IEEE.

[Dong et al., 2018] Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018). Style aggregated network for facial landmark detection. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 379–388.

[Duo and Song, 2010] Duo, S. and Song, L. (2010). An e-learning system based on affective computing. *Physics Procedia*, 24.

[Ekman et al., 1978] Ekman, P., Friesen, W., and Hager, J. (1978). A technique for the measurement of facial action. *Consulting, Palo Alto*.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

[El Ayadi et al., 2011] El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

[Eyben et al., 2016] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the* 18^{th} *ACM International Conference on Multimedia*, pages 1459–1462, Firenze, Italy. ACM.

[Fasel and Luettin, 2003] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275.

[Fine et al., 1998] Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62.

[Gan et al., 2015] Gan, Q., Guo, Q., Zhang, Z., and Cho, K. (2015). First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks. pages 1–13.

[Garciarena et al., 2018] Garciarena, U., Santana, R., and Mendiburu, A. (2018). Expanding variational autoencoders for learning and exploiting latent representations in search distributions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO 18, pages 849–856, New York, NY, USA. Association for Computing Machinery.

[Georghiades et al., 2001] Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660.

[Gopalan et al., 2018] Gopalan, N., Bellamkonda, S., and Chaitanya, V. S. (2018). Facial expression recognition using geometric landmark points and convolutional neural networks. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1149–1153. IEEE.

[Gordon et al., 2018] Gordon, D., Farhadi, A., and Fox, D. (2018). Re 3 : Re al-Time Re current Re gression Networks for Object Tracking. *Ieee Robotics and Automation Letters*, 3(2):788–795.

[Goswami et al., 2018] Goswami, G., Ratha, N., Agarwal, A., Singh, R., and Vatsa, M. (2018). Unravelling robustness of deep learning based face recognition against adversarial attacks.

[Gower, 1975] Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

[Greenspan et al., 2016] Greenspan, H., van Ginneken, B., and Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159.

[Greff et al., 2017] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232.

[Gu et al., 2017] Gu, J., Yang, X., Mello, S. D., and Kautz, J. (2017). Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1531–1540.

[Gulcehre et al., 2018] Gulcehre, C., Chandar, S., Cho, K., and Bengio, Y. (2018). Dynamic neural turing machine with continuous and discrete addressing schemes. *Neural computation*, 30(4):857–884.

[Gunes and Pantic, 2010] Gunes, H. and Pantic, M. (2010). Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *International conference on intelligent virtual agents*, pages 371–377. Springer.

[Gunes and Piccardi, 2005] Gunes, H. and Piccardi, M. (2005). Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE SYS MAN CYBERN*, volume 4, pages 3437–3443. IEEE.

[Guo et al., 2016] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer.

[Han et al., 2019] Han, J., Zhang, Z., and Schuller, B. (2019). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine*, 14(2):68–81.

[Handrich et al., 2020] Handrich, S., Dinges, L., Al-Hamadi, A., Werner, P., and Al Aghbari, Z. (2020). Simultaneous prediction of valence/arousal and emotions on affectnet, aff-wild and afew-va. *Procedia Computer Science*, 170:634–641.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE CVPR*, pages 7132–7141.

[Huang et al., 2018] Huang, R., Pedoeem, J., and Chen, C. (2018). Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2503–2510. IEEE.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE CVPR*, pages 1125–1134.

[Kazemi and Sullivan, 2011] Kazemi, V. and Sullivan, J. (2011). Face alignment with part-based modeling. In *2011 22nd British Machine Vision Conference, BMVC 2011, Dundee, United Kingdom, 29 August 2011 through 2 September 2011*, pages 27–1. British Machine Vision Association, BMVA.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874.

[Kim et al., 2018] Kim, C., Li, F., and Rehg, J. M. (2018). Multi-object tracking with neural gating using bilinear lstm. In *The European Conference on Computer Vision (ECCV)*.

[Kim et al., 2017] Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *34th ICML 70*, pages 1857–1865. JMLR. org.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[Kollias et al., 2020] Kollias, D., Schulc, A., Hajiyev, E., and Zafeiriou, S. (2020). Analysing affective behavior in the first abaw 2020 competition. *arXiv preprint arXiv:2001.11409*.

[Kollias et al., 2019] Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. (2019). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23.

[Kollias and Zafeiriou, 2018] Kollias, D. and Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*.

[Kollias and Zafeiriou, 2019] Kollias, D. and Zafeiriou, S. (2019). Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.

[Kossaifi et al., 2020] Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T. M., and Pantic, M. (2020). Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6060–6069.

[Kossaifi et al., 2017] Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.

[Kossaifi et al., 2019] Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *arXiv preprint arXiv:1901.02839*.

[Kumar et al., 2019] Kumar, A., Kaur, A., and Kumar, M. (2019). Face detection techniques: a review. *Artificial Intelligence Review*, 52(2):927–948.

[Kumar et al., 2017] Kumar, A., Sattigeri, P., and Fletcher, T. (2017). Semi-supervised learning with gans: Manifold invariance with improved inference. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5534–5544. Curran Associates, Inc.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[Lee et al., 2005] Lee, K.-C., Ho, J., and Kriegman, D. J. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):684–698.

[Li et al., 2020] Li, C., Bao, Z., Li, L., and Zhao, Z. (2020). Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Information Processing & Management*, 57(3):102185.

[Li and Deng, 2018] Li, S. and Deng, W. (2018). Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348.

[Liu et al., 2008] Liu, C., Conn, K., Sarkar, N., and Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE T Robot*, 24:883 – 896.

[Luong et al., 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[Lv et al., 2017a] Lv, J.-J., Shao, X., Xing, J., Cheng, C., and Zhou, X. (2017a). A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3700.

[Lv et al., 2017b] Lv, J.-J., Shao, X., Xing, J., Cheng, C., and Zhou, X. (2017b). A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. *2017 IEEE CVPR*, pages 3691–3700.

[Ma et al., 2019] Ma, J., Tang, H., Zheng, W.-L., and Lu, B.-L. (2019). Emotion recognition using multimodal residual lstm network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 176–183.

[Mao et al., 2016] Mao, X., Shen, C., and Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2802–2810. Curran Associates, Inc.

[McFee et al., 2020] McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Mason, J., Ellis, D., Battenberg, E., Seyfarth, S., Yamamoto, R., Choi, K., viktorandreevichmorozov, Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.-R., Friesch, P., Weiss, A., Vollrath, M., and Kim, T. (2020). librosa/librosa: 0.8.0.

[McKeown et al., 2010] McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. In *2010 IEEE Int Con Multi*, pages 1079–1084. IEEE.

[Michael et al., 2019] Michael, J., Labahn, R., Grüning, T., and Zöllner, J. (2019). Evaluating sequence-to-sequence models for handwritten text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1286–1293. IEEE.

[Mitenkova et al., 2019] Mitenkova, A., Kossaifi, J., Panagakis, Y., and Pantic, M. (2019). Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE FG 2019*, pages 1–7. IEEE.

[Mnih et al., 2016] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937.

[Mollahosseini et al., 2015] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2015). Affectnet: A database for facial expression. *Valence, and Arousal Computing in the Wild Department of Electrical and Computer Engineering, University of Denver, Denver, CO*, 80210.

[Nada et al., 2018] Nada, H., Sindagi, V. A., Zhang, H., and Patel, V. M. (2018). Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. *CoRR*, abs/1804.10275.

[Nam and Han, 2016] Nam, H. and Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Nicolaou et al., 2011] Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE T Affect Comput*, 2(2):92–105.

[Odena, 2016] Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

[Pantic and Rothkrantz, 2000] Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. *Procedings of the British Machine Vision Conference 2015*, (Section 3):41.1–41.12.

[Peng et al., 2018] Peng, X., Feris, R. S., Wang, X., and Metaxas, D. N. (2018). Red-net: A recurrent encoder–decoder network for video-based face alignment. *International Journal of Computer Vision*, 126(10):1103–1119.

[Poria et al., 2019] Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

[Povolny et al., 2016] Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., Wood, I., Robin, C., and Lamel, L. (2016). Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC 16, pages 75–82, New York, NY, USA. Association for Computing Machinery.

[Qu et al., 2015] Qu, C., Gao, H., Monari, E., Beyerer, J., and Thiran, J.-P. (2015). Towards robust cascaded regression for face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9.

[Radford et al., 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[Rajamanoharan and Cootes, 2015] Rajamanoharan, G. and Cootes, T. F. (2015). Multi-view constrained local models for large head angle facial tracking. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 971–978.

[Ringeval et al., 2019] Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (2019). Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE FG*, pages 1–8.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

[Sagonas et al., 2013] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3d face analysis. In *European workshop on biometrics and identity management*, pages 47–56. Springer.

[Schmitt et al., 2019] Schmitt, M., Cummins, N., and Schuller, B. (2019). Continuous emotion recognition in speech–do we need recurrence? *Training*, 34(93):12.

[Sha et al., 2016] Sha, L., Chang, B., Sui, Z., and Li, S. (2016). Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879.

[Shen et al., 2015] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011.

[Snoek et al., 2005] Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Szegedy et al., 2016] Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

[Tang et al., 2020] Tang, Y., Serdan, T. D., Masi, L. N., Tang, S., Gorjao, R., and Hirabara, S. M. (2020). Epidemiology of covid-19 in brazil: using a mathematical model to estimate the outbreak peak and temporal evolution. *Emerging microbes & infections*, 9(1):1453–1456.

[Tao and Tan, 2005] Tao, J. and Tan, T. (2005). Affective computing: A review. In Tao, J., Tan, T., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, pages 981–995, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Tellamekala and Valstar, 2019] Tellamekala, M. K. and Valstar, M. (2019). Temporally coherent visual representations for dimensional affect recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

[Thies et al., 2016] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

[Tian et al., 2001] Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.

[Triantafyllidou and Tefas, 2016] Triantafyllidou, D. and Tefas, A. (2016). Face detection based on deep convolutional neural networks exploiting incremental facial part learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3560–3565.

[Triantafyllidou and Tefas, 2016] Triantafyllidou, D. and Tefas, A. (2016). Face detection based on deep convolutional neural networks exploiting incremental facial part learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3560–3565.

[Trumble et al., 2018] Trumble, M., Gilbert, A., Hilton, A., and Collomosse, J. (2018). Deep autoencoder for combined human pose estimation and body model upscaling. In *ECCV*.

[Tu et al., 2017] Tu, Y.-H., Du, J., Wang, Q., Bao, X., Dai, L.-R., and Lee, C.-H. (2017). An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech. *Computer Speech Language*, 46:517 – 534.

[Uricár et al., 2015] Uricár, M., Franc, V., and Hlavác, V. (2015). Facial landmark tracking by tree-based deformable part model based detector. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 963–970.

[Valstar et al., 2010] Valstar, M., Martinez, B., Binefa, X., and Pantic, M. (2010). Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Vỳrost et al., 2015] Vỳrost, T., Lyócsa, Š., and Baumöhl, E. (2015). Granger causality stock market networks: Temporal proximity and

preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 427:262–276.

[Wang et al., 2012a] Wang, Q., Chen, F., Xu, W., and Yang, M. (2012a). Online discriminative object tracking with local sparse representation. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 425–432.

[Wang et al., 2012b] Wang, Y., Guan, L., and Venetsanopoulos, A. N. (2012b). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3):597–607.

[Wu et al., 2017] Wu, Y., Gou, C., and Ji, Q. (2017). Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. *arXiv preprint arXiv:1709.08130*.

[Wu and Ji, 2015] Wu, Y. and Ji, Q. (2015). Shape augmented regression method for face alignment. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 979–985.

[Wu and Ji, 2018] Wu, Y. and Ji, Q. (2018). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, pages 1–28.

[Wu et al., 2013] Wu, Y., Lim, J., and Yang, M.-H. (2013). Online object tracking: A benchmark. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2411–2418, Washington, DC, USA. IEEE Computer Society.

[Xia et al., 2020] Xia, Y., Braun, S., Reddy, C. K. A., Dubey, H., Cutler, R., and Tashev, I. (2020). Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 871–875.

[Xiao et al., 2015] Xiao, S., Yan, S., and Kassim, A. A. (2015). Facial landmark detection via progressive initialization. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 986–993.

[Xiaohua et al., 2019] Xiaohua, W., Muzi, P., Lijuan, P., Min, H., Chunhua, J., and Fuji, R. (2019). Two-level attention with two-stage multitask learning for facial emotion recognition. *Journal of Visual Communication and Image Representation*, 62:217–225.

[Xie et al., 2016] Xie, J., Girshick, R. B., and Farhadi, A. (2016). Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV 2016*, pages 842–857.

[Xiong and la Torre, 2015] Xiong, X. and la Torre, F. D. (2015). Global supervised descent method. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673.

[Yan et al., 2018] Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., and Zong, Y. (2018). Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 309:27–35.

[Yan et al.,] Yan, Y., Naturel, X., Chateau, T., Duffner, S., Garcia, C., and Blanc, C. A survey of deep facial landmark detection.

[Yang et al., 2015] Yang, J., Deng, J., Zhang, K., and Liu, Q. (2015). Facial shape tracking via spatio-temporal cascade shape regression. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 994–1002.

[Yang and Hirschberg, 2018] Yang, Z. and Hirschberg, J. (2018). Predicting arousal and valence from waveforms and spectrograms using deep neural networks. In *INTERSPEECH*, pages 3092–3096.

[Yao and Guan, 2018] Yao, L. and Guan, Y. (2018). An improved lstm structure for natural language processing. In *2018 IEEE International*

Conference of Safety Produce Informatization (IICSPI), pages 565–569. IEEE.

[Ye et al., 2018] Ye, H., Li, G. Y., Juang, B.-H. F., and Sivanesan, K. (2018). Channel agnostic end-to-end learning based communication systems with conditional gan. In *2018 IEEE Globecom Workshops (GC Wkshps)*, pages 1–5. IEEE.

[Yeasin et al., 2006] Yeasin, M., Bullot, B., and Sharma, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508.

[Yi et al., 2014] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *CoRR*, abs/1411.7923.

[Yuille et al., 1992] Yuille, A. L., Hallinan, P. W., and Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111.

[Zadeh et al., 2017] Zadeh, A., Chong Lim, Y., Baltrusaitis, T., and Morency, L.-P. (2017). Convolutional experts constrained local model for 3d facial landmark detection. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

[Zafeiriou et al., 2018] Zafeiriou, S., Chrysos, G. G., Roussos, A., Ververas, E., Deng, J., and Trigeorgis, G. (2018). The 3D Menpo Facial Landmark Tracking Challenge. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018-Janua:2503–2511.

[Zafeiriou et al., 2017a] Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaioannou, A., Zhao, G., and Kotsia, I. (2017a). Aff-wild: Valence and arousal 'in-the-wild'challenge. In *IEEE CVPRW, 2017*, pages 1980–1987. IEEE.

[Zafeiriou et al., 2017b] Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., and Shen, J. (2017b). The menpo facial landmark localisation

challenge: A step towards the solution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2116–2125.

[Zeng et al., 2009] Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T Pattern Anal*, 31(1):39–58.

[Zhang et al., 2018a] Zhang, H., Li, Q., Sun, Z., and Liu, Y. (2018a). Combining data-driven and model-driven methods for robust facial landmark detection. *IEEE Transactions on Information Forensics and Security*, 13(10):2409–2422.

[Zhang et al., 2016a] Zhang, K., Zhang, Z., Li, Z., Member, S., Qiao, Y., and Member, S. (2016a). Joint Face Detection and Alignment using Multi - task Cascaded Convolutional Networks. *Spl*, (1):1–5.

[Zhang et al., 2018b] Zhang, W., Huang, H., Schmitz, M., Sun, X., Wang, H., and Mayer, H. (2018b). Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling. *Remote Sensing*, 10(1):52.

[Zhang and Yang, 2017] Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

[Zhang et al., 2016b] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016b). Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(5):918–930.

[Zheng et al., 2013] Zheng, S., Sturgess, P., and Torr, P. H. S. (2013). Approximate structured output learning for constrained local models with application to real-time facial feature detection and tracking on low-power devices. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

[Zhou et al., 2019] Zhou, T., Ruan, S., and Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004.

[Zhou et al., 2018] Zhou, Y., Liu, D., and Huang, T. (2018). Survey of face detection on low-quality images. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 769–773.

[Zhu et al., 2015a] Zhu, S., Li, C., Loy, C. C., and Tang, X. (2015a). Face alignment by coarse-to-fine shape searching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006.

[Zhu et al., 2015b] Zhu, X., Lei, Z., Yan, J., Yi, D., and Li, S. Z. (2015b). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796.

[Zhu et al., 2017] Zhu, X., xiaoming Liu, Lei, Z., and Li, S. Z. (2017). Face Alignment In Full Pose Range: A 3D Total Solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14.