Leveraging learning-based computer vision for vibration informed structural diagnosis

Ricard Lado Roigé

http://hdl.handle.net/10803/689664

Data de defensa: 19-12-2023

UNIVERSITAT RAMON LLULL

DOCTORAL THESIS

Title	Leveraging Learning-Based Computer Vision for Vibration-Informed Structural Diagnosis
Presented by	Ricard Lado Roigé
Centre	IQS School of Engineering
Department	Industrial Engineering
Directed by	Dr. Marco Antonio Pérez Martínez

This thesis is dedicated to all the open source developers who have generously shared their knowledge, skills and code with the world. They have inspired me to pursue my passion for computer science and to contribute to the advancement of this field. I am deeply grateful for their guidance, support and collaboration throughout my academic journey.

This thesis has been partially funded with the support of the Catalan Agency for Management of University and Research Grants (AGAUR) through the project 2019 LLAV 00020 and the Catalan Agency for Business Competitiveness (ACCIÓ) through the project INNOTEC ACE034/21/000041 ISAPREF 2021.

Acknowledgements

Primerament, m'agradaria agrair a IQS l'oportunitat d'haver realitzat el doctorat, així com el finançament rebut. No imaginava pas que la meva etapa formativa per IQS em portaria fins aquí.

De la mateixa manera, vull donar les gràcies al meu director, el Dr. Marco Antonio Pérez, pel seu acompanyament i per la seva confiança, fins i tot en els moments on he optat per descarrilar la tesi i explorar nous camins. Aquesta ha estat una experiència enriquidora en la qual he crescut personalment i professional. Gràcies per enganyar-me a fer la tesi.

També vull agrair als professors que m'han acompanyat durant aquesta etapa. Especialment, al Dr. Andrés García pel seu esforç per impulsar activitats pensades de cara als estudiants, així com per la seva confiança en les meves habilitats tècniques. També agrair al Sr. Sauro Yagüe les facilitats que m'ha donat per col·laborar en la seva assignatura i les excel·lents converses tècniques que hem tingut durant el meu pas per IQS. Al Dr. Joaquín Menacho, pel seu ajut desinteressat en els meus projectes personals. I finalment, al Dr. Giovanni Gómez, pel seu consell durant les reunions periòdiques del grup d'investigació.

Molt especialment vull agrair també a tots els companys que m'han acompanyat durant aquests últims tres anys pels molt bons moments que hem compartit dins i fora d'IQS. Albert, Andrea, Ariadna, Dani, Ferran, Guillermo, Héctor, Josep i Laia, he après moltíssim de vosaltres i ha estat un plaer treballar al vostre costat.

Cal destacar, que amb l'Albert hem treballat en diversos projectes més enllà dels nostres doctorats i junts hem construït projectes extraordinaris. Gràcies pel teu entusiasme i la teva constància, sóc millor enginyer gràcies a tu. També volia agrair a l'Ariadna els bons moments que hem passat anant a escalar plegats així com el seu bon humor. Per últim, vull agrair-li al Josep el seu refinat sentit de l'humor i la seva disciplina de treball que ha inspirat la temàtica d'aquesta tesi.

Finalment, vull agrair a la meva família, i molt especialment als meus pares, el seu suport i estima durant tots aquests anys. Gràcies per fer aquesta tesi possible.

Abstract

Novel vision-based monitoring systems present a compelling value proposition within the realm of Structural Health Monitoring (SHM) applications. They serve as a non-intrusive alternative to conventional contact-based sensors, capable of capturing comprehensive measurements across the entire structural field, even when placed at considerable distances from the structure. This inherent characteristic renders vision-based techniques highly cost-effective when compared to their contact-based counterparts. However, these techniques, although promising, still exhibit some technical limitations due to their relative immaturity.

The primary objective of this thesis is to develop, enhance, and validate a cutting-edge computer vision methodology grounded in learning-based video motion magnification. This methodology is specifically designed to facilitate the implementation of SHM strategies in complex structures, employing optical means to capture full-field data.

The research conducted in this thesis establishes the robust capability of learning-based video motion magnification techniques in accurately and consistently amplifying intricate structural movements. Consequently, this allows for the precise detection of subtle changes in the dynamic response of the structure under scrutiny. Furthermore, the integration of Eulerian video motion magnification with advanced deep learning techniques, such as visual transformers and convolutional neural networks, significantly elevates both accuracy and efficiency beyond what is currently considered state-of-the-art in video motion magnification image quality.

In addition to these technical achievements, this work also provides a comprehensive methodology and software toolset, which can prove invaluable for researchers and engineers engaged in the efficient and effective monitoring and maintenance of complex structures.

Resum

Els nous sistemes de monitoratge basats en la visió presenten una proposta de valor convincent dins de l'àmbit de les aplicacions de monitoratge de la salut estructural (SHM). Serveixen com una alternativa no intrusiva als sensors convencionals basats en el contacte, capaços de capturar mesures exhaustives a tot el camp estructural, fins i tot quan es col·loquen a distàncies considerables de l'estructura. Aquesta característica inherent fa que les tècniques basades en la visió siguin molt rendibles en comparació amb les seves contraparts basades en el contacte. No obstant això, aquestes tècniques, encara que prometedores, encara presenten algunes limitacions tècniques degudes a la seva relativa immaduresa.

L'objectiu principal d'aquesta tesi és desenvolupar, millorar i validar una metodologia d'última generació de visió per ordinador basada en l'amplificació del moviment de vídeo basada en l'aprenentatge. Aquesta metodologia està dissenyada específicament per facilitar la implementació d'estratègies SHM en estructures complexes, utilitzant mitjans òptics per capturar dades de camp complet.

La recerca realitzada en aquesta tesi estableix la capacitat robusta de les tècniques d'amplificació del moviment de vídeo basades en l'aprenentatge per amplificar amb precisió i consistència els moviments estructurals intricats. En conseqüència, això permet la detecció precisa de canvis subtils en la resposta dinàmica de l'estructura sotmesa a escrutini. A més, la integració de l'amplificació del moviment de vídeo eulerià amb tècniques avançades d'aprenentatge profund, com ara els transformadors visuals i les xarxes neuronals convolucionals, eleva significativament tant la precisió com l'eficiència més enllà del que actualment es considera l'estat de l'art en la qualitat d'imatge de l'amplificació del moviment de vídeo.

A més d'aquests assoliments tècnics, aquest treball també proporciona una metodologia completa i un conjunt d'eines de programari que poden resultar inestimables per als investigadors i enginyers que es dediquen al monitoratge i manteniment eficaços i eficients d'estructures complexes.

Resumen

Los nuevos sistemas de monitoreo basados en la visión presentan una propuesta de valor convincente en el ámbito de las aplicaciones de monitoreo de la salud estructural (SHM). Sirven como una alternativa no intrusiva a los sensores convencionales basados en el contacto, capaces de capturar mediciones exhaustivas en todo el campo estructural, incluso cuando se colocan a distancias considerables de la estructura. Esta característica inherente hace que las técnicas basadas en la visión sean muy rentables en comparación con sus contrapartes basadas en el contacto. Sin embargo, estas técnicas, aunque prometedoras, aún presentan algunas limitaciones técnicas debido a su relativa inmadurez.

El objetivo principal de esta tesis es desarrollar, mejorar y validar una metodología de vanguardia de visión por ordenador basada en la amplificación del movimiento de video basada en el aprendizaje. Esta metodología está diseñada específicamente para facilitar la implementación de estrategias SHM en estructuras complejas, utilizando medios ópticos para capturar datos de campo completo.

La investigación realizada en esta tesis establece la capacidad robusta de las técnicas de amplificación del movimiento de video basadas en el aprendizaje para amplificar con precisión y consistencia los movimientos estructurales intrincados. En consecuencia, esto permite la detección precisa de cambios sutiles en la respuesta dinámica de la estructura sometida a escrutinio. Además, la integración de la amplificación del movimiento de video euleriano con técnicas avanzadas de aprendizaje profundo, como transformadores visuales y redes neuronales convolucionales, eleva significativamente tanto la precisión como la eficiencia más allá de lo que actualmente se considera el estado del arte en la calidad de imagen de la amplificación del movimiento de video.

Además de estos logros técnicos, este trabajo también proporciona una metodología completa y un conjunto de herramientas de software que pueden resultar invaluables para los investigadores e ingenieros que se dedican al monitoreo y mantenimiento eficientes y efectivos de estructuras complejas.

Publications

Journal publications

Lado-Roigé, R., Font-Moré, J. & Pérez, M. A., Learning-based video motion magnification approach for vibration-based damage detection. Meas. 206, 112218. doi:10.1016/j.measurement.2022.112218 (2023).

Lado-Roigé, R. & Pérez, M. A., STB-VMM: Swin Transformer based Video Motion Magnification. Knowl.-Based Syst., 110493. doi:10.1016/j.knosys.2023.110493 (2023).

Lado-Roigé, R. & Pérez, M. A., ViMag: A Visual Vibration Analysis Toolbox. Journal of Open Source Software 8, 5491. doi:10.21105/joss.05491 (2023).

Lado-Roigé, R. & Pérez, M. A., Vision-based dynamic response analysis of engineering structures. *(Under review)*

Scientific communications

Lado-Roigé, R. & Pérez, M. A., A novel deep learning vision-based approach to structural damage detection. First Conference on Durability, Repair and Maintenance of Structures (DRMS 2023), Porto (Portugal), March 16-17, 2023. *(Oral presentation)*

Lado-Roigé, R. & Pérez, M. A., Detección de daño en estructuras a partir de la respuesta dinámica, mediante técnicas de deep-learning y visión artificial. XXIV Congreso Nacional de Ingenieria Mecanica (CNIM 2023), Las Palmas de Gran Canaria (Spain), October 25-27, 2023. *(Oral presentation)*

Other publications

Font-Moré, J., Reyes-Carmenaty, G., **Lado-Roigé, R.** & Pérez, M. A., Performance analysis of vibration-based damage indicators under low-modal information structures. MSSP 190, 110166. doi:10.1016/j.ymssp.2023.110166 (2023).

Reyes-Carmenaty, G., Font-Moré, J., **Lado-Roigé, R.**, & Pérez, M. A., Cross-domain Transfer Learning for Vibration-based Structural Damage Classification via Convolutional Neural Networks. Eng. Appl. Artif. Intell. *(Under review)*

Contents

List of figures

List of tables

List of abbreviations

ANN	**A**rtificial **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
CV	**C**omputer **V**ision
FCNN	**F**ully **C**onnected **N**eural **N**etwork
FRF	**F**recuency **R**esponse **F**unction
IQA	**I**mage **Q**uality **A**ssesment
LB-VMM	**L**earning-**B**ased **V**ideo **M**otion **M**agnification
MLP	**M**ulti-**L**ayer **P**erceptron
MSA	**M**ulti-head **S**elf **A**ttention
MUSIQ	**Mu**lti-**S**cale **I**mage **Q**uality Transformer
NLP	**N**atural **L**anguage **P**rocessing
RNN	**R**ecurrent **N**eural **N**etwork
RSTB	**R**esidual **S**win **T**ransformer **B**lock
ODS	**O**perating **D**eflection **S**hapes
SHM	**S**tructural **H**ealth **M**onitoring
STB-VMM	**S**win **T**ransformer-**B**ased **V**ideo **M**otion **M**agnification
STL	**S**win **T**ransformer **L**ayer
SWIN	**S**hifted **Win**dow
ViT	**Vi**sion **T**ransformer
VMM	**V**ideo **M**otion **M**agnification
VSHM	**V**ibration-based **S**tructural **H**ealth **M**onitoring

1 Introduction

This initial chapter presents the underpinning motivation for this thesis and outlines its central hypotheses and objectives. The chapter also delivers a brief overview of the outline of the contents of this work.

1.1 Motivation

Infrastructure collapses, although rare, happen all over the world. These incidents have serious consequences, including significant economic losses, delays, injuries, and even loss of life. Unfortunately, these events are more frequent than most people realise. For example, in July 2023, a bridge undergoing renovations collapsed near Patras (Greece), killing one person and injuring four more. Earlier in June, a bridge crossing over the Yellowstone River near Columbus, Montana (USA) also collapsed, causing several cars of a freight train carrying hazardous materials to fall into the water below. Further, in October 2022, a bridge in Morbi, India, collapsed killing 135 people and injuring many more.

Infrastructure plays a key role in society, and therefore ensuring its safety is of vital importance. For this reason, regular monitoring and maintenance is key to ensure optimal operation. Structural Health Monitoring (SHM) is a field of research that provides a set of tools to civil engineers for ensuring the integrity and safety of in-service structures. SHM systems monitor and assess structures by collecting data from sensor networks that allow for the tracking, identification, and measurement of structural responses that have powerful applications for operational cost reduction, maintenance, and public safety [1].

SHM systems are designed to continuously monitor the condition of a structure in order to detect defects or damage that may be difficult to identify during routine visual inspections. Early defect detection can help prevent catastrophic failure and reduce maintenance costs by allowing proper maintenance scheduling, allowing the performance of maintenance duties and repairs before damage propagates or becomes more severe. SHM provides a layer of improved safety against structural degradation on top of significant cost and time savings due to a decrease in unplanned service disruptions. Moreover, SHM can help ensure compliance with safety standards and regulations, as many structures are required to undergo regular inspections.

Nevertheless, the use of SHM systems in civil infrastructure is not ubiquitous as instrumenting those structures is generally a costly process. Sensors need to be installed, calibrated, and maintained to keep data flowing to the signal processing systems. In addition, the installation costs significantly increase with structural complexity, as larger and more complex structures require greater sensor coverage, thus requiring larger and costlier communications systems.

Novel vision-based monitoring systems offer an interesting value proposition as they are a non-intrusive alternative to contact-based sensors, being able to obtain full-field measurements with a single sensor placed far away from the structure. This makes vision-based techniques more cost-effective than their contact counterparts, nevertheless, these techniques are still immature and consequently present some technical shortcomings.

Vision-based techniques are more sensitive to environmental conditions, changes in lighting, weather, or temperature can affect the quality of the images captured, and in turn, the measurement's accuracy. Another shortfall, is the amount of data generated by these systems, as vision-based systems generate large amounts of data which can be challenging to manage and analyse. Effective data management and processing techniques are essential to ensure that the information generated is used effectively. Moreover, vision-based techniques may not be suitable for some applications due to visibility limitations, harsh environments, or privacy concerns.

Despite the challenges associated with vision-based approaches, they remain a promising option for lowering the barrier of entry to the use of SHM systems. The ubiquity of cost-effective cloud solutions, combined with the recent advancements in computer vision and deep learning, allow for the storage and processing of large amounts of data, thus enabling the exploration of new technical frontiers.

1.2 Hypotheses and research objectives

After analysing the context that inspired this work, five main hypotheses were formulated to guide the research presented and establish testable predictions about the expected outcome of this thesis.

- Learning-based video motion magnification techniques could be capable of accurately and reliably magnifying the complex movements of a structure, enabling the detection of subtle changes in its structural dynamic response.
- Combining Eulerian video motion magnification with deep learning techniques such as visual transformers and convolutional neural networks could significantly improve accuracy and efficiency over the state-of-the-art video motion magnification image quality.
- Incorporating advanced techniques such as procedural data generation, visual transformers, and convolutional neural networks into learning-based vision models could improve their accuracy and efficiency for structural dynamic response identification in complex monitoring systems, outperforming the state-of-the-art.
- Displacement-over-time signals obtained from video motion magnification could be precise enough to detect shifts in a structure's natural frequencies caused by changes in stiffness or mass.
- An automatic structural health monitoring video processing pipeline could be established using learning-based video motion magnification techniques to obtain displacement-over-time signals with high accuracy. Therefore, enabling the development of automated tools for vision-based monitoring.

Following these guidelines, this thesis' objectives are to develop, improve, and validate a learning-based Video Motion Magnification (VMM) methodology for the vibration-based monitoring of structures. This research focuses on showing how deep learning techniques can be applied in structural damage detection using optical means. This work also aims to produce practical tools for scientists and engineers who need to monitor and maintain structures efficiently and effectively by providing the methodology and the software described within it.

Based on the aforementioned hypotheses, the following specific objectives establish the milestones for the development of this research.

- Adapt and validate the use of machine learning for the detection of structural dynamic responses through optical means by developing a novel algorithm that combines video motion magnification with deep learning techniques.
- Improve the state-of-the-art learning-based vision models for the purpose of damage detection in complex monitoring systems by incorporating advanced techniques such as procedural data generation, visual transformers, and convolutional neural networks.
- Benchmark other learning-based vision systems against the proposed model by conducting experiments on a dataset of structural videos.
- Combine and integrate the advances in machine learning vision systems achieved in this thesis to develop a tool for the detection of structural performance degradation.
- Validate and verify the performance of the resulting tool on complex structures in a laboratory environment, using a comprehensive set of metrics and benchmarks to demonstrate its effectiveness and reliability.

1.3 Thesis outline

This document is structured in six chapters that summarise the theoretical framework, methodology, results, and conclusions of this thesis exploring novel vision-based approaches for the obtention of the dynamic response of structures.

- Chapter 1 covers the motivation, hypotheses and research objectives of this thesis.
- Chapter 2 establishes a basic theoretical foundation of deep learning concepts required for understanding this work and its methods.
- Chapter 3 reviews the most influential works in the three areas of knowledge related to this thesis. This chapter first presents an overview of the history and the state-of-the-art of deep learning in the computer vision space. Next, it discusses the context

around academic research on video motion magnification. And finally, introduces the topic of structural health monitoring along with its benefits and drawbacks, presenting the intersection with motion magnification techniques and their potential benefits.

- Chapter 4 presents the methodology and results of the work undertaken to understand and surpass state-of-the-art learning-based video motion magnification techniques and benchmarks the technical advances developed in the context of this thesis.
- Chapter 5 presents experimental work on using learning-based video motion magnification for damage detection purposes. It includes preliminary tests to validate the usage of the learning-based methodology, as well as validation of the improved motion magnification methodology and software tools developed in this thesis.
- Chapter 6 summarises the most relevant contributions and insights of this work and reflects on the future avenues of research for video motion magnification based techniques and their applications in mechanical and civil engineering.

2

Fundamentals of deep learning

The aim of this chapter is to establish a basic theoretical foundation to support the contents of this work, presenting the inner workings of modern deep learning models.

2.1 Introduction

Artificial neural networks (ANNs) [2, 3] are computational models inspired by the structure and function of biological neurons. They consist of interconnected units called artificial neurons that process information through weighted connections and activation functions. ANNs can learn from data by adjusting their weights and biases using various learning algorithms. ANNs have been widely used for various tasks such as classification, regression, clustering, dimensionality reduction, natural language processing, computer vision, speech recognition, etc.

Deep-learning is a branch of machine learning that focuses on building and training ANNs with multiple layers of artificial neurons. These layers can extract features from raw data at different levels of abstraction and complexity, enabling ANNs to handle large-scale and high-dimensionality data with minimal human intervention or domain knowledge. Note that, machine learning is sometimes considered a synonym for deep learning, however this is an inaccurate use of the term. Machine learning encompasses multiple techniques besides deep learning, such as decision trees, support vector machines, k-means clustering, among others.

Three types of learning behaviour [4] can generally be differentiated based on the form of training employed:

- **Supervised learning** refers to the class of algorithms and techniques that train a model (or agent) by using input/output pairs as examples. This implies providing the algorithm with both the problem and the intended solution so that it can learn the underlying logic or function.
- **Unsupervised learning** algorithms work with datasets that only provide inputs, requiring the discovery of patterns in the data. This means that unsupervised algorithms learn from data that has not been assigned any labels, categories, or classes. Unsupervised learning algorithms recognise common features in the data and respond according to whether these features are present or absent in each new data point.
- **Reinforcement learning** is a methodology that enables agents to perform actions in a specified environment with the aim of maximising a predefined cumulative reward. It is typically used in real-time applications, where the agent's actions on the environment elicit positive or negative feedback from the agent, who is programmed to optimise the acquisition of reward. Usually, the most challenging aspect of reinforcement learning is accurately defining rewards and penalties to facilitate a fast learning rate while forcing the agent to learn what is expected of it.

In recent years, deep learning has achieved remarkable results in many domains such as image recognition [5], natural language processing [6], machine translation [7, 8], self-driving cars [9],

etc. That makes deep learning one of the most relevant and influential technologies in the present and for the coming years.

2.2 Overview

ANNs are mathematical models that can learn from data to perform various tasks [4]. They are regarded as universal function approximators, provided that they have sufficient neurons and layers in their configuration. ANNs typically consist of linear algebraic operations interspersed with non-linear activation functions.

The learning process, or training, involves optimising the weights and biases of the mathematical model based on the discrepancy between the model's output and the desired outcome for a given input. A loss function is employed to measure the performance of the network during training. Subsequently, the gradient of the loss with respect to each weight is computed to determine the contribution of each weight to a specific result; this process is also known as backpropagation. Then, an optimiser algorithm adjusts the weights accordingly to minimise loss, and this process is repeated until the loss function attains a value within an acceptable margin of error. Once the system has been trained the models are generally tested, validated, and verified before their final deployment in production systems.

Training data constitutes a crucial component in any deep learning model, as it serves as the foundation upon which the model is trained and evaluated. The quality, diversity, and quantity of data that is employed to train a deep learning model can considerably influence its performance and accuracy. Data structures and algorithms are also imperative for effectively storing and processing large volumes of data, which is indispensable for training and deploying machine learning models.

2.3 Fully connected neural networks

Fully connected or dense neural networks are a type of artificial neural network that has no particular assumptions about the input data structure, and sometimes are also referred as Multi-Layer Perceptrons (MLP) [10, 11]. They consist of a series of fully connected layers, where each layer is a function from $m \in \mathbb{R}$ to $n \in \mathbb{R}$. Each neuron in a layer is connected to every other neuron in the next layer through a weights matrix. In addition to weights, biases are additional parameters added to each neuron in a layer, except for the input layer. Biases are used to shift the activation function of each node by a constant value. This can help the network learn more complex patterns and avoid underfitting.

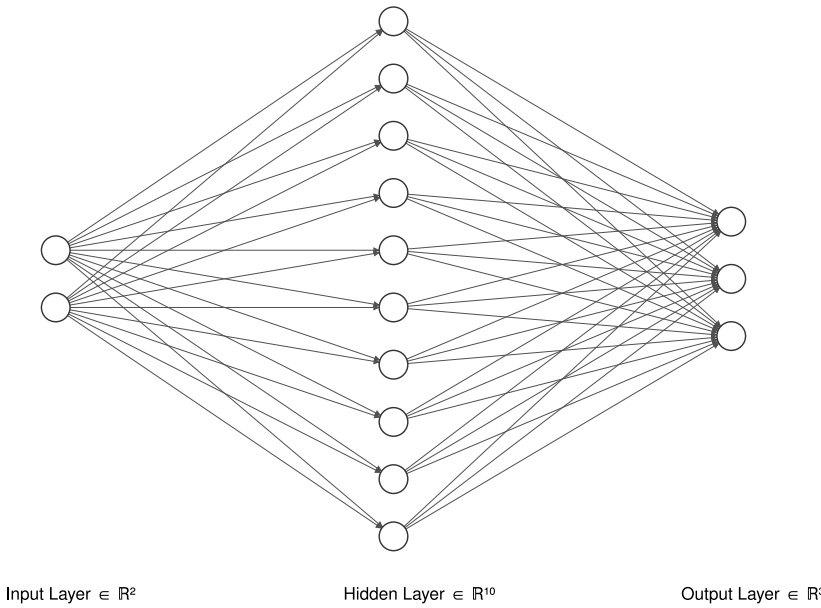


Figure 2.1: A diagram of a simple fully connected neural network, with 2 input neurons, 10 hidden layer neurons, and 3 output neurons

To calculate the output of a layer in a FCNN, first multiply the input values by the matrix of weights, then add the biases and apply the activation function.

$$y = f(W \times x + b) \tag{2.1}$$

where:

- y : is the result of the forward pass of a single layer
- f : stands for the activation function
- W : is the weights matrix
- x : is the input vector
- b : are biases

Taking, for example, the diagram in Figure 2.1 the forward calculation of the input layer towards the hidden layer would be:

$$HiddenLayer = f\left(\begin{bmatrix} i_1 & i_2 \end{bmatrix} \times \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} & w_{16} & w_{17} & w_{18} & w_{19} & w_{110} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} & w_{26} & w_{27} & w_{28} & w_{29} & w_{210} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 & b_8 & b_9 & b_{10} \end{bmatrix}\right) \tag{2.2}$$

And continuing with the example the final output of the network would be calculated like so:

$$Output = f\left(HiddenLayer \times \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \\ w_{61} & w_{62} & w_{63} \\ w_{71} & w_{72} & w_{73} \\ w_{81} & w_{82} & w_{83} \\ w_{91} & w_{92} & w_{93} \\ w_{101} & w_{102} & w_{103} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}\right) \tag{2.3}$$

This calculation completes the forward pass of the architecture described in Figure 2.1. To train a model like this one, after the evaluation of the forward pass the obtained results would be compared against known-good results, attributing result deviations to the trainable parameters (weights and biases) by using backpropagation. With these attributions calculated, an optimisation strategy, of which many exist, would be used to change those trainable parameters to minimise result deviation with respect to ground truth. This process would repeat until the calculated results of the forward pass would be within acceptable margins of error.

As can be inferred from this example, infinite architectures are possible, with more or less layers and neurons arranged following this simple structure. In the coming sections, strategies for choosing said structures and architectures to address concrete problems will be discussed.

2.4 Activation functions

Activation functions are used in neural networks to introduce non-linearity into the output of each neuron. Without them, neural networks would be limited to linear transformations, and consequently would be unable to model complex relationships between input and output data. Non-linear activation functions allow neural networks to learn and model complex and abstract patterns, which is crucial for many real-world applications such as image recognition, natural language processing, and speech recognition.

Some examples amongst the most popular activation functions used in deep learning are the Sigmoid, Tanh, ReLU, and Softmax functions. However, many more activation functions exist or can be invented [12], these four functions do not come close to being an exhaustive list of the available possibilities.

The sigmoid function was a commonly used activation function in the early days of neural networks, but it has largely been replaced by other functions due to its tendency to saturate and slow down learning. The Tanh function is another popular choice, which is similar to the Sigmoid function but with a larger output range. The ReLU (Rectified Linear Unit) function is currently one of the most commonly used activation function, as it has been shown to provide excellent performance and faster training times compared to other functions [12]. ReLU is a simple and fast function, attractive for many applications. Some authors have implemented modifications over ReLU such as the LeakyReLU to address the issues of "dead neurons" by allowing small negative values to be returned. The Softmax function on the other hand, is a specialised function commonly used in the output layer of neural networks performing classification tasks, that converts the output of each neuron into a probability distribution over the multiple outputs of the final layer.

Sigmoid:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Tanh:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

ReLU:

$$\mathrm{ReLU}(x) = \max(0, x)$$

Softmax:

$$\mathrm{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

When choosing an activation function, stability is an important consideration. Stability refers to the ability of an activation function to maintain consistent and reliable outputs over different inputs and network parameters. Some activation functions are more stable than others, depending on their output range, gradient behaviour, and numerical properties. For example, Sigmoid and Tanh functions are less stable than ReLU or other more modern functions such as ELU, as they are more prone to saturation, clipping, and numerical issues. ReLU and ELU functions are more stable than Sigmoid and Tanh functions, as they have unbounded output ranges and non-zero gradients for negative inputs. However, they can also cause instability if their inputs or outputs become too large or too small. Therefore, it is advisable to use proper initialisation and regularisation techniques to ensure the stability of the activation functions. Additionally, activation functions need to be differentiable in order for backpropagation to be used for training the neural network. If an activation function is not differentiable, the gradients cannot be computed, and the backpropagation algorithm cannot be used.

2.5 Skip-connections

Skip-connections shift the output of one or more layers as the input of one or more of the following non-consecutive layers. These connections help to avoid overfitting, increase model capacity, and preserve spatial information.

One of the most common types of skip-connections is the residual connection, which was initially introduced by the ResNet [13] architecture. The residual connection allows the output of one layer to be added to the input of another layer, bypassing some arbitrary amount of intermediate layers. This can be expressed as $y = F(x) + x$, where $F(x)$ is the function learned by the intermediate layers, and x is the original input. The idea is that $F(x)$ learns the residual or difference between the desired output and the input, rather than learning the output directly. This can help to prevent the degradation of the gradient signal and improve the accuracy of deep neural networks. Residual connections are widely used in computer vision tasks, such as image classification, object detection, and semantic segmentation.

Another type of skip-connection is the dense connection, which was introduced by DenseNet [14]. A dense connection concatenates the output of one layer to the input of all subsequent layers, creating a densely connected network. This can be expressed as $y = [x, F(x)]$, where $[x, F(x)]$ denotes the concatenation operation. The idea is that each layer receives information from all preceding layers and passes on its own information to all subsequent layers. This can help to reduce the number of parameters, increase feature diversity, and enhance feature propagation. Dense connections are also used in computer vision tasks, such as image classification and segmentation.

A third type of skip-connection, which is most commonly used in encoder-decoder networks, is the u-shaped connection (long-connection) introduced by Long *et al.* [15] and popularised by the U-Net [16] architecture. A U-shaped connection connects the encoder and decoder parts of a network, preserving the spatial information across different scales. This can be expressed as $y = F(x) + G(x)$, where $F(x)$ is the function learned by the encoder part, $G(x)$ is the function learned by the decoder part, and $+$ denotes an element-wise addition or concatenation operation. The idea is that each layer in the decoder part receives information from a corresponding layer in the encoder part with the same resolution, and combines it with its own information to reconstruct the output. This can help to improve the localisation accuracy and recover fine-grained details. U-shaped connections have been used in medical image analysis tasks, such as image segmentation and reconstruction [16–19].

As for many structures in deep learning there are no hard rules indicating when and where skip-connections should be used to improve performance. Skip-connections can have different forms and functions depending on the architecture and task, and are generally known to help

avoid overfitting, increase model capacity, preserve spatial information, and improve overall accuracy.

2.6 Convolutional neural networks

Convolutional neural networks (CNNs) [5, 20] are a specialised architecture for processing structured data, being very commonly employed in computer vision and image processing applications. CNNs exploit their mathematical structure to reduce the computational cost of processing ordered data. Unlike FCNNs, that require a one-dimensional fixed-size vector as input, convolutional layers can handle different-size inputs and harness the structure of the data to extract meaningful information. For instance, in images, neighbouring pixels are more informative than distant pixels for capturing the local context.

The convolution operation is the key defining characteristic that differentiates CNNs from other architectures. It involves sliding a filter or kernel over the input matrices and computing the dot product between the filter and the input at each position. This results in a feature map that highlights the regions of the input that match the filter. By using multiple trainable filters and stacking multiple convolutional layers, CNNs can learn complex and hierarchical patterns from data. In contrast, fully connected neural networks (FCNNs) require a one-dimensional fixed-size vector as input, which breaks the two-dimensional nature of images, making them less efficient and suitable for image processing than CNNs, which preserve the multidimensional structure of images and thus process them more effectively.

Mathematically, the general form convolution operation can be expressed as shown in equation 2.4. Note that * denotes convolution and is different from matrix multiplication.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} * \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} x_{(m-i)(n-j)} y_{(1+i)(1+j)} \tag{2.4}$$

Mechanically, convolution is performed by placing an arbitrary size filter or kernel over a matrix, multiplying each of the masked elements of the matrix and summing the results. Next, the calculated result is placed in the location where the centre element of the kernel was located, and the kernel is moved to the next element to repeat the process until the full matrix has been convoluted. However, an obvious issue appears at the edges of the matrix, where the kernel overflows outside the bounds. Multiple criteria exist for handling edges, for example, padding the outer border of the matrix with zeros, padding with a mirror image of the inner elements, cropping the kernel, etc. It is common for neural networks to use zeros padding to maintain

the dimensions of the input tensors stable. A useful formula for padding calculations is shown below (Eq. 2.5).

$$O = \frac{W - K + 2P}{S} + 1 \tag{2.5}$$

where:

O : Output height/width
W : Input height/width
K : Kernel size
P : Padding
S : Stride

Convolutional operations are commonly used to implement various image processing techniques such as blurring, sharpening, edge detection, etc. For example, to blur an image, a kernel that assigns equal weights to each pixel and its neighbours would result in smoothing out the variations in pixel values, thus blurring the image in the process. Using this reasoning, we can intuitively understand how CNNs function. CNNs are generally formed of multiple layers of trainable filters stacked together and interspersed with activation functions. Their training strategy is very similar to the one employed with FCNNs. The model is presented with an input, the output and loss are calculated, backwards propagation is performed, and then filters are optimised to minimise loss.

Convolutional layers can be found in both fully convolutional neural networks or in hybrid models. For example, it is common to find architectures for image classification that start with a few convolutional layers to extract position-sensitive features and then flatten their feature maps into a vector for a final stage of classification.

The context-aware architecture makes CNNs ideal for capturing the local context around data, this however, becomes a problem if a holistic understanding of a large input matrix is required. For that purpose, multiple strategies exist and will be further discussed in chapter 3. Most of these strategies transform the input matrix to multiple scales to extract the local context in multiple levels of resolution, effectively capturing the global context of the processed data. To alter the resolution it is common to see the application of pooling operations, such as max or average pooling, as well as convolutions with strides larger than one. Nevertheless, using a pooling layer after the activation function of the previous convolutional layer is generally considered the more robust option when downsampling on a convolutional model.

Pooling is a constant operation and needs no optimisation, this means that it does not require any learning parameters. The two most common pooling operations are average pooling and max pooling. Average pooling calculates the average value for each $n \times n$ patch on the feature map, while max pooling returns the maximum value for each patch. The result of using a

pooling layer and creating downsampled feature maps is a summarised version of the features detected in the input. Which can help some models become invariant to local translations and small-scale changes.

2.7 Recurrent neural networks

Recurrent neural networks (RNNs) are a class of neural network that specialise in processing sequential data [21], such as time series or natural language. In contrast to feedforward neural networks, which treat each input independently, RNNs retain a "memory" of previous inputs and utilise it to inform the processing of the following inputs.

The main idea behind RNNs is that they use the same set of parameters at each time step, enabling them to learn a function that maps input sequences to output sequences. This makes them particularly suitable for tasks such as language modelling, where the objective is to predict the next word in a sequence given the previous words.

One of the difficulties of training RNNs is the vanishing or exploding gradient problem, where the gradients used to update the parameters of the network become very small or very large as they propagate back in time, making it hard for the network to learn long-term dependencies. To address this problem, several variants of RNNs have been proposed, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which use specialised gating mechanisms to control the flow of information through the network and mitigate the vanishing gradient problem. However, RNNs have been mostly superseded by the Transformer.

2.8 Transformers

Transformers are a very versatile neural network architecture that has been used to great effect in language models, signal processing, time series forecasting, or computer vision [22]. The transformer architecture is based on the concept of self-attention, which focuses the model on different parts of the input sequence and weights them differently based on their relevance to the task at hand.

The transformer approach avoids the limitations of traditional recurrent neural networks that process the input in a sequential manner. In recurrent models, the computation depends on the symbol positions of both the input and output sequences, where, each position corresponds to a computation step, which produces a hidden state h_t as a function of the previous hidden state h_{t-1} and the current input. The sequential nature of recurrent neural networks computation prevents parallelisation within each training example, which becomes essential for longer sequence lengths, as memory limitations restrict batching across examples. At the same time,

the transformer architecture avoids the exploding or vanishing gradient problem altogether by processing all of the input data at once.

Attention mechanisms are components that manage and quantify the interdependence between different parts of an input or an output sequence. Self-attention is a particular case of an attention mechanism that relates different positions of a single sequence to compute a weighted representation of that sequence. This mechanism can capture long-range dependencies within sequences without using recurrence or convolution. An attention function maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output of such function is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Vaswani et al. [23] describe the particular self-attention mechanism used by the transformer, named "Scaled Dot-Product Attention", as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.6}$$

where:

Q : Queries vector

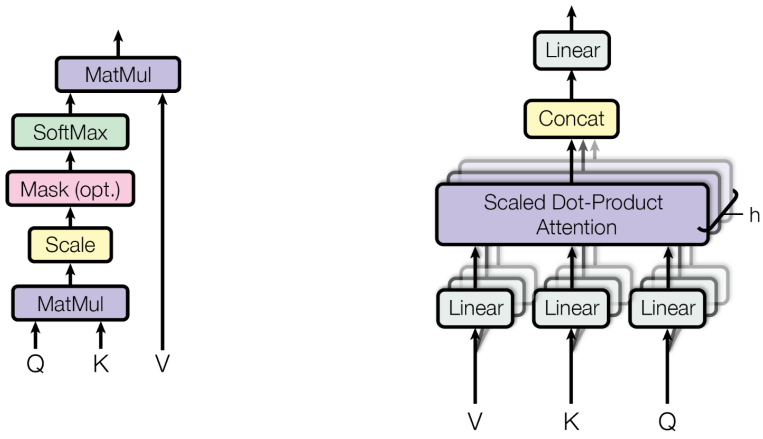
K : Keys vector

V : Values vector

d_k : Queries and keys dimension

Additive attention and dot-product or multiplicative attention are other common attention functions in neural networks. Additive attention computes the compatibility function using a feed-forward network with a single hidden layer. On the other hand, dot-product attention is identical to the previously described scaled dot-product attention except for the scaling factor. While the two attention methods are similar in theoretical complexity, multiplicative attention is preferred since it can be implemented using optimised matrix multiplication routines, therefore being the fastest and most efficient option in practice. Nevertheless, additive attention outperforms dot product attention for larger values of d_k [23], this effect appears because the softmax function is pushed to small gradient regions when dot products grow large in magnitude. For this reason, the scaled dot-product attention scales by $\frac{1}{\sqrt{d_k}}$.

The multi-head self-attention mechanism is the core component that enables the transformer to process data efficiently, allowing the model to jointly attend to information from different representation subspaces at different positions. Instead of performing a single attention function with keys, values, and queries multi-head attention runs h parallel attention functions with different learned projection functions. Once calculated, the resulting vectors are concatenated and projected once again.



(a) Scaled dot-product attention **(b)** Multi-head attention

Figure 2.2: Attention mechanisms [23].

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.7}$$

where:
$head_i : Attention(QW_i^Q, KW_i^K, VW_i^V)$

And where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$.

In Figure 2.3 the full transformer architecture is shown as described in [23]. The architecture is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. The transformer consists of an encoder and decoder each made up of N blocks, that process the input sequence to generate an output sequence in the decoder. Each block in the encoder and decoder consists of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network (eq. 2.8). A residual connection [24] is employed around each of the two sub-layers, followed by layer normalisation [25]. Additionally, the decoder has a sub-layer which performs multi-head attention over the output embeddings. The multi-head attention mechanism, that appears throughout the architecture, allows the model to jointly attend to information from different representation subspaces at different positions.

$$FFN(x) = max(0, xW_1 + b_1) * W_2 + b_2 \tag{2.8}$$

Similarly to other sequence transduction models, embeddings are used to convert the input tokens and output tokens to vectors of dimension d_{model}. However, since no recurrence or convolution is used, positional encodings [23] are employed to maintain the order of the

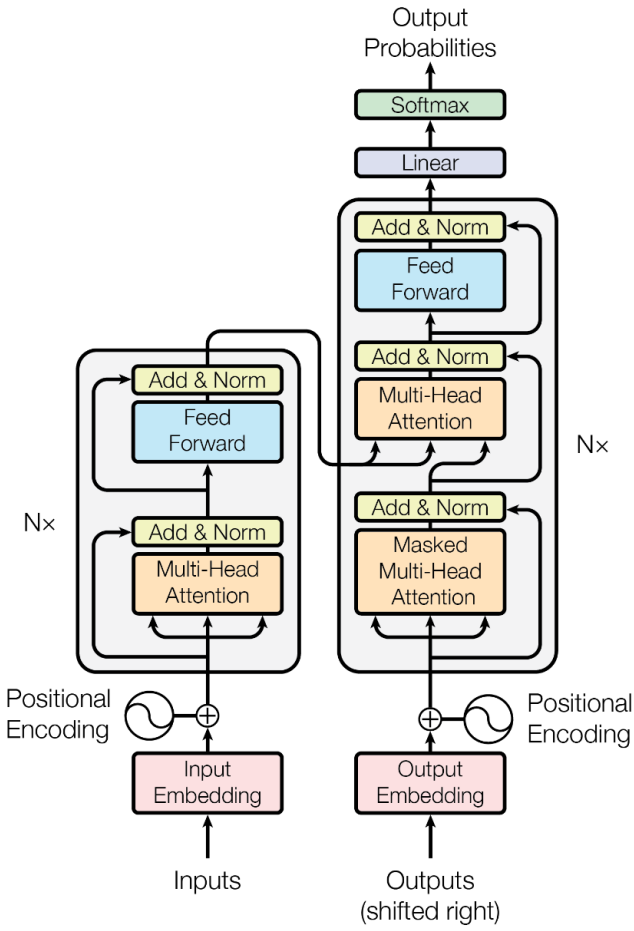


Figure 2.3: The transformer model architecture [23].

sequence. A well-known strategy is to use sine and cosine functions of different frequencies so that each dimension of the positional encoding corresponds to a sinusoid as such:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{2.9}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2.10}$$

where:

pos : position

i : dimension

At the time of writing, the transformer has become one of the most prominent and successful architectures in the natural language processing landscape, furthermore, it has also shown excellent results in other areas such as in vision or signal processing [22, 26].

2.9 Transformers in vision

As discussed, the transformer has been a key concept in the field of natural language processing, however, applying the transformer directly to images is not feasible within the constraints of present computing capabilities. The naive application of self-attention to images would require that each pixel attends to every other pixel, scaling quadratically with image size.

The Vision Transformer (ViT) [27] minimally modified the transformer architecture to accommodate image processing by breaking up images into fixed-size patches and linearly embedding them along with positional encodings, effectively treating each patch as a token. Then the resulting patches are fed into a standard transformer encoder. With this strategy the ViT achieved state-of-the-art performance on multiple image recognition benchmarks, albeit, given enough training data. Transformers lack some of the inductive biases inherent to CNNs, such as locality or translation equivariance, and therefore do not generalise well when trained on insufficient data.

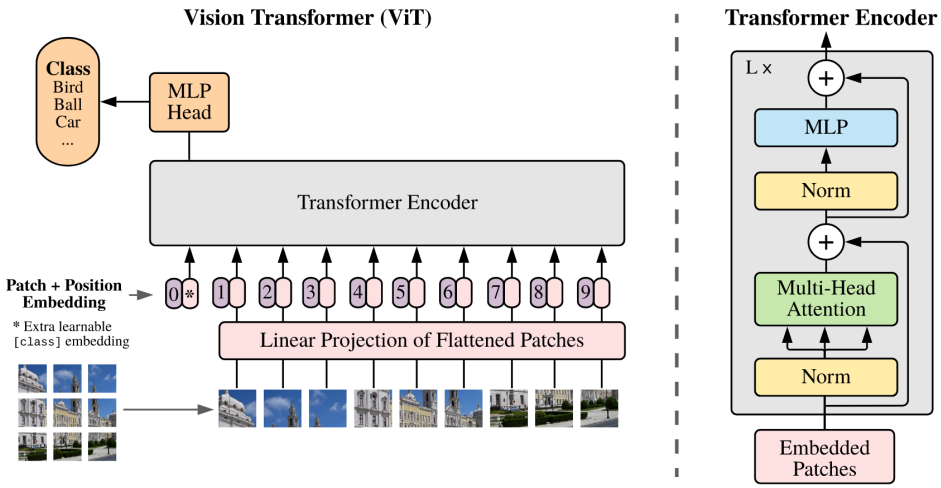


Figure 2.4: Vision Transformer model overview [27].

In Figure 2.4 an overview of the ViT structure is depicted. The standard transformer encoder requires a sequence of unidimensional token embeddings, however, images are naturally structured as bidimensional arrays. To produce such unidimensional sequences the image patches are flattened from $x \in \mathbb{R}^{H \times W \times C}$ into $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where H, W is the resolution of the original image, C the number of channels, P is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. At the same time, N also represents the input sequence length for the transformer. Then, data is position embedded in a similar manner as described in the previous section and fed into a standard transformer encoder. From there, the architecture can be derived to perform multiple image-processing tasks.

Subsequent iterations on the idea brought forth by the ViT have introduced new optimisations and performance gains. For example, the Swin Transformer [28] builds a hierarchical group of feature maps that break the image up into multiple small ViT-like processing areas to capture local information that builds back up to the full frame. This structure has the advantage of presenting linear computational complexity with respect to input image size, therefore allowing the processing of bigger images or the use of deeper networks.

2.10 Model performance optimisation

One of the main challenges of deep learning is the design and optimisation of models and their hyperparameters to optimise their performance on a given task. Hyperparameters are parameters not learned by the model, but are set by the user before training, such as optimiser learning rate and momentum, batch size, number of layers, etc. Tuning is essential for achieving good results but is also a very challenging [29, 30] and time-consuming task that is not yet fully understood.

Common pitfalls when training deep learning models include:

- Lack of training data: Deep learning models require large amounts of labelled data to learn complex patterns and generalise well to new inputs. However, obtaining and annotating such data might not be feasible in some cases. Time, cost or the physical impossibility of obtaining data compels the use of techniques such as data augmentation, transfer learning, semi-supervised learning, or synthetic data generation.
- Imbalanced data: Some applications display skewed distributions towards some classes or categories, while other classes are underrepresented or rare. This can cause the model to be biased towards the majority class and perform poorly when shown less common scenarios.
- Overfitting: Overfitting occurs when a particular model performs well on the training dataset but fails to generalise when tested with previously unseen data. This can result from having too many parameters relative to the amount of data, or from having noisy or irrelevant features in the data.
- Vanishing/exploding gradients: As a result of training, deep learning models can suffer from vanishing or exploding gradients as they get calculated during backpropagation becoming either excessively big or small. These problems are especially common in very deep neural networks, and result in unstable models or models that learn very slowly or not at all. To reduce these problems, normalisation techniques such as batch or layer normalisation are often used.

- Poor performance: General poor performance can be caused by multiple factors or combinations of multiple factors, from insufficient network parameters and bad network initialisation to suboptimal optimiser settings. Approaching these scenarios is a difficult task that often requires an in-depth methodical study of the model hyperparameters and the task to be performed. However, not many authors share their optimisation procedure in favour of a more streamlined presentation of their models, nevertheless, some efforts exist to establish some scientific guidelines for model optimisation [31].

3

Literature review

This chapter reviews the history and state-of-the-art of the scientific fields related to this thesis and points out the gaps in the literature this work addresses.

3.1 Deep learning in computer vision

3.1.1 Origin

Deep learning is a branch of machine learning framed within the larger artificial intelligence field. Deep learning has proven to be a powerful technology that is transforming many computer science disciplines by providing powerful insights into data to improve translations, speech recognition, image and pattern recognition, or generating new content.

The roots of deep learning date back to 1943 when Warren McCulloch, a neurophysiologist, and Walter Pitts, a mathmatician, presented their mathematical bio-inspired model of a neural network [32]. The model, known as the MCP neural model, was a linear step function triggered by linearly interpolated weighted data. This model presents notable similarities to modern neural networks, however it also presents significant differences with respect to modern perceptron models. The MCP model was built as an electrical circuit with fixed non-trainable weights, a key feature of modern ML models. Moreover, the model also presented inhibitory inputs that could completely prevent neuron activation at any time, another feature not present in modern models.

Despite this modelisation effort, the foundation for modern neural networks would not be established until six years later when Donald Hebb, considered by some as the father of modern neural networks [2], introduced the Hebbian Learning Rule [33], inspired by previous work in neural groupings [34]. The Hebbian Learning Rule describes how neuronal activity influences the connections between neurons; stating that when the axon of a cell is close enough to excite a second cell and takes part in its activation in a repetitive and persistent way, the connection between them becomes stronger. This aspect of the rule is also known as the synaptic plasticity of Hebbian learning.

Later, in 1958, inspired by the success of the MCP model and the Hebbian Learning Rule, Frank Rosenblatt introduced the perceptron [10]. Rosenblatt constructed an electronic device that showed the capacity of learning by associationism. This is unlike previous theorists like Hebb that would focus on the biological system in the natural environment. The perceptron was fundamentally a linear function of input signals, and therefore it was limited to only represent linear operations. This limitation was criticised by other researchers [35, 36] leading to a stagnation period in the field, further accentuated by the limited computing power available at the time, that lasted until the 1980s. Nevertheless, the evolution from perceptrons to artificial neural networks was only a small step away. Placing multiple perceptrons together into layers and stacking those layers together would yield a Multi-Layer Perceptron or MLP [37], that is widely considered as a universal function approximator used in modern deep learning architectures.

The 1980s saw significant works improving, understanding, and inventing new neural network architectures such as the Hopefield Network [3], a neural network with bidirectional weights, the Boltzmann Machine [38], that introduced the concept of hidden layers in the Hopefield network, or the Restricted Boltzmann Machine [39] and the first Recurrent Neural Network (RNN) [40].

However, an especially significant development for computer vision would come in 1989 when Yann LeCun introduced LeNet, the first convolutional neural network [20]. Inspired by the Neocognitron [41] and the visual cortex, LeCun pioneered the Convolutional Neural Network (CNN) and further popularised it by using it for handwritten character recognition [20]. Nevertheless, the kickstart to the modern era of deep learning would come later in the 2010s, inspired by the introduction of Deep Belief Networks by Ruslan Salakhutdinov and Geoffry Hinton in 2006 [42] and the concept of dropout in 2012 [43].

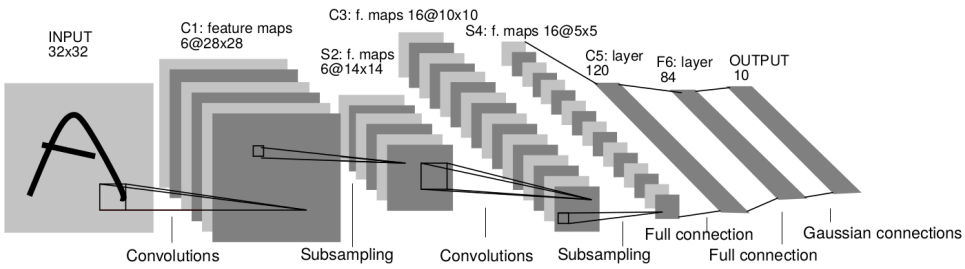


Figure 3.1: LeNet-5 model architecture [20].

If LeNet pioneered the convolutional neural network, AlexNet [5, 44] brought them into the mainstream. Invented by Krizhevsky et al. in 2012, AlexNet was the first CNN that proved this technology could perform well on historically difficult challenges such as the ImageNet [45] dataset. AlexNet was a remarkable jump forward both in terms of structure and training, using GPU compute to significantly improve training times, enabling the use of larger training sets and enabling faster design iteration. However, despite its great performance, AlexNet did not have a solid theoretical base behind some of its architectural choices [2] and for this reason it is difficult to point out which aspects of the model contributed the most to its success. Subsequent architectures would explore many different configurations of layers, connections, and kernel sizes; the rest of this section will focus on some of the most relevant findings.

One of such interesting findings would come in 2014 with the Visual Geometry Group model or VGG [46], which demonstrated how smaller kernel sizes, like 3×3, could perform as well or better than larger kernels used at the time like 5×5 or 11×11. This showed experimentally that the parallel assignment of these small-size filters could produce the same influence in the receptive field as larger filters, with the added benefit of reducing the computational cost. In addition, the VGG network regulated its complexity by inserting 1×1 convolutions in the middle of its convolutional layers. This allowed the network to learn linear groupings of

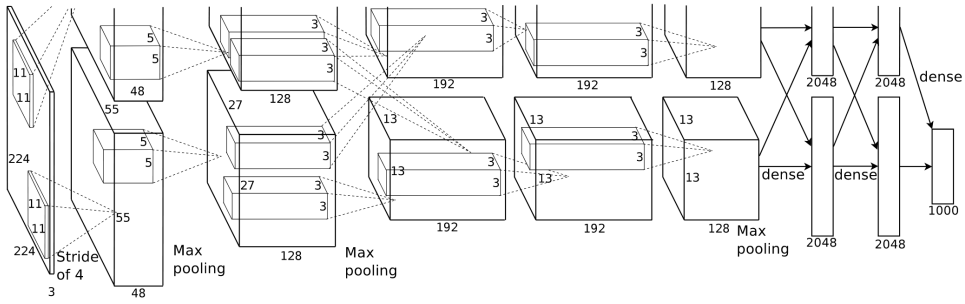
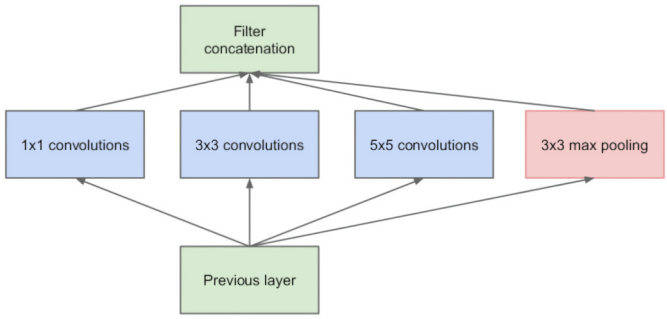


Figure 3.2: The AlexNet architecture is designed to divide its calculations between two GPUs. One GPU processes the top part of the figure, while the other handles the bottom part of the figure, as illustrated in [44].

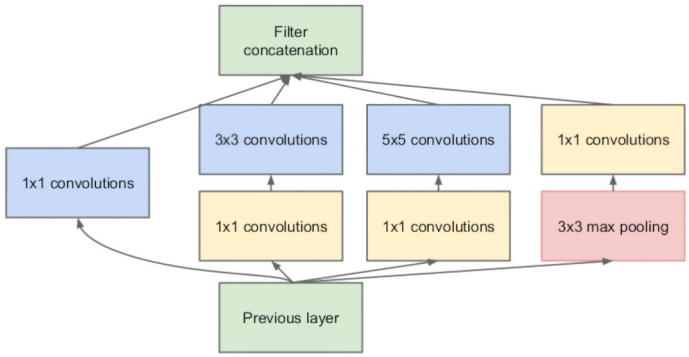
subsequent feature maps. Moreover, to fine-tune the network, a max pooling layer was added after each convolutional layer, using padding to preserve the spatial resolution. Overall, VGG achieved impressive results in both image classification and localisation tasks. However, VGG did not win the ImageNet competition that year, but planted the seeds for highly influential models like Residual Net [24].

The winner of the ImageNet 2014 challenge was not the VGG model but instead GoogLeNet [47], named after LeNet [20], which introduced the concept of the inception module, later used on the R-CNN [48, 49] object detection model. The inception module or inception block combined multiple-scale convolutional transformations by employing merge, transform, and split operations for feature extraction, as illustrated by Figure 3.3. The main hallmark of this architecture was the improved utilisation of the computing resources inside the network. This was achieved by a carefully crafted design that allowed for increasing the depth and width of the network while keeping the computational budget under control. This architecture was 22 layers deep but had an order of magnitude less trainable parameters than AlexNet [5], which had 14 layers less.

Then, following some of the ideas previously established by VGG, ResNet or Residual Net [24] explored the usage of deeper architectures with simpler layers. At the time, researchers found that deeper networks usually performed better than shallower ones, however there were diminishing returns passed a certain point, even leading to performance degradation. ResNet, however, was ten times deeper than other networks used at the time but managed to achieve better results. This was made possible by the residual connection, which allowed ResNet to avoid losing too much of the original input information in its multiple convolutional layers while creating a deeper and more performant model. As illustrated in Figure 3.4, the residual connection skips data over several layers and adds it to the output of subsequent layers, enabling higher-level features to be preserved further down the model. This idea has been highly influential in the field of deep learning, and residual connections are now commonly found in many top-performing models.



(a) Inception module, naïve version



(b) Inception module with dimension reductions

Figure 3.3: Inception block architecture [47].

During the last decade, many researchers have worked on particular applications of deep learning that have enriched the literature and provided inspiration for newer architectures and applications. The following table presents a summary of some of the most influential models and milestones in the deep-learning field.

Table 3.1: References on the evolution and growth of the deep-learning techniques in image-related tasks.

Reference	Contribution
Bain, 1873 [34]	Introduced neural groupings, a term used to refer to neural networks, thus laying down the preliminary basis for contemporary neural networks.
McCulloch & Pitts, 1943 [32]	Created the first mathematical approximation to a bio-inspired model of a neural network.

Table 3.1: Continued from previous page.

Reference	Contribution
Hebb, 1949 [33]	Defined the Hebbian learning rule and established the foundations for modern artificial neural networks. The Hebbian rule specifies a variable weight in the connection between two neurons in proportion to the product of their activation.
Rosenblatt, 1958 [10]	Introduced the first perceptron.
Werbos, 1994 [50]	Introduced the backpropagation algorithm.
Fukushima, 1980 [41]	Defined the Neocognitron model.
Hopfield, 1982 [3]	Crated the Hopefield Network.
Ackley *et al.*, 1985 [38]	Created the Boltzmann Machine.
Smolensky, 1986 [39]	Created the harmonium, later known as the Restricted Boltzmann Machine.
Jordan, 1986 [40]	Defined and introduced the Recurrent Neural Network (RNN).
Lecun *et al.*, 1989 [20]	Introduced LeNet and pioneered the use of Convolutional Neural Networks (CNNs).
Hochreiter & Schmidhuber, 1997 [51]	Introduced LSTM, and solved the problem of vanishing gradients in RNN
Hinton *et al.*, 2006 [52]	Introduced Deep Belief Networks and layer-wise pretraining.
Salakhutdinov & Hinton, 2009 [42]	Introduced the Deep Boltzmann Machines.
Hinton *et al.*, 2012 [43]	Defined the Dropout mechanism to mitigate overfitting and improve training efficiency.
Krizhevsky *et al.*, 2012 [44]	Introduced AlexNet winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [45, 53] and popularising the usage of dropout, ReLU and GPU compute. This work had a very remarkable impact in the field of computer vision and AI, proving the capabilities of CNNs.
Kingma & Welling, 2022 [54]	Introduced the Variational Autoencoder (VAE), a type of generative model that uses neural networks to learn a compact representation of data. It consists of two main components: an encoder network that maps the input data to a latent representation, and a decoder network that reconstructs the input data from the latent representation.
Simonyan & Zisserman, 2015 [46]	Experimented with deeper neural networks and smaller kernels that reduced the computational cost of their VGG model. This work would go on to inspire other very influential architectures and ideas.

Table 3.1: Continued from previous page.

Reference	Contribution
Szegedy *et al.*, 2014 [47]	Introduced GoogLeNet along with the concept of blocks with heterogeneous layers and operations. Moreover, it inspired several improved versions of the architecture [55–57].
Goodfellow *et al.*, 2014 [58]	Introduced the Generative Adversarial Network (GAN), a class of generative model that uses two neural networks to generate new data samples that are similar to a training dataset. The two networks compete to improve their performance, where one generates new images similar to the dataset, and the other trains to detect those newly generated images. These networks create a feedback loop that improves the generative capabilities of the resulting model.
He *et al.*, 2015 [24]	Introduced the residual connection and ResNet, enabling deep-learning models to go deeper.
Szegedy *et al.*, 2016 [57]	Combined the inception modules with ResNet, introducing the concept of residual links.
Vaswani *et al.*, 2017 [23]	Introduced the Transformer architecture, establishing the foundational mechanism for many well-known natural language processing (NLP) models, including BERT [7] and GPT [6, 59, 60].
Huang *et al.*, 2018 [61]	Introduced DenseNet and the concept of dense connectivity, where each layer is connected to every other layer in a feed-forward fashion.
Howard *et al.*, 2017 [62]	Presented MobileNet a class of efficient models designed for mobile and embedded vision applications.
Wang *et al.*, 2020 [63]	Created HRNet, a convolutional neural network that maintained several high-resolution representations throughout the whole network, allowing the model to better capture fine-grained details in images.
Dosovitskiy *et al.*, 2020 [27]	Introduced the Vision Transformer (ViT), using the well-known transformer architecture for image processing.
Liu *et al.*, 2021 [28]	Modified the Vision Transformer by using a shifted window approach to reduce computational cost while improving the receptive field and performance of the original model.
Liu *et al.*, 2022 [64]	Introduced ConvNeXt, a new convolutional architecture competitive with the transformer-based models.

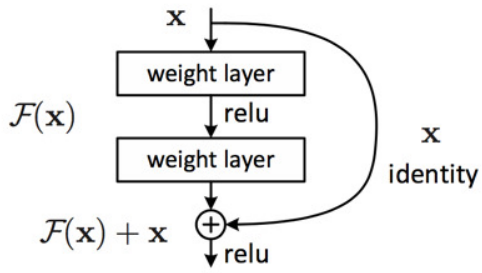


Figure 3.4: Diagram depicting a residual connection [24].

3.1.2 Transformers in vision: Modern deep-learning computer vision

Computer vision tasks during the 2010's decade have experienced great advancements, massively improving the performance of tasks such as image classification, segmentation, colourisation, reconstruction, among many others. These developments have in the majority of cases made use of CNNs, which have been one of the more dominant architectural fixtures in the space. During the same period, the field of natural language processing saw a similar evolution mostly relying on recurrent neural networks. However, in 2017, the development of the Transformer [23] would mark a significant architectural milestone and become the dominant backbone architecture for language processing tasks. This development has sparked the creation of models such as BERT [7] or GPT [6, 59, 60] that have made a significant cultural impact on the mainstream public.

Remarkably, despite the disparity of tasks, the transformer architecture was adapted for vision tasks in the 2020 article titled "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [27], demonstrating the great flexibility of its architecture. This adaptation left most of the original [23] transformer structure intact, only adding a preliminary encoding step to convert images to tokens. The Vision Transformer (ViT) [27] was proven to perform on par or better than contemporary CNN-based models and spawned many derivative works [28, 65–77] adapting the technique for particular applications or proposing improvements at the architectural level.

Except for the initial image encoding stage, in which the image is cut into patches and then treated as tokens, the Transformer architecture introduces no image-specific inductive bias and is capable of analysing scenes regardless of scale. These facts enable the transformer to outperform standard ResNets by a significant margin given a large enough model and dataset. At the same time, not having these inherent biases generally means that Transformers require larger datasets to be trained well enough to outperform the more specialised CNN-based

models.

However, the Transformer has not completely taken over the computer vision field, and some CNN-based or hybrid models have been introduced to rival the performance of transformer-based models [64, 78]. Moreover, the evolution of the transformer-based models has advanced towards hierarchical transformers [28] that introduce ideas resembling CNN's that make vision transformers viable backbones for vision tasks other than image classification. Convolutional neural networks still offer very compelling properties and are very unlikely to disappear.

In the last few years since the ViT introduction the debate [79–84] to determine which backbone is better for image tasks has been active with authors claiming one or the other architecture is the superior choice. Many of these comparisons show intense competition with slim victories, which indicates that the debate is far from settled.

3.2 Video motion magnification

Video Motion Magnification (VMM) is a task included in the field of computer vision whose goal is to detect and magnify movements in video sequences. These techniques have been used to detect physiological parameters on the human body [85–89] or animals [90], to detect and analyse vibrations in machines or buildings [91–96], to enhance movements under a microscope [97, 98], or even to read microexpressions and detect AI generated videos [99, 100]. These techniques are often described as microscopes for motion [101] as they produce videos in which very subtle motions and deformations become clearly visible aiding in the visualisation of motions that otherwise could have gone unnoticed.

Unlike traditional microscopes reliant on optics, motion magnification techniques are software solutions that harness the inherent properties of video sequences. Thus allowing the application of these techniques without the need for highly specialised equipment beyond a video camera. In this sense, motion magnification techniques offer a cost-effective and low-maintenance solution for the diagnostic of dynamic scenes.

Techniques for motion magnification can be approached from two different theoretical frames of reference, the Lagrangian and the Eulerian approaches. On the one hand, the Lagrangian approaches use optical flow techniques to track movement in the scene and extrapolate those movements to generate a magnified video. Optical flow techniques [102] are algorithms used to estimate the motion of objects in visual scenes by analysing the changes in brightness patterns over time. These techniques generate a description of both the area of the image experiencing motion and the velocity of that motion. Optical flow calculates and tracks the movement of each pixel between successive frames, providing a feasible way to obtain the trajectory of objects. On the other hand, Eulerian techniques measure and amplify variations over time

based on pixel-wise changes with fixed spatial locations. This approach focuses on changes at specific locations in the video frame along the temporal dimension, and is well-suited for visualising smaller movements and changes. These approaches are named after the Eulerian and Lagrangian descriptions of motion in fluid dynamics, which focus on describing the behaviour of specific locations in space or individual fluid parcels, respectively.

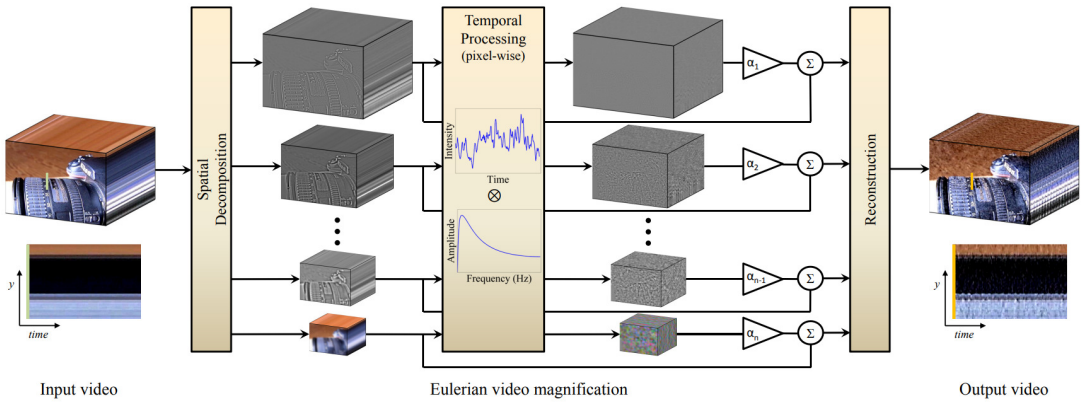


Figure 3.5: Overview of the Eulerian video motion magnification framework, adapted from [103]. The Eulerian Video Magnification framework processes an input video sequence by decomposing it into distinct spatial frequency bands. A temporal filter is applied uniformly to all bands, which are then amplified by a factor of α and recombined with the original signal. The resulting output video is generated by collapsing the amplified bands. The selection of temporal filter and amplification factors can be customised to accommodate various applications.

The concept of video motion magnification was introduced in 2005 by Liu *et al.* [101]. Their work presented a Lagrangian perspective to amplify subtle motions in video sequences, following particles over time and analysing feature point trajectories. The technique proposed a procedure to automatically group several pixel clusters based on position, intensity, and motion. Then, using a motion similarity measure, motion groups are defined based on their trajectories over time and instantaneous velocities, allowing the system to determine which clusters should be magnified and by how much.

This technique pioneered motion magnification techniques, however presents two notable shortcomings with respect to more modern Eulerian approaches. The first is that optical flow techniques tend to be computationally expensive with respect to the more modern linear and phase-based Eulerian approaches. And the second flaw is that magnification quality is not that good compared to Eulerian techniques, especially struggling in areas with object occlusion or complex motion. This often results in blurry magnification and frequent visual artefacting. Optical flow is not optimal for tracking small and sub-pixel motions, and it is often better employed in object detection and tracking, image dominant plane extraction, movement detection, robot navigation, or visual odometry [102, 104–108]. For this reason, subsequent developments in motion magnification evolved towards Eulerian approaches.

Within the Eulerian classification, three different types of techniques can be found: linear, phase-based, or learning-based, which were developed chronologically in that order. The first linear video motion magnification approach was introduced by Wu *et al.* (2012) [103] and offered an efficient and straightforward method for magnification based on image filtering. The structure of the model is shown in Figure 3.5. This technique involves decomposing the input video into spatial frequency bands using a Laplacian image pyramid, applying a temporal filter to all bands to remove frame noise, and magnifying the filtered spatial bands by a user-specified magnification factor. The resulting magnified signal is then combined with the original signal obtained during the decomposition step, and the spatial bands are collapsed to render the magnified output video.

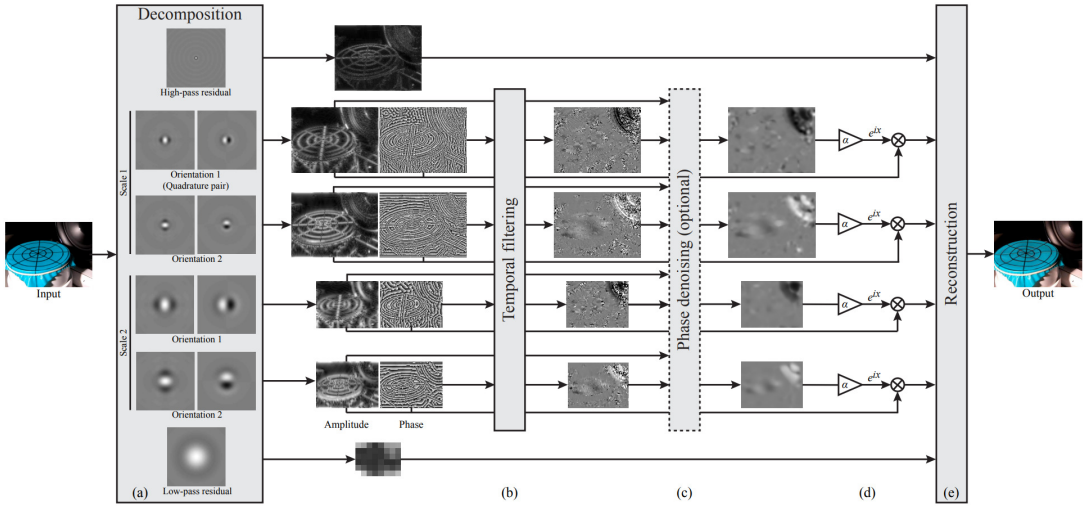


Figure 3.6: The phase-based approach manipulates motion in videos by analysing the signals of local phase over time in different spatial scales and orientations. The video is decomposed using complex steerable pyramids to separate the amplitude of local wavelets from their phase (a). The phases are then independently temporally filtered at each location, orientation, and scale (b). Optionally amplitude-weighted spatial smoothing can be applied to increase the phase signal-to-noise ratio, which has been empirically shown to improve results. Then temporally-bandpassed phases are amplified or attenuated (d) before the video is reconstructed (e). Adapted from [109].

This linear approach is characterised by using first-order Taylor expansion, which reduces the computational cost with respect to the Lagrangian method and improves performance. However, this method still presents some downsides, as linear Eulerian approaches tend to suffer when dealing with high magnification factors or large input motions. These limitations may lead to incorrect noise filtering causing noise magnification, and ultimately provoking a significant degradation in output image quality. Zhang *et al.* [110] proposed an alternative approach to address the limitation posed by large motions in video magnification by utilising a temporal processing technique based on the second-order Taylor series expansion, effectively converting the model from motion magnification to acceleration magnification. While this

approach somewhat alters the goal, this technique may be used for similar applications and thus it is worth mentioning.

On the other hand, phase-based approaches address some of these limitations presented by the linear techniques. Wadhwa *et al.* [111] introduced a new approach offering better outputs and a more robust behaviour against noise. This new technique applies filtering to the signal's phase at various spatial orientations and scales. Unfortunately, this improved method presents the downside of being more computationally expensive than the linear solutions. Figure 3.6 illustrates the frame decomposition process using complex steerable pyramids, followed by the temporal band-pass filtering, the phase manipulation process, and finally the video reconstruction step. The goal of this process is to reconstruct a video where the phase has been magnified but the magnitude is left unaltered. Subsequent developments by Wadhwa *et al.* [109] would focus on reducing the computational cost of the phase-based method while keeping image quality performance constant. These improvements would allow phase-based methods to be applied in real-time applications.

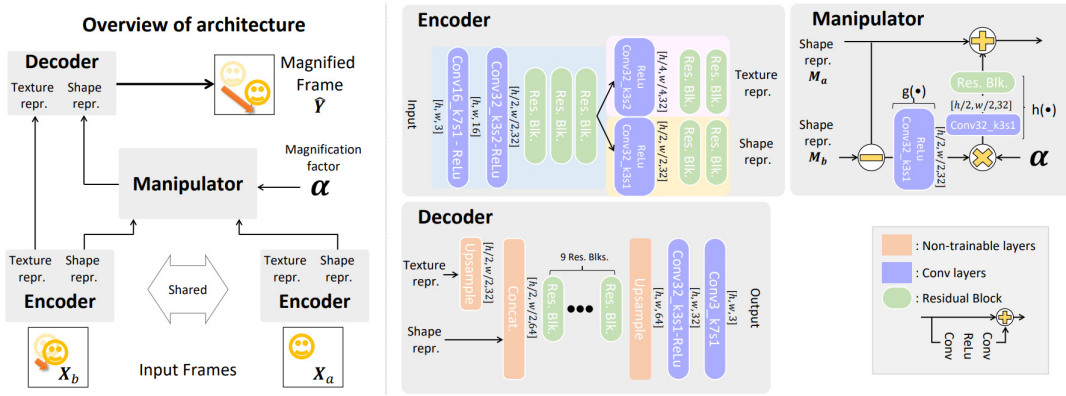


Figure 3.7: Oh *et al.*'s learning-based architecture, with a model overview pane on the left and its detailed components on the right. Adapted from [112].

Finally, the last branch of Eulerian techniques to be developed were the learning-based techniques [112]. Instead of defining custom application-specific filters, these techniques rely on learning-based approaches such as CNNs or Transformers to process the video sequences, as shown in Figure 3.7. In other words, learning-based techniques learn the filters from motion magnification data instead of implementing hand-designed filters out-of-the-box. This enables state-of-the-art output image quality without the explicit need for temporal band-pass filters that restrict the frequency of the magnified motions. Moreover, these techniques retain the capability of using a user-configurable magnification factor. However, these models are computationally expensive and can not be used in real-time applications.

3.3 Structural health monitoring

Structural Health Monitoring (SHM) is a discipline that aims to monitor the condition and performance of structures, such as bridges, buildings, wind turbines, pipelines, railways, etc. By employing sensors, data acquisition systems, data analysis methods, and software tools, SHM detects and identifies damage, deterioration, anomalies, or changes in the structural behaviour. The main objectives of SHM techniques is to enhance the safety, reliability, and efficiency of structures and infrastructure. For this reason, the main key issues of concern for the SHM discipline can be broken down into data acquisition, data processing, damage detection, damage diagnosis, and damage prognosis.

SHM aims to design automated processes that protect an entire structure or a significant part of it automatically or with very little human intervention, and thus, these systems must be capable of damage detection. Damage detection can be loosely defined as the capacity to detect variations in the nominal operating parameters of a structure. Nevertheless, modern understandings of SHM tend to rely on the definitions formalised by Rytter [113] that defined four levels of damage identification. Level 1 establishes the capacity to yield a qualitative indication of the presence of damage in a structure, also known as damage detection. Level 2 methods, should at least be able to provide information about the probable location of the damage in addition to detecting it. Level 3 must additionally provide information about the size or magnitude of the damage. Finally, level 4 methods inform about the actual safety of the structure given the reported state. The four levels can be further summarised as detection, localisation, assessment, and consequence prediction.

Other authors [114] have expanded this definition by adding two additional levels for self-diagnosis and self-healing, meaning that the detection, localisation and quantification would be done by the structure itself as well as being able to repair itself in case of damage.

Further, Farrar *et al.* [115, 116] defined the SHM process in terms of a four-step statistical pattern recognition paradigm. Being defined as a process consisting of four stages:

1. Operational evaluation
2. Data acquisition, normalisation and cleansing
3. Feature selection and information condensation
4. Statistical model development for feature discrimination

Operational evaluation is the process by which limitations are set on what will be monitored and how the monitoring will be accomplished. This evaluation tailors the damage identification

process to features that are unique to the system being monitored and tries to take advantage of unique features of the damage that is to be detected.

Data acquisition involves the selection of the excitation methods, the sensor types, and the data acquisition storage and transmittal hardware. This process is yet again application specific, and economic considerations play a major role in these decisions. This is also one of the major focuses of this thesis, proposing a cost-efective alternative for vision-based data acquisition.

Feature extraction and information condensation is the area of SHM that receives the most attention, defining the characteristic markers and data processing pipelines [117] to reliably diagnose a structure.

Finally, the fourth stage consists of the statistical model development for discrimination between features from the undamaged and damaged structures [118]. Statistical model development is concerned with the implementation of the algorithms that operate on the extracted features to quantify the damage state of the structure. And is closely related to the levels of damage defined by [113].

SHM systems not only aim to provide advantages in safety but additionally can reduce maintenance costs and inform engineers about the actual performance of structures so that they can be contrasted against design assumptions and inform improvements or new constructive techniques. Reductions in maintenance costs are linked to the capacity of these systems to provide early warning for deterioration and damage, allowing for proper maintenance planning and scheduling. Predictive maintenance can significantly cut down costs by reducing the frequency in which large costly repairs must be performed, since addressing problems early avoids unnecessary deterioration. Moreover, proper maintenance planning reduces the costs associated with emergency maintenance operations, and of course increases safety.

SHM applications target many subfields of engineering such as the civil, aeronautical, or automotive sectors, presenting several challenges and opening a wide field of research for measurement technologies and damage detection algorithms. One of the most popular approaches to structural diagnosis are vibration-based SHM techniques.

3.3.1 Vibration-based structural health monitoring

Vibration-based Structural Health Monitoring (VSHM) [115, 119–137] is a subfield within SHM that encompasses a range of non-destructive and non-invasive techniques for monitoring the health of structures. These techniques collect and analyse the dynamic responses of structures to determine their inherent properties and assess their current state. Structures and objects have natural modes of vibration that are affected by their geometry and material properties. These mechanical properties can be easily measured and provide valuable information for SHM purposes.

When the vibrational signature of a structure changes, it can be inferred that either the structure itself or the boundary conditions have changed. This allows VSHM methods to detect damage even when there is no sensor near the damaged zone. However, this is not a trivial task, as variations in the loading of the structure or changes in the ambient temperature will also affect its dynamic response [138–140]. Therefore, researchers in the field of VSHM have developed multiple methods for capturing and analysing these dynamic responses.

VSHM methodologies can be classified into three different domains: time, frequency, and the modal domain. These approaches offer a trade-off between post-processing and the physical input data required [137].

Time domain methods examine the vibration response of a structure to identify the presence of damage. These methods can be further subdivided into two groups: non-parametric methods, based on correlation functions of the vibration responses of the structure [141–145]; and parametric methods which generate a model from parameters obtained by the vibration responses. An example of a parametric method is [146], based on the use of transmissibility in the time domain. This method takes the discrepancy of transmissibility of the structure response before and after damage as the basis for finite element model updating. Then, the damage is located and quantified through iteration by minimising the difference between the measurements at the sensor location and the reconstruction response extrapolated by a finite element model.

Frequency domain methods [136, 147–155] on the other hand, use the frequency response of the structure directly as a feature for damage identification, correlating changes in the Frequency Response Function (FRF) with physical structural variations. Therefore, comparing the different damage scenarios with the undamaged conditions results in damage-sensitive features.

Finally, modal domain methods extract modal parameters, such as natural frequencies, mode shapes, and damping ratios, by means of classical techniques. The modal parameters can be extracted by input-output or only-output measurements [133]. The main idea behind modal-based damage detection is that damage causes changes in the stiffness of a structure, which in turn affects its dynamic response. When damage occurs, a frequency shift appears, therefore damage can be detected by relative changes in the natural frequencies [156–162].

3.3.2 Video motion magnification for structural health monitoring

SHM systems are very powerful tools to ensure the safety and reliability of infrastructure. These systems often employ contact sensors installed on the target structures to detect changes in the material and geometric properties, providing key information about structural performance.

However, despite their benefits, most buildings and civil infrastructure are not instrumented and continuously monitored.

The adoption of SHM systems is hindered by several drawbacks. The first barrier to adoption is complexity, SHM systems are complex and multidisciplinary systems that require expertise to install, maintain, and operate. On top of that, a lack of standardisation for these systems, further complicates the task to select the suitable equipment for a particular deployment, in part because of the different monitoring needs a target structure might present depending on their design specifications. Moreover, the lack of awareness of the existence of SHM systems and their benefits also limits the adoption of these systems in certain sectors.

However, one of the most recognized drawbacks is cost. Not only the initial deployment of a full-scope and long-term SHM system incurs in significant costs, but also its operational and maintenance costs play an important role in the decision to adopt the use of these systems. It is generally the case that a correlation exists between the cost of an SHM system and its monitoring scope, the larger the structure, the costlier the system. Moreover, the costs are not limited to the sensors and their installation, peripheral systems must also be considered, such as the hardware and software to process and analyse measured data, or the storage infrastructure to safely archive data long-term for future reference and inspections. In this sense, the cost of instrumentation is only a fraction of the overall costs, as the installation and long-term maintenance of a remote monitoring system can represent upwards of six times the cost of instrumentation [163].

Currently, there is a consensus that the benefits of SHM exceed its costs [163], but in many cases these paybacks are difficult to estimate, further disincentivising their use by infrastructure operators.

Non-contact methods of measurement could help to avoid some of these drawbacks and are being researched for the purposes of modal analysis. Vision-based methods are one of such methods of interest, offering the possibility of using video cameras to monitor structures at a distance, with the possibility to cover large structures with much fewer sensors than using contact-based techniques. Using video cameras also means that a wide range of options regarding specifications, optics, and prices are available to satisfy different applications.

Motion can be quantified using a number of different image processing techniques such as edge detection or feature detection. However, these methods are most often not accurate enough for VSHM. In this sense, motion magnification offers an interesting research path towards the usage of vision-based techniques for vibration-based monitoring [91, 93, 164–191]. In 2014 Chen *et al.* [181], introduced such idea using phase-based video motion magnification on a laboratory experiment where a cantilever beam was instrumented with accelerometers and laser vibrometers and compared against the results of the motion magnification based system. A very similar follow-up study [93] also described the process of visualising the operating

deflection shapes of the test structure, demonstrating the capabilities of their proposed system. Other authors also experimented with the capacity of motion magnification techniques for obtaining vibration modes on more complex structures such as the well-known three-story building benchmark [116, 186], or with structures outside of the laboratory, using motion magnification techniques to measure full-scale buildings and bridges [94, 95, 166, 192].

3.4 Literature gaps

This thesis examines the intersection of deep learning, motion magnification, and VSHM to add value to the existing work in the field. While some authors have used motion magnification to diagnose structural issues, none attempted [193] to use the latest state-of-the-art learning-based method introduced by Oh *et al.* [112]. Accordingly, the aim of this thesis is to develop and validate an algorithm built on learning-based video motion magnification techniques for use in structural condition assessment and damage detection, harnessing the superior image quality of state-of-the-art motion magnification.

Since the introduction of learning-based video motion magnification in 2018 the computer vision field has made major strides in several image-related tasks such as generation [194], super-resolution [195, 196], denoising [197], or inpainting [198]. These improvements have made use of the latest architectural stategies such as the Vision Transformer (ViT) [27] or the Swin Transformer [28]. The application of new state-of-the-art ideas to learning-based video motion magnification has the potential to yield superior results applicable to VSHM in the form of new deep-learning models with superior magnification quality.

Moreover, the vision-based VSHM literature lacks publicly available tooling for ensuring replicability and reproducibility of the existing studies. Based on the objectives of this thesis and the mentioned literature gaps, there is a clear need for the development of open-source tools that leverage advanced learning-based techniques.

Finally, the leveraging of these gaps in the literature should create the need for the validation and verification of the developed algorithms and tools for their intended purpose of obtaining the dynamic response, thus aligning with the stated objectives of this thesis.

4

Shifted window transformer based video motion magnification

This chapter presents the technical advances in the field of video motion magnification methods developed in the context of this thesis and compares the results within the current video motion magnification landscape.

As explained before, Video Motion Magnification (VMM) is a computer vision technique that magnifies small motions in a video to reveal previously invisible or unseen movement. These techniques can be divided into two general approaches regarding movement tracking, Lagrangian methods and Eulerian methods. The Lagrangian approaches rely on motion tracking or optical flow to isolate motion prior to magnification, while Eulerian methods observe changes in fixed regions of pixels instead of tracking features in time and space. It is generally understood that Eulerian methods require less computing power and produce fewer visual errors, especially on regions with occlusion boundaries or when magnifying complex motions [103].

VMM can be used in structural health monitoring to detect small movements or deformations in structures that may not be visible to the naked eye. By magnifying these small motions, potential issues can be identified before they become major problems. For example, VMM could be used to detect cracks in a bridge or monitor the movement of a building to ensure its structural integrity.

More advanced techniques, such as the ones discussed in the following chapter, use VMM to extract the dynamic response of structures. Those dynamic responses can then be analysed using advanced signal processing techniques, such as damage indicators [136], to diagnose structural problems in areas where no cracks or defects might be visible with or without magnification. In those applications, a sensible choice is to use Eulerian-based methods because they produce fewer artefacts than a Lagrangian approach. When using VMM for SHM the goal is always to yield the best quality signal possible for the subsequent analysis.

This chapter presents the technical advances developed in this thesis in deep-learning-based video motion magnification with respect to prior art, therefore it is important to establish the context of these techniques before building on those concepts. As previously mentioned, structural health monitoring techniques require good quality and reliable signals to produce accurate damage detection and diagnosis. Consequently, the choice of a method is conditioned to a technique capable of yielding the best image quality possible when dealing with complex movements.

Learning-based video motion magnification (LB-VMM) is an Eulerian magnification technique that abandoned the previous approach of manually designing filters for spacial decomposition and motion isolation in favour of learning those filters using CNNs. This technique was the subject of study in the early stages of this thesis to verify its reliability for SHM applications [193], as it was the top-performing video motion magnification technique in terms of output image quality [112]. Nevertheless, its reliance on additional temporal filtering to improve image quality occasionally produced errors in magnification. While it is possible to obtain acceptable results without temporal filtering, its use improves the output frame's clarity as it removes some noise and motion prior to learning-based magnification.

The performance degradation is especially pronounced when using temporal filters over small motions, especially when using large magnification factors. In some cases, the method appears to be blind to small motions resulting in patchy magnification, creating an effect where some patches get occasionally magnified as soon as their motions get large enough to cross the threshold for the model to detect them. This issue limits the usefulness of learning-based magnification methods employing temporal filtering, and therefore should not be used when attempting to obtain reliable dynamic response signals.

Advances in deep learning and computer vision have enabled the improvement of the learning-based technique to obtain better quality results without the need to use temporal filtering. This technique was developed as part of this thesis and has been named Shifted Window Transformer Based Video Motion Magnification (STB-VMM). This technique relies on exploiting the latest advances in vision-based transformers to replace the fully convolutional structure of its predecessor.

4.1 Related work

CNNs have been a mainstay of the Computer Vision (CV) field for the last decade, with many of the top-performing models having made extensive use of them [5, 20, 199]. This period roughly started after Krizhevsky et al. [5] won the ImageNet Large Scale Visual Recognition Challenge [45, 53] (ILSVRC) on September 30^{th} 2012, which led to a surge in publications that employed CNNs and GPUs to accelerate deep learning. CNNs use filters to generate feature maps that summarise the most relevant parts of an image. These filters capture important local information through the convolution operation and, when combined with multi-scale architectures [63, 200], produce rich feature maps that can effectively represent an image's content, both in a local and global context.

However, the recent irruption of the transformer architecture, which has demonstrated its massive potential in tasks such as natural language processing, with applications like ChatGPT [6, 7, 59, 60], has changed the computer vision landscape as well. The Vision Transformer (ViT) [27], which, employs the attention mechanism has demonstrated state-of-the-art performance in many CV tasks.

The attention mechanism can be described as mapping from a query and a set of key-value pairs into an output. The output, represented in vector format, is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function taking into account the query and the corresponding key [23]. The transformer was the first model to rely exclusively on self-attention to compute representations of its input and output without using sequence-aligned recursive neural networks or convolution operations. Unlike CNNs, transformers do not present translation invariance or a locally-restricted receptive field, instead

they offer permutation invariance. Said feature enabled NLP models to infer relations between words and ideas much further into a text than previous recurrent models could. However, CV applications require the processing of grid-structured data, which can not trivially be processed by a transformer. The ViT [27] overcame the challenge of processing grid-structured data by converting it into sequential data by dividing the image into patches. These patches are then flattened into vectors and embedded into a lower dimension before being summed with positional embeddings and fed as a sequence to a standard transformer encoder. Image patches essentially become sequence tokens just like words are when working in NLP, in fact, ViT uses the exact same encoder described in [23].

Microsoft researchers later improved upon the ViT with the SWIN transformer [28], a hierarchical vision transformer that uses shifted windows. This work further refined the solution to adapt the original transformer from language to vision. The SWIN transformer addressed issues caused by large discrepancies in the scale of visual entities while limiting self-attention computation to non-overlapping local windows while still allowing for cross-window interaction. This limitation on the scope of self-attention significantly reduced computational complexity, which scales quadratically with respect to image size, allowing for processing higher-resolution images that were previously unmanageable. Further developments in the CV field have implemented the SWIN transformer for various tasks achieving state-of-the-art performance [65, 201, 202].

Inspired by the recent prominence of the transformer and its success in many CV problems such as image classification [27, 28, 203–207], object detection [208–210], segmentation [204, 211, 212], crowd counting [213, 214] and image restoration [215–217], Liang et al. [65] proposed a new state-of-the-art image restoration model based on the Swin transformer [28]. The SwinIR model consists of three modules: a shallow feature extractor, a transformer-based deep feature extractor, and a high-quality image reconstruction module. This structure provides excellent performance in various image restoration tasks such as image super-resolution, JPEG compression artifact reduction, and image denoising. These applications are of particular interest when working with VMM because current state-of-the-art methods can be negatively affected by noisy input images, resulting in noisier and blurrier results at large magnification rates. This phenomenon occurs due to noise not being properly filtered beforehand, causing it to be magnified along with the motion in the magnification stage.

STB-VMM borrows its architecture from the blueprint established by previous Eulerian techniques (see Table 4.1), which are divided into three subsystems in charge of spatial decomposition, motion isolation and manipulation, and representation denoising. However, it also introduces new ideas that make use of the Residual Swin Transformer Block (RSTB) [65] to improve the filtering on those subsystems. The RSTB (see Figure 4.1) is one of the main building blocks of the feature extractor and image reconstructor subsystems of the model that combines multiple Swin Transformer Layers (STL) [28] with convolutional layers

to compound the benefits of the spatially invariant filters of the convolutional layers with the residual connections that allow for multilevel feature processing.

Table 4.1: Motion magnification techniques summary table. Adapted from [66, 112].

Method	Liu et al. [101] (2005)	Wu et al. [103] (2012)	Wadhwa et al. [111] (2013)	Wadhwa et al. [109] (2014)	Zhang et al. [110] (2017)	Oh et al. [112] (2018)	Lado et al. [66] (2023)
Spatial decomposition	Tracking, optical flow	Laplacian pyramid	Steerable filters	Riesz pyramid	Steerable filters	Deep convolution layers	Swin Transformer
Motion isolation	-	Temporal bandpass filter	Temporal bandpass filter	Temporal bandpass	Temporal bandpass filter (2_{nd} order derivative)	Subtraction or bandpass filter	Subtraction
Representatic denoising	Expectation-Maximisation	-	Amplitude-weighted Gaussian filtering	Amplitude-weighted Gaussian filtering	Amplitude-weighted Gaussian filtering	Trainable convolution	Swin Transformer

The Swin transformer layer partitions an image of dimensions $H \times W \times C$ into non-overlapping $\frac{HW}{M^2}$ local windows using an MxM sliding window that then computes its local attention, effectively reshaping the input image into $\frac{HW}{M^2} \times M^2 \times C$. The main difference with respect to the original transformer layer [23] lies in the local attention and the shifted window mechanism. For a local window feature $F \in \mathbb{R}^{M^2 \times C}$, the query, key, and value matrices Q, K, and $V \in \mathbb{R}^{M^2 \times d}$ are computed as

$$Q = FW_Q; \quad K = FW_k; \quad V = FW_V \tag{4.1}$$

where W_Q, W_K, and W_V are the learnable parameters shared across different windows, and d is the dimension of Q, K, and V. Therefore, the attention matrix is computed for each window as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}} + P)V \tag{4.2}$$

where P is the learnable relative positional encoding. Computing the attention mechanism multiple times yields the results of the Multi-head Self Attention (MSA), which are then passed on to a Multi-Layer Perceptron (MLP). Therefore, the whole STL process can be summed up like so

$$F = MSA(LayerNorm(F)) + F \tag{4.3}$$

then

$$F = MLP(LayerNorm(F)) + F \tag{4.4}$$

where the MLP is formed by two fully connected layers with a GELU activation layer in between.

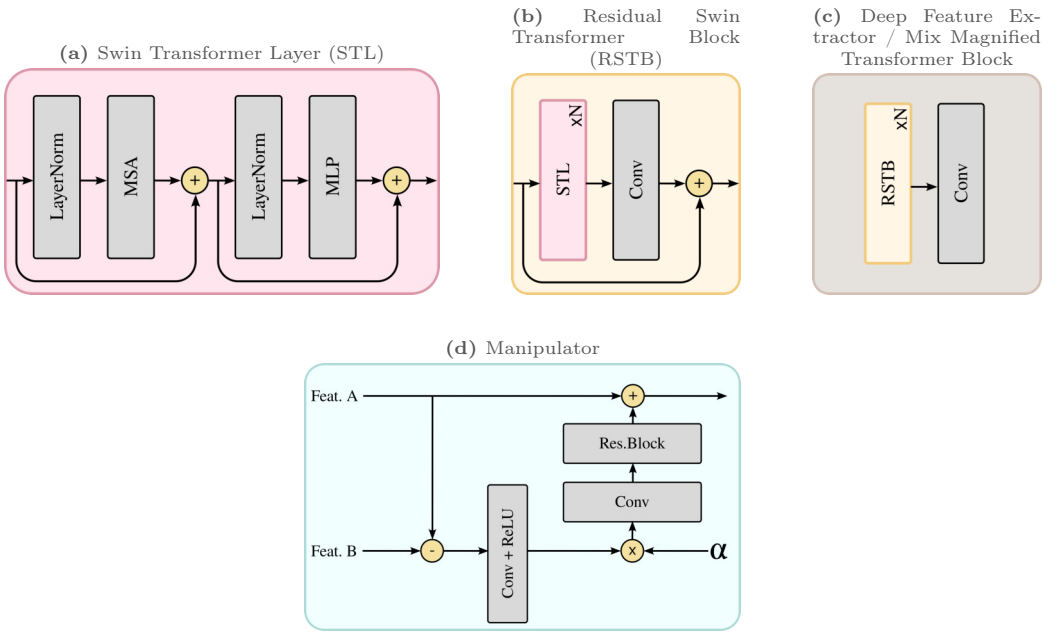


Figure 4.1: STB-VMM's fundamental architectural building blocks [66].

4.2 Architecture overview

STB-VMM's architecture consists of three big functional blocks named: feature extractor, manipulator, and reconstructor. A general overview of the model is shown in Figure 4.2, where the flow of data through the model can be observed. Like Oh et al.'s learning-based method [112], STB-VMM takes in two frames and returns a single magnified frame.

At a high level, the architecture works by sequentially applying each of the three aforementioned functional blocks. The feature extractor takes in two frames (I_A, I_B), and is tasked with decomposing the input images into high-quality abstract representations of those frames. These abstract representations include intrinsic properties of the images such as shape, texture, lighting, edges, etc. The extraction is carried out in two stages, first with a shallow feature extractor and then a deep feature extractor that refines the results. In between both stages, a residual connection is placed so that the shallow feature extractor can contribute with less processed data further down the network, avoiding information loss. Next, the manipulator receives the abstract representations of the input frames and magnifies the motion by multiplying the difference between the two feature spaces by the user-selectable magnification factor α. Finally, the reconstructor takes the conjoined manipulated feature space that results from the manipulator and recomposes it back into a frame, effectively undoing the encoding produced by the feature extractor and obtaining a magnified 3 channel image.

Taking a closer look at the model, it can further be described by following the sequential

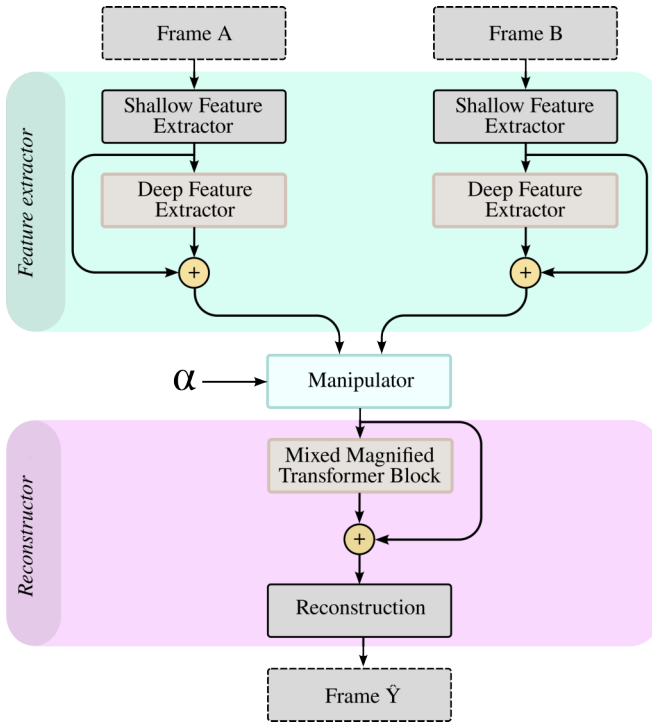


Figure 4.2: Architecture overview of the proposed model [66].

transformations of two sample frames of a target sequence $[I_A, I_B] \in \mathbb{R}^{H \times W \times C_{in}}$, where H is the height of the image, W is the width of the image and C_{in} represents the number of input channels. The convolutional shallow feature extractor (G_{SF}) maps high-level features into a higher dimensional feature space, thus providing early local feature extraction (F_{AS}, F_{BS}) and leading to a more stable optimisation and better results [218].

$$[F_{AS}, F_{BS}] = G_{SF}([I_A, I_B]) \tag{4.5}$$

Next, the deep feature extraction module (G_{DF}) takes the features from the previous step and processes them further to obtain the final motion representations. This module has N Residual Swin Transformer Blocks (RSTB) connected in series.

$$[F_{AD}, F_{BD}] = G_{DF}([F_{AS}, F_{BS}]) \tag{4.6}$$

The manipulator (G_M) receives the resulting motion representations of both frames and calculates the difference between the feature spaces, magnifying them by multiplying by the user-defined magnification factor α.

$$G_M(F_{AS}+F_{AD},F_{BS}+F_{BD}) = (F_{AS}+F_{AD}) + h(\alpha \cdot t(((F_{BS}+F_{BD})-(F_{AS}+F_{AD})))) \quad (4.7)$$

Note that $t(\cdot)$ is a 3×3 convolution followed by a ReLU activation function, and $h(\cdot)$ is a 3×3 convolution followed by a 3×3 residual block.

$$F_M = G_M(F_{AS}+F_{AD},F_{BS}+F_{BD}) \quad (4.8)$$

The conjoined manipulated feature space of both frames is then moved to the Mixed Magnified Transformer Block (MMTB) (G_{MMTB}) formed by N RSTB blocks. This stage enables the attention mechanism to affect the combined magnified features of both frames, resulting in a more coherent result after reconstruction.

$$F_{MMTB} = G_{MMTB}(F_M) \quad (4.9)$$

Finally, a convolutional block (G_R) performs the inverse operation of the initial feature mapping, conducted by the shallow feature extractor (G_{SF}), and restores the image representation to a magnified frame ($I_{\hat{Y}}$).

$$I_{\hat{Y}} = G_R(F_M + F_{MMTB}) \quad (4.10)$$

4.3 Modes of operation

The STB-VMM model was designed to perform inference on a full video sequence, with the modes of operation describing the order in which the input frames I_A and I_B are fed into the model to obtain the desired results. The approach can be applied to any video sequence with two or more frames, regardless of the time scale between them. Sequences can be processed in either static or dynamic mode.

The static magnification mode uses a reference frame against which all subsequent frames in the sequence are magnified. This mode adheres more closely to the traditional definition of motion magnification and uses the first frame of the sequence as a reference for all subsequent movements. It can be mathematically expressed as $model(I_0, I_t)$, where t is the frame number that increases sequentially with time. In contrast, the dynamic mode magnifies the difference between two consecutive frames $[model(I_t, I_{t+1})]$, thus magnifying the velocity between each frame.

These modes do not require any changes to the network; instead, they change the order in which the input frames are fed to the model. It is important to note that the meaning of the magnification factor α differs between the two modes.

Additionally, Oh et al. [112] in its learning-based video motion magnification model, proposed one more operation mode that used temporal filtering to mitigate the effects of unwanted motion and noise. The filtering was applied in the manipulator to produce temporally-filtered motion-magnified frames similar to those produced by classical techniques. However, this temporal mode can cause blindness to small motions, resulting in patchy and inaccurate magnification. While it is theoretically possible to incorporate a temporal mode into the STB-VMM model, its magnification results do not suffer from excessive noise or blurring, rendering temporal filtering unnecessary.

4.4 Training

The model is trained end-to-end using an L1 loss function between the network output $\hat{I}$ and the ground-truth magnified frame I. An additional L1 loss function is also used as an intermediate regularisation loss to improve the robustness of the feature extraction process. This loss is calculated by comparing the altered c frames from the chosen dataset against the unperturbed frames once both have undergone feature extraction. The resulting regularisation loss is then added to the overall end-to-end loss of the entire network with a weight coefficient λ set to 0.1.

However, obtaining magnified image pairs in the real world is a very difficult task, if not impossible. Fortunately, this problem was already solved by Oh et al. in their 2018 work [112] in which they proposed a procedurally generated dataset.

To obtain a suitable method for generating training data for motion magnification applications, Oh et al. [112] procedurally generated 200,000 images that were carefully constructed using images from the MS COCO dataset [219] over which 7,000 segmented objects from the PASCAL VOC dataset [220] were overlaid. The segmented objects were scaled, rotated, moved and occluded with careful consideration for consistency between magnified and unmagnified pairs.

Each training sample of the resulting dataset contains between 7 and 15 foreground objects randomly distributed and scaled from their original size. Additionally, variants with altered contrast, blurriness, and complex motion are also present to improve model generalisation on difficult magnification scenarios.

This dataset has been used to train and benchmark STB-VMM. The results in the following section show the differences in image quality between STB-VMM and LB-VMM [112]. Using

Figure 4.3: Sample frames of the training dataset [112]

the same training dataset, this allows for a better comparison and justifies the use of pre-trained weights made available by Oh *et al.*

4.5 Results

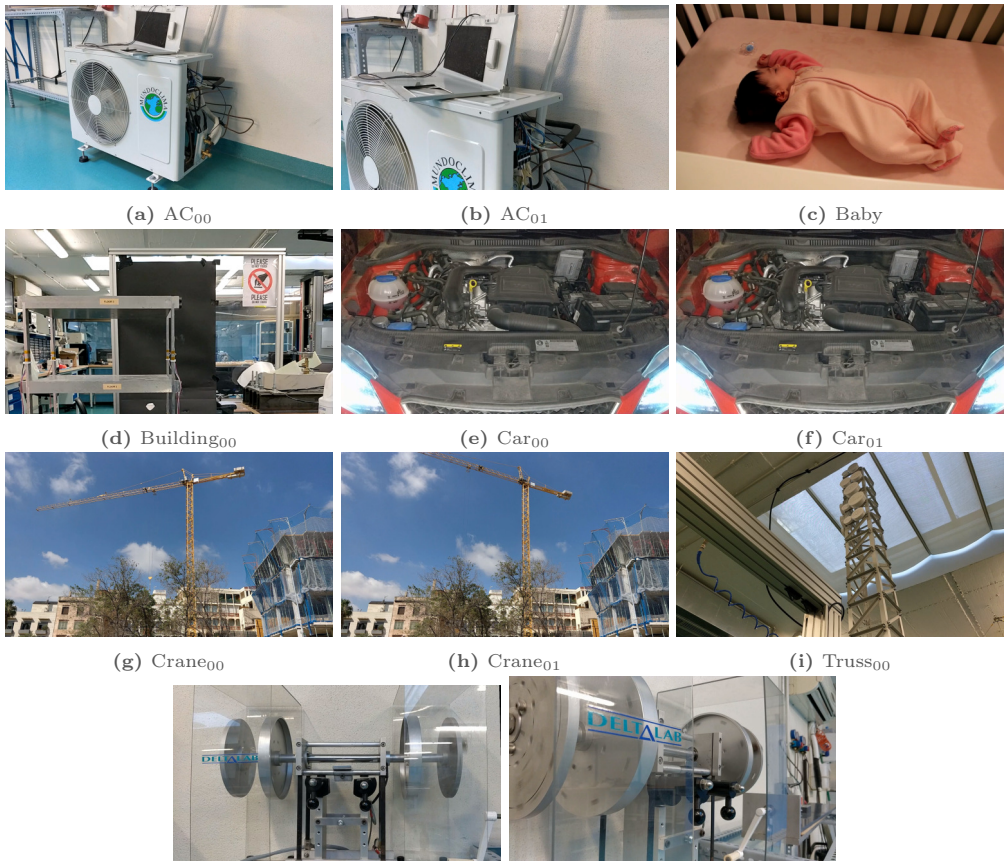
STB-VMM was designed to improve upon the image quality of previous techniques by providing better noise filtering, resulting in higher-quality magnified images with fewer artefacts and blurring while maintaining more stable edges. These features are critical for reliable tracking in SHM applications.

This section presents a comparison between STB-VMM and LB-VMM in terms of image quality. The comparison is limited to these two methods for two reasons. First, only LB-VMM produces results comparable to STB-VMM because both methods are capable of full-band motion magnification, meaning they can magnify the entire motion spectrum in a single inference pass. In contrast, traditional methods require limiting the motion bandwidth to filter noise prior to magnification. Therefore, those methods can not produce comparable results to the learning-based methods. Since no other methods besides the ones based on deep learning are capable of full-band magnification, comparisons with other methods would not accurately reflect actual performance. Second, since LB-VMM has already been proven to be the state-of-the-art model in video motion magnified image quality, there is no need to test against other methods.

The comparison is divided into two parts: a qualitative comparison and a quantitative comparison. The qualitative comparison involves visually comparing the results of the two methods to assess their performance in terms of image quality, noise filtering, and preservation of edges. The quantitative comparison, on the other hand, uses objective metrics to evaluate the aesthetic and technical qualities of the magnified images.

The quantitative comparison of image quality, or Image Quality Assessment (IQA), is a complex topic that involves many variables and methodologies. IQA can be divided into three main categories based on whether or not they require a reference image: full-reference, reduced-reference, and no-reference. Referenced algorithms [221–223] require a pristine sample to assess the quality of a degraded image, while no-reference algorithms [224–227] produce an image score without any reference. Since we cannot generate an ideally magnified frame, our comparison method must be based on a no-reference assessment method.

The Multi-scale Image Quality Transformer (MUSIQ) [227, 228] is the chosen method for the quantitative comparison, allowing us to evaluate the aesthetic and technical qualities of images. MUSIQ was developed by researchers at Google Research to address some of the limitations of existing IQA methods. It can process native-resolution images with varying sizes and aspect ratios and uses a multi-scale image representation to capture image quality at different levels of context, size, and scale. This makes MUSIQ a powerful tool for assessing the quality of videos that have undergone motion magnification.



(a) AC_{00} **(b)** AC_{01} **(c)** Baby

(d) $Building_{00}$ **(e)** Car_{00} **(f)** Car_{01}

(g) $Crane_{00}$ **(h)** $Crane_{01}$ **(i)** $Truss_{00}$

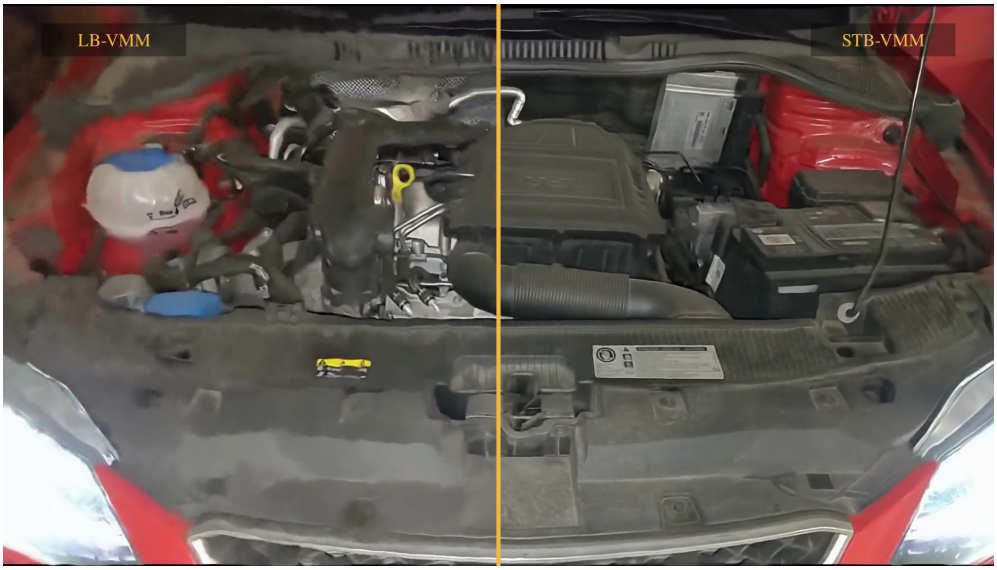
(j) $Wheel_{00}$ **(k)** $Wheel_{01}$

Figure 4.4: Sample frames of the eleven video sequences used as a benchmark

The subject videos for this evaluation are eleven different benchmark sequences that showcase interesting motion patterns and examples. These videos include a sequence from the test set provided by [103] to ground the comparison against other publications. The sequences, of which a representative frame is shown in Figure 4.4, were captured at 1080p 60fps using a mid-range smartphone to demonstrate the potential of STB-VMM even when using accessible video equipment (except in the case of the *Baby* sequence).

4.5.1 Qualitative comparison

This section compares the visual quality differences that can be observed in the benchmark sequences presented in Figure 4.4. The comparisons focus on the image quality, with special attention to edge stability and blurriness, which are the most common downsides introduced by motion magnification techniques.



(a) Split frame showing LB-VMM on the left and STB-VMM on the right

(b) LB-VMM **(c)** STB-VMM **(d)** LB-VMM **(e)** STB-VMM

Figure 4.5: Qualitative comparison of the car sequence. Highlighted in the bottom row of the figure, the car's coolant reservoir, engine cover, and ventilation slits demonstrate that STB-VMM results are noticeably sharper and less distorted.

Starting with the Car_{00} sequence, Figure 4.5 shows a split frame with LB-VMM on the left and STB-VMM on the right. It can be observed that STB-VMM resolves more detail in the image while providing the same level of magnification. This is especially apparent in the vertical grill below the windshield and in the highlighted details below the main image. In the STB-VMM magnified frame, letters on the decorative engine cover are visible, while in the LB-VMM version of the frame, those same letters are obscured by intense blurring. This example presents a particularly challenging situation due to sub-optimal lighting conditions, which can cause digital images to appear grainier or noisier due to limitations in the camera's dynamic range. STB-VMM demonstrates exceptional performance in filtering out noise to produce a noticeably higher-quality magnified image.

Figure 4.6 on the other hand, seems to be recorded in a low-light environment but does not seem to show the same blurring issues. Indeed, low light performance depends on multiple factors such as the camera sensor, the optics, exposure time or digital post-processing. In

Figure 4.6: Qualitative comparison of the Baby sequence [103]

this case, the lower resolution of the source video helps average out the noise, thus trading resolution for a clearer image. Nevertheless, the performance of STB-VMM is still visibly superior in areas like the feet, which appear better defined, or in the less wobbly edges of the crib when the video is in motion. This last point can be observed statically in the edge of the crib's wall closest to the bottom side of the image, where STB-VMM yields a straight line, while LB-VMM creates small artefacting around it. Additionally, the difference in sharpness between the two methods is evident in the pacifier, which appears clearer and more well-defined in the STB-VMM magnified frame.

If we now take a look at a benchmark recorded in much more favourable lighting conditions, such as the Truss_{00} sequence, the trend observed so far continues, albeit less pronounced. Nevertheless, STB-VMM still results sharper overall, resolving certain areas more convincingly than its learning-based counterpart. The tubing and wiring on the ceiling are one such example,

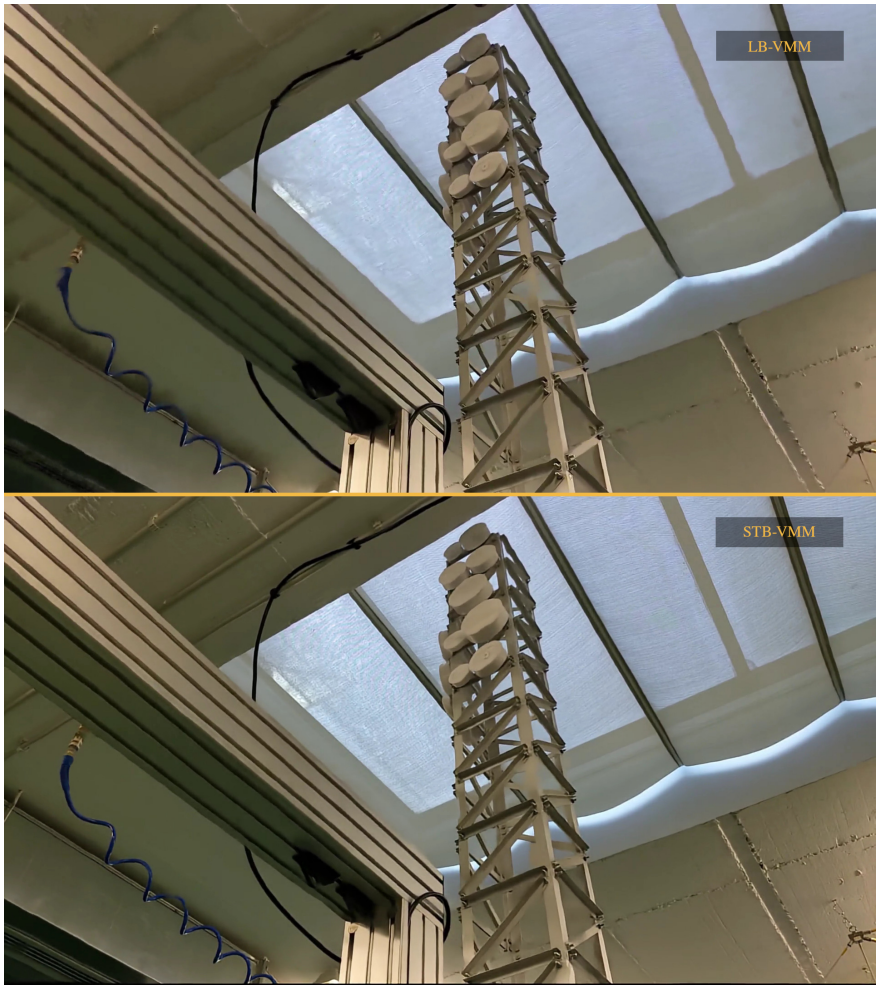


Figure 4.7: Qualitative comparison of the Truss_{00} sequence

furthermore the differences can also be seen in more subtle details such as in the carabiner that holds a laboratory test structure to the ceiling on the lower right side of the image, or in the edge of the curtains placed under the skylight.

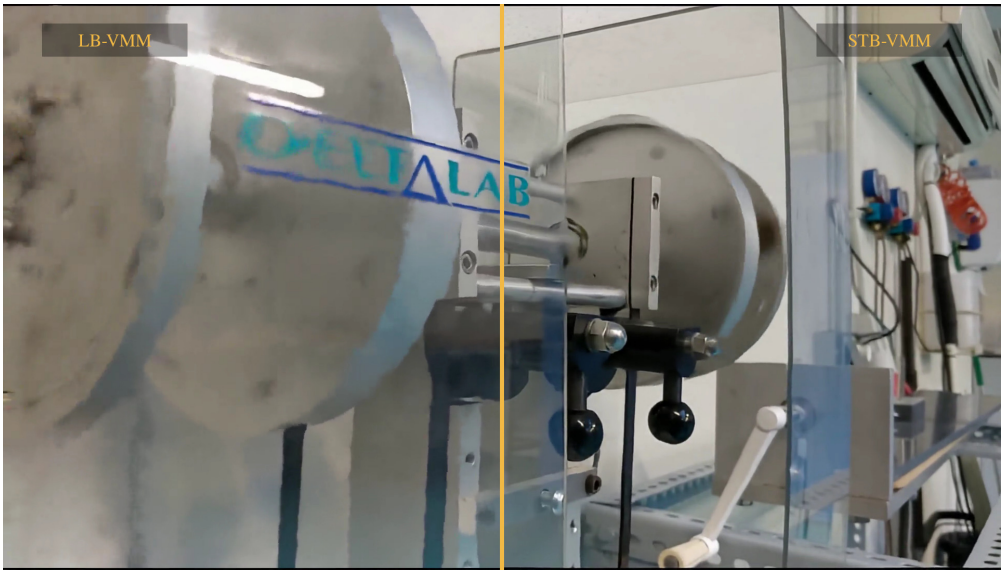
Motion magnification also has interesting applications for condition monitoring [229, 230], in Figure 4.8 the exterior unit of an air conditioning system is shown without its side cover which is placed on top of the unit. The image comparison shows substantial differences between LB-VMM and STB-VMM. Shadows on STB-VMM appear much sharper and more well-defined. Also, thin elements such as tubes and wires are displayed in better definition, this might be especially relevant for those applications that rely on monitoring thin elements of a structure that could easily be confused with a blurry background. If we take an overview of the two images it is quite apparent that the texture of the wall, the insulating black foam, or the background metallic structure look noticeably sharper when using STB-VMM.

Figure 4.8: Qualitative comparison of the AC_{01} sequence

Similar observations can be made about the *wheel* test structure shown in Figure 4.9. Letters on the transparent plastic shielding are much sharper using the method designed in this thesis, as well as the background elements that appear clearer and better defined.

From this qualitative comparison, it can be concluded that the higher the noise in the source image, the greater the quality differential will be on magnification. That is not to say that STB-VMM is not useful in low-noise video sequences, as it has demonstrated better edge stability in motion and less blurriness and artefacting.

All digital image sensors introduce small amounts of random noise into images. While this noise may be imperceptible to the naked eye, motion magnification techniques must be able to filter out frame defects to prevent magnified sequences from becoming unusable due to magnified noise. Using a higher-quality magnification method is particularly important when relying on magnification for measurement purposes, such as in Structural Health Monitoring



(a) Split frame showing LB-VMM on the left and STB-VMM on the right

(b) LB-VMM **(c)** STB-VMM **(d)** LB-VMM **(e)** STB-VMM

Figure 4.9: Qualitative comparison of the Wheel_{01} sequence. STB-VMM displays sharper letters and a better-defined background with respect to LB-VMM.

(SHM), because it is often not possible to control or alter ambient conditions in the field. A more robust method ensures that the results will be usable regardless of the conditions under which they are recorded.

4.5.2 Quantitative comparison

To reinforce the previous observations with tangible and reproducible values, each of the eleven video sequences shown in Figure 4.4 have been magnified, split into individual frames, and evaluated using the MUSIQ indicator. The following results display the average of the image quality scores for each of the benchmarks.

Table 4.2 displays the average, 1st, and 99th percentile MUSIQ scores for the benchmark sequences tested on both the LB-VMM and STB-VMM models. For reference, the score of the original unmagnified source sequences has also been included. It is important to note that the

original sequence has not been magnified, so the comparison is not valid. However, it provides a useful frame of reference for understanding the MUSIQ scores.

Table 4.2: Comparative MUSIQ scores of the original sequence, the sequence magnified using Learning-Based Video Motion Magnification (using the *o3f_hmhm2_bg_qnoise_mix4_nl_n_t_ds3* checkpoint), and the proposed method. (×20)

	Original			**LB-VMM**			**STB-VMM**		
	Avg.	η_1	η_{99}	Avg.	η_1	η_{99}	Avg.	η_1	η_{99}
AC_{00}	72.11	69.65	72.75	55.73	49.61	58.69	62.45	61.05	63.29
AC_{01}	69.15	68.30	70.05	48.35	34.07	51.22	59.27	57.72	60.96
Baby	74.39	69.71	74.87	55.51	53.26	59.95	57.12	54.41	62.90
$Building_{00}$	66.84	66.01	75.45	52.46	49.51	62.75	52.30	50.07	56.43
Car_{00}	52.55	50.65	54.41	31.40	18.27	35.50	43.37	23.28	48.06
Car_{01}	55.81	54.77	57.01	33.51	30.67	64.99	50.28	48.08	52.07
$Crane_{00}$	75.26	74.86	75.57	56.92	52.70	65.02	59.13	56.19	62.89
$Crane_{01}$	75.09	74.63	75.44	51.05	45.25	57.37	54.93	51.11	64.70
$Truss_{00}$	66.94	65.92	67.49	55.90	52.65	57.98	56.27	54.93	57.61
$Wheel_{00}$	72.84	71.87	73.38	51.04	28.82	54.40	57.04	36.41	61.19
$Wheel_{01}$	52.15	50.23	53.55	34.84	31.12	59.03	46.21	43.68	48.48
Total avg.	**66.13**	**51.25**	**75.45**	**48.05**	**32.32**	**60.09**	**54.42**	**45.68**	**63.29**
% dev. to avg.	*13.58%*	*22.50%*	*14.09%*	*20.34%*	*32.75%*	*25.04%*	*10.70%*	*16.07%*	*16.28%*

The results in Table 4.2 demonstrate that STB-VMM produces better quality results than LB-VMM . On average, STB-VMM scores are 9.63% higher while also showing much higher 1% lows, implying that the quality of the magnification is substantially more consistent during the entire length of the sequence.

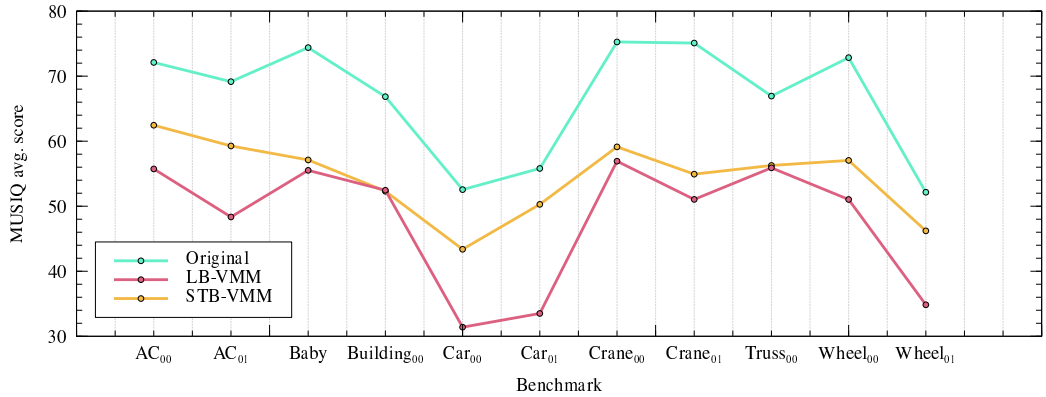


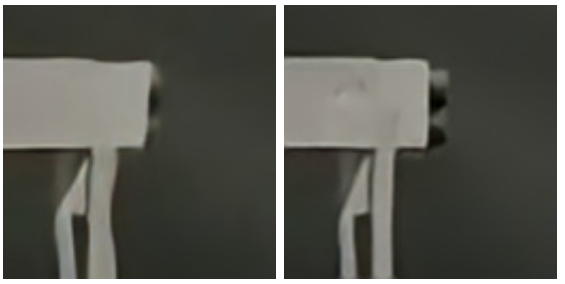
Figure 4.10: Graphic representation of the average MUSIQ scores per test sequence magnified ×20.

To aid in comparing these values and finding the underlying trends Table 4.3 and Figure 4.10 summarise the results in a more visual manner. It can be observed how STB-VMM is remarkably stable in its output quality, maintaining the scores firmly above LB-VMM. Only one standout case positions LB-VMM slightly over STB-VMM on the $building_{00}$ benchmark, where LB-VMM scores higher by a margin of a 0.23% on average. Nevertheless, when we qualitatively inspect the comparison between the two methods, STB-VMM still exhibits its

Table 4.3: MUSIQ score difference between STB-VMM and LB-VMM.

	Avg. (%)	η_1 (%)	η_{99} (%)
AC_{00}	9.32	16.42	6.32
AC_{01}	15.79	34.64	13.91
Baby	2.15	1.65	3.95
$Building_{00}$	-0.23	0.86	-8.38
Car_{00}	22.79	9.89	23.08
Car_{01}	30.06	31.78	-22.65
$Crane_{00}$	2.93	4.67	-2.81
$Crane_{01}$	5.17	7.85	9.71
$Truss_{00}$	0.55	3.46	-0.55
$Wheel_{00}$	8.24	10.56	9.26
$Wheel_{01}$	21.81	25.01	-19.72
Total	**9.63**	**26.07**	**4.24**

characteristic sharper look and stable magnification. Figure 4.11 shows the difference in the top floor bolted joint detail of the sequence.



(a) LB-VMM **(b)** STB-VMM

Figure 4.11: Qualitative comparison of the top floor joint in the $Building_{00}$ sequence

Besides this abnormal datum, none of the magnified results fall above the original's score as expected. Nevertheless, original and magnified scores follow very similar trends, implying that low-quality inputs will result in lower-quality outputs. Despite this observation, STB-VMM maintains a more stable trend than LB-VMM, not declining as sharply when input quality is reduced. This phenomenon occurs thanks to the superior noise-filtering capabilities of the transformer-based architecture.

4.6 Limitations

Overall, the results exhibited in this comparison are very satisfactory and achieve the design goals set for STB-VMM. The new technique is a solid alternative that achieves better image quality than prior motion magnification methods while still using the same training data. However, despite the favourable comparisons, LB-VMM still has a significant advantage over STB-VMM in terms of computing time. Using our hardware setup[1], LB-VMM can magnify

[1]AMD Ryzen 9 5950X; Nvidia RTX 3090

the baby sequence [103], consisting of 300 960x576 frames, in approximately 76 seconds. In contrast, STB-VMM takes almost twice as long, clocking in at 130 seconds for the same sequence. Nevertheless, future software optimisations and improvements in hardware may help to reduce STB-VMM's computing time. It is also important to note that one of the design goals for STB-VMM was to provide a more reliable magnification technique for scientific measurement, where a performance penalty is acceptable within practical limits.

5

Video motion magnification for vibration-based damage detection

This chapter presents and validates the deep learning video motion magnification based technique developed to assess a structure's condition based on their dynamic response

Regular structural maintenance is crucial for ensuring the safety and reliability of civil structures. Suboptimal maintenance can lead to structural defects and other issues that compromise the structural integrity, increasing the risk of a catastrophic failure. The collapse of an infrastructure can result in injury, loss of life, and economic damages. For these reasons, periodic inspection and monitoring are necessary to address structural problems before they become critical.

Structural monitoring provides several benefits that can help ensure the safety and longevity of structures. These systems can warn of early signs of structural degradation, allowing engineers to take corrective action before the problem worsens. In cases where immediate action is not practical, early warning systems also enable better maintenance planning, reducing emergency maintenance actions that disrupt normal operations and create additional costs.

However, the installation of instrumentation on structures is generally a costly process that must be assumed upfront while generally providing no immediate benefits. The instrumentation process involves deciding the placement and principle of operation of sensors, the deployment of communications infrastructure to receive and centralise data collection, and the provisioning of computing infrastructure to process captured data and alert of anomalous behaviour. It is often the case, that the sensors used, such as accelerometers or strain gauges, require to be in contact with the structure, which creates the necessity for a more complex communications infrastructure with data lines, repeaters, or independent power sources for wireless sensors.

Vision-based monitoring methods attempt to mitigate these issues by providing contactless measurements, inferring the displacement of structural elements by tracking pixels or features over time. Vision-based techniques provide full field measurements which reduce the number of sensors required for the same amount of information. This makes vision-based techniques very attractive as a lower-cost alternative to contact-based monitoring systems. This thesis explores and validates the use of learning-based video motion magnification techniques for SHM purposes.

5.1 Methodology

5.1.1 Preliminary tests

Vibration-based damage detection techniques measure the vibration response of a structure to identify changes in its dynamic properties that may indicate the presence of damage. All objects and structures have natural modes of vibration which are a direct consequence of their material and geometrical properties. This implies that when the structure's properties change, the natural frequencies change as well, reflecting the geometric or material changes.

Vibration-based damage detection techniques offer several advantages for structural health monitoring as they provide continuous information on the health state of the structure at a global level without the need to access the damaged elements or know their location. Furthermore, damage can be identified using the data from sensors that are not necessarily located in the immediate vicinity of the damage.

Some vision-based techniques work by detecting geometry changes and cracks, however it is generally unfeasible to have full visual coverage of all structural elements. Consequently, the focus of this work is on combining the use of vision-based techniques with vibration-based damage detection. In order to do that, video motion magnification is used to reveal and measure the generally invisible or imperceptible dynamic response of structures.

As discussed in previous chapters, multiple approaches and techniques exist for motion magnification. Amongst them, the learning-based motion magnification techniques are interesting to our application due to their full-band motion magnification capacity as well as their superior image quality. Nevertheless, as previously mentioned in chapter 3 these techniques remained unproven in terms of obtaining reliable displacement measurements. Consequently, part of this thesis has been dedicated to the validation of learning-based techniques for SHM applications.

A set of preliminary experiments was designed to test the capacities and sensibility of Oh et al.'s [112] learning-based video motion magnification method, to test whether a trained model is actually capable of faithfully magnifying complex motion patterns.

The validation tests were conducted on a scaled-down benchmark structure subjected to different simulated damage scenarios. The chosen benchmark was the three-story building benchmark from Los Alamos National Laboatory (LANL) [231]. A structure is made of four square aluminium plates ($305 \times 305 \times 25mm$) connected by four prismatic aluminium columns measuring $177 \times 25 \times 6mm$. Each of the four plates is connected to the rest by four columns, forming a laboratory-scale three-story building (see Figure 5.1). The joints between the structural elements are all bolted together, allowing a quick replacement of the load-bearing elements to simulate damage.

The structure is connected to an electrodynamic shaker[1] that provides lateral excitation to the centre of the base floor, which is mounted over linear rails that allow for movement in one axis. Both the building and the shaker are then mounted over a rigid steel frame isolated from other sources of vibration.

For validation purposes, the three-story building benchmark has been instrumented with piezoelectric accelerometers[2] placed on each floor of the test structure that serve as the ground-truth reference for the simulated damage experiments. On the other hand, the structure was

[1] Brüel & Kjær electrodynamic shaker model 4824
[2] Brüel & Kjær type 4519-003

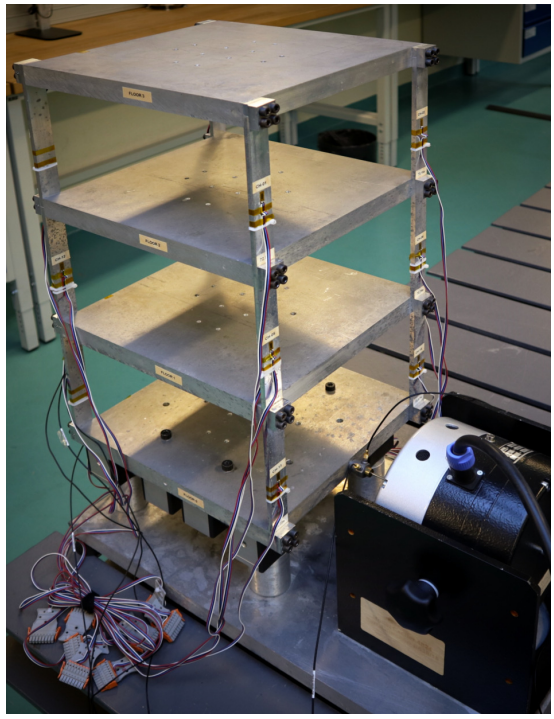


Figure 5.1: Three-story building benchmark setup [193].

also being recorded with a monochrome global shutter high-speed camera[3] placed in front of the structure so that the field of view covered the full height of the building. The video recording sampling rate was set to 200 frames per second (fps) to avoid aliased data in the experiments with excitation frequencies ranging 0 to 100 Hz.

To measure the vibration response of a structure, the electrodynamic shaker was used to produce different patterns of excitation. At the same time, a high-speed camera captured a 5-second video sequence of the structure. In order to obtain the dynamic response using the recorded video sequence it was magnified using the LB-VMM model to produce a series of motion-magnified frames. The magnification factor was chosen to be high enough to clearly capture the movement of the target structure while avoiding excessive magnification distortion.

Next, after the magnification stage, a single-pixel-wide linear region was manually selected from the captured video to track its displacement over time. Once selected, the linear regions were rearranged chronologically for each magnified frame, forming an image consisting of columns of pixels placed in chronological order. This image is known as a temporal slice and represents the displacement of a physical area over time. It is important to select high-contrast areas so that the displacement is clearly visible in the resulting slice.

[3] iX Cameras i-Speed 220 equipped with a 12.5-75mm f/1.8 2/3" format lens

Finally, to discretise the slice into a processable signal, a colour threshold was applied to the upper and lower sections of the time slice to find the boundary between the measured object and the background. Depending on the input data, it may also be helpful to apply a binary threshold to eliminate low-contrast areas. Once the boundaries were identified, their pixel coordinates were averaged to obtain the displacements over time in pixel units. Then, the Fast Fourier Transform (FFT) could be applied to the pixel displacement data to obtain a frequency-domain signal that can be used for damage detection in this test setup.

The preliminary tests to validate the viability of the previously described procedure were designed to address three main scenarios: single-frequency detection, multi-frequency detection, and damage detection. In all the tests, excitation amplitude was kept minimal so that the structure did not visibly move or vibrate without the aid of motion magnification. Meanwhile, both measurement systems, contact accelerometers and LB-VMM, were simultaneously used to measure the dynamic response of the subject structure.

The single-frequency tests were performed to assess the model's ability to accurately detect a very simple monofrequency sine excitation signal acting upon the structure. Sine waves at 20, 50, and 70 Hz were used for these tests, knowing that those frequencies roughly coincide with the three-story building benchmark's natural frequencies. The goal was to detect those frequencies and verify that they matched the input excitation frequencies.

The next tested scenario, were the multi-frequency excitation tests that were performed to evaluate the system's performance under more realistic conditions. The structure was excited using random noise or a periodic chirp signal within the range of 0 to 100Hz. A periodic chirp signal is a signal that continuously sweeps through a defined frequency range, in this case 0 to 100 Hz. In these experiments the motion magnification system was expected to identify the natural frequencies of the structure, as those frequencies get naturally amplified by the structural resonance. The visually obtained frequency response was expected to have a higher noise floor compared to the piezoelectric accelerometers, but still provide useful information about the frequency response.

Finally, to test the damage detection capability of the proposed method, six scenarios were assessed under a 100mV random noise excitation signal: the undamaged state, a 1.2 kg load affixed to the top floor, and four column damage scenarios. The column damage scenarios reduced the flat section of a single column of the structure by 10%, 20%, and 50%, which translates to approximately a 20%, 50%, and 85% stiffness reduction of the subject column, respectively. In the remaining fourth column damage scenario, the column was removed completely. These changes were expected to significantly impact the structure's stiffness, therefore altering its natural frequencies.

5.1.2 Tooling development for video motion magnification based SHM

The preliminary tests showed promising results with learning-based magnification being capable of detecting structural damage in the tested scenarios. However, the video processing methodology required a long and manual process to define pixel coordinates, crop frames to magnify, and rebuild slices from those magnified frames. This process, discouraged fine-tuning of the measurements and was a big inconvenience for the adoption of the technology. Along with the improvements to video motion magnification techniques presented in the previous chapter, building tools for its application to SHM was key in the usability of this technology.

A tool named ViMag [232] was built to provide an easy-to-use graphical user interface aimed at extracting time-series signals of vibrating machinery and structure videos. This software enables the visualisation of videos, the selection of one or more magnification areas, and the signal discretisation and processing. However, ViMag does not offer signal analysis or SHM capabilities, its goal is to extract the dynamic response signals from videos so that other specialised techniques and tools [233] can be used to further process the data.

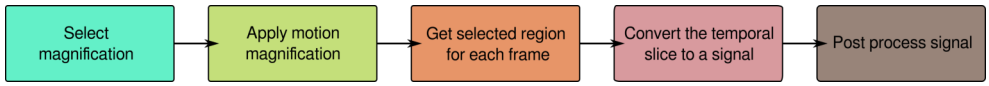


Figure 5.2: Video sequence to discrete signal pipeline [232]

Figure 5.2 illustrates the process of converting a video sequence into a discrete signal using ViMag [232]. The software employs an algorithm similar to that utilised in the preliminary tests, but facilitates the process through the use of graphical tools and automatic implementation of the algorithm's individual steps. The process commences when the user selects a linear region for measurement, which is subsequently cropped and magnified using STB-VMM. The linear regions of pixels from each magnified frame are then placed sequentially on an image to form a temporal slice, as depicted in Figure 5.3. The newly generated image depicts the displacement over time of the selected region of interest, which can be transformed into a discrete signal through the application of relatively simple image processing techniques. Once the signal has been discretised, a variety of signal processing techniques, such as the Fourier transform or other advanced damage detection algorithms, may be employed to extract additional information from the data.

ViMag can be configured to use any motion magnification backend, however it only officially supports STB-VMM. Nevertheless, when comparisons between STB-VMM and LB-VMM are presented in the results section the comparisons have been made using a modified version of ViMag that uses LB-VMM. This modification allows the comparisons to use the same project files and regions to provide reliable and directly comparable results.

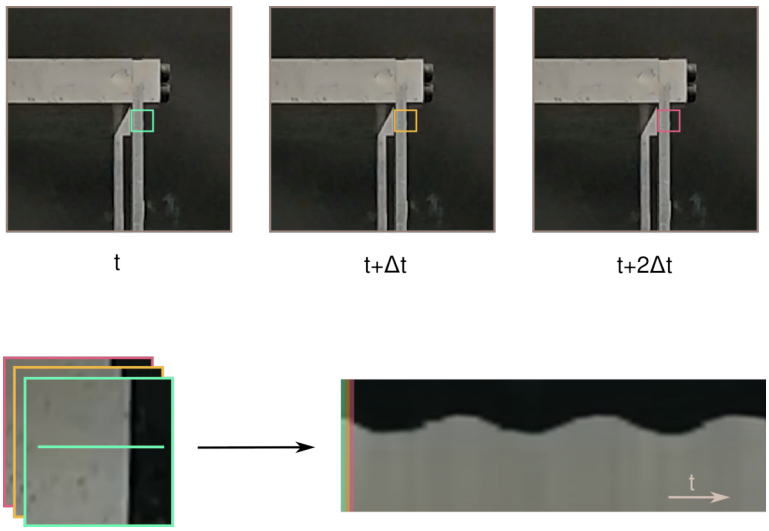


Figure 5.3: Video sequence transformation to temporal slice [232]

5.1.3 Model validation

After developing the STB-VMM technique and software tooling, two experimental cases were used to validate the new tools and compare them to LB-VMM. The first case was conducted, once again, on the three-story building benchmark [231], while the second case employed third-party data of a large-scale cast-in-place reinforced concrete frame structure [234] from Tongji University.

5.1.3.1 Case 1: Three-story building benchmark

In the first case, the experiments were conducted on the three-story building benchmark, a well-known reference structure made of aluminium also used in the preliminary tests. In this second round of experimental testing, the structure was instrumented again with piezoelectric accelerometers placed on each floor of the test structure serving as a ground truth reference. In addition, a new global shutter industrial camera[4] was recording the experiments at 500 fps in 8-bit depth monochrome. The resulting video sequences of this recordings were processed using ViMag with both STB-VMM and LB-VMM as magnification backends.

The studied damages detection scenarios for these tests consisted in reducing the longitudinal cross-section of the lower left column of the benchmark structure by 10%, 20%, 50% and 100% (column removal). These cross-section reductions impacted the stiffness of the individual column by 20%, 50%, 85%, and 100%, respectively. Those alterations on a component of the test structure translate into a shift in the global natural frequencies of the structure and

[4] Allied Vision Alvium 1800 U-158c

therefore, mark a clear indicator of damage to be detected by the vision-based techniques. On top of these damage scenarios, a moderate loading scenario was also reproduced adding a 1.2 kg mass to the top floor of the building in its undamaged configuration.

The excitation for these experiments was limited between 0 and 100 Hz using two different patterns, random noise and a periodic chirp. The frequency sweep generally produces cleaner dynamic response signals, however, it is an unrealistic scenario for large civil structure subjected to ambiental and operational loadings. For this reason, random noise excitation is also used, as it provides more realistic excitation to test damage detection algorithms despite producing dirtier signals.

5.1.3.2 Case 2: Reinforced concrete frame structure

The second case study is based on the data provided by [234], which designed a vision system to detect and track structural displacements and rotations on a 6 m tall four-story reinforced concrete structure. Each of the stories of the test building were 1.5 m tall with a 3×1 m footprint and were loaded with 2700 kg each except for the roof, which was loaded with 2580 kg. The test building was placed over a 4×4 m 25-ton shaker table that excited the structure using recordings of real earthquakes. Figure 5.4 shows the structure being tested on the shaker table.

Figure 5.4: Video frame corresponding to the tests performed on the reinforced concrete frame. The left-hand side corresponds to the original unaltered frame, while on the right, a $\times 20$ magnified version of the frame shows its operating deflection shape.

The building was instrumented with a non-contact high-speed motion capture system[5] with an acquisition rate of 128 Hz. At the same time, the experiments were recorded with a consumer-grade camera[6] equipped with wide-angle optics and placed at approximately 3 m in front of the reinforced concrete structure. The camera recordings were performed with a resolution of 1920×1080 and 25 fps. In the following results section, the data from the high-speed motion capture system will be compared against the motion magnification based vision systems.

5.2 Results

The following section presents the results of the experimental campaigns previously described. First, the results of the preliminary tests are presented and then followed by the comparison and validation tests for STB-VMM and the visual vibration toolbox.

5.2.1 Preliminary tests

5.2.1.1 Single frequency performance

The tests for single-frequency performance reveal the capacity of the learning-based model to detect simple and precise frequencies. Figure 5.5 compares the motion-magnified results of different excitation frequencies at different amplitudes.

It can be observed that in both amplitude scenarios the structure's excitation frequencies are clearly reflected on the graphs, as should be expected. However, different frequency peaks appear with different heights despite an equal excitation amplitude. This phenomenon occurs due to the natural resonance of the structure, the 50 Hz frequency is the closest to one of the natural frequencies of the structure and for this reason it gets mechanically amplified further than the other signals tested. Exploiting this principle, we can find the natural frequencies of a structure by using a wider spectrum of test signals.

5.2.1.2 Multi-frequency performance

The multi-frequency tests check the capacity of the vision-based method to magnify and track complex motion. Using either a periodic signal or random noise, these tests reveal the natural frequencies of the structure. Figure 5.6 plots the frequency domain results obtained using LB-VMM comparing different excitation signals. The figure also shows the comparison between the LB-VMM method and the reference accelerometers.

[5] NDI Optotrack Certus
[6] Sony HDR-PJ220

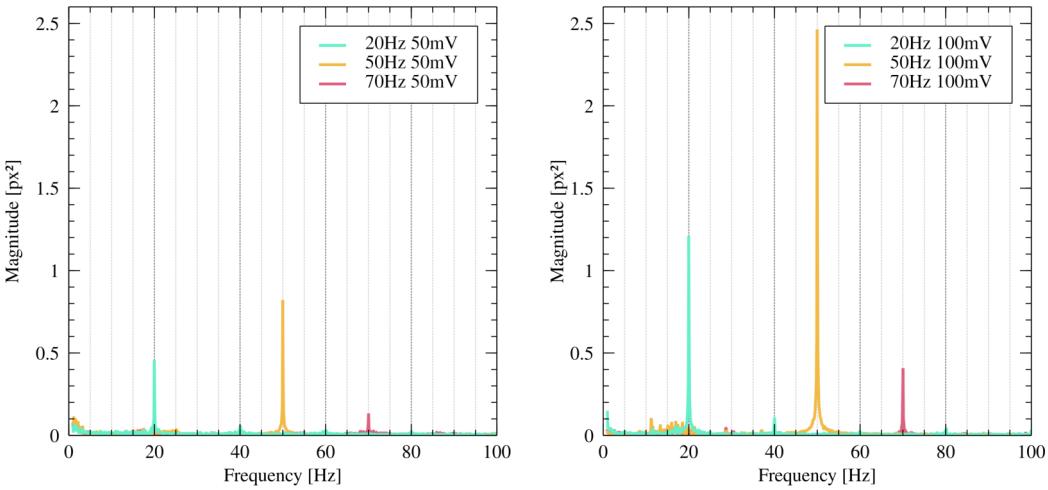


Figure 5.5: Frequency domain response for three different tests using a 20, 50, and 70 Hz sine wave for excitation at two different levels of amplitude: 50mV and 100mV.

Each of the categories shown in the legend of the figure display multiple curves that correspond to the different sensors placed on each floor of the three-story building benchmark. Notice how some curves display less prominent peaks than others as a result of the point of measurement being closer or further to a node of that particular mode shape. In the case of the three-story building benchmark, three different bending modes are found between 0 and 100Hz (see Figure 5.7).

Figure 5.6 demonstrates that the LB-VMM approach is capable of detecting multiple frequencies at once with suitable accuracy, as the response inferred using the visual method coincides with the piezoelectric accelerometers. However, as anticipated, the LB-VMM approach suffers from higher noise floors than the accelerometers. Nevertheless, the following set of experiments proved that this was not an obstacle to accurate damage detection.

5.2.1.3 Damage detection

The damage detection capability of the LB-VMM approach was evaluated through the implementation of five distinct damage scenarios, each subjected to a 100mV random noise excitation. The first scenario, which was the most severe, involved the removal of a column from the lowermost story of the building, specifically on the side closest to both the shaker and camera. Subsequently, this same column was subjected to three additional damage scenarios of lesser severity, achieved through reductions in column thickness by 10%, 20%, and 50%, resulting in corresponding reductions in column stiffness of approximately 20%, 50%, and 85%. The fifth scenario entailed affixing a test mass of 1.2 kg to the uppermost floor of the

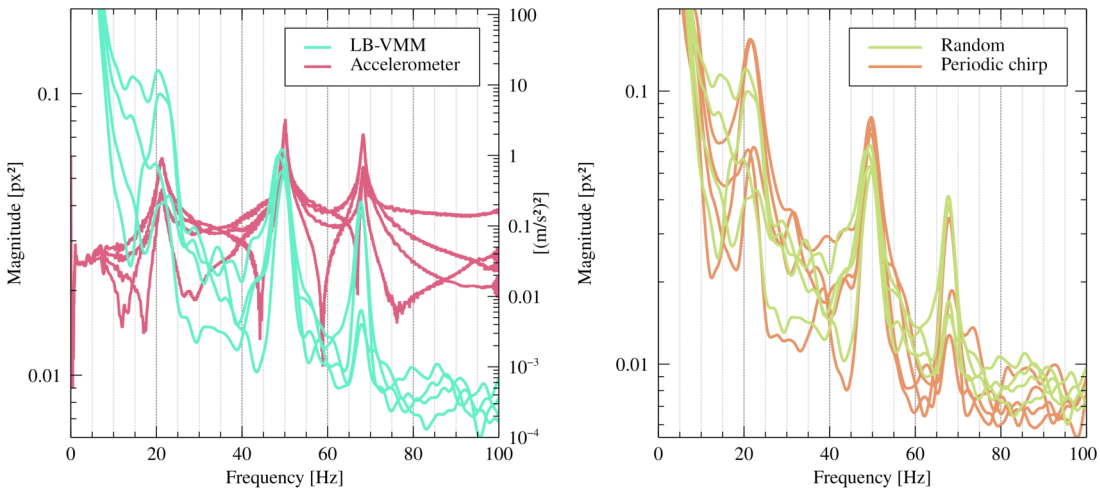


Figure 5.6: Comparison of multi-frequency performance test results inferred from the LB-VMM approach (×20, 100 mV) between the type of excitation (right) and acquisition sensor (left).

undamaged building. These scenarios were reproduced in accordance with the methodology detailed in the original reference [231].

Overall, the results demonstrate small yet significant shifts in the natural frequencies of the building (see Figure 5.8), losing stiffness as the damage increases in severity, as anticipated. For the sake of simplicity, Figure 5.8, only shows the signal corresponding to an accelerometer located on the top floor to represent the frequency response function. The results obtained through LB-VMM and accelerometer data exhibit highly similar shifts in frequency, although it should be noted that the third mode is masked on the LB-VMM spectrum. This can be attributed to two primary factors: firstly, an inadequate selection of tracking line placement coinciding with the position of the upper floor accelerometer, which was situated in a region characterised by low mobility of the operating deflection shape; and secondly, due to the elevated noise floors exhibited by LB-VMM in comparison to those of the accelerometer.

However, this result should not be interpreted as undermining the potential utility of the LB-VMM approach; as demonstrated in the previous point (see Figure 5.6), where different tracking lines were employed and the third mode was satisfactorily identified. In other words, in order to capture a greater number of modes, it is necessary to determine response data from various areas along the structure, analogous to the spatial distribution of accelerometers.

Figure 5.8 presents a comparative analysis of the results for each damage scenario. Figure 5.8a illustrates the frequency response difference between an undamaged structure and an identical structure missing a column on its ground floor. The LB-VMM approach effectively demonstrates its suitability for detecting such damage; however, it should be noted that

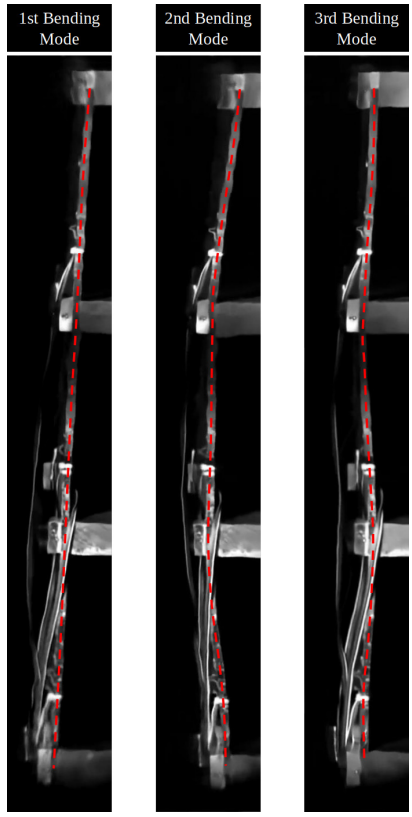


Figure 5.7: ODS of the first three bending modes of the benchmarked structure [193].

complete column failure represents a somewhat unrealistic scenario. Meanwhile, Figure 5.8b presents a comparison between three distinct column thickness reductions for both accelerometer and LB-VMM data. The graphic exhibits consistent results between both spectra, thereby confirming the capability for damage detection utilising LB-VMM techniques. It is also noteworthy that this system proves capable of detecting modifications as small as a 10% difference in thickness for a single column. Finally, Figure 5.8c displays the comparison between the undamaged unloaded structure and the same undamaged structure with a 1.2 kg mass affixed to its uppermost floor. Once again, it can be observed that changes to the natural frequencies are minimal and approach the resolution limit for the LB-VMM setup; however, these changes remain present and verifiable against accelerometer data. Notice that in this case, the resolution limit of the vision-based system is determined by the camera's acquisition frame rate.

For a more detailed validation of the system's capacity, Table 5.1 includes the comparative numerical data of the frequency peaks identified in each damage scenario with each sensing method, including the calculated percentage errors. These same results have been plotted in Figure 5.9 for easier visualisation.

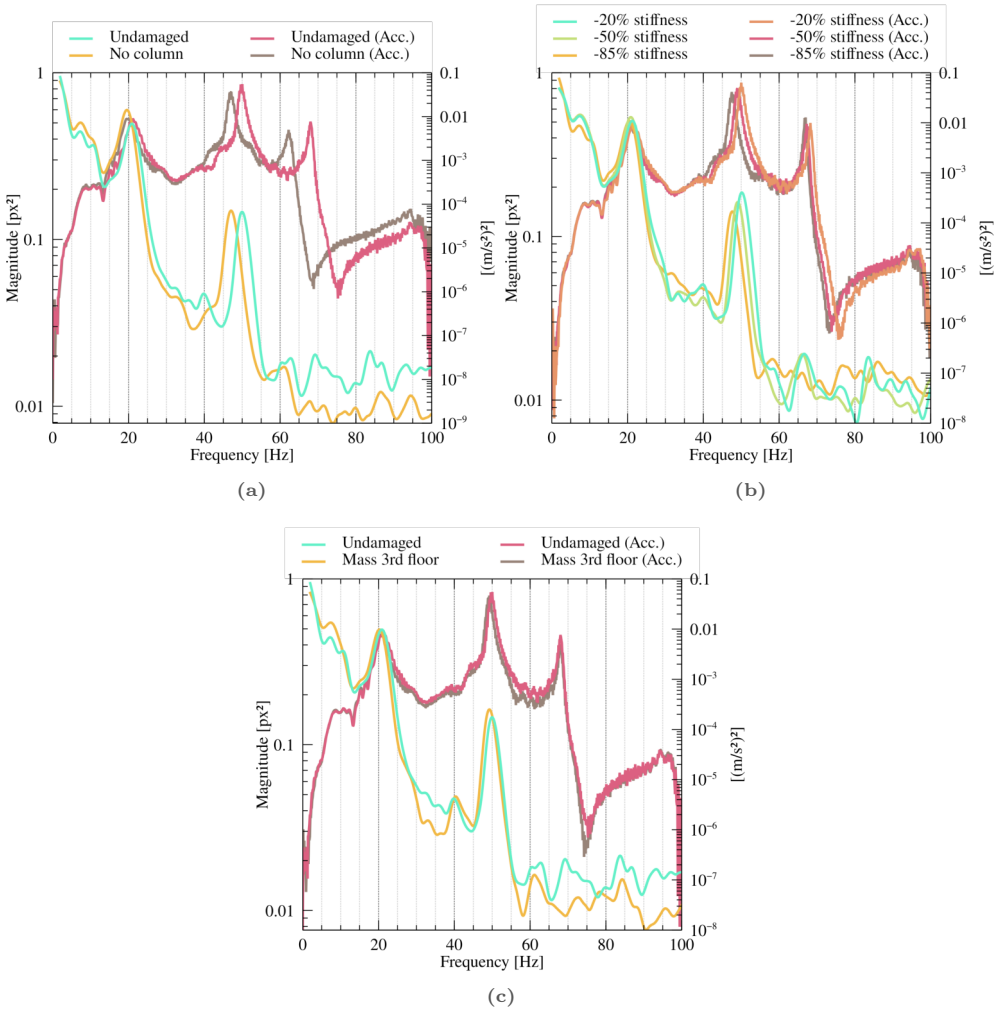


Figure 5.8: Comparison between frequency-domain data from the LB-VMM model and piezoelectric accelerometers, showing the frequency response differences between the undamaged structure and different damage scenarios [193].

The overall results of these preliminary tests indicate that the learning-based video motion magnification method offers suitable outputs for SHM purposes. Therefore, the results justified the time investment into the improvement of the learning-based magnification techniques and their tools.

5.2.2 Model validation and comparison

The following are the results of the experiments conducted using the newly developed tools and methods that compare STB-VMM and LB-VMM on their displacement detection capabilities.

Table 5.1: Comparative of the mode frequencies obtained from each damage scenario with each sensing method [193]. (See Figure 5.9)

	Accelerometer (Hz)		LB-VMM (Hz)		Error (%)	
	1st order frequency	2nd order frequency	1st order frequency	2nd order frequency	1st order frequency	2nd order frequency
Pristine	21.125	50.000	20.859	49.945	1.26	0.11
Mass	20.500	49.250	20.272	49.259	1.11	0.02
20%	21.250	48.875	21.251	50.141	0.01	2.59
50%	21.125	49.000	20.859	49.064	1.26	0.13
85%	21.125	47.500	20.468	47.692	3.11	0.41
No column	19.500	47.125	19.586	47.007	0.44	0.25

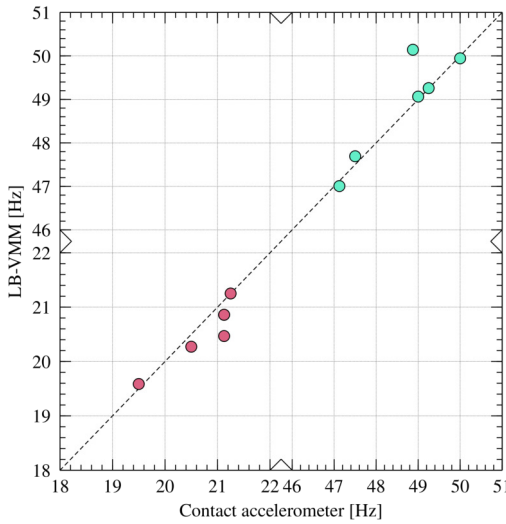


Figure 5.9: Comparative of the mode frequencies obtained from each damage scenario with each sensing method [193]. (See Table 5.1.)

5.2.2.1 Case 1: Three-story building benchmark

This section presents and analyses the results for the three-story building benchmark and its various configurations. Figure 5.10 compares the autospectra obtained using STB-VMM with those from reference piezoelectric accelerometers. As expected, the vision-based method has a higher noise floor but only requires a single sensor to obtain an array of signals, in this case, four signals were captured to match the four accelerometers placed in the test structure. Additionally, the visual method, as a side-effect of being based on displacement, is more sensitive to lower frequencies. In fact, STB-VMM registers frequencies lower than the reference piezoelectric accelerometers. This discrepancy can be explained due to the low-frequency start of the periodic chirp signal used to excite the structure, that displaces the whole structure creating a visible peak on the autospectrum. It must be noted then, that these peaks do not correspond with any vibration mode of the structure. Instead, this indicates the frequency of the displacement of the structure when the excitation cycle begins. On the other hand, higher frequencies are better captured using accelerometers, which are better suited for

this task and have much higher acquisition rates than industrial or general-purpose cameras. This trade-off might be acceptable for civil infrastructures as they generally exhibit frequencies under 15 Hz.

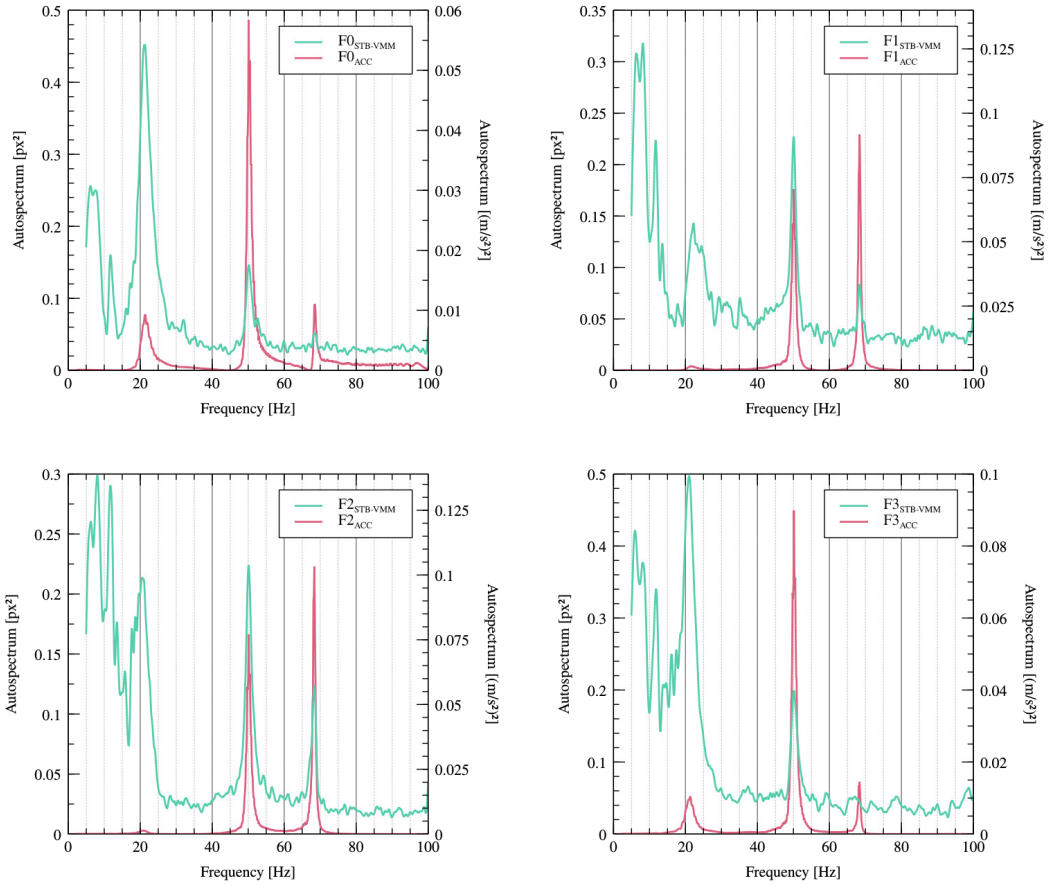


Figure 5.10: Frequency response functions for each floor of the three-story building benchmark in its undamaged configuration. The structure was excited using a periodic chirp signal ranging from 0 to 100 Hz.

Focusing the attention back on Figure 5.10 it can be seen that not all floors of the structure have the same magnitude for their natural frequencies, this is due to their physical position in the building. The magnitude difference for each frequency is proportional to the distance between the measured point and the closest node of the modal shape.

After verifying that STB-VMM can correctly detect frequencies in the undamaged conditions, the following results evaluate its sensitivity to damage. Figures 5.11a, 5.11b, and 5.11c show the results of the damage scenarios performed on the lower column of the three-story building benchmark, as described in section 5.1.3.1. The results are split into three figures for readability.

Figure 5.11a shows results for undamaged, 1.2 kg added mass, and missing column cases.

Table 5.2: Results obtained from periodic excitation

	Accelerometer (Hz)			STB-VMM (Hz)			LB-VMM (Hz)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	21.38	50.12	68.38	21.10	50.33	68.27	21.10	50.36	68.30
Undamaged$_{Repr.}$	21.38	50.12	68.38	21.26	50.11	68.37	21.34	50.15	68.29
Mass	21.25	49.00	68.00	20.12	49.16	68.00	20.47	49.01	67.88
20%	21.38	49.38	67.50	21.36	49.34	67.70	21.02	49.34	67.77
50%	21.38	48.62	67.62	20.86	48.59	67.52	20.78	48.55	67.48
85%	20.62	46.62	66.75	20.42	46.76	66.60	20.40	46.76	66.64
100%	19.88	45.88	62.62	19.80	45.82	62.47	19.72	45.89	62.62

The difference between the undamaged and the missing column scenarios is most noticeable, while the structure loading experiment produces a more subtle frequency shift visible in all three methods tested. Figures 5.11a and 5.11b present the data obtained using STB-VMM, LB-VMM, and accelerometers in three separate side-by-side graphs. Comparing the different measurements, the vision-based methods, which are based on displacement, present a more significant sensitivity for lower frequencies and a noticeably higher noise floor with respect to the contact accelerometers. Nevertheless, damage sensitivity is similar for all methods, with LB-VMM presenting a slightly higher noise floor than STB-VMM.

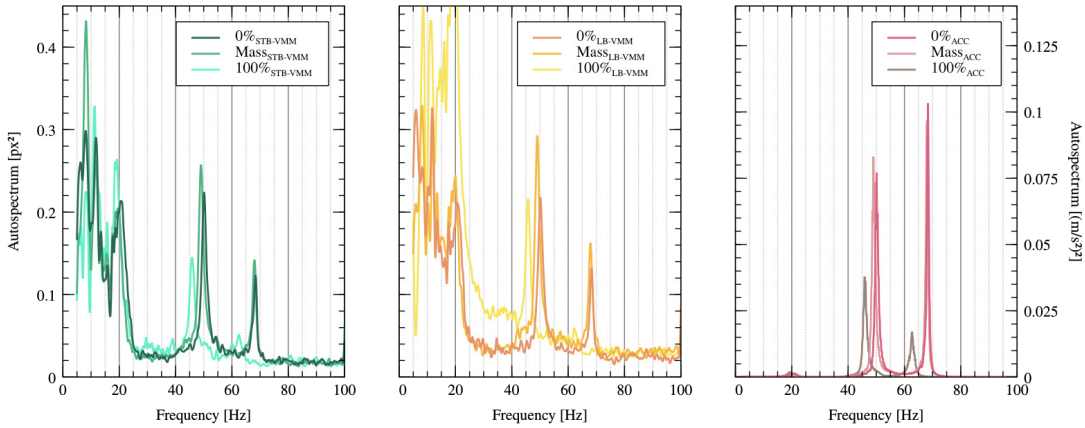
The measurements and magnification for vision-based graphs were performed on the exact same coordinates for both methods and then post-processed individually to improve signal quality as much as possible. The differences in magnification quality between the methods often resulted in variations in slice contrast due to blurring and artefacting, making uniform post-processing unfeasible.

Figure 5.11b continues with the trend previously observed, showing that all three methods can discern between the tested damage scenarios, where a single column loses 10%, 20%, and 50% of its longitudinal section. These experiments pose a challenging test as the resulting frequency spectrums are very close to each other.

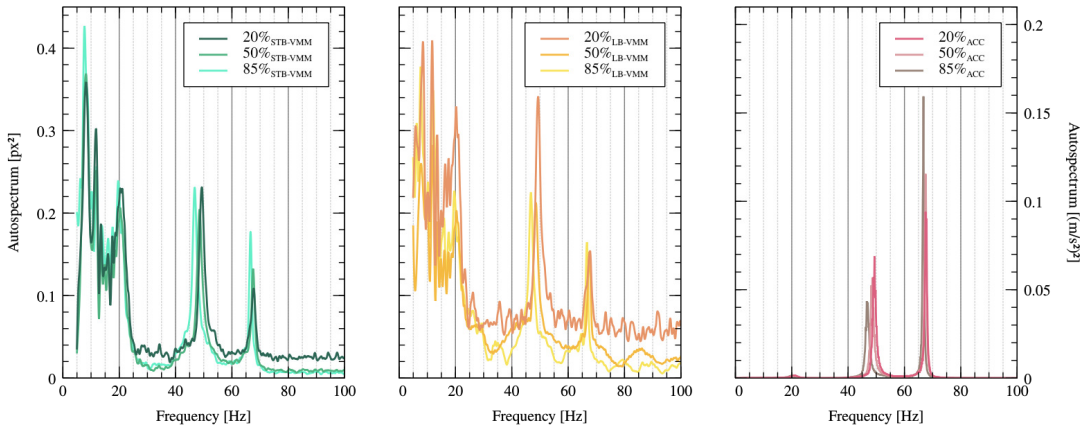
When these same tests are performed using random noise excitation instead of a periodic sinusoidal signal, the overall conclusions remain the same, but with a visible increase in noise for all methods, as observed in Figure 5.11c.

Table 5.3: Relative error to the reference acceleration data when using periodic excitation

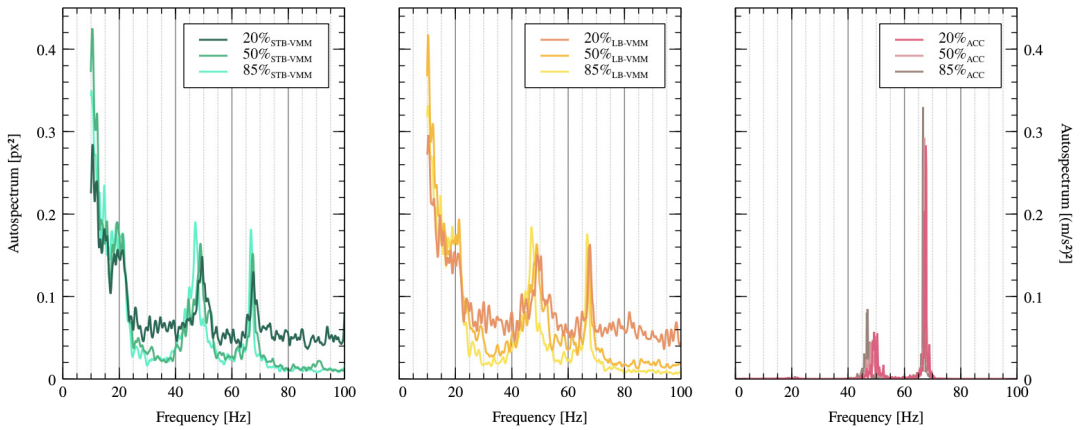
	Error STB-VMM (%)			Error LB-VMM (%)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	1.31%	0.42%	0.16%	1.31%	0.48%	0.12%
Undamaged$_{Repr.}$	0.56%	0.02%	0.01%	0.19%	0.06%	0.13%
Mass	5.32%	0.33%	0.00%	3.67%	0.02%	0.18%
20%	0.09%	0.08%	0.30%	1.68%	0.08%	0.40%
50%	2.43%	0.06%	0.15%	2.81%	0.14%	0.21%
85%	0.97%	0.30%	0.22%	1.07%	0.30%	0.16%
100%	0.40%	0.13%	0.24%	0.80%	0.02%	0.00%



(a) Undamaged, loaded, and missing column damage scenarios performed under periodic excitation ranging from 0 to 100Hz.



(b) Column damage scenarios performed under periodic excitation ranging from 0 to 100Hz.



(c) Column damage scenarios performed under random noise excitation.

Figure 5.11: Damage sensibility study of the damage scenarios presented in section 5.1.3.1. Each signal corresponds to the second floor of the three-story building benchmark measured under different damage conditions or using different measurement techniques.

Tables 5.2-5.5 present the numerical results of the experiments performed on the three-story building benchmark as well as the errors of visual methods compared to reference accelerometers. On average, the errors of STB-VMM compared to accelerometer data are 0.64% for periodic excitation and 1.03% for random excitation. Similarly, the errors for LB-VMM are 0.66% and 1.38%, respectively, slightly worse than the STB-VMM results.

Table 5.4: Results obtained from random excitation

	Accelerometer (Hz)			STB-VMM (Hz)			LB-VMM (Hz)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	22.25	50.38	68.88	22.12	50.35	68.76	21.16	50.25	68.71
Undamaged$_{Repr.}$	22.50	50.38	68.75	22.08	49.66	68.58	22.15	49.98	68.77
Mass	22.25	48.75	67.75	21.16	48.40	67.64	21.16	48.49	67.64
20%	22.25	49.25	67.62	21.35	49.39	67.63	22.18	50.35	67.69
50%	22.25	48.75	67.12	21.23	48.80	67.43	21.47	48.80	67.43
85%	21.12	46.88	66.62	21.01	47.03	66.79	21.50	47.03	66.79
100%	19.75	45.75	62.75	19.73	45.40	62.71	20.90	45.47	62.71

Table 5.5: Relative error to the reference acceleration data when using random excitation

	Error STB-VMM (%)			Error LB-VMM (%)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	0.58%	0.06%	0.17%	4.90%	0.26%	0.25%
Undamaged$_{Repr.}$	1.87%	1.43%	0.25%	1.56%	0.79%	0.03%
Mass	4.90%	0.72%	0.16%	4.90%	0.53%	0.16%
20%	4.04%	0.28%	0.01%	0.31%	2.23%	0.10%
50%	4.58%	0.10%	0.46%	3.51%	0.10%	0.46%
85%	0.52%	0.32%	0.26%	1.80%	0.32%	0.26%
100%	0.10%	0.77%	0.06%	5.82%	0.61%	0.06%

5.2.2.2 Case 1: Visual noise resilience

To evaluate the effect of visual noise on VMM measurements, white noise was added to the same video source files used in the previous experiments. STB-VMM, with its intensive use of image filtering, is expected to perform better than LB-VMM in high-noise situations. Visual noise is very disruptive for motion magnification as it masks sub-pixel motion and destroys key information when the dynamic range is low, or noise is sufficiently pronounced. Unfortunately, noise is a common occurrence in all image sensors that becomes more pronounced in poorly lit scenes.

Figure 5.12 shows the difference in measured frequency response when noise is added to original recordings. The altered video files have an average Structural Similarity Index Measure (SSIM) [223] of 0.45, indicating a significant amount of noise while remaining legible for human observers. However, the information loss is pronounced, causing higher frequencies to disappear in the noise and reduce their magnitude.

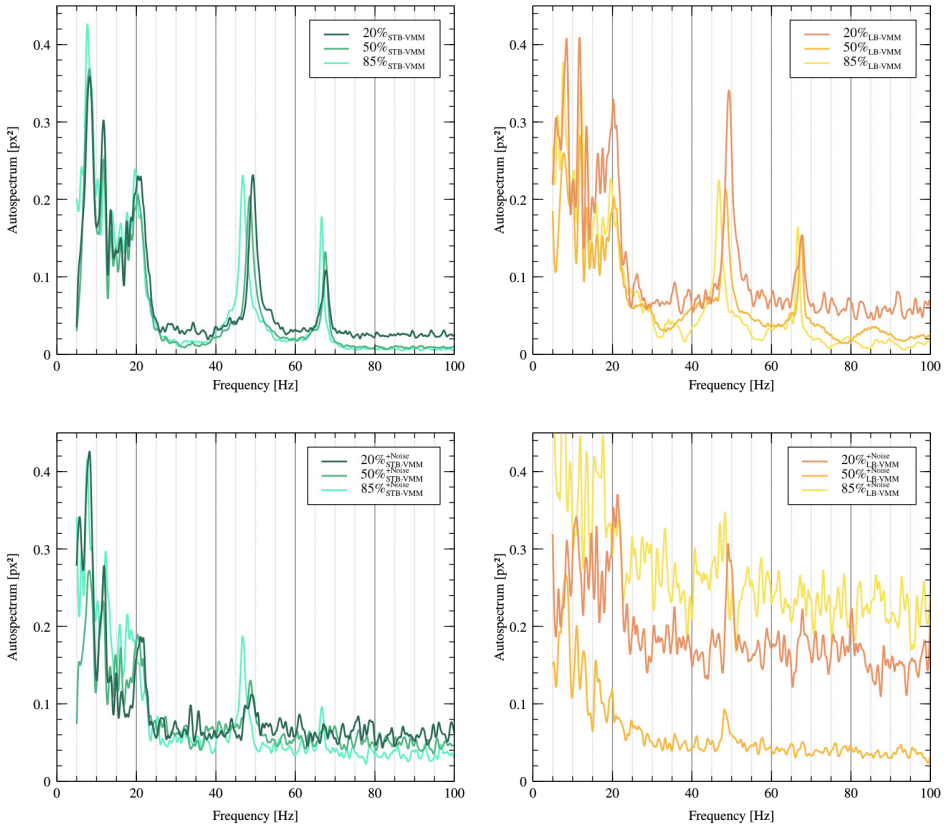


Figure 5.12: Comparison between the frequency responses obtained from the undamaged video files on top, and the white noise altered ones, at the bottom.

As expected, STB-VMM maintained a more stable behaviour than LB-VMM, however its results are still far from the noiseless measurements previously obtained. In less extreme situations, STB-VMM is expected to yield better results, especially in unfavourable conditions, or allow for the use of more economical image sensors with little or no penalty.

5.2.2.3 Case 1: Resolution study

The effect of reducing the resolution of source video files was also studied. Like noise, a reduction in resolution deletes information present in the video, reducing key features from a few pixels or sub-pixels to variations indistinguishable from noise. This is analogous to moving the image sensor further away from a target subject or reducing optical zoom, as the resolution in pixels of a particular target feature gets reduced. However, unlike adding noise, no spurious results are expected as no new false information is added to the image sequences.

Figure 5.13 shows results of reducing the original horizontal resolution by 20%, 50%, and 80%. The videos are scaled to maintain the aspect ratio, resulting in the following video

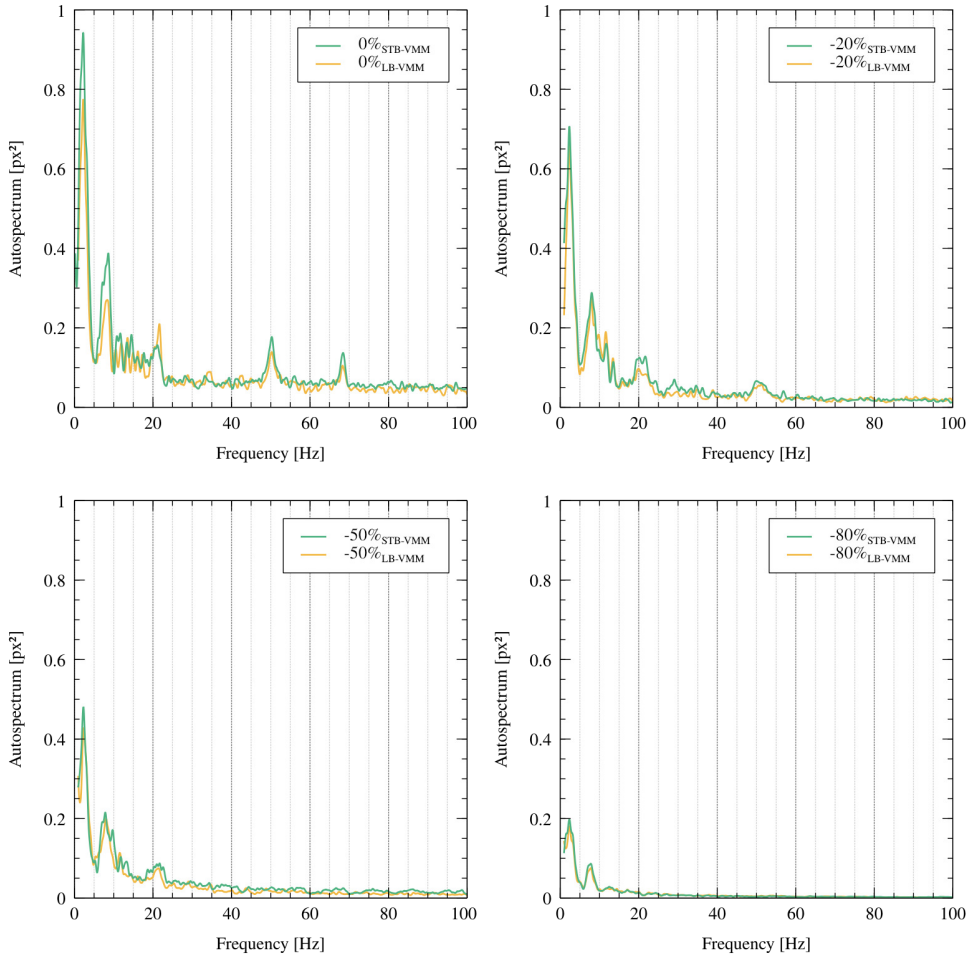


Figure 5.13: Frequency response graphs of decreasing resolution experiments, from the original sized measure to a reduction of 80% in the horizontal resolution. These measurements correspond to the second floor of the undamaged condition building under periodic excitation.

resolutions: original (1456×384), -20% horizontal resolution (1164×307), -50% horizontal resolution (728×192), and -80% horizontal resolution (291×77). In terms of number of pixels, these sizes contain 36.1%, 75%, and 96% fewer pixels than the original.

The results show a clear decrease in magnitude as resolution decreases, losing higher frequency peaks progressively to low-resolution blur. This makes sense as higher frequencies manifest as lower amplitude displacements because excitation power is constant in all experiments. When losing pixels to represent movement, shorter displacements end up getting lost beyond what motion magnification can recover.

As shown in Figure 5.13, differences between STB-VMM and LB-VMM are not significant when working with reduced resolution. In this case, better filtering cannot recover missing

information, and no significant improvement is registered.

5.2.2.4 Case 2: Reinforced concrete frame structure

The reinforced concrete frame structure offers fewer configuration options and variability than the three-story building benchmark, but its scale makes it a relevant test for the scalability of visual-based measurement systems. Additionally, these videos were recorded with a consumer-grade camera by independent third-party researchers holding no bias on their setup for accommodating motion magnification.

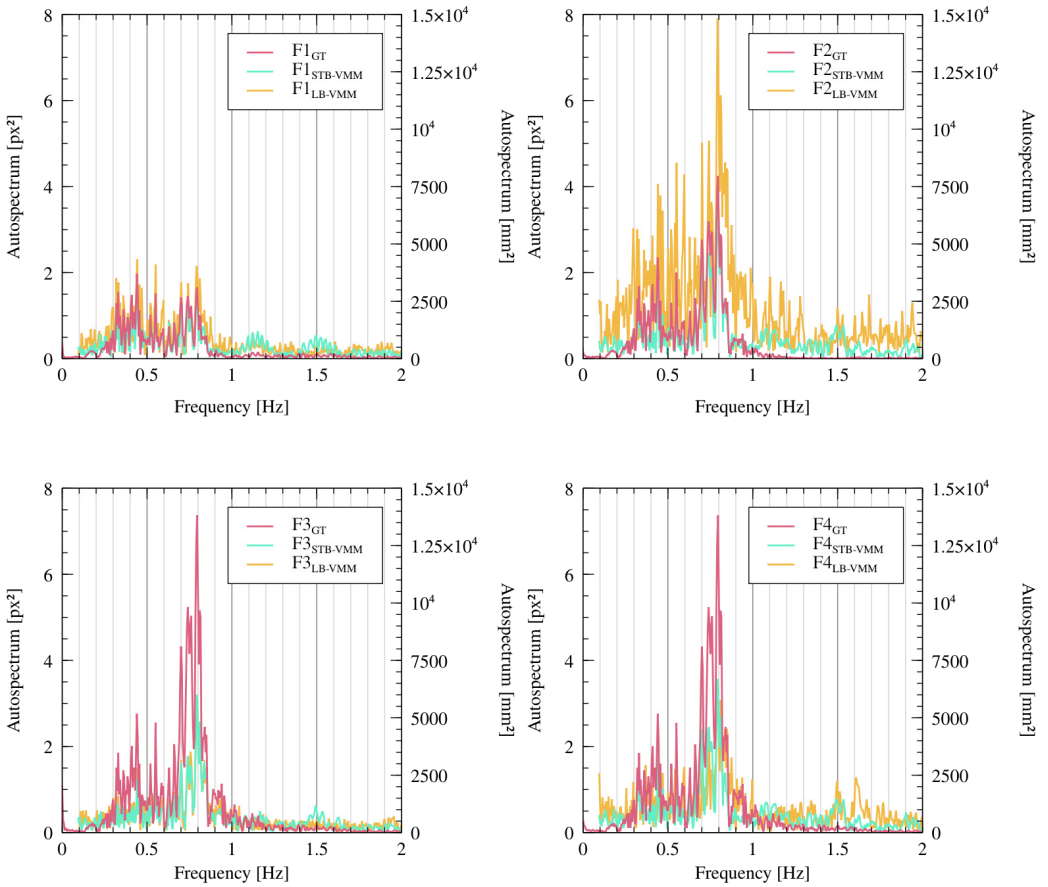


Figure 5.14: Graphs comparing each floor measurement against experiment 1's ground truth. (Raw data provided by Wang et al. [234].)

Figures 5.14 and 5.15 show two of the experiments shared by Wang et al. [234] and processed using STB-VMM and LB-VMM techniques. Experiments 1 and 2 show vision-based measurements remarkably close to the ground-truth data provided, albeit with more noise present. The spectra are represented from 0 to 2 Hz as the most relevant frequencies are concentrated between 0.6 and 0.9 Hz depending on the floor, which is expected for a structure

of its size and rigidity. No notable performance differences appear between STB-VMM and LB-VMM results besides some minor differences in noise levels.

Considering the temporal resolution difference between the signal produced by the optical vibrometers at 128 Hz and the camera at 25 Hz, the results are considered very satisfactory. It should be noted that the camera is not optimally positioned for these measurements, instead it should be placed directly in front and perpendicular to the measurement plane, nor is it equipped with lenses that do not excessively distort the image. Despite these limitations, the resulting autospectra provides relevant information on the structure's frequencies.

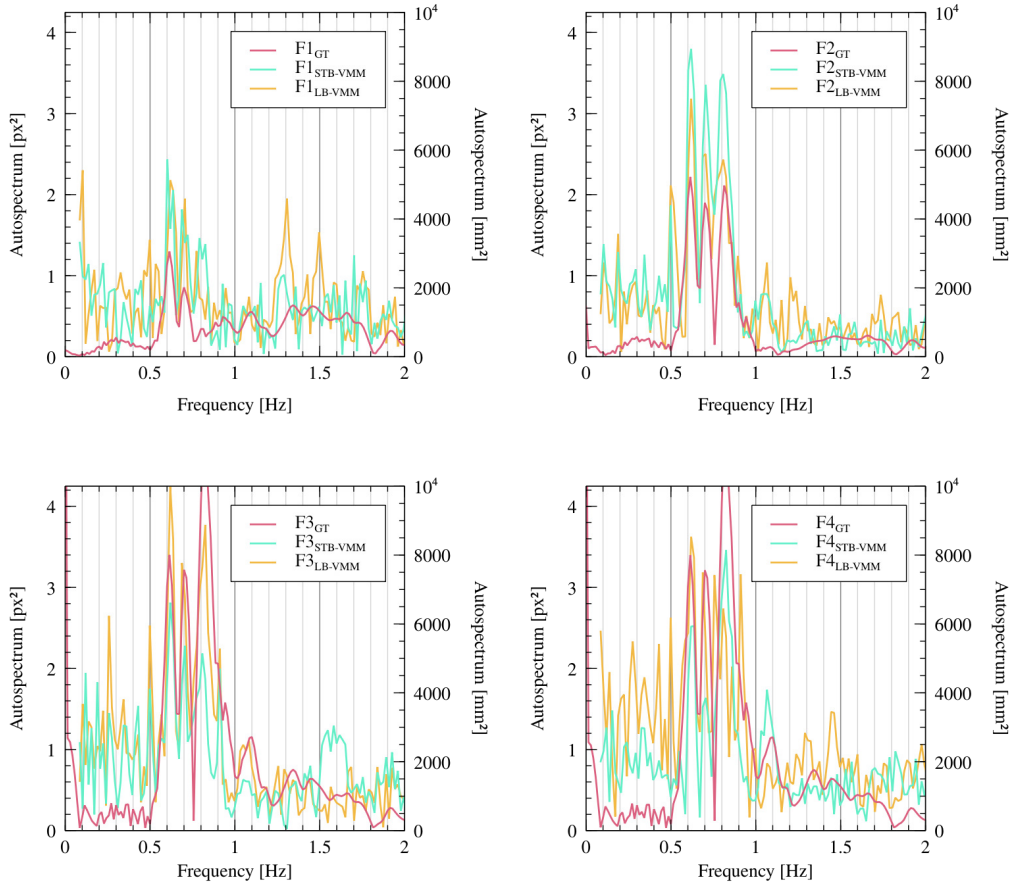


Figure 5.15: Graphs comparing each floor measurement against experiment 2's ground truth. (Raw data provided by Wang et al. [234].)

The overall results of this study validate the use of STB-VMM and its tools for the purpose of SHM. The experimental work has shown that these systems are capable of identifying a structure's natural frequencies and operating deflection shapes, thus revealing their potential for contactless full-field damage detection.

6

Conclusions and future works

This chapter summarises the most relevant contributions and insights of this work and reflects on the future avenues of research for video motion magnification based techniques and their applications in mechanical and civil engineering

6.1 General conclusions

The objective of this thesis is to develop, improve, and validate a computer vision methodology based on learning-based video motion magnification for the purpose of using structural health monitoring techniques in complex structures through full-field optical means. Based on the results of this work, the following conclusions can be drawn:

- Learning-based video motion magnification enhances the sensitivity of vision-based vibration analysis, allowing for more detailed and accurate measurements. This approach provides several benefits, including full-field measurements that simultaneously reduce sensor installation complexity while providing information on multiple targets. The proposed methodology has been proven effective for dynamic response identification in structures subjected to complex excitation patterns.
- The combination of Eulerian video motion magnification with deep learning techniques such as visual transformers and convolutional neural networks proposed has significantly improved video motion magnification image quality. This quality enhancement has been proven both qualitatively, with side-by-side comparisons, and quantitatively, by using a well-known image quality assessment algorithm. The Shifted Window Transformer Based Video Motion Magnification (STB-VMM) method has been demonstrated to outperform the prior state-of-the-art learning-based magnification model in terms of image quality, displaying better noise tolerance characteristics, less blurry outputs, and better edge stability, resulting in clearer and less noisy magnification with minimal artefacting.
- STB-VMM is a rather computationally expensive model, limiting its use in real-time applications of motion magnification. Nevertheless, applications that require precise magnification, like vibration monitoring, can benefit from its superior output image quality, making this model a good fit for structural health monitoring.
- The preliminary tests performed on the three-story building benchmark constituted the first validation tests of the usage of learning-based video motion magnification techniques for the purpose of structural health monitoring and damage detection. This work proved that the learning-based model (LB-VMM) yields reliable results even in complex motion scenarios, identifying the test structure's natural frequencies and operating deflection shapes. Moreover, the dynamic response measurements were accurate enough to detect several challenging simulated damage scenarios.
- Employing video motion magnification for the study and monitoring of structural and mechanical systems allows for the reconstruction and visualization of operating deflection shapes irrespective of the vibration source. This approach highlights abnormal patterns in structural contacts, cracks, or excessive displacements, thereby facilitating visual inspection.

- Despite using learning-based models trained on an unrelated synthetic motion magnification dataset that was not specifically designed for structural measurements, the learning-based models yielded outstanding magnification results accurate to the ground truth measurements.

- The validation tests for STB-VMM and ViMag successfully proved the validity of the developed software, and established an automatic video processing pipeline that favourably compares to the LB-VMM methodology initially developed. STB-VMM presents clear advantages when processing noisy data with respect to LB-VMM, however, in more favourable image conditions, these improvements become marginal. This contrasts with the results of the comparison between STB-VMM and LB-VMM as video motion magnification algorithms, where STB-VMM clearly showed superior image quality results. The most probable cause for this discrepancy is that the applied dynamic response measurement methodology just uses unidimensional strips of pixels, negating some of the improvements in texture resolution and edge stability of the new model. Nevertheless, STB-VMM should remain the preferred method for measurement since conditions on the field are rarely ideal.

- The proposed motion magnification approach for SHM requires consistent lighting and a stable camera that maintains a stationary frame of reference. Additionally, measurements may present higher noise floors compared to contact-based solutions. Nevertheless, these limitations depend on the particular boundary conditions and requirements of each use case. Visual methods, although presenting higher noise floors than contact sensors or laser techniques, open up new possibilities for applications and research in structural health monitoring. Newer and cost-effective techniques, such as vision-based monitoring, combined with the transition towards higher digitalisation in civil engineering, are continuously increasing safety and reducing the costs associated with unplanned maintenance and service disruptions. Overall, motion magnification in vision-based vibration analysis offers a significant cost advantage for structural health monitoring and damage detection.

6.2 Research contributions

The results achieved during this thesis contribute to the following scientific findings:

- A novel methodology based on Eulerian learning-based video motion magnification was developed and validated for measuring the dynamic response of structures. This methodology provides a state-of-the-art, non-contact approach to structural health monitoring and damage detection.

- A new state-of-the-art model for video motion magnification was designed and developed, improving output visual quality over prior art with better noise tolerance, less blur, and improved edge stability. This model sets a new benchmark for video motion magnification research.
- The newly developed state-of-the-art video motion magnification model has been open-sourced and made publicly available along with pre-trained weights.
- A software toolbox for vibration-based monitoring has been developed to provide a graphical user interface for the dynamic response measurement methodology described in this thesis. ViMag improves the user experience for taking measurements, automating tedious manual work that previously required the use of multiple command-line tools. The simplification of the process not only reduces the barrier of entry to the technology but also allows for obtaining better results in less time, easing experimentation and discovery of feature-rich areas of the subject video sequence.
- The newly developed video motion magnification model and tools were validated for use in structural health monitoring applications, demonstrating their capability for contactless full-field damage detection. This validation provides strong evidence for the practical applicability of the developed methodology and tools.

6.3 Future research perspectives

Based on the aforementioned conclusions, several promising avenues for future research could be taken:

- Testing the developed methodology and tools on in-service structures to further validate their practical applicability and effectiveness in real-world scenarios.
- Conducting combined studies on the use of video motion magnification and other damage indicators or evaluation methods to investigate the potential synergies and benefits of combining these approaches.
- Investigating the potential applications of the developed methodology and tools in condition monitoring, comparing their performance against existing commercial solutions.
- Exploring the use of event-based cameras [235] for video motion magnification to potentially improve the performance and capabilities of the methodology presented in this thesis. Event-based cameras are bio-inspired sensors that asynchronously measure per-pixel brightness changes instead of capturing images at a fixed rate like conventional cameras. This operating principle allows event-based cameras to achieve very high

temporal resolution, high dynamic range, and reduced motion blur, all of them being critical factors for VSHM.

- Continuing to improve the video motion magnification model to further enhance its performance and capabilities by exploring newer image processing techniques in related fields of computer vision or studying the effects of different procedural techniques for generating a training dataset.

- Reducing the computational cost of the STB-VMM model to enable its use in real-time applications and improve its usability. Studying different model configurations and the effects of the model meta-parameters on output quality could optimize the computational cost regarding the resulting image quality. Additionally, substantial performance gains could be achieved over the current model by migrating to a newer version of Pytorch or porting the model to C++.

7

Appended publications

Publication I:

Learning-based video motion magnification approach for vibration-based damage detection

10.1016/j.measurement.2022.112218

Measurement 206 (2023) 112218

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Learning-based video motion magnification approach for vibration-based damage detection

Ricard Lado-Roigé, Josep Font-Moré, Marco A. Pérez *

IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain

ARTICLE INFO

Keywords:
Computer vision
Convolutional Neural Network
Damage identification
Deep learning
Frequency response functions
Natural frequency
Operating deflection shape
Structural assessment
Structural health monitoring
Vibration testing

ABSTRACT

This paper explores the viability of using learning-based state-of-the-art video motion magnification to extract vibrational signatures for damage detection in structures. Unlike previous research, the proposed model uses learning-based video motion magnification rather than implementing hand-designed filters. This change allows the presented approach to detect more subtle sub-pixel movement and thus allows for greater sensibility to vibration. This novel approach is validated and verified on a laboratory structural benchmark under different damage scenarios. Although the learning-based model was trained on a synthetic and non-related image dataset, the experimental results prove that the system is suitable for identifying natural frequencies and operating deflection shapes, thus enabling damage detection algorithms to identify structural damage reliably. The results demonstrate the feasibility and suitability of this novel monitoring technique and thus open an avenue for further research regarding deep learning and its applications to structural health monitoring.

1. Introduction

Structural health monitoring (SHM) is the set of analysis and assessment techniques applied to autonomously determine the integrity and durability of engineering structures. This set of processes aims to track the operational status, assess the condition, and alert to the presence of damage [1,2]. Generally understood as damage are any undesired changes in the geometric or material properties that can affect a structure's overall performance, safety, reliability, and operational life [3,4]. According to this definition, damage does not necessarily imply the complete failure of a system but a comparative deterioration causing suboptimal performance.

The trend towards digitalization has fostered the interest in extensive monitoring of civil structures to ensure the safety and reliability of critical infrastructure. In this regard, SHM systems provide information about structural performance under different operating conditions, thus enabling maintenance operations optimization and the reduction of costs by anticipating structural issues before severe damages occur.

The most basic approach for damage diagnosis of civil structures is based on visual inspection. However, this approach presents several issues in practice. First, the size of civil structures is relatively large, making the inspection process laborious and time-consuming. Second, since this inspection method relies on human judgment, it requires skilled and highly-trained operators. Moreover, and third, it is not always possible to visually inspect load-bearing elements since they might be difficult to access or obstructed by non-structural members or coverings.

In contrast, more sophisticated approaches, such as vibration-based methods, provide a systematic, feasible, and consistent way of identifying, locating, and quantifying the severity of damage based on the vibration response of the monitored structure [5–7]. To this end, sensors are deployed to feed algorithms with structural responses [8] to determine if any damage or change has occurred. Accelerometers are the most commonly used type of sensor, however their installation, calibration, and maintenance makes them unsuitable for most applications, especially in those environments that may require resistance to harsh climatic conditions. Moreover, wired sensors might be inconvenient to run over long distances and suffer from signal degradation, while their wireless counterparts have only recently become viable but still need development on their power consumption. These shortcomings have spurred interest in non-contact measurement methods [9] such as approaches based on vision.

Vision-based methods infer the displacement of structures by tracking the motion of pixels or features in a sequence of images [10–13]. These techniques generally allow for spatially denser measurements using fewer sensors, however this benefit comes at the cost of less precise data compared to contact techniques. Digital image correlation or optical flow are two such techniques currently used in other fields with very positive results. Vision-based SHM has successfully been tested both in laboratory experiments as well as in civil infrastructure [14–18].

* Corresponding author.
E-mail address: marcoantonio.perez@iqs.url.edu (M.A. Pérez).

https://doi.org/10.1016/j.measurement.2022.112218
Received 20 September 2022; Received in revised form 10 November 2022; Accepted 15 November 2022
Available online 24 November 2022

Recent developments in computer vision have brought a new technique called video motion magnification which is capable of identifying and magnifying almost eye-imperceptible movements in video data [19–24]. This set of techniques has proved effective in applications, such as producing visual representations of an object's operating deflection shapes [25] or recovering sound from a room behind a soundproof glass [26]. Hence, interesting new avenues for research open up combining video motion magnification (VMM) and SHM to identify operating deflection shapes (ODS) and resonant frequencies of large civil structures [27–38]. However, these techniques produce higher signal noise floors compared to contact accelerometers by up to several orders of magnitude. Consequently, further research is needed for a more in-depth understanding of these techniques, expanding the range of use cases and improving its cutting-edge approach.

Based on the above, this work explores the viability of using machine learning-based approaches combined with video motion magnification (VMM) techniques to autonomously extract vibrational signatures with the final aim of detecting damage in complex structures. Said learning-based approach to VMM focuses on learning the filters from the data instead of relying on the manually tuned filters of conventional VMM approaches. Recently, learning-based methods have achieved excellent results compared to other VMM approaches yielding higher-quality magnification and noise characteristics, with fewer artifacts and blurring [39]. Thus, its implementation into a strategy for estimating vibrational response might bring better quality measurements, leading to a lower damage threshold detection.

This paper validates the use of a novel learning-based video motion magnification approach and verifies its application on a structural benchmark under different damage scenarios. The following section describes the deep learning model implemented for damage detection. Section 3 describes the test setup details, the test procedure, and the reproduced damage scenarios. The fourth section presents the results of the verification tests conducted by comparing against high-resolution contact accelerometers. Section 4 also includes a parametric analysis to evaluate the technique's resolution as well as an assessment of the performance of the proposed approach for the detection of damage on a benchmark structure. Finally, the conclusions of the study are summarized in Section 5.

2. Learning-based approach for video motion magnification

In mathematical terms, given an image $I(x,t) = f(x+\delta(x,t))$, where $\delta(x,t)$ represents the motion field as a function of position and time, the goal of motion magnification is to amplify pixel movement so that the magnified image $\hat{I}$ becomes $\hat{I}(x,t) = f(x+(1+\alpha)*\delta(x,t))$ where α is the magnification factor [21,23]. To archive this goal, motion magnification techniques decompose video sequences into motion representations by using complex filters, then these motion representations are selectively magnified and reassembled into magnified frames.

Motion magnification techniques can be approached from either a Lagrangian or an Eulerian point of view. Lagrangian approaches explicitly extract the motion field by tracking pixel movement directly, using a set of techniques often referred as optical flow. On the other hand, the Eulerian approach decomposes a video sequence into representations that allow for motion manipulation without the need for explicit tracking. The distinction between Lagrangian and Eulerian is not dissimilar from the same terms used in fluid dynamics, where Lagrangian methods track a volume of fluid through a flow while Eulerian-based approaches study the evolution of flow in a fixed volume.

Eulerian motion magnification approaches consist of three stages: decomposition, manipulation, and reconstruction. The first stage, decomposition, yields multiple features or motion representations of a video sequence which are then selectively manipulated in the second stage to archive motion magnification. Proper decomposition is key to obtaining good results as filtering out noise at this first stage avoids its magnification in the subsequent manipulation stage. The third and final stage must reconstruct the sequence from the manipulated representations generated by the first and second stages. While Eulerian techniques are suitable for revealing subtle motion, they are prone to noise and excessive blurring due to an imperfect filter design. Learning-based approaches aim to mitigate this drawback from Eulerian classical approaches by learning the filters instead of hand-tuning them using a deep convolutional neural network.

The learning-based model for video motion magnification (LB-VMM) implemented in this work consists of three main parts, as depicted in Fig. 1: encoder (G_E), manipulator (G_M), and decoder (G_D). In Fig. 1 the nomenclature used to represent layers works as follows: Conv c denotes a convolutional layer of c output channels, with a k×k kernel size, and stride s; following that an activation function may be attached to the layer, indicating its application at the end of the layer computation. For example, *Conv_16_k7s1_ReLU* means: convolutional layer with 16 output channels with a 7 × 7 kernel size, stride of 1 and a ReLU activation function. This model has been adopted from Oh et al. proposed architecture [39], and has been implemented in Pytorch [40] open source machine learning framework. The goal of this neural network architecture is to extract a motion representation (M), which can be manipulated by a simple multiplication and then reconstructed into a magnified displacement frame ($\hat{I}$).

The encoder block (G_E) acts as a spatial decomposition filter that extracts a motion representation for two input frames (A and B).

$$G_E(A) = [M_A, V] \tag{1}$$

$$G_E(B) = [M_B] \tag{2}$$

Then, the manipulator block multiplies the difference of these representations, which is then passed on to the decoder that reconstructs the motion representation into a new motion magnified frame, aided by an unmanipulated visual representation (V) previously yielded by the encoder. These three elements are constituted by fully convolutional layers, allowing for flexible learnable filters at multiple input/output resolutions.

$$G_M(M_A, M_B, \alpha) = M_A + h(\alpha \cdot g(M_B - M_A)) \tag{3}$$

$$G_M(M_A, M_B, \alpha) = M \tag{4}$$

Where $g(\cdot)$ is represented by a 3 × 3 convolution followed by ReLU, and $h(\cdot)$ is a 3 × 3 convolution followed by a 3 × 3 residual block.

$$G_D(M, V) = \hat{I} \tag{5}$$

To train this deep learning model a fully synthetic dataset [39] was used, since it is impossible to obtain real motion-magnified video pairs. It is important to stress the relevance of representative high-quality training data to a model's performance, in fact, machine learning models, in general, tend to be as good as their training data. The synthetic dataset was built by moving segmented objects from the PASCAL VOC [41] dataset over background images taken from the MS COCO [42] dataset, carefully considering texture, magnification factor, and subpixel motion generation. Even though the contents of the images are entirely unrelated to the final application, once trained, the motion magnification model yields excellent performance, comparable or superior to previously existing methods.

As stated in [39], the model is trained end-to-end using l_1-*loss* between the network output $\hat{I}$ and the ground-truth magnified frame I. Additionally, regularization losses are imposed between V_B and V, V_A and V_B, and M'_B and M_B, where M'_B is the result of the additional intensity-perturbed frame variants provided by the dataset. Then the model is optimized using ADAM with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a batch size of 10, and set the learning rate to 10^{-5}. The learning-based model is trained with 100,000 training examples with different α values contained between 0 and 100 where $\alpha \in \mathbb{R}$. Consequently, after training the model, it is possible to run inference on a sequence using

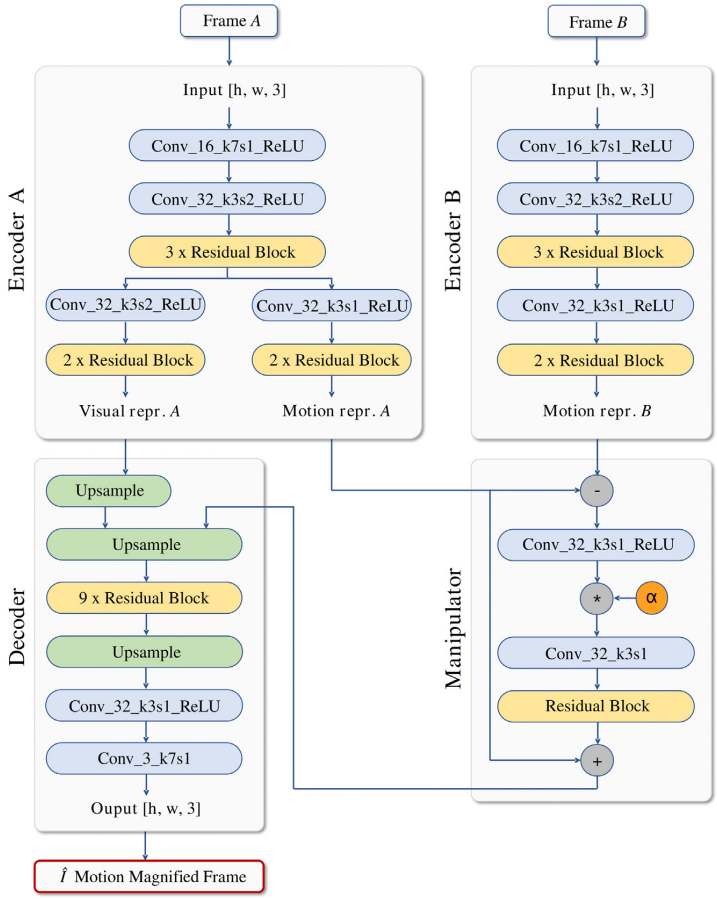


Fig. 1. Machine learning-based video motion magnification (LB-VMM) architecture implemented, adapted from [39]. α stands for magnification factor, Conv c denotes a convolutional layer of c channels, k×k kernel size, and stride s. A residual block is composed of a Conv_32_k3_s1_ReLU + Conv_32_k3_s1 in parallel with its input. An upsampling block upsamples by ×2 using 2D nearest neighbor.

an arbitrary magnification factor within the 0 to 100 range without any further training.

By applying LB-VMM it is possible to obtain the time response of a subject structure inferring from a sequence of motion magnified frames. Following magnification, a vibration response signature can be obtained by applying a Fast Fourier Transform (FFT) to the pixel displacement data. Fig. 2 shows an example of a temporal signal reconstructed using two different video magnification approaches: the phase-based Riesz pyramids [22] method on the top frame and the learning-based approach implemented in this work on the bottom. The example video was obtained from the benchmark test structure, as presented in the following section. The time sequences correspond to the magnification of the slice, marked in red on the left-hand side, located on the lower level of the test structure. Figs. 2 and 3 illustrate how the LB-VMM approach yields less noisy outputs with fewer blurring and ringing artifacts, thus resulting in higher-quality magnification compared to the phase-based methods. While no significant differences are observed in the frequencies detected, the learning-based approach has a lower noise floor thus enabling the detection of a third peak. This extra sensibility may be useful in cases where image resolution is low or movement amplitude is small, enabling the learning-based vision technique to detect damage where the phase-based techniques could not.

3. Methodology

3.1. Experimental setup

The validation tests have been conducted on a scaled-down structure under different damage scenarios using the three-story building benchmark, as seen in Fig. 4, proposed by Los Alamos National Laboratory (LANL) [43]. The structure is made of aluminum columns (177 × 25 × 6 mm) and plates (305 × 305 × 25 mm) attached using bolted joints forming a four degree-of-freedom system. The structure is connected to an electrodynamic shaker[1] that provides lateral excitation to the center of the base floor, which is mounted over linear rails allowing for movement on one axis. Both the building and shaker are mounted together on a rigid steel frame. For comparison and validation purposes, the structure is instrumented with four piezoelectric accelerometers[2] placed centered on each floor and opposed to the shaker's

[1] Brüel & Kjær electrodynamic shaker model 4824.
[2] Brüel & Kjær accelerometers type 4519-003.

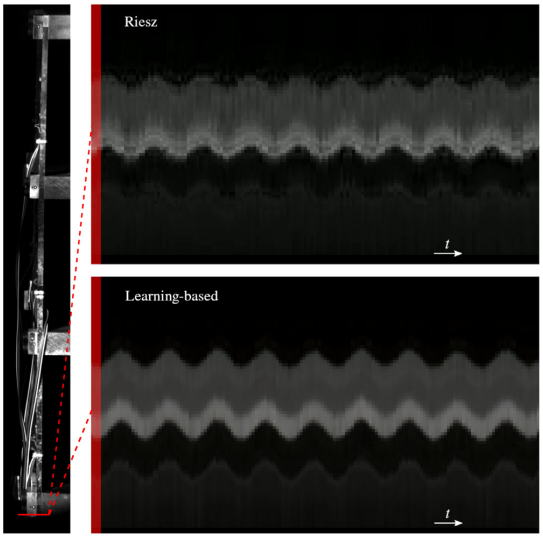


Fig. 2. Temporal signal (200 ms) reconstructed using two different video magnification approaches: the Riesz [22] pyramids and the learning-based approach implemented in this work.

coupling point. The accelerometers and shaker then are connected to a Brüel & Kjær data acquisition module.[3] Induced shaker excitation was set over the range of 0 to 100 Hz. And a monochrome high-speed camera[4] was placed 2 meters in front of the structure to record the vibration tests. The video recording sampling rate was set to 200 fps to avoid aliased data, with a resolution of 192 by 1358 px. The camera is set up to capture a corner of the building where each of the three floors are clearly visible (see Fig. 2), additionally LED light was used to improve image contrast.

3.2. Data acquisition

To obtain the vibration response, the structure was excited using an electrodynamic shaker producing different excitation types. Concurrently a video sequence of 5 s was captured using a high-speed camera. Then, the resulting video sequence was run through the trained LB-VMM model, yielding a sequence of motion magnified frames. The magnification factor α is chosen high enough to clearly display the target's movement while avoiding distorting the video sequence with excessive magnification.

Following motion magnification, a manually selected single-pixel wide linear region of the captured video was selected for tracking displacement over time (see Fig. 2). The criteria imposed for selecting this regions was to pick a high contrast region on every floor of the building, preferably close to the accelerometers. However, this manual selection process could be automated, for example: by placing identifiable markers on the structure or using other vision techniques to automatically select objects or regions of interest. Once selected, the linear regions were saved for every magnified frame, forming an image consisting of columns of pixels placed in chronological order. The resulting image is a temporal slice that condenses the dynamic response of a physical point over time. It is important to select high-contrast areas so that the displacement is clearly visible. A color

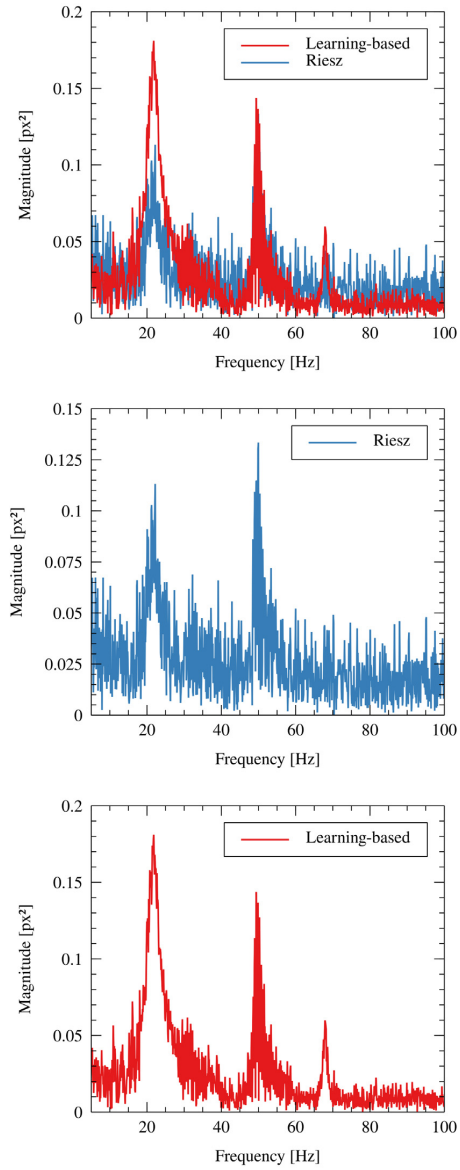


Fig. 3. Frequency-domain data comparing the phase-based [22] and the learning-based approaches. The frequencies detected in both cases are very similar, however, the learning-based method yields a clearer signal.

threshold was then applied to the upper and lower sections of the time slice to find the boundary between the measured object and the background (see Fig. 5). Optionally, depending on the input data, it is often helpful to apply a binary threshold to the time slice with the intent to eliminate low contrast areas. Once the boundaries were identified, the pixel values were used to average both boundaries and obtain the displacements over time in pixels. Finally, the frequency-domain signal was obtained using the Fast Fourier Transform (FFT) on the pixel displacement data.

[3] DAQ Brüel & Kjær Lan-XI DAQ 3160-A-4/2 and 3053-A-12/0.

[4] iX Cameras i-Speed 220 equipped with a 12.5–75 mm f/1.8 2/3" format lens.

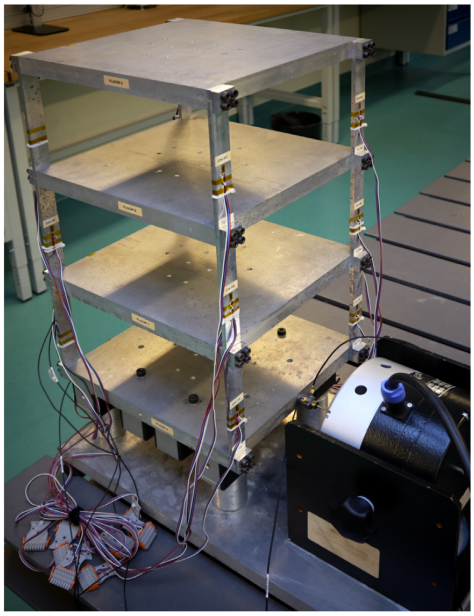


Fig. 4. Three-story building benchmark setup for validation test.

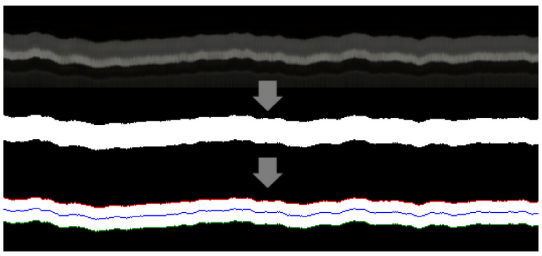


Fig. 5. Example of a temporal response (665 ms) inferred through the trained LB-VMM model from a random excitation. The frequency-domain data of this time sequence is depicted in Fig. 8. This figure illustrates the process of obtaining numerical data from an image resulting of a slice.

Fig. 5 shows an example of a temporal response inferred through the trained LB-VMM model from a random excitation. The uppermost image is the original magnified sequence, the image below corresponds to the binary thresholded sequence, and finally the lower one shows the boundary identification with their average. The frequency-domain data of this time sequence is depicted in Fig. 8.

3.3. Validation tests

To assess the validity of the LB-VMM approach on SHM applications, the following tests were devised to address three main issues: single frequency detection performance, multi-frequency detection performance, and damage detection. Tests were performed with minimal excitation amplitude so that the test structure was not visibly moving or vibrating without the help of motion magnification. In addition, all tests were performed and simultaneously measured with both contact accelerometers and LB-VMM so that results could be reliably compared.

Testing for single-frequency performance reveals the model's capacity to detect a singular frequency vibration acting upon a structure

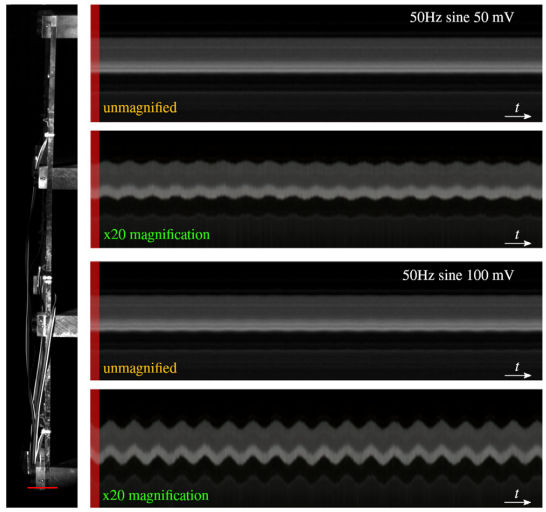


Fig. 6. Comparison between magnified (α=20) and unmagnified slices at two different excitation amplitudes.

precisely. The structure was excited using a sine wave at 20, 50, and 70 Hz individually, which roughly coincided with the structure's natural frequencies. The goal was to detect said frequencies and check that they matched the input excitation frequencies.

The multi-frequency tests assess the capacity of the system to work in slightly more realistic conditions. By exciting the structure using random noise or a periodic chirp, i.e., a signal that is continuously swept through a defined frequency range, the proposed motion magnification system should identify the natural frequencies of the structure. The visually obtained frequency response is expected to have a much higher noise floor compared to the piezoelectric accelerometers but still provides helpful information regarding the frequency response. These tests will focus on the range of frequencies between 0 and 100 Hz.

Finally, to test the damage detection capability of the proposed method, six possible scenarios were assessed under a 100 mV random noise excitation: an undamaged state, a 1.2 kg load fixed on the top floor, and four column damage scenarios. The column damage scenarios consist of a longitudinal reduction of a single column's flat section by 10%, 20%, and 50%, which in terms of column stiffness translates to approximately 20%, 50% and 85% stiffness reductions. Lastly, the fourth scenario completely removes the column from the structure. These changes were expected to impact the structure's stiffness significantly, thus altering its natural frequencies.

4. Results and discussion

4.1. Single frequency performance

Fig. 6 compares the motion magnification results between a 50 mV and a 100 mV sinusoidal monotone signal acting upon the structure. On the left-hand side of the figure, the left column of the benchmark structure is displayed at its full height. Marked in red, a line on the bottom floor highlights the region selected for tracking displacement over time. This linear region of size 1×80 px is vertically displayed for every frame of the video on the right-hand side of the figure, forming a row of pixels placed in chronological order, which represents the temporal response of the marked area.

Note that the unmagnified video shows imperceptibly small, seemingly static movements, and only careful examination of the temporal

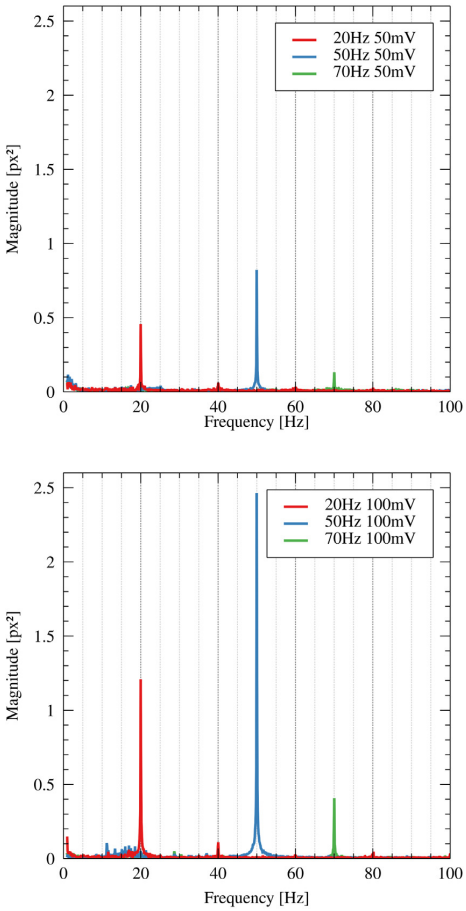


Fig. 7. Frequency-domain data obtained from the LB-VMM approach (α=20) for 20, 50 and 70 Hz sine waves excitation at 50 mV (top) and 100 mV (bottom) amplitudes.

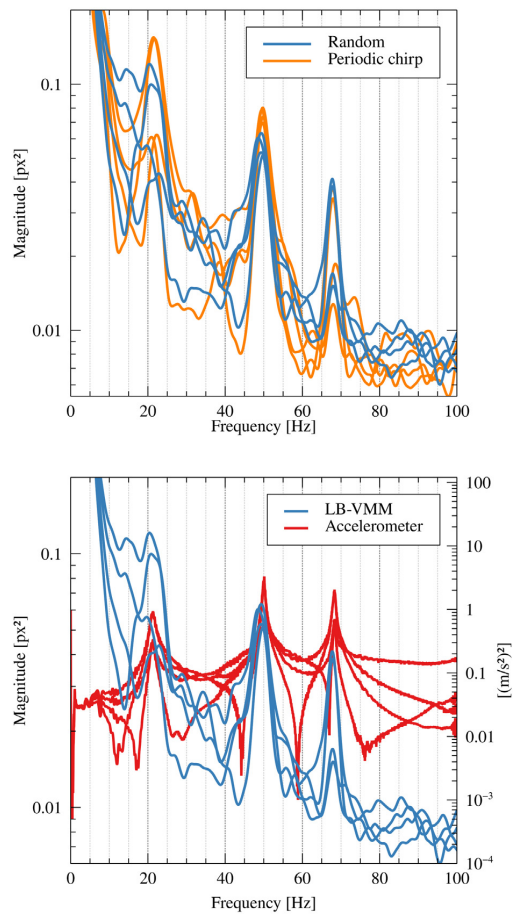


Fig. 8. Comparison of multi-frequency performance test results inferred from the LB-VMM approach (α=20, 100 mV) between the type of excitation (top) and acquisition sensor (bottom).

slices reveals pixel or subpixel shifts indicating movement. However, the magnified results display clearly visible movement allowing the identification of the waveforms. In addition, the effect of the amplitude of the excitation signal is also evident.

Fig. 7 depicts the frequency-domain data results obtained from the LB-VMM approach for both of the chosen excitation amplitudes. As shown, both tests allow for the detection of single frequency signals acting upon the structure since the three estimated frequencies match the chosen input excitation frequencies of 20, 50, and 70 Hz.

4.2. Multi-frequency performance

The multi-frequency test assessment was conducted by exciting the benchmark structure with both random noise and periodic chirp signals ranging from 0 to 100 Hz. Fig. 8-top plots the LB-VMM results obtained in the frequency domain, showing a comparison between the types of excitation. Each curve corresponds to the results inferred from the tracking points located on each floor. Fig. 8-bottom, on the other hand, compares the results obtained by the LB-VMM approach against the four piezoelectric accelerometers.

As demonstrated, the LB-VMM approach can detect multiple frequencies at once with suitable accuracy. The response inferred matches

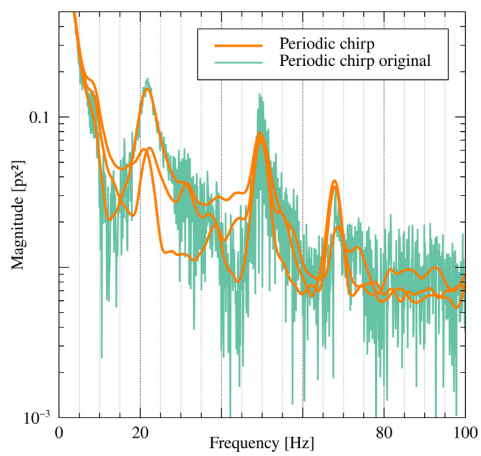


Fig. 9. Smooth of FFTs resulting from LB-VMM model, using a cubic spline approximation to filter out noise.

Table 1
Comparative mode frequencies obtained from each damage scenario with each sensing method. (See Fig. 12).

	Accelerometer (Hz)		LB-VMM (Hz)		Error (%)	
	1st order frequency	2nd order frequency	1st order frequency	2nd order frequency	1st order frequency	2nd order frequency
Pristine	21.125	50.000	20.859	49.945	1.26	0.11
Mass	20.500	49.250	20.272	49.259	1.11	0.02
20%	21.250	48.875	21.251	50.141	0.01	2.59
50%	21.125	49.000	20.859	49.064	1.26	0.13
85%	21.125	47.500	20.468	47.692	3.11	0.41
No column	19.500	47.125	19.586	47.007	0.44	0.25

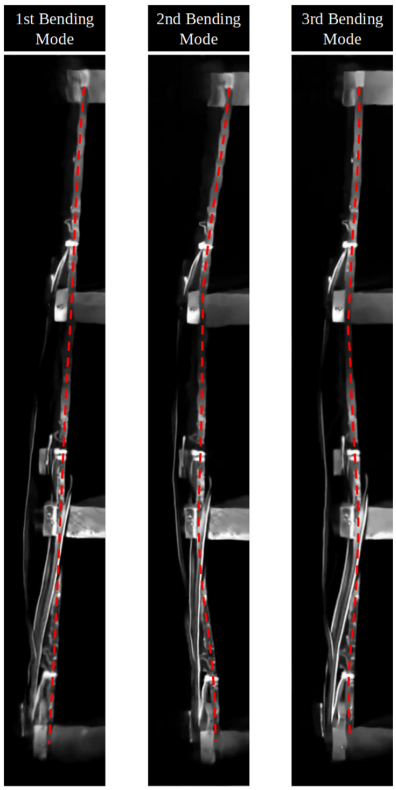


Fig. 10. ODS of the first three bending modes of the benchmark structure.

the accelerometer's data and displays the structure's natural frequencies when excited using either a periodic chirp signal or random noise. However, as anticipated, the LB-VMM approach suffers from a much higher noise floor than the piezoelectric accelerometer's output. Hence, to improve the graphs readability, the FFTs resulting from VMM have been smoothed using a cubic spline approximation to filter out noise (Fig. 9).

In addition, the ODS have also been determined by taking advantage of the magnification of the full-field displacements. Fig. 10 shows the modal shapes of the first three bending modes of the structure, whose frequencies correspond to the peaks shown in Fig. 8. Videos of the operating deflection shapes are attached as supplementary material alongside this article.

4.3. Damage detection

As stated before, to assess the damage detection capability of the LB-VMM approach, five damage scenarios were evaluated under a 100 mV random noise excitation. The first scenario, and the most severe one, consists of removing a column from the lowermost story of the building, on the side closest to the shaker and the camera. Then, the same column was subjected to three less severe damage scenarios by reducing its thickness by 10%, 20%, and 50%, thus reducing the column stiffness by approximately 20%, 50%, and 85%. Finally, the fifth scenario was conducted fixing a test mass of 1.2 kg on top of the uppermost floor of the building. These scenarios have been reproduced as detailed in the original Ref. [43].

Fig. 11 showcases the comparative results of each damage scenario presented. Overall, the results depict small but significant shifts in the building's natural frequencies, which decrease along with the damage severity, as expected. For the sake of simplicity, only an accelerometer located on the top floor is used to represent the frequency response function. It can be observed that the results between LB-VMM and the accelerometer present very similar shifts in frequency, however, on the LB-VMM spectrum the third mode is masked. This fact is attributed to two reasons: on the one hand to an inadequate selection of the tracking line placement coincident with the position of the accelerometer of the upper floor, placed in a region of low mobility of the operating deflection shape (see Fig. 10); on the other hand, due to the high noise floors offered by LB-VMM compared to the accelerometer. In any case, this result should not undermine the potential of the LB-VMM approach since the results of the previous section (see Fig. 8), in which different tracking lines are used, the third mode has been satisfactorily identified. In other words, to capture a higher number of modes, it is necessary to determine the response from different areas along the structure, analogous to the spatial distribution of accelerometers.

Fig. 11-top shows the response frequency difference between the undamaged structure and the same structure missing a column on the ground floor. The LB-VMM approach demonstrates its suitability to detect such damage, however, a complete column failure is a somewhat unrealistic scenario. Meanwhile, Fig. 11-middle depicts the comparison between the three column thickness reductions for both accelerometer and LB-VMM data. The graphic shows consistent results between both spectrums, thus confirming the capability for damage detection using LB-VMM techniques. It is also remarkable that the system proves capable of detecting modifications as small as a 10% difference in a single column's thickness. Finally, the third comparative scenario (Fig. 11-bottom) displays the difference between the undamaged unloaded structure and the same undamaged structure with a 1.2 kg mass bolted on the uppermost floor. Again, it can be observed that the changes to the natural frequencies are very small and near the limit of the LB-VMM setups resolution, however still present and verifiable against accelerometer data.

For a more detailed validation of the system's capacity, Table 1 includes the comparative numerical data of the frequency peaks identified in each damage scenario with each sensing method. The table

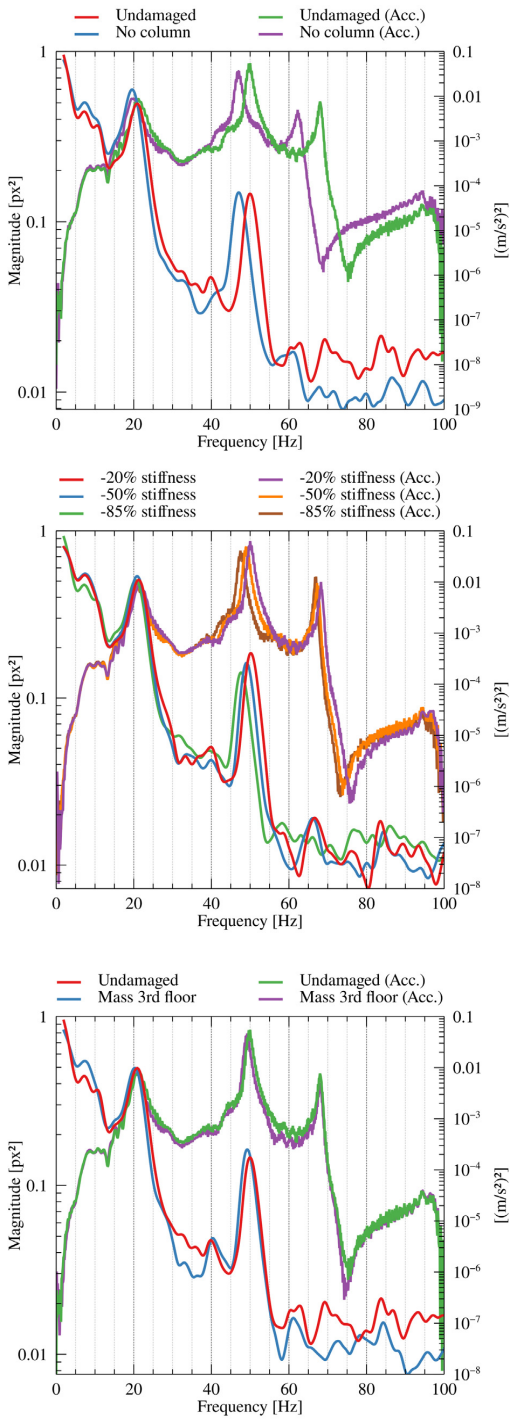


Fig. 11. Comparison between frequency-domain data from the LB-VMM model and piezoelectric accelerometers, showing the frequency response differences between the undamaged structure and different damage scenarios.

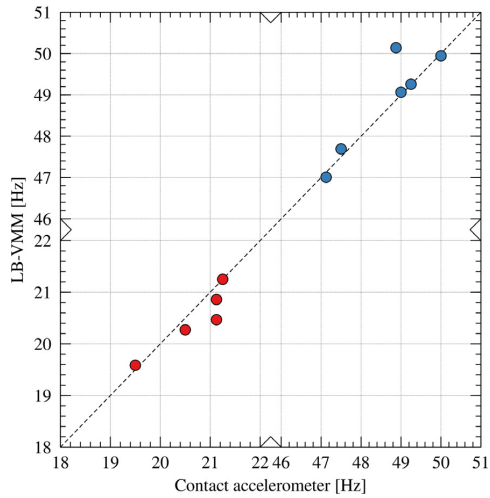


Fig. 12. Comparative mode frequency obtained from each damage scenario with each sensing method. See Table 1.

also includes the calculated percentage errors. These same results have been plotted in Fig. 12 for better illustration. Hence, the quantitative results are experimental evidence of the technique's potential, whose errors are within the resolution errors of the LB-VMM technique itself. Please note that the accelerometers are taken as ground truth, however, both methods are subject to some amount of error. Nevertheless, both measurement techniques display the same trend, therefore both reflecting the structural modifications introduced. To accurately assess the minimum detectable error, a more extensive statistical study should be performed.

5. Conclusions

This work validates the use of a deep learning-based video motion magnification (LB-VMM) approach and verifies its application on a laboratory structural benchmark under different damage scenarios. Although the model was trained on synthetic and non-related image data, the result of the experimental validation proves that the LB-VMM system is suitable for identifying the structure's natural frequencies and ODS, enabling algorithms to identify structural damage reliably. The proposed method yields a temporal response signal that can later be further processed to obtain parameters of interest for the application of more sophisticated strategies of damage detection based on the dynamic response [5,44]. Furthermore, these techniques could be compounded with environmental and operational factor mitigation [45–47] to decouple the effects of environmental conditions such as wind or temperature from the effects of structural damage.

In contrast to previous research, the implemented learning-based model gives a higher quality magnification than other VMM methods that use hand-designed filters instead of learning the filters directly from the data, thus obtaining the complete temporal vibration response all at once. Despite this fact, the LB-VMM shows no relevant drawbacks for the intended vibration testing applications. Therefore the preferred solution to obtain the individual frequency components of the complete temporal response is the application of a straightforward Fourier transform. However, LB-VMM is not incompatible with the use of temporal filters should they be required.

The use of motion magnification allows for greater sensibility when using vision-based vibration analysis. However, this increased sensibility comes with some limitations, such as the need for a stationary

camera and the need for consistent inter-frame lighting. Further shortcomings of the method include higher noise floors than contact or laser techniques, which, compounded with the fact that higher frequencies often display smaller amplitudes, might make the method unsuitable for some applications. Nevertheless, civil structures often display low frequencies due to their environmental excitation, making them suitable candidates for VMM techniques.

The methodology described in this work opens an avenue for research regarding structural monitoring since this technique offers a cost-effective tool for monitoring multiple structure points using just one camera. Further research will address adapting this deep learning-based technique to civil infrastructure monitoring and work to improve the processing time of the LB-VMM model for real-time applications.

CRediT authorship contribution statement

Ricard Lado-Roigé: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Software, Writing – original draft. **Josep Font-Moré:** Methodology, Validation, Investigation. **Marco A. Pérez:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to gratefully acknowledge the support and funding of the Catalan Agency for Management of University and Research Grants (AGAUR) through the project 2019 LLAV 00020 and Catalan Agency for Business Competitiveness (ACCIÓ) through the project INNOTEC ISAPREF 2021. Furthermore, the first and second authors would like to thank the Doctoral Scholarships from IQS. Also, the second author would like to thank the Doctoral Scholarships from Col·legi d'Enginyers Industrials de Catalunya - Fundació Caixa d'Enginyers.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.measurement.2022.112218.

References

[1] C.R. Farrar, K. Worden, An introduction to structural health monitoring, Phil. Trans. R. Soc. A. 365 (1851) (2007) 303–315, http://dx.doi.org/10.1098/rsta.2006.1928.

[2] W.-J. Yan, M.-Y. Zhao, Q. Sun, W.-X. Ren, Transmissibility-based system identification for structural health monitoring: Fundamentals, approaches, and applications, Mech. Syst. Signal Process. 117 (2019) 453–482, http://dx.doi.org/10.1016/j.ymssp.2018.06.053.

[3] D.M. Frangopol, J.P. Curley, Effects of damage and redundancy on structural reliability, J. Struct. Eng. 113 (7) (1987) 1533–1549, http://dx.doi.org/10.1061/(ASCE)0733-9445(1987)113:7(1533).

[4] E. Cosenza, G. Manfredi, Damage indices and damage measures, Prog. in Struct. Eng. Mater. S 2 (1) (2000) 50–59, http://dx.doi.org/10.1002/(SICI)1528-2716(200001/03)2:1<50::AID-PSE7>3.0.CO;2-S.

[5] M.A. Pérez, R. Serra-López, A frequency domain-based correlation approach for structural assessment and damage identification, Mech. Syst. Signal Process. 119 (2019) 432–456, http://dx.doi.org/10.1016/j.ymssp.2018.09.042.

[6] M.A. Pérez, J. Font-Moré, J. Fernández-Esmerats, Structural damage assessment in lattice towers based on a novel frequency domain-based correlation approach, Eng. Struct. 226 (2021) http://dx.doi.org/10.1016/j.engstruct.2020.111329.

[7] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, D.J. Inman, A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications, Mech. Syst. Signal Process. 147 (2021) 107077, http://dx.doi.org/10.1016/j.ymssp.2020.107077.

[8] R. Hou, Y. Xia, Review on the new development of vibration-based damage identification for civil engineering structures: 2010–2019, J. Sound Vib. 491 (2021) 115741, http://dx.doi.org/10.1016/j.jsv.2020.115741.

[9] C.-Z. Dong, F.N. Catbas, A review of computer vision–based structural health monitoring at local and global levels, Struct. Health Monit. 20 (2) (2021) 692–743, http://dx.doi.org/10.1177/1475921720935585.

[10] D. Feng, M.Q. Feng, Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection – A review, Eng. Struct. S 156 (2018) 105–117, http://dx.doi.org/10.1016/j.engstruct.2017.11.018.

[11] S. Patsias, W.J. Staszewskiy, Damage detection using optical measurements and wavelets, Struct. Health Monit. 1 (1) (2002) 5–22, http://dx.doi.org/10.1177/147592170200100102.

[12] A. Wahbeh, J. Caffrey, S. Masri, A vision-based approach for the direct measurement of displacements in vibrating systems, Smart Mater. Struct. 12 (5) (2003) 785–794, http://dx.doi.org/10.1088/0964-1726/12/5/016.

[13] J.J. Lee, M. Shinozuka, A vision-based system for remote sensing of bridge displacement, NDT E Int. 39 (5) (2006) 425–431, http://dx.doi.org/10.1016/j.ndteint.2005.12.003.

[14] J.G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W.T. Freeman, O. Buyukozturk, Modal identification of simple structures with high-speed video using motion magnification, J. Sound Vib. 345 (2015) 58–71, http://dx.doi.org/10.1016/j.jsv.2015.01.024.

[15] J.G. Chen, N. Wadhwa, F. Durand, W.T. Freeman, O. Buyukozturk, Developments with motion magnification for structural modal identification through camera video, in: J. Caicedo, S. Pakzad (Eds.), Dyn. Civil Struct. S (2), 2015, pp. 49–57, http://dx.doi.org/10.1007/978-3-319-15248-6_5.

[16] J.G. Chen, A. Davis, N. Wadhwa, Video camera-based vibration measurement for condition assessment of civil infrastructure, 2015, http://dspace.mit.edu/handle/1721.1/7582.

[17] M. Lu, Y. Chai, Q. Liu, Observation of tower vibration based on subtle motion magnification, IFAC-PapersOnLine 52 (24) (2019) 346–350, http://dx.doi.org/10.1016/j.ifacol.2019.12.434.

[18] J.G. Chen, A. Davis, N. Wadhwa, F. Durand, W.T. Freeman, O. Büyüköztürk, Video camera–based vibration measurement for civil infrastructure applications, J. Infrastruct. Syst. 23 (3) (2017) B4016013, http://dx.doi.org/10.1061/(ASCE)IS.1943-555X.0000348.

[19] C. Liu, A. Torralba, W.T. Freeman, F. Durand, E.H. Adelson, Motion magnification, in: ACM SIGGRAPH 2005 Papers, 2005, pp. 519–526, http://dx.doi.org/10.1145/1186822.1073223.

[20] Y. Zhang, S.L. Pintea, J.C. van Gemert, Video acceleration magnification, 2017, http://dx.doi.org/10.48550/arXiv.1704.04186, arXiv:1704.04186.

[21] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, Phase-based video motion processing, ACM Trans. Graph. 32 (4) (2013) 1–10, http://dx.doi.org/10.1145/2461912.2461966.

[22] N. Wadhwa, M. Rubinstein, F. Durand, W. Freeman, Riesz pyramids for fast phase-based video magnification, in: 2014 IEEE ICCP, 2014, pp. 1–10, http://dx.doi.org/10.1109/ICCPHOT.2014.6831820.

[23] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. (2012) http://dx.doi.org/10.1145/2185520.2185561.

[24] N. Wadhwa, H.-Y. Wu, A. Davis, M. Rubinstein, E. Shih, G.J. Mysore, J.G. Chen, O. Buyukozturk, J.V. Guttag, W.T. Freeman, F. Durand, Eulerian video magnification and analysis, Commun. ACM 60 (1) (2016) 87–95, http://dx.doi.org/10.1145/3015573.

[25] A. Davis, K.L. Bouman, J.G. Chen, M. Rubinstein, O. Büyüköztürk, F. Durand, W.T. Freeman, Visual vibrometry: Estimating material properties from small motions in video, IEEE PAMI 39 (4) (2017) 732–745, http://dx.doi.org/10.1109/TPAMI.2016.2622271.

[26] A. Davis, M. Rubinstein, N. Wadhwa, G.J. Mysore, F. Durand, W.T. Freeman, The visual microphone: Passive recovery of sound from video, ACM Trans. Graph. 33 (4) (2014) 79:1–79:10, http://dx.doi.org/10.1145/2601097.2601119.

[27] B.K. Oh, J.W. Hwang, Y. Kim, T. Cho, H.S. Park, Vision-based system identification technique for building structures using a motion capture system, J. Sound Vib. 356 (2015) 72–85, http://dx.doi.org/10.1016/j.jsv.2015.07.011.

[28] Y.-J. Cha, J. Chen, O. Büyüköztürk, Output-only computer vision based damage detection using phase-based optical flow and unscented Kalman filters, Eng. Struct. 132 (2017) 300–313, http://dx.doi.org/10.1016/j.engstruct.2016.11.038.

[29] A. Molina-Viedma, L. Felipe-Sesé, E. López-Alba, F. Díaz, High frequency mode shapes characterisation using Digital Image Correlation and phase-based motion magnification, Mech. Syst. Signal Process. 102 (2018) 245–261, http://dx.doi.org/10.1016/j.ymssp.2017.09.019.

[30] A. Molina-Viedma, L. Felipe-Sesé, E. López-Alba, F. Díaz, 3D mode shapes characterisation using phase-based motion magnification in large structures using stereoscopic DIC, Mech. Syst. Signal Process. 108 (2018) 140–155, http://dx.doi.org/10.1016/j.ymssp.2018.02.006.

[31] A. Sarrafi, Z. Mao, C. Niezrecki, P. Poozesh, Vibration-based damage detection in wind turbine blades using phase-based motion estimation and motion magnification, J. Sound Vib. 421 (2018) 300–318, http://dx.doi.org/10.1016/j.jsv.2018.01.050.

[32] K. Holak, A motion magnification application in video-based vibration measurement, in: Adv. Mech. Mach. Sci., 2019, pp. 4135–4144, http://dx.doi.org/10.1007/978-3-030-20131-9_412.

[33] M. Civera, L. Zanotti Fragonara, P. Antonaci, G. Anglani, C. Surace, An experimental validation of phase-based motion magnification for structures with developing cracks and time-varying configurations, Shock Vib. 2021 (2021) http://dx.doi.org/10.1155/2021/5518163.

[34] D.P. Rohe, P.L. Reu, Experimental modal analysis using phase quantities from phase-based motion processing and motion magnification, Exp. Tech. 45 (3) (2021) 297–312, http://dx.doi.org/10.1007/s40799-020-00392-7.

[35] M. Eitner, B. Miller, J. Sirohi, C. Tinney, Effect of broad-band phase-based motion magnification on modal parameter estimation, Mech. Syst. Signal Process. 146 (2021) 106995, http://dx.doi.org/10.1016/j.ymssp.2020.106995.

[36] F. Cosco, J. Cuenca, W. Desmet, K. Janssens, D. Mundo, Towards phase-based defect detection: A feasibility study in vibrating panels, J. Sound Vib. 537 (2022) 117196, http://dx.doi.org/10.1016/j.jsv.2022.117196.

[37] N.A. Valente, A. Sarrafi, Z. Mao, C. Niezrecki, Streamlined particle filtering of phase-based magnified videos for quantified operational deflection shapes, Mech. Syst. Signal Process. 177 (2022) 109233, http://dx.doi.org/10.1016/j.ymssp.2022.109233.

[38] L. Felipe-Sesé, A. Molina-Viedma, M. Pastor-Cintas, E. López-Alba, F. Díaz, Exploiting phase-based motion magnification for the measurement of subtle 3D deformation maps with FP + 2D-DIC, Meas. 195 (2022) 111122, http://dx.doi.org/10.1016/j.measurement.2022.111122.

[39] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W.T. Freeman, W. Matusik, Learning-based video motion magnification, 2018, http://dx.doi.org/10.48550/arXiv.1804.02684, arXiv:1804.02684.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. (2019) 8024–8035, http://dx.doi.org/10.5555/3454287.3455008.

[41] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, http://dx.doi.org/10.1007/s11263-009-0275-4.

[42] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common Objects in Context, 2015, http://dx.doi.org/10.48550/arXiv.1405.0312, arXiv:1405.0312.

[43] E. Figueiredo, G. Park, J. Figueiras, C. Farrar, K. Worden, Structural Health Monitoring Algorithm Comparisons Using Standard Data Sets, Tech. Rep. LA-14393, 961604, 2009, http://dx.doi.org/10.2172/961604.

[44] R. Sampaio, N. Maia, R. Almeida, A. Urgueira, A simple damage detection indicator using operational deflection shapes, Mech. Syst. Signal Process. 72–73 (2016) 629–641, http://dx.doi.org/10.1016/j.ymssp.2015.10.023.

[45] D. García Cava, L.D. Avendaño-Valencia, A. Movsessian, C. Roberts, D. Tcherniak, On explicit and implicit procedures to mitigate environmental and operational variabilities in data-driven structural health monitoring, in: A. Cury, D. Ribeiro, F. Ubertini, M.D. Todd (Eds.), Structural Health Monitoring Based on Data Science Techniques, Springer International Publishing, 2022, pp. 309–330, http://dx.doi.org/10.1007/978-3-030-81716-9_15.

[46] W.-H. Hu, A. Cunha, E. Caetano, R. Rohrmann, S. Said, J. Teng, Comparison of different statistical approaches for removing environmental/operational effects for massive data continuously collected from footbridges, Struct. Control Health Monit. 24 (8) (2017) e1955, http://dx.doi.org/10.1002/stc.1955.

[47] G. Oliveira, F. Magalhães, A. Cunha, E. Caetano, Vibration-based damage detection in a wind turbine using 1 year of data, Struct. Control Health Monit. 25 (11) (2018) e2238, http://dx.doi.org/10.1002/stc.2238.

Publication II:

STB-VMM: Swin Transformer based Video Motion Magnification

10.1016/j.knosys.2023.110493

Knowledge-Based Systems 269 (2023) 110493

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

STB-VMM: Swin Transformer based Video Motion Magnification

Ricard Lado-Roigé, Marco A. Pérez *

IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain

ARTICLE INFO

Article history:
Received 12 July 2022
Received in revised form 22 December 2022
Accepted 18 March 2023
Available online 27 March 2023

Dataset link: https://doi.org/10.17632/76s26nrcpv.1, https://github.com/RLado/STB-VMM

Keywords:
Computer vision
Deep learning
Swin Transformer
Motion magnification
Image quality assessment

ABSTRACT

The goal of video motion magnification techniques is to magnify small motions in a video to reveal previously invisible or unseen movement. Its uses extend from bio-medical applications and deepfake detection to structural modal analysis and predictive maintenance. However, discerning small motion from noise is a complex task, especially when attempting to magnify very subtle, often sub-pixel movement. As a result, motion magnification techniques generally suffer from noisy and blurry outputs. This work presents a new state-of-the-art model based on the Swin Transformer, which offers better tolerance to noisy inputs as well as higher-quality outputs that exhibit less noise, blurriness, and artifacts than prior-art. Improvements in output image quality will enable more precise measurements for any application reliant on magnified video sequences, and may enable further development of video motion magnification techniques in new technical fields.

1. Introduction

Video Motion Magnification (VMM) is a computer vision task consistent in magnifying small motions in a video sequence, having several uses in many fields from bio-medical applications [1–3] and deepfake detection [4] to structural modal analysis [5] and condition monitoring. These techniques act like a microscope for motion, to reveal previously invisible or unseen movements. Despite this simple premise, discerning small motions from noise is a complex task, especially when attempting to magnify very subtle, often sub-pixel movement. As a result, motion magnification techniques generally suffer from noisy and blurry outputs. Therefore, multiple authors have explored techniques to remediate these shortcomings and improve magnification quality and performance.

Early motion magnification algorithms, such as [6], used a Lagrangian approach, reliant on motion tracking or optical flow, to isolate motion prior to magnification. However, this approach is very computationally expensive and difficult to execute artifact-free, especially in regions affected by occlusion boundaries and complex motion. On the other hand, more modern techniques [7–11] have relied on Eulerian approaches, which observe the changes in a fixed region of pixels instead of tracking features in time and space. These Eulerian approaches are less computationally expensive, perform better with small motions, and generally yield better magnification results. Nevertheless, these approaches still display noticeable blurring and artifacting due to the complex challenge of designing filters for noise removal, which at the same time, do not interfere with motion magnification. For this reason, Oh et al. [12] proposed a novel learning-based approach to VMM. Learning-based motion magnification departs from the use of hand-designed filters in favor of learning those filters using Convolutional Neural Networks (CNN) instead. This method achieved higher-quality magnification yielding fewer ringing artifacts and showing better noise characteristics than previously published methods. However, its reliance on additional temporal filtering to improve image quality sometimes produces errors in magnification. While it is possible to obtain fairly clear results with no temporal filtering, the image quality generally improves when filtering is applied as it removes unwanted motion and noise before learning-based magnification.

The method presented in this work improves on the learnable filters and abandons temporal filtering to ensure correct magnification outputs. Resulting in a novel architecture capable of producing state-of-the-art results in terms of magnified image quality. The main contributions of this work are:

(a) A novel motion magnification architecture based on the SWIN transformer.
(b) A discussion, comparison, and validation of learning-based VMM techniques, both in a quantitative and qualitative sense.
(c) The proposed novel architecture outperforms relevant VMM techniques in both quantitative evaluation and

* Corresponding author.
E-mail address: marcoantonio.perez@iqs.url.edu (M.A. Pérez).

https://doi.org/10.1016/j.knosys.2023.110493
0950-7051/

Table 1
Motion magnification techniques summary table.
Source: Adapted from [12].

Method	Liu et al. [6]	Wu et al. [7]	Wadhwa et al. [8]	Wadhwa et al. [9]	Zhang et al. [11]	LB-VMM [12]	STB-VMM
Spatial decomposition	Tracking, optical flow	Laplacian pyramid	Steerable filters	Riesz pyramid	Steerable filters	Deep convolution layers	Swin transformer
Motion isolation	–	Temporal bandpass filter	Temporal bandpass filter	Temporal bandpass	Temporal bandpass filter (2_{nd} order derivative)	Subtraction or bandpass filter	Subtraction
Representation denoising	Expectation–Maximization	–	Amplitude weighted Gaussian filtering	Amplitude weighted Gaussian filtering	Amplitude weighted Gaussian filtering	Trainable convolution	Swin transformer

observed output quality, offering higher-quality magnification, less blurry frame reconstruction, better noise tolerance, and fewer artifacts than prior-art.

The following section summarizes previous influential works and their relation to the development of the presented model. Section three describes in detail the model's architecture and its training process. The fourth section presents results and comparisons of the model's performance, focusing on magnification and image quality with respect to prior work. Finally, the conclusions of this paper are summarized in section five.

2. Related work

2.1. Learning-based video motion magnification

Eulerian approaches to video motion magnification function by decomposing video sequences into motion representations that can later be manipulated mathematically and then reconstructed into magnified frames. On the other hand, Lagrangian approaches explicitly track a pixel or feature's movement throughout a video sequence. This distinction between Lagrangian and Eulerian approaches is not dissimilar to the same terms used in fluid dynamics, where Lagrangian methods [6] track a volume of fluid through the flow, while Eulerian approaches [7–9] study the evolution of flow in a fixed volume in space. Eulerian-based methods generally have the upper hand when processing small motion but produce blurry results when encountering large motion. The technique presented in this paper belongs to the Eulerian approach and is inspired by the work of Oh et al.'s learning-based video motion magnification [12].

Eulerian techniques generally consist of three stages: spatial decomposition, motion isolation and manipulation, and representation denoising. From this blueprint, different authors have proposed increasingly sophisticated techniques to improve magnification quality and performance as reflected in Table 1. In technical terms, the motion magnification problem can be summarized as follows. Given a signal $I(x,t)$ representing image intensity at position x and time t, and $\delta(t)$ representing translational motion in time such that

$$I(x,t) = f(x+\delta(t));\ I(x,0) = f(x) \tag{1}$$

The goal of motion magnification is to synthesize the signal

$$\hat{I}(x,t) = f(x+(1+\alpha)\cdot\delta(t)) \tag{2}$$

for some amplification factor α. In practice, only certain frequencies of motion $\delta(t)$ are useful to motion magnification, so a selector $T(\cdot)$ is applied to $\delta(t)$, which is typically a temporal bandpass filter.

Prior to learning-based VMM (LB-VMM), magnification techniques relied on multi-frame temporal filtering to isolate motions of interest from random noise [7–9,11,13]. By contrast, the learning-based approach [12] directly employs CNNs to both filter noise and extract features, achieving comparable or better quality than prior-art without using temporal filtering. The LB-VMM model is composed of three stages: encoder, manipulator, and decoder. Said model is designed to accept two frames and return a single motion-magnified frame. The goal of the encoder is to extract relevant features from each of the two input frames and yield a visual and a motion representation. The motion representation of both input frames is then passed to the manipulator, which will subtract both representations and magnify the result by an arbitrary parameter α defined by the user. Finally, the results of the manipulator and the previously-obtained visual representation enter the decoder, where the motion and visual components are reconstructed into a motion-magnified frame. These three CNN-based components allow for flexible learnable filters that are better suited to the task of motion magnification and thus yield better quality magnification results.

To train the model and given the impossibility of obtaining motion magnified video pairs, Oh et al. generated and used a fully-synthetic dataset for training their model, built by moving segmented objects from the PASCAL VOC [14] dataset over background images taken from MS COCO [15]. Careful consideration to the generation of the dataset was paid to ensure accurate pixel and sub-pixel motion as well as learnability. The dataset learning examples are parametrized to make sure they are within a defined range. Specifically, the dataset's magnification is upper-limited to an α magnification factor of 100, and input motion is sampled so that magnified motion does not exceed 30 pixels.

2.2. Transformers as a Computer Vision tool

CNNs have been a staple of the Computer Vision (CV) field in the last few years, with many of the top-performing models having made extensive use of them [16–18]. This period roughly started after Krizhevsky et al. [17] won the ImageNet Large Scale Visual Recognition Challenge [19,20] (ILSVRC) on September 30th 2012, and spurred many publications employing CNNs and GPUs to accelerate deep learning. Through the use of filters, these networks generate feature maps that summarize an image's most relevant parts. These filters capture relevant local information by the very nature of the convolution operation, which, combined with multi-scale architectures [21,22] result in rich feature maps that can efficiently obtain a representation of an image's content, both in a local and global context. Recently, the CV field has been revolutionized yet again by the Vision Transformer (ViT) [23], which, employing the attention mechanism has demonstrated state-of-the-art performance in many CV tasks. The attention mechanism was first popularized in the field of Natural Language Processing (NLP) by Vaswani et al. [24], where the transformer architecture has become the de-facto standard.

The attention mechanism can be described as mapping from a query and a set of key–value pairs into an output. The output, represented in vector format, is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function taking into account the query and the corresponding key [24]. The transformer was the first model which

exclusively relied on self-attention to compute representations of its input and output without using sequence-aligned recursive neural networks or convolution operations. Unlike CNNs, transformers lack translation invariance and a locally-restricted receptive field, in its place transformers offer permutation invariance. Said feature enabled NLP models to infer relations between words and ideas much further into a text than previous recurrent models could. However, CV applications require the processing of grid-structured data which cannot trivially be processed by a transformer. The ViT [23] overcame this burden by mapping grid-structured data into sequential data by splitting the image into patches. Patches are then flattened into vectors and embedded into a lower dimension. These flattened patches are then summed with positional embeddings and fed as a sequence to a standard transformer encoder. Image patches essentially become sequence tokens just like words are when working in NLP, in fact, ViT uses the exact same encoder described in [24].

Later, Microsoft researchers improved on the ViT publishing the SWIN transformer, a hierarchical vision transformer using shifted windows [25]. This work further refined the solution to adapt the original transformer from language to vision. The SWIN transformer solved issues caused by large discrepancies in the scale of visual entities at the same time that limited self-attention computation to non-overlapping local windows, yet still allowing for cross-window interaction. The introduced limitation on the scope of self-attention significantly reduced the computational complexity, which scales quadratically with respect to image size, allowing for the processing of higher-resolution images that were previously unmanageable. Further developments in the CV field have implemented the SWIN transformer for various tasks achieving state-of-the-art performance [26–28].

2.3. Swinir image restoration

Inspired by the recent prominence of the transformer and its success in many CV problems such as image classification [23,25,29–33], object detection [34–36], segmentation [30,37,38], crowd counting [39,40] and image restoration [41–43], Liang et al. [27] proposed a new state-of-the-art image restoration model based on the Swin transformer [25]. The SwinIR model consists yet again of three modules: a shallow feature extractor, a transformer-based deep feature extractor and a high-quality image reconstruction module. This structure offers excellent performance in various image restoration tasks such as image super-resolution, JPEG compression artifact reduction, and image denoising. These applications are very interesting when working with VMM, as current state-of-the-art methods can be negatively affected by noisy input images, causing much noisier and blurrier results, especially at large magnification rates. This occurs as a result of noise not being properly filtered beforehand, therefore as the motion gets magnified, the noise gets magnified as well.

3. Methodology

3.1. Residual swin transformer block

The Residual Swin Transformer Block (RSTB) [27] is used as one of the fundamental building blocks of the proposed architecture, appearing in parts of both the feature extractor and the reconstructor. The RSTB is a residual block combining multiple Swin Transformer Layers (STL) [25] and convolutional layers, compounding the benefits of the spatially invariant filters of the convolutional layers with the residual connections that allow for multilevel feature processing.

The Swin transformer layer shown in Fig. 2 partitions an $H \times W \times C$ image into non-overlapping $\frac{HW}{M^2}$ local windows using an MxM sliding window and then computing its local attention, effectively reshaping the input image into $\frac{HW}{M^2} \times M^2 \times C$. The main difference with respect to the original transformer layer [24] lies in the local attention and the shifted window mechanism. For a local window feature $F \in \mathbb{R}^{M^2 \times C}$, the query, key, and value matrices Q, K, and $V \in \mathbb{R}^{M^2 \times d}$ are computed as

$$Q = FW_Q; \quad K = FW_k; \quad V = FW_V \tag{3}$$

where W_Q, W_K, and W_V are the learnable parameters shared across different windows, and d is the dimension of Q, K, and V. Therefore, the attention matrix is computed for each window as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}} + P)V \tag{4}$$

where P is the learnable relative positional encoding. Computing the attention mechanism multiple times yields the results of the Multi-head Self Attention (MSA), which are then passed on to a Multi-Layer Perceptron (MLP). Therefore, the whole STL process can be summed up like so

$$F = MSA(LayerNorm(F)) + F \tag{5}$$

then

$$F = MLP(LayerNorm(F)) + F \tag{6}$$

where the MLP is formed by two fully-connected layers with a GELU activation layer in between.

3.2. Network architecture

The proposed model architecture, shown in Fig. 1, consists of three main functional blocks: the feature extractor, the manipulator, and the reconstructor. The feature extractor is further subdivided into the shallow and deep feature extractors, and their job is to extract a high-quality representation of an input frame. Next, the manipulator, using the features from two frames, magnifies the motion by multiplying the difference between the two feature spaces by a user-selected magnification factor α. Finally, the reconstructor converts the resulting manipulated feature space back into a magnified frame.

Given two frames of a target sequence $[I_A, I_B] \in \mathbb{R}^{H \times W \times C_{in}}$ (where H is the height of the image, W is the width of the image and C_{in} represents the number of input channels) the convolutional shallow feature extractor (G_{SF}) maps high-level features into a higher dimensional feature space, thus providing early local feature extraction (F_{AS}, F_{BS}) and leading to a more stable optimization and better results [44].

$$[F_{AS}, F_{BS}] = G_{SF}([I_A, I_B]) \tag{7}$$

Then, the features extracted in the previous step are further processed in the deep feature extraction module (G_{DF}), which consists of N Residual Swin Transformer Blocks (RSTB).

$$[F_{AD}, F_{BD}] = G_{DF}([F_{AS}, F_{BS}]) \tag{8}$$

After feature extraction, both frames' feature spaces are then sent to the manipulator [12] (G_M), which works by taking the difference of both frames' feature spaces and directly multiplying by a magnification factor α.

$$\begin{aligned} G_M(F_{AS} + F_{AD}, F_{BS} + F_{BD}) &= (F_{AS} + F_{AD}) \\ &+ h(\alpha \cdot t(((F_{BS} + F_{BD}) - (F_{AS} + F_{AD})))) \end{aligned} \tag{9}$$

where $t(\cdot)$ is a 3×3 convolution followed by a ReLU activation, and $h(\cdot)$ is a 3×3 convolution followed by a 3×3 residual block.

$$F_M = G_M(F_{AS} + F_{AD}, F_{BS} + F_{BD}) \tag{10}$$

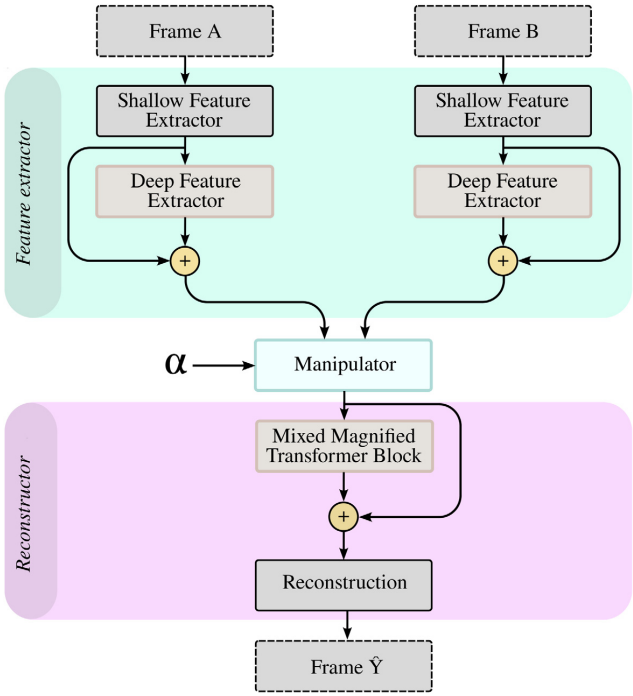


Fig. 1. Architecture overview of the proposed model.

The conjoined manipulated feature space of both frames is then processed by the Mixed Magnified Transformer Block (MMTB) (G_{MMTB}) formed by N RSTB blocks. This stage enables the attention mechanism to affect the combined magnified features of both frames, resulting in a more coherent result after reconstruction.

$$F_{MMTB} = G_{MMTB}(F_M) \tag{11}$$

Finally, reconstruction is dealt with a convolutional block (G_R) that inverts the initial feature mapping, done in the shallow feature extractor, back onto a frame ($I_{\hat{Y}}$).

$$I_{\hat{Y}} = G_R(F_M + F_{MMTB}) \tag{12}$$

Further detail on the architecture can be found in Fig. 2 along with a graphical representation of the Swin Transformer Layer (STL) and the Residual Swin Transformer Block (RSTB).

3.3. Training

The whole network is trained end-to-end using the dataset provided by [12], which allows the results comparison to depend exclusively on network architecture. Nevertheless, in addition to enabling fair comparison, the dataset has proven [12] to produce good-quality models capable of generalizing trained scenarios and returning excellent-quality magnified videos on scenes totally unrelated to the dataset. These reasons led to the adoption of the dataset as the only source of training data.

The L1-Loss cost function was chosen for end-to-end training and placed between the network's output $I_{\hat{Y}}$ and the ground truth frame I_Y. Additionally, in order to improve the feature extraction and make a more robust system, the perturbed c frames provided by the dataset were compared against their non-perturbed counterparts after feature extraction, using yet again L1-Loss. The resulting regularization loss was then added to the end-to-end loss of the whole network with a λ weight coefficient set to 0.1.

Finally, the optimizer of choice for training the model was ADAM [45] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size set to 5 and a learning rate of 10^{-5} with no weight decay.

3.4. Modes of operation

The proposed approach, STB-VMM, can be applied to any input video sequence containing two frames or more, regardless of the time scale between the two frames. Sequences can be treated in one of two modes, static or dynamic, borrowed from [12]. No changes to the network are made for these modes. Instead, the modes refer to the order in which the input frames are fed to the model. The static mode, which follows more closely the classical definition of motion magnification, uses the first frame of the sequence as reference. In terms of computation, the static mode can be expressed like so: $model(I_0, I_t)$, where the t is the frame number increasing sequentially with time. On the other hand, the dynamic mode magnifies the difference between two consecutive frames $[model(I_t, I_{t+1})]$, therefore magnifying velocity between each frame. Note that in each of the modes, the magnification factor α has different meanings.

Oh et al. [12] proposed one additional operation mode with temporal filtering to mitigate the effects of undesired motion and noise. The filtering was applied in the manipulator to produce temporarily-filtered motion-magnified frames similar to those of classical techniques. On the downside, the temporal mode appears to cause blindness to small motions, resulting in patchy magnification. This phenomenon occurs because motion amplitude crosses the threshold to be large enough to be detected and causes some regions to be suddenly magnified mid-sequence. This performance degradation gets worst when the magnification

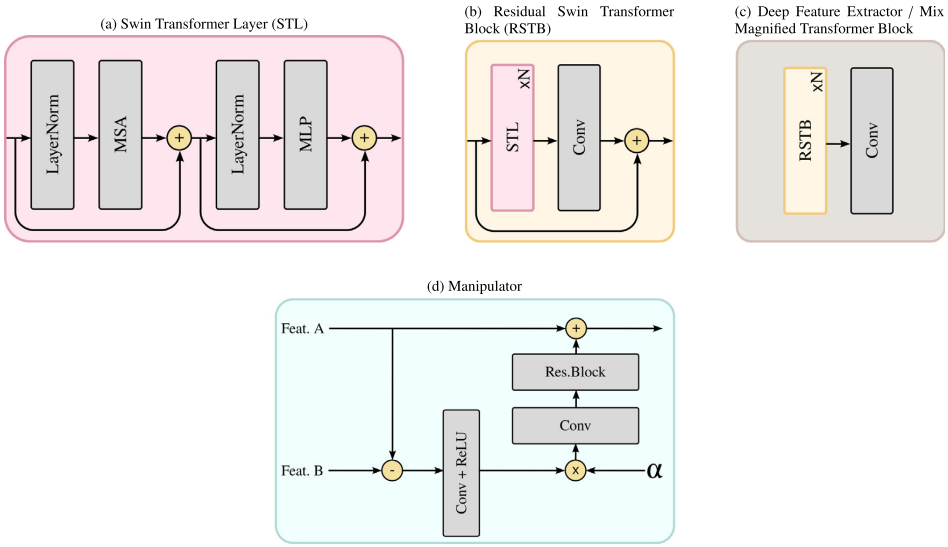


Fig. 2. Architectural details.

factor is high and motion is small. While theoretically possible to incorporate a temporal mode into the proposed model, the magnification results do not suffer from excessive noise or blurring, therefore, temporal filtering is unnecessary and the full spectrum of frequencies is magnified all at once producing good results.

4. Results and discussion

In the following section, the results yielded by the STB-VMM model are compared to the current state-of-the-art learning-based video motion magnification model [12]. Performance is measured quantitatively and qualitatively, showing that our model improves on the previous state-of-the-art in magnification quality and clarity. The video versions of all the comparisons are available in the supplementary materials (see Data availability section).

Quantitative comparison of image quality or Image Quality Assessment (IQA) is a complex topic involving many variables and methods. Said methods are divided into three main categories: full-reference, reduced-reference, and no-reference. A referenced algorithm [46–48] requires a pristine sample to assess the quality of a degraded image, while no-reference methods [49–52] produce an image score without the need of any reference. When evaluating VMM it is impossible to obtain a pristine motion-magnified frame. Therefore, to evaluate the results presented in the following section the MUSIQ [52,53] algorithm was chosen to compare the models' performance.

The following results comparative analyzes the performance of Oh et al.'s Learning-Based Video Motion Magnification (LB-VMM) model and STB-VMM on ten different video benchmarks that showcase interesting motion magnification examples. In addition, a comparison against the baby [7] sequence is added to provide a fair point of comparison. The sequences were captured at 1080p 60fps on a mid-range smartphone to demonstrate the potential of STB-VMM with accessible video equipment.

4.1. Quantitative comparison

Table 2 shows the average, 1st, and 99th percentile average MUSIQ scores for the tested benchmark sequences ran on the Learning-based Video Motion Magnification model and the STB-VMM model. The values presented in the table are calculated for each individual frame of the full sequences and then summarized on an average score. The original sequences are also added as control, and scores are expected to be higher than both of the magnification methods.

The results on Table 2 demonstrate that STB-VMM produces better results than LB-VMM. On average, the scores obtained by STB-VMM are 9.63% higher, and boast a much higher 1% lows, implying that the quality of magnification is noticeably more consistent throughout the sequence. This trend can be observed in Table 3 and Fig. 3, where STB-VMM shows remarkable stability on its output quality. LB-VMM only manages a single higher score than STB-VMM in the building benchmark by a difference of 0.23% (Building_{00}). However, in the authors' opinion, STB-VMM produces better quality magnification with more stable edges and less blurry patches.

On the other hand, none of the magnified scores fall above the original's, as expected. Nevertheless, magnified and original scores follow the same trend, implying that low-quality source videos produce worse outputs. However, STB-VMM is much more capable of dealing with low-quality input images, even closing the quality gap with respect to the original when input quality declines. The sharp quality declines seen in both car sequences can be, in part, attributed to the poor low light performance of the camera employed.

4.2. Qualitative comparison

To reinforce the previous section's scores and claims, this section presents a few qualitative comparisons that demonstrate the effectiveness of our proposed network against the current state-of-the-art in terms of resulting image quality.

Fig. 4 shows the same frame chosen at random from the Car_{00} sequence using both models. STB-VMM, shown on the right, yields a much superior result in terms of image clarity that can be appreciated in both edges and texture.

The car sequence recording was filmed in a rather low light environment, thus yielding noisier/grainier video than otherwise could have been archived. This highlights one of the main benefits

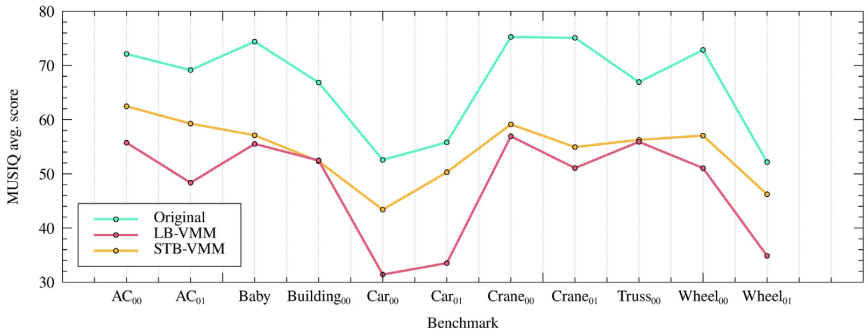
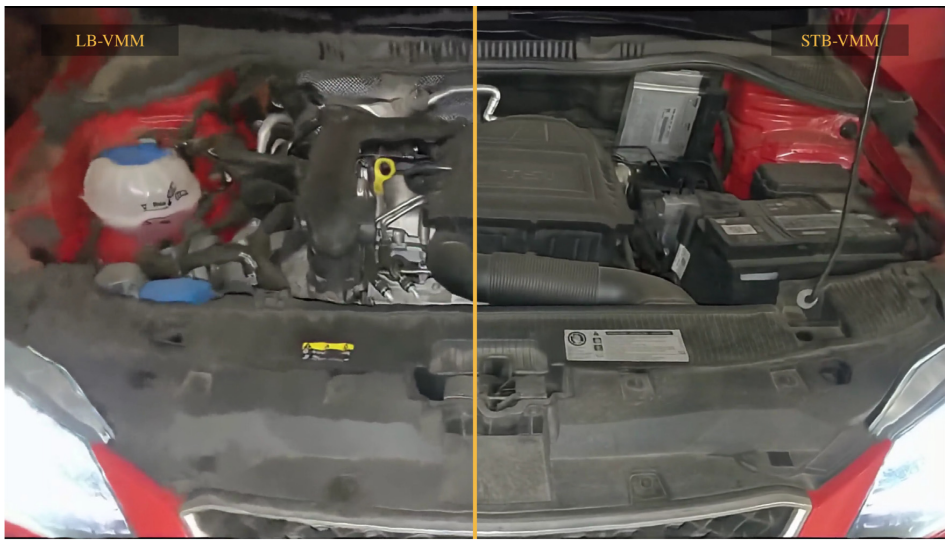


Fig. 3. Graphic representation of the average MUSIQ scores per test sequence magnified x20.



(a) Split frame showing LB-VMM on the left and STB-VMM on the right

(b) LB-VMM (c) STB-VMM (d) LB-VMM (e) STB-VMM

Fig. 4. Qualitative comparison of the car sequence. Highlighted in the bottom row of the figure the car's coolant reservoir, engine cover, and ventilation slits demonstrate that STB-VMM results are noticeably sharper and less distorted.

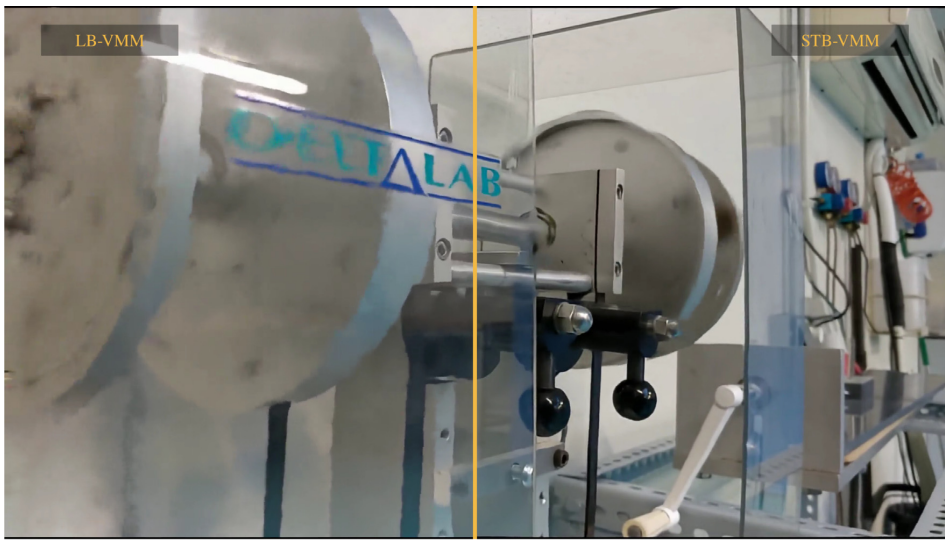
of the proposed architecture, which is a much better tolerance to noisy input. Regardless of clarity, both models perform well on motion magnification with very few artifacts, if any.

The next example, shown in Fig. 5, was filmed in better lighting conditions, yet the quality score of the unmagnified video is no better. This might have been caused, in part, due to the framing of the sequence, which keeps only parts of the image in focus. Regardless of the base score set by the original, STB-VMM clearly outperforms LB-VMM, with better-defined letters and a much more clear background. In terms of motion magnification, both methods display good quality magnification.

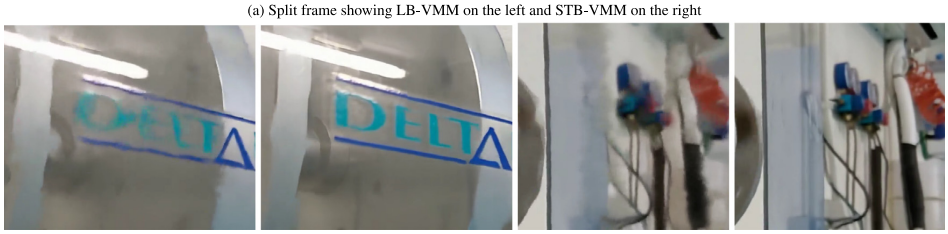
On the other hand, the building sequence (Building_{00}) is the only benchmark where LB-VMM outperforms on average STB-VMM. Nevertheless, the better edge stability offered by STB-VMM enables the authors to obtain better frequency readings from the magnified video. Such application is interesting in technical fields where vibration needs to be monitored, such as in structural health monitoring [5,54–56]. Fig. 6 shows the cropped upper right

Table 2
Comparative MUSIQ scores of the original sequence, the sequence magnified using learning-based video motion magnification (*o3f_hmhm2_bg_qnoise_mix4_nl_n_t_ds3* checkpoint), and the proposed method. (x20).

	Original			LB-VMM			STB-VMM		
	Avg.	η_1	η_{99}	Avg.	η_1	η_{99}	Avg.	η_1	η_{99}
AC_{00}	72.11	69.65	72.75	55.73	49.61	58.69	62.45	61.05	63.29
AC_{01}	69.15	68.30	70.05	48.35	34.07	51.22	59.27	57.72	60.96
Baby	74.39	69.71	74.87	55.51	53.26	59.95	57.12	54.41	62.90
$Building_{00}$	66.84	66.01	75.45	52.46	49.51	62.75	52.30	50.07	56.43
Car_{00}	52.55	50.65	54.41	31.40	18.27	35.50	43.37	23.28	48.06
Car_{01}	55.81	54.77	57.01	33.51	30.67	64.99	50.28	48.08	52.07
$Crane_{00}$	75.26	74.86	75.57	56.92	52.70	65.02	59.13	56.19	62.89
$Crane_{01}$	75.09	74.63	75.44	51.05	45.25	57.37	54.93	51.11	64.70
$Truss_{00}$	66.94	65.92	67.49	55.90	52.65	57.98	56.27	54.93	57.61
$Wheel_{00}$	72.84	71.87	73.38	51.04	28.82	54.40	57.04	36.41	61.19
$Wheel_{01}$	52.15	50.23	53.55	34.84	31.12	59.03	46.21	43.68	48.48
Total avg.	**66.13**	**51.25**	**75.45**	**48.05**	**32.32**	**60.09**	**54.42**	**45.68**	**63.29**
% dev. to avg.	*13.58%*	*22.50%*	*14.09%*	*20.34%*	*32.75%*	*25.04%*	*10.70%*	*16.07%*	*16.28%*



(a) Split frame showing LB-VMM on the left and STB-VMM on the right



(b) LB-VMM (c) STB-VMM (d) LB-VMM (e) STB-VMM

Fig. 5. Qualitative comparison of the wheel sequence. STB-VMM displays sharper letters and a better-defined background with respect to LB-VMM.

corner of the building [57] and the slice used for frequency measuring. Below, in Fig. 6(d), the FFTs obtained from the movement of the sequences are plotted. While both sequences detect a peak at 14.25 Hz, STB-VMM produces a much cleaner signal. During the experiment, the building was intentionally excited with an electrodynamic shaker reproducing a 14.25 Hz sine wave.

The authors acknowledge that image quality can be a somewhat subjective metric and recommend watching the comparison videos attached in the supplementary materials.

4.3. Limitations

In spite of the favorable comparisons, LB-VMM still has a significant advantage in computing time over STB-VMM. With our hardware setup,[1] LB-VMM magnifies the baby [7] sequence, consisting of 300 frames at a resolution of 960 × 576, in approximately 76 s. Meanwhile, STB-VMM almost doubles the compute time, clocking in at 130 s for the exact same sequence. Software optimizations combined with upcoming improvements in

[1] AMD Ryzen 9 5950X; Nvidia RTX 3090.

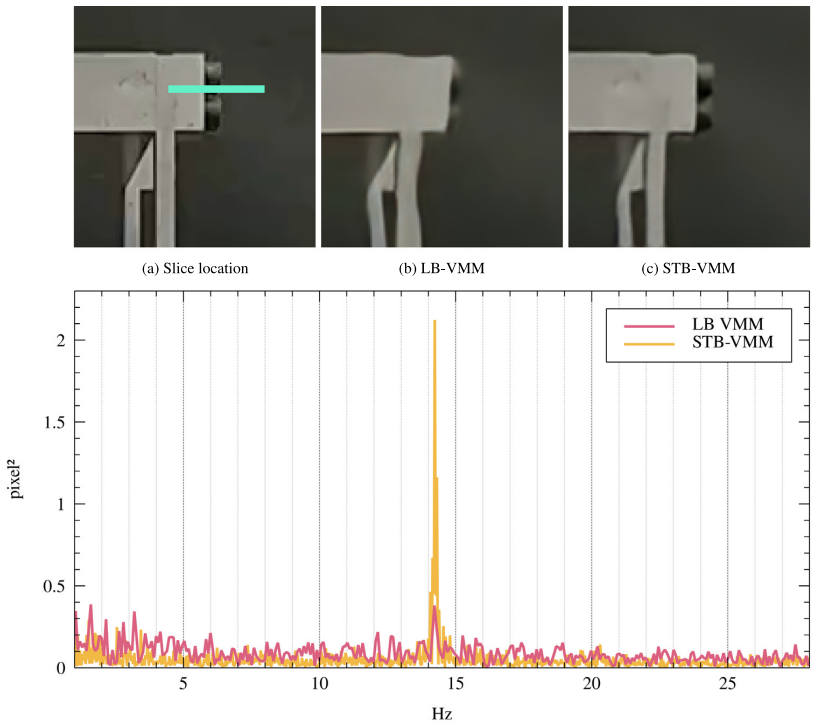


Fig. 6. Vibration readings on the Building_{00} sequence. While the noise floor remains the same on both readings, the FFT obtained using STB-VMM displays a much more prominent peak at 14.25 Hz.

Table 3
MUSIQ score difference between STB-VMM and LB-VMM.

	Avg. (%)	η_1 (%)	η_{99} (%)
AC_{00}	9.32	16.42	6.32
AC_{01}	15.79	34.64	13.91
Baby	2.15	1.65	3.95
Building_{00}	−0.23	0.86	−8.38
Car_{00}	22.79	9.89	23.08
Car_{01}	30.06	31.78	−22.65
Crane_{00}	2.93	4.67	−2.81
Crane_{01}	5.17	7.85	9.71
Truss_{00}	0.55	3.46	−0.55
Wheel_{00}	8.24	10.56	9.26
Wheel_{01}	21.81	25.01	−19.72
Total	**9.63**	**26.07**	**4.24**

hardware might help mitigate STB-VMM's compute time shortcomings.

5. Conclusions

This work presents a new state-of-the-art model for video motion magnification based on the Swin Transformer that has been shown to outperform previous state-of-the-art learning-based models. The new model displays better noise tolerance characteristics, a less blurry output image, and better edge stability, resulting in clearer and less noisy magnification with very few, if any, artifacts.

On the downside, the new model requires more computing resources than previous models and cannot be run in real-time like phase-based methods [8]. Nevertheless, applications that require precise magnification for vibration monitoring [5] could greatly benefit from improvements in the technology. Further work will address the integration of this model in specific applications that require precise vibration monitoring and could benefit from a full-field solution like a camera instead of installing and wiring multiple contact sensors such as accelerometers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The video sources for the comparisons presented in this work are available at: https://doi.org/10.17632/76s26nrcpv.1.

The official implementation of STB-VMM is available at: https://github.com/RLado/STB-VMM.

Acknowledgments

The authors would like to gratefully acknowledge the support and funding of the Catalan Agency for Business Competitiveness (ACCIÓ) through the project INNOTEC ISAPREF 2021. Furthermore, the first author would like to acknowledge a Doctoral Scholarship from IQS. Finally, the authors would like to thank Dr. Eduardo Blanco from the University of Arizona and Dr. Ariadna Chueca de Bruijn for their help.

References

[1] O. Shabi, S. Natan, A. Kolel, A. Mukherjee, O. Tchaicheeyan, H. Wolfenson, N. Kiryati, A. Lesman, Motion magnification analysis of microscopy videos of biological cells, PLOS ONE 15 (11) (2020) 1–18, http://dx.doi.org/10.1371/journal.pone.0240127.

[2] A.J. McLeod, J.S. Baxter, S. de Ribaupierre, T.M. Peters, Motion magnification for endoscopic surgery, in: Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 9036, SPIE, 2014, pp. 81–88, http://dx.doi.org/10.1117/12.2043997.

[3] H. Lauridsen, S. Gonzales, D. Hedwig, K.L. Perrin, C.J. Williams, P.H. Wrege, M.F. Bertelsen, M. Pedersen, J.T. Butcher, Extracting physiological information in experimental biology via Eulerian video magnification, BMC Biol. 17 (1) (2019) 1–26, http://dx.doi.org/10.1186/s12915-019-0716-7.

[4] J. Fei, Z. Xia, P. Yu, F. Xiao, Exposing AI-generated videos with motion magnification, Multimed. Tools Appl. 80 (20) (2021) 30789–30802, http://dx.doi.org/10.1007/s11042-020-09147-3.

[5] R. Lado-Roigé, J. Font-Moré, M.A. Pérez, Learning-based video motion magnification approach for vibration-based damage detection, Measurement 206 (2023) 112218, http://dx.doi.org/10.1016/j.measurement.2022.112218.

[6] C. Liu, A. Torralba, W.T. Freeman, F. Durand, E.H. Adelson, Motion magnification, in: ACM SIGGRAPH 2005 Papers, 2005, pp. 519–526, http://dx.doi.org/10.1145/1186822.1073223.

[7] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. (2012) http://dx.doi.org/10.1145/2185520.2185561.

[8] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, Phase-based video motion processing, ACM Trans. Graph. 32 (4) (2013) 1–10, http://dx.doi.org/10.1145/2461912.2461966.

[9] N. Wadhwa, M. Rubinstein, F. Durand, W. Freeman, Riesz pyramids for fast phase-based video magnification, in: 2014 IEEE ICCP, 2014, pp. 1–10, http://dx.doi.org/10.1109/ICCPHOT.2014.6831820.

[10] N. Wadhwa, H.-Y. Wu, A. Davis, M. Rubinstein, E. Shih, G.J. Mysore, J.G. Chen, O. Buyukozturk, J.V. Guttag, W.T. Freeman, F. Durand, Eulerian video magnification and analysis, Commun. ACM 60 (1) (2016) 87–95, http://dx.doi.org/10.1145/3015573.

[11] Y. Zhang, S.L. Pintea, J.C. van Gemert, Video acceleration magnification, in: CVPR, 2017, http://dx.doi.org/10.48550/arXiv.1704.04186.

[12] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W.T. Freeman, W. Matusik, Learning-based video motion magnification, 2018, http://dx.doi.org/10.48550/arXiv.1804.02684.

[13] M.A. Elgharib, M. Hefeeda, F. Durand, W.T. Freeman, Video magnification in presence of large motions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4119–4127, http://dx.doi.org/10.1109/CVPR.2015.7299039.

[14] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, http://dx.doi.org/10.1007/s11263-009-0275-4.

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common objects in context, 2015, http://dx.doi.org/10.48550/arXiv.1405.0312, arXiv:1405.0312.

[16] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, http://dx.doi.org/10.1162/neco.1989.1.4.541.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Vol. 60, New York, NY, USA, 2017, pp. 84–90, http://dx.doi.org/10.1145/3065386.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, http://dx.doi.org/10.48550/ARXIV.1512.03385.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, IEEE, 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, 2014, http://dx.doi.org/10.48550/ARXIV.1409.0575.

[21] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015, http://dx.doi.org/10.48550/ARXIV.1505.04597.

[22] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, 2019, http://dx.doi.org/10.48550/ARXIV.1908.07919.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, http://dx.doi.org/10.48550/ARXIV.2010.11929.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/ARXIV.1706.03762.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021, http://dx.doi.org/10.48550/ARXIV.2103.14030.

[26] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. (2021) http://dx.doi.org/10.1145/3505244.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, 2021, http://dx.doi.org/10.48550/ARXIV.2108.10257.

[28] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, L. Van Gool, VRT: A video restoration transformer, 2022, http://dx.doi.org/10.48550/ARXIV.2201.12288.

[29] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, 2019, http://dx.doi.org/10.48550/ARXIV.1906.05909.

[30] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, 2020, http://dx.doi.org/10.48550/ARXIV.2006.03677.

[31] Y. Li, K. Zhang, J. Cao, R. Timofte, L. Van Gool, LocalViT: Bringing locality to vision transformers, 2021, http://dx.doi.org/10.48550/ARXIV.2104.05707.

[32] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, L. Van Gool, Vision transformers with hierarchical attention, 2021, http://dx.doi.org/10.48550/ARXIV.2106.03180.

[33] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, J. Shlens, Scaling local self-attention for parameter efficient visual backbones, 2021, http://dx.doi.org/10.48550/ARXIV.2103.12731.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, 2020, http://dx.doi.org/10.48550/ARXIV.2005.12872.

[35] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep Learning for Generic Object Detection: A Survey, Int. J. Comput. Vis. 128 (2) (2020) 261–318, http://dx.doi.org/10.1007/s11263-019-01247-4.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers and distillation through attention, 2020, http://dx.doi.org/10.48550/ARXIV.2012.12877.

[37] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr, L. Zhang, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2020, http://dx.doi.org/10.48550/ARXIV.2012.15840.

[38] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, 2021, http://dx.doi.org/10.48550/ARXIV.2105.05537.

[39] D. Liang, X. Chen, W. Xu, Y. Zhou, X. Bai, TransCrowd: weakly-supervised crowd counting with transformers, Sci. China Inf. Sci. 65 (6) (2022) 160104, http://dx.doi.org/10.1007/s11432-021-3445-y.

[40] G. Sun, Y. Liu, T. Probst, D.P. Paudel, N. Popovic, L. Van Gool, Boosting crowd counting with transformers, 2021, http://dx.doi.org/10.48550/ARXIV.2105.10926.

[41] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-trained image processing transformer, 2020, http://dx.doi.org/10.48550/ARXIV.2012.00364.

[42] J. Cao, Y. Li, K. Zhang, L. Van Gool, Video super-resolution transformer, 2021, http://dx.doi.org/10.48550/ARXIV.2106.06847.

[43] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general U-shaped transformer for image restoration, 2021, http://dx.doi.org/10.48550/ARXIV.2106.03106.

[44] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, 2021, http://dx.doi.org/10.48550/ARXIV.2106.14881.

[45] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015, http://dx.doi.org/10.48550/arXiv.1412.6980.

[46] O. Keleş, M.A. Yılmaz, A.M. Tekalp, C. Korkmaz, Z. Dogan, On the computation of PSNR for a set of images or video, 2021, http://dx.doi.org/10.48550/ARXIV.2104.14868.

[47] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[48] L. Wang, A survey on IQA, 2021, http://dx.doi.org/10.48550/ARXIV.2109.00347.
[49] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the Spatial domain, IEEE Trans. Image Process. 21 (2012) 4695–4708, http://dx.doi.org/10.1109/TIP.2012.2214050.
[50] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212, http://dx.doi.org/10.1109/LSP.2012.2227726.
[51] N. Venkatanath, D. Praneeth, M.C. Bh, S.S. Channappayya, S.S. Medasani, Blind image quality evaluation using perception based features, in: IEEE NCC, IEEE, 2015, pp. 1–6, http://dx.doi.org/10.1109/NCC.2015.7084843.
[52] J. Ke, Q. Wang, Y. Wang, P. Milanfar, F. Yang, MUSIQ: Multi-scale image quality transformer, 2021, http://dx.doi.org/10.48550/ARXIV.2108.05997.
[53] Pytorch toolbox for image quality assessment, 2022, URL https://pypi.org/project/pyiqa/.
[54] J. Font-Moré, G. Reyes-Carmenaty, R. Lado-Roigé, M.A. Pérez, Performance analysis of vibration-based damage indicators under low-modal information structures, Mech. Syst. Signal Process. 190 (2023) http://dx.doi.org/10.1016/j.ymssp.2023.110166.
[55] M.A. Pérez, J. Font-Moré, J. Fernández-Esmerats, Structural damage assessment in lattice towers based on a novel frequency domain-based correlation approach, Eng. Struct. 226 (2021) http://dx.doi.org/10.1016/j.engstruct.2020.111329.
[56] M.A. Pérez, R. Serra-López, A frequency domain-based correlation approach for structural assessment and damage identification, Mech. Syst. Signal Process. 119 (2019) 432–456, http://dx.doi.org/10.1016/j.ymssp.2018.09.042.
[57] E. Figueiredo, G. Park, J. Figueiras, C. Farrar, K. Worden, Structural Health Monitoring Algorithm Comparisons Using Standard Data Sets, Tech. Rep. LA-14393, 2009, 961604, http://dx.doi.org/10.2172/961604.

Publication III:

ViMag: A Visual Vibration Analysis Toolbox

10.21105/joss.05491

The Journal of Open Source Software

ViMag: A Visual Vibration Analysis Toolbox

Ricard Lado-Roigé [1] **and Marco A. Pérez** [1¶]

1 IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain ¶ Corresponding author

DOI: 10.21105/joss.05491

Software

- Review
- Repository
- Archive

Editor: Kevin M. Moerman

Reviewers:

- @jankoslavic
- @iamtsac

Submitted: 21 February 2023
Published: 31 July 2023

Summary

Recent developments in computer vision have brought about a new set of techniques called Video Motion Magnification, which are capable of identifying and magnifying eye-imperceptible movements in video data. These techniques have proved effective in applications such as producing visual representations of an object's operating deflection shapes or recovering sound from a room behind soundproof glass. Our research explores the new possibilities of motion magnification applied to Structural Health Monitoring (SHM) and vibration testing, harnessing the latest advances in deep-learning to achieve state-of-the-art results.

Vision-based damage detection techniques can reduce sensor deployment costs while providing accurate, useful, and full-field readings of structural behavior. We present a new video processing approach that allows the treatment of video data to obtain vibrational signatures of complex structures. This approach enables the identification of very light structural damage in a controlled lab environment. The presented software is based on the use of state-of-the-art deep-learning video motion magnification techniques to offer an easy-to-use, effective, full-field tool for SHM at a fraction of the cost of contact-based techniques.

Related work

This work is based on the method developed by Lado-Roigé et al. (2022) for vibration-based damage detection and on the Swin Transformer Based Video Motion Magnification (STB-VMM) method (Lado-Roigé & Pérez, 2023), which improves on the previous motion magnification backend (Oh et al., 2018) in terms of image quality.

Other researchers have used similar techniques for vibration testing (Eitner et al., 2021; Molina-Viedma et al., 2018). However, to the authors' knowledge, non have released a software tool to go along with their publications. ViMag offers a simple interface to replicate some of these experiments using state-of-the-art learning-based video motion magnification.

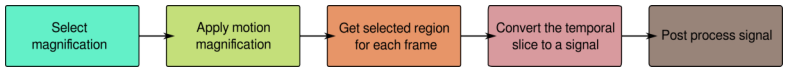


Figure 1: Video sequence to signal pipeline

Motion magnification is a video processing technique that consists on the transformation of input frames to exaggerate motion. The goal of these algorithms is to amplify subtle motions in a video sequence, allowing the visualization of vibrations and deformations that would otherwise be invisible. Video motion magnification was first developed by C. Liu et al. (2005) and opened a new range of possibilities for research, however, this first approach produced numerous visual artifacts on top of being computationally expensive. Years later,further developments by Wu et al. (2012) introduced a novel Eulerian approach to magnification that produced much cleaner

results with less computational cost, paving the way for newer and more refined algorithms that produced increasingly better results such as (Wadhwa et al., 2014), (Oh et al., 2018) or (Lado-Roigé & Pérez, 2023).

Statement of need

ViMag provides an easy-to-use graphical user interface aimed at extracting time-series signals of vibrating machinery and structure videos. This software enables the visualization of videos, selection of magnification area, and signal processing. Consequently, it facilitates and automates the technique developed by Lado-Roigé et al. (2022) and allows machine learning layman to obtain reliable results without having to apply a manual multistage image processing pipeline. Therefore, this software facilitates the use of a camera as a functional replacement for an accelerometer by employing STB-VMM as the motion magnification backend.

The intended use of ViMag is to support the assessment of mechanical systems' performance, such as machines or structures. Researchers and engineers should consider employing condition monitoring or SHM methodologies on the outcomes yielded by ViMag. Such techniques are defined as the set of analysis and assessment tools applied to autonomously determine the integrity and durability of engineering structures. These techniques are aimed at tracking the operational status, assessing the condition, and alerting to the changes in the geometric or material properties that can affect a structure's overall performance, safety, reliability, and operational life (Cosenza & Manfredi, 2000; Frangopol & Curley, 1987).

However, the use for ViMag might not be constrained to mechanical engineering exclusively, and some other interesting applications could also benefit from the software, such as medical applications (Janatka et al., 2020) or miscellaneous technical demos like recovering sound from video (Davis et al., 2014).

Video processing workflow

Figure 1 presents a graph depicting the process of converting a video sequence to a discrete signal. To begin the signal extraction process, the user is asked to select a linear region of interest, preferably on a high-contrast area of the frame. Then, the area surrounding the selected region of interest is magnified using STB-VMM throughout the target video sequence's length. The motion-magnified result is then converted into a single image that represents movement in the temporal domain, achieved by extracting the selected linear region in each of the frames and stacking them horizontally as shown in Figure 2. Finally, an edge detection algorithm is run over the temporal slice to determine the discrete temporal signal and convert it into an array of values over time. From this point on, existing signal processing techniques, such as the Fourier transform, can be used to extract further information.

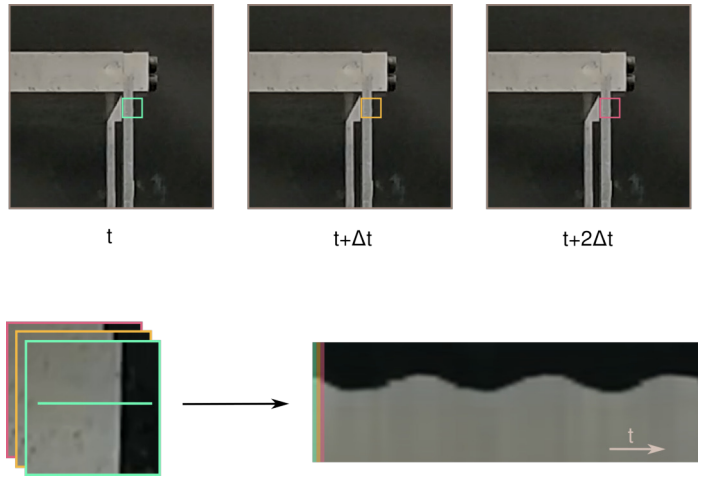


Figure 2: Video sequence transformation to temporal slice

Motion magnification acts like a microscope for motion, magnifying tiny movements on video sequences to retrieve seemingly invisible or almost imperceptible movements. Consequently, motion magnification may allow the naked eye to see a structure's operating deflection shapes as they happened in real operating conditions. The STB-VMM model consists of three main functional blocks that extract features from frames, manipulate those features and finally reconstruct the frames. Implemented in PyTorch (Paszke et al., 2019), STB-VMM borrows ideas from Dosovitskiy et al. (2020), Vaswani et al. (2017), and Z. Liu et al. (2021) to improve the image quality offered by prior motion magnification methods at the cost of some performance. The lack of temporal filtering and the higher image quality offered by STB-VMM play an important role in applications that require precise magnification for vibration monitoring, as less-noisy images produce clearer signals that highlight abnormal behaviors sooner.

Acknowledgements

The authors would like to gratefully acknowledge the support and funding of the Catalan Agency for Business Competitiveness (ACCIÓ) through the project INNOTEC ISAPREF 2021. Additionally, the first author would like to acknowledge a Doctoral Scholarship from IQS.

References

Cosenza, E., & Manfredi, G. (2000). Damage indices and damage measures. *Prog. In Struct. Eng. Mater.s*, *2*(1), 50–59. https://doi.org/10.1002/(SICI)1528-2716(200001/03)2:1%3C50::AID-PSE7%3E3.0.CO;2-S

Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G. J., Durand, F., & Freeman, W. T. (2014). The visual microphone: Passive recovery of sound from video. *ACM Trans. Graph.*, *33*(4). https://doi.org/10.1145/2601097.2601119

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. https://doi.org/10.48550/ARXIV.2010.11929

Eitner, M., Miller, B., Sirohi, J., & Tinney, C. (2021). Effect of broad-band phase-based motion magnification on modal parameter estimation. *Mechanical Systems and Signal Processing*, *146*, 106995. https://doi.org/10.1016/j.ymssp.2020.106995

Frangopol, D. M., & Curley, J. P. (1987). Effects of damage and redundancy on structural reliability. *J. Struct. Eng.*, *113*(7), 1533–1549. https://doi.org/10.1061/(ASCE)0733-9445(1987)113:7(1533)

Janatka, M., Marcus, H. J., Dorward, N. L., & Stoyanov, D. (2020). Surgical video motion magnification with suppression of instrument artefacts. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 353–363. https://doi.org/10.48550/arXiv.2009.07432

Lado-Roigé, R., Font-Moré, J., & Pérez, M. A. (2022). Learning-based video motion magnification approach for vibration-based damage detection. *Measurement*, 112218. https://doi.org/10.1016/j.measurement.2022.112218

Lado-Roigé, R., & Pérez, M. A. (2023). STB-VMM: Swin transformer based video motion magnification. *Knowledge-Based Systems*, 110493. https://doi.org/10.1016/j.knosys.2023.110493

Liu, C., Torralba, A., Freeman, W. T., Durand, F., & Adelson, E. H. (2005). Motion magnification. *ACM SIGGRAPH 2005 Papers*, 519–526. https://doi.org/10.1145/1186822.1073223

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin transformer: Hierarchical vision transformer using shifted windows*. https://doi.org/10.48550/ARXIV.2103.14030

Molina-Viedma, A. J., Felipe-Sesé, L., López-Alba, E., & Díaz, F. (2018). High frequency mode shapes characterisation using digital image correlation and phase-based motion magnification. *Mech. Syst. Sig. Process.*, *102*, 245–261. https://doi.org/10.1016/j.ymssp.2017.09.019

Oh, T.-H., Jaroensri, R., Kim, C., Elgharib, M., Durand, F., Freeman, W. T., & Matusik, W. (2018). *Learning-based video motion magnification*. https://doi.org/10.48550/arXiv.1804.02684

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. https://doi.org/10.48550/ARXIV.1706.03762

Wadhwa, N., Rubinstein, M., Durand, F., & Freeman, W. (2014). Riesz pyramids for fast phase-based video magnification. *2014 IEEE ICCP*, 1–10. https://doi.org/10.1109/ICCPHOT.2014.6831820

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* https://doi.org/10.1145/2185520.2185561

Publication IV:

Vision-based dynamic response analysis of engineering structures

Vision-based dynamic response analysis of engineering structures

Ricard Lado-Roigé[a], Marco A. Pérez[a,*]

[a]*IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain*

Abstract

This paper presents a comprehensive evaluation of the STB-VMM motion magnification algorithm for use in structural health monitoring applications. The video motion magnification system's ability to accurately identify a structure's natural frequencies and operating deflection shapes is demonstrated through experimental testing, thus confirming the approach as a reliable structural damage detection method. A comparative analysis of the performance of STB-VMM and prior methods using learning-based video motion magnification under varying image conditions reveals the advantages of STB-VMM when processing noisy data, as well as its limitations in more favourable image conditions. The potential for motion magnification to increase the sensitivity of vision-based vibration analysis is explored, discussed, and compared to existing contact and non-contact sensors.

Keywords: Computer Vision, Deep Learning, Damage Identification, Motion Magnification, Structural Health Monitoring

1. Introduction

Structural Health Monitoring (SHM) [1, 2] is the research field that studies and analyses the behaviour of mechanical systems over time in order to detect structural performance degradation. The goal of SHM is to provide updated information regarding the ability of engineering structures to perform their intended function in light of the inevitable degradation of the building materials and systems, often resulting from factors such as fatigue, wear, or corrosion present in their operational environments. SHM provides the framework to determine the adequate use of monitoring equipment as well as the processes involved in the analysis of captured signals. The SHM process involves several steps, from data aquisition and feature extraction to EOV[1] mitigation [3, 4] and decision-making. This work focuses on the data aquisition process which generally involves selecting excitation methods, sensor types and their positions, and the data acquisition and transmittal hardware. More precisely, this paper focuses on vision-based methods for cap-

*Corresponding author.
Email address: marcoantonio.perez@iqs.url.edu (Marco A. Pérez)

[1]EOV: Environmental and Operational Variabilities

turing the necessary signals for the appliacation of mathematical SHM tools to detect, diagnose and predict damage [5, 6].

To directly monitor the state of a system, it is necessary to identify features in the acquired data that allow the distinction between the undamaged and damaged structure [7, 8]. One of the most common feature extraction methods is based on correlating measured system response quantities, such as vibration amplitude or frequency, with observations of the degraded system.

The growing trend towards digitalization has positively impacted the practice of mechanical and civil engineering, providing the means and infrastructure to collect, store, and process operational data for analysis [9, 10]. Structural monitoring is key for learning about engineering structures and to enable engineers to validate their calculation assumptions as well as to detect suboptimal performance due to wear, corrosion, or other factors. In turn, this improves the visibility of incipient structural problems and allows the use of predictive maintenance strategies to prevent further damage as well as prevent unexpected service interruptions and safety risks. Therefore, reducing maintenance costs and improving operational safety [11, 12].

Despite these benefits, SHM has not yet been widely adopted in the civil engineering field, and thus, the most common approach for evaluating the operational condition of civil infrastructure is for a trained operator to perform a visual assessment of the structure. However, this presents some notable disadvantages, such as the fact that visual inspections are time-consuming, require the attention of qualified engineers and technicians, and not all load-bearing elements might be inspectable due to occlusion by other elements or decorative coverings.

SHM systems reinforce the information obtained by periodic inspections, and is an excellent diagnostic tool to detect damage in hard-to-reach areas, allows continuous monitoring, and issues alerts when detecting anomalous behaviours, therefore helping reduce repair costs. However, the downside of SHM systems tends to be the upfront cost of instrumenting the structure and maintaining the monitoring systems operational.

For the purpose of instrumenting a structure, multiple types of sensors [13, 14] are used that measure displacement, such as strain gauges or fibre optics sensors, acceleration, like piezoelectric or MEMS accelerometers, or inclination, with sensors like inclinometers or tiltmeters. The obvious downside of these example sensors is that they must be installed on the structure and must have some means of transmitting a signal, either wired or wirelessly. On the one hand, wired sensors provide good quality signals and do not require extra power sources, but routing the cables might not be feasible on hard-to-access or moving elements. Moreover, long cable runs might affect the signal quality and might require of additional hardware. On the other hand, wireless sensors require an independent power source for transmitting data, which might be unfeasible in some instances.

Vision-based monitoring methods [15–18] attempt to mitigate these issues by providing data without the need

for contact with the structure, inferring the displacement of structural elements by tracking pixels or features over time. Vision-based techniques provide full field measurements reducing the number of sensors that need to be deployed. These advantages make vision-based options very attractive in spite of generally lower accuracy than their contact-based counterparts. Techniques such as digital image correlation, optical flow, and video motion magnification (VMM) [19–23] have been successfully implemented in multiple fields and applications.

This work intends to validate the novel STB-VMM motion magnification algorithm for SHM applications by using the method described in [24, 25]. Furthermore, this paper compares the results obtained using the learning-based video motion magnification model presented by Oh et al. against the results of STB-VMM to highlight the differences between these methods in the SHM context. This work also includes a comparison of both learning-based magnification methods to determine their tolerance to unfavourable imaging conditions, that are often found in the field. Finally, this paper expands on previous tests [24] by including experiments on a large-scale structure.

The following section introduces the theoretical background regarding video motion magnification as well as its application to SHM. The third section details the testing methodology and image processing steps. Next, the fourth section presents and discusses the results to highlight the differences between VMM methods and ground truth. Finally, the fifth section summarizes the conclusions of this work.

2. Related work

Motion magnification is an area of study in computer vision that involves the amplification of very small motions present in video sequences, making them easily visible to the naked eye. These techniques find applications in diverse fields such as biomedical research [26–28], deepfake detection [29], structural modal analysis [19–24], or condition monitoring [30]. Acting as a microscope for motion, these methods unveil previously invisible or imperceptible movements. However, despite its seemingly straightforward objective, the separation of subtle motions from noise poses a complex challenge, particularly when attempting to enhance extremely delicate movements that often occur at sub-pixel levels. Consequently, conventional motion magnification techniques commonly encounter issues such as noisy and blurred outputs due to the difficulty involved in designing filters that eliminate noise without removing critical sub-pixel data. Hence, researchers have devoted considerable efforts to addressing these limitations and enhancing the quality and performance of magnification.

Earlier approaches to motion magnification, such as Liu et al. [31], relied on a Lagrangian approach that used motion tracking or optical flow estimation to isolate the desired motion from noise prior to magnification. However, this methodology proved computationally demanding and prone to artifacts, especially in regions affected by occlusion boundaries and intricate motion patterns. On the other hand, subsequent developments [32–34, 38, 35] adopted an

Method	Liu et al. [31] (2005)	Wu et al. [32] (2012)	Wadhwa et al. [33] (2013)	Wadhwa et al. [34] (2014)	Zhang et al. [35] (2017)	Oh et al. [36] (2018)	Lado et al. [37] (2023)
Spatial decomposition	Tracking, optical flow	Laplacian pyramid	Steerable filters	Riesz pyramid	Steerable filters	Deep convolution layers	Swin Transformer
Motion isolation	-	Temporal bandpass filter	Temporal bandpass filter	Temporal bandpass	Temporal bandpass filter (2_{nd} order derivative)	Subtraction or bandpass filter	Subtraction
Representation denoising	Expectation-Maximization	-	Amplitude weighted Gaussian filtering	Amplitude weighted Gaussian filtering	Amplitude weighted Gaussian filtering	Trainable convolution	Swin Transformer

Table 1: Motion magnification techniques summary table. Adapted from [36, 37].

Eulerian approach that focused on observing changes within a fixed region of pixels over time, rather than tracking specific features across space and time. The distinction between Lagrangian and Eulerian methods is not dissimilar from the same distinction used in the field of fluid dynamics. These Eulerian-based methods offer advantages such as lower computational requirements and improved performance with small motions, resulting in enhanced magnification outcomes. Nevertheless, these approaches still exhibit noticeable blurring and artifacts due to the inherent challenge of designing noise removal filters that do not interfere with the motion magnification process.

Most Eulerian approaches function by decomposing video sequences into motion representations that can later be manipulated mathematically and then reconstructed back into magnified video frames. These stages are present in the different approaches taken by many researchers in the field and are summarized in Table 1.

2.1. Learning-based video motion magnification

To address the challenges presented by prior motion magnification algorithms, Oh et al. [36] introduced a novel learning-based approach to video motion magnification (LB-VMM). Departing from conventional techniques that employed hand-designed filters, this method leverages Convolutional Neural Networks (CNNs) to learn the filters required for motion magnification. The proposed approach achieved superior magnification quality, with fewer ringing artifacts and improved noise characteristics compared to previously published methods. However, this technique relies on additional temporal filtering to enhance image quality, which occasionally introduces errors in the magnification process. Although satisfactory results can be obtained without temporal filtering, the application of filtering improves image quality by eliminating undesired motion and noise prior to learning-based magnification.

The learning-based model as described by [36] consists of three main parts, the encoders (G_E), manipulator (G_M), and decoder (G_M), which are constituted by fully convolutional layers, allowing for flexible learnable filters at multiple input/output resolutions. The goal of the model is to generate a motion representation (M), which can be manipulated by a simple scalar product to be then reconstructed into a magnified displacement frame ($\hat{I}$). In this architecture, the encoder block acts as a learnable spatial decomposition filter that extracts a motion representation for two input frames

(A and B).

$$G_E(A) = [M_A, V] \tag{1}$$

$$G_E(B) = [M_B] \tag{2}$$

Then, the manipulator block multiplies the difference of these representations by a scalar value that serves as the magnification factor α chosen by the user. This result is then passed on to the decoder that reconstructs the motion representation into a new motion magnified frame ($\hat{I}$), aided by an unmanipulated visual representation (V) previously yielded by the encoder.

$$G_M(M_A, M_B, \alpha) = M_A + h(\alpha \cdot g(M_B - M_A)) \tag{3}$$

$$G_M(M_A, M_B, \alpha) = M \tag{4}$$

Where $g()$ is represented by a 3×3 convolution followed by ReLU, and $h()$ is a 3×3 convolution followed by a 3×3 residual block.

$$G_D(M, V) = \hat{I} \tag{5}$$

The model is trained end-to-end using *l1-loss* between the network output $\hat{I}$ and the ground-truth magnified frame I. However, obtaining real-world motion-magnified video pairs for training is challenging, if not impossible. For this reason, Oh et al. devised a procedural strategy for generating synthetic data involving several considerations [36] to create a suitable dataset.

The dataset is comprised of 200,000 images from the MS COCO dataset [39] used as backgrounds over which 7,000 segmented objects from the PASCAL VOC dataset [40] were carefully moved. Especial attention to detail was paid to ensure that conflictive areas, such as the ones that should appear occluded on the magnified frame actually do. Each training sample contains between 7 and 15 foreground objects randomly distributed and scaled from their original size. Additionally, variants with altered contrast, blurriness, and complex motion were generated to improve model generalization on difficult magnification scenarios.

2.2. Learning-based video motion magnification approach to vibration-based damage detection

Prior studies [24] have researched the possibility of using learning-based video motion magnification for structural damage detection, concluding that the method is viable based on laboratory-scale tests. In [24] the authors outlined a

methodology for extracting dynamic responses from recorded footage by delimiting a linear region within the video that will be magnified. This linear region is extracted from each individual magnified frame and sequentially ordered from oldest to newest onto an image, effectively converting the unidimensional linear regions into a bidimensional image containing the displacement over time of the selected region. For this method to work, it requires linear regions to be placed over elements of contrasting colour or brightness. By extracting the magnified displacement in the described manner, the authors could calculate the frequency response autospectra of the displacement-over-time signals to obtain frequency and magnitude graphs. Using a well-known structural benchmark [41] the authors compared different damage scenarios and registered the frequency shifts using both contact piezoelectric accelerometers and the LB-VMM method. The authors concluded that the method based on LB-VMM provided suitable accuracy for damage detection, presenting errors no bigger than 0.9% on average with respect to the accelerometers.

Later, the displacement extraction algorithm previously described in [24] was implemented in a graphical tool [25] that allows for fast and consistent results with respect to the original command line coordinate-based approach used in the original paper. However, this tool uses a different motion magnification backend based on an improved version of LB-VMM named STB-VMM or Shifted Window Transformer Based Video Motion Magnification, a yet unproven algorithm for use in SHM applications.

2.3. STB-VMM

STB-VMM [37] is a state-of-the-art model for video motion magnification based on the Swin Transformer [42] proven to outperform prior learning-based methods in terms of image quality. STB-VMM presents superior input noise tolerance, less blurry results, and better edge stability, resulting in clearer and less noisy magnification with fewer artifacts. Moreover, STB-VMM is capable of producing good quality results without temporal filtering, thus eliminating the limitations this feature imposed on LB-VMM.

STB-VMM mostly abandons CNNs in favour of the Visual Transformer (ViT) [43], a novel architecture based on the attention mechanism, first popularized in the field of natural language processing. The attention mechanism maps a query and a set of key-value pairs into an output. The output, represented in vector format, is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function between the query and the keys [44]. The transformer was the first model which solely used self-attention to compute representations of its input and output without using sequence-aligned recursive neural networks or convolutional operations. Unlike CNNs, transformers lack translation invariance and a locally restricted receptive field, in its place, transformers offer permutation invariance. This feature allowed natural language processing models to infer relations deep within a text, thus enabling a stronger contextualization of the subject matter, giving birth to wildly popular models such as BERT [45] or GPT [46–48]. However, computer vision applications and their natural predisposition

to using large grid structured data presented problems when trivially trying to apply the self-attention strategy, as the full self-attention layer would be too costly for the current computing capabilities. The ViT overcame this issue by dividing the image data into a grid to flatten and encode smaller image tiles that were later treated as tokens. This tokenization of images allowed the ViT to directly use the original structure of the transformer first proposed by [44] on top of significantly reducing computational costs. Further improvements on the ViT led to the Shifted Window Transformer (Swin Transformer) [42], a hierarchical vision transformer using shifted windows to better capture image context while keeping computing resource consumption under control.

As shown in Table 1 and Figure 1, STB-VMM has a similar structure to Eulerian methods, which consists of three parts: the feature extractor, the manipulator, and the reconstructor. The feature extractor has two components: the shallow and deep feature extractors, which produce a high-quality representation of an input frame. The manipulator then amplifies the motion by multiplying the difference between the features of two frames by a user-defined factor α. Finally, the reconstructor transforms the manipulated features back into a magnified frame.

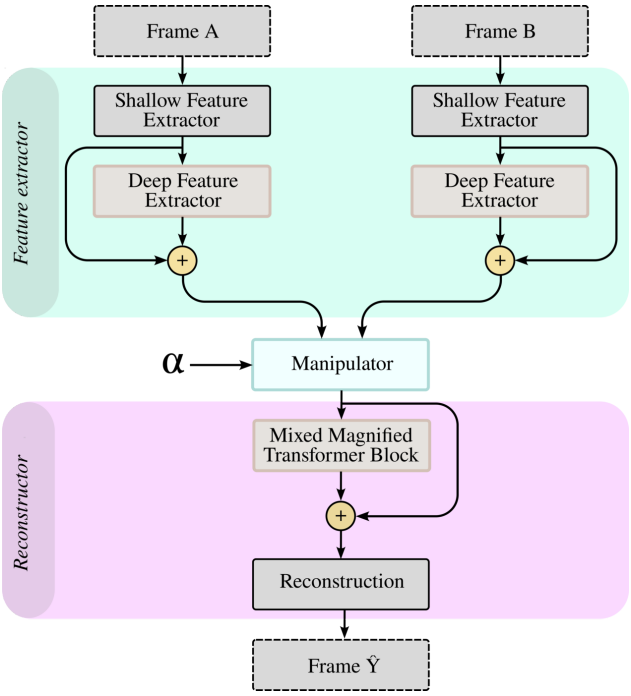


Figure 1: Architecture overview of STB-VMM [37]

The convolutional shallow feature extractor (G_{SF}) takes two frames of a target sequence $[I_A, I_B] \in \mathbb{R}^{H \times W \times C_{in}}$ (where H is the height of the image, W is the width of the image and C_{in} represents the number of input channels) and

maps them to a higher dimensional feature space, where high-level features are extracted (F_{AS}, F_{BS}). The use of an initial convolutional layer is a deliberate choice since it provides a more stable optimization and better results when later combined with transformers [49].

$$[F_{AS}, F_{BS}] = G_{SF}([I_A, I_B]) \tag{6}$$

Next, the deep feature extraction module (G_{DF}) takes the features from the previous step and processes them further to obtain the final motion representations. This module has N Residual Swin Transformer Blocks (RSTB) connected in series.

$$[F_{AD}, F_{BD}] = G_{DF}([F_{AS}, F_{BS}]) \tag{7}$$

The manipulator (G_M) receives the resulting motion representations of both frames and calculates the difference between the feature spaces, magnifying them by multiplying them by the user-defined magnification factor α.

$$G_M(F_{AS} + F_{AD}, F_{BS} + F_{BD}) = (F_{AS} + F_{AD}) + h(\alpha \cdot t(((F_{BS} + F_{BD}) - (F_{AS} + F_{AD})))) \tag{8}$$

Where $t(\cdot)$ is a 3×3 convolution followed by a ReLU activation, and $h(\cdot)$ is a 3×3 convolution followed by a 3×3 residual block.

$$F_M = G_M(F_{AS} + F_{AD}, F_{BS} + F_{BD}) \tag{9}$$

The conjoined manipulated feature space of both frames is then moved to the Mixed Magnified Transformer Block (MMTB) (G_{MMTB}) formed by N RSTB blocks. This stage enables the attention mechanism to affect the combined magnified features of both frames, resulting in a more coherent result after reconstruction.

$$F_{MMTB} = G_{MMTB}(F_M) \tag{10}$$

Finally, a convolutional block (G_R) performs the inverse operation of the initial feature mapping, conducted by the shallow feature extractor (G_{SF}), and restores the image representation to a magnified frame ($I_{\hat{Y}}$).

$$I_{\hat{Y}} = G_R(F_M + F_{MMTB}) \tag{11}$$

STB-VMM is a more complex and consequently a more computationally expensive model than LB-VMM, never-

theless, the quality increase of the unfiltered magnified videos is notably better when having both models trained with Oh et al.'s original synthetic dataset [36]. The authors speculate that the performance differential shown with respect to LB-VMM might benefit its SHM applications. However, STB-VMM is yet untested as a suitable replacement for LB-VMM in SHM applications that attempt frequency-based damage identification.

3. Methodology

3.1. Visual dynamic response capture

To obtain the displacement-over-time responses of the experiments that follow, this paper uses the method described in [24] by using a graphical tool implementation [25]. The method implemented follows the methodology described in section 2.2, by which a linear region of measure is user-selected to be magnified. Next, the selection area is processed, and for each of the frames of the magnified video sequence, the linear regions of pixels are placed sequentially on an image (see fig. 3). The resulting image contains the variation-over-time of the selected unidimensional region, which can be transformed into a discrete displacement-over-time signal using image processing techniques. Once a discrete signal is obtained, well-known signal processing techniques [5] can be used to analyze its properties and characteristics.

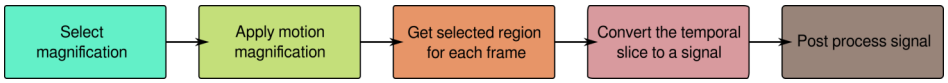


Figure 2: Video sequence to discrete signal pipeline [25]

Figure 2 shows a chart summarizing the signal extraction process. Note that no specific VMM algorithm is required since magnification is the basis for subsequent image treatment. However, non-learning-based methods require frequency filtering; consequently, only a narrow band of movement can be processed simultaneously. On the other hand, learning-based approaches can magnify the complete frequency spectrum at once, enabling the visualization of the true displacement of the target structure.

For all comparisons between STB-VMM and LB-VMM, the same software tool [25] and project files have been used to ensure that the regions measured match strictly between both vision-based methods. In the case of LB-VMM the graphical tool has been modified, replacing the STB-VMM backend with the LB-VMM implementation and weights used in [24].

3.2. Experimental configuration

This work presents two experimental cases on which the results of STB-VMM are validated and compared against LB-VMM. The first case is conducted on a laboratory-scale structure like the one described by [41] from Los Alamos

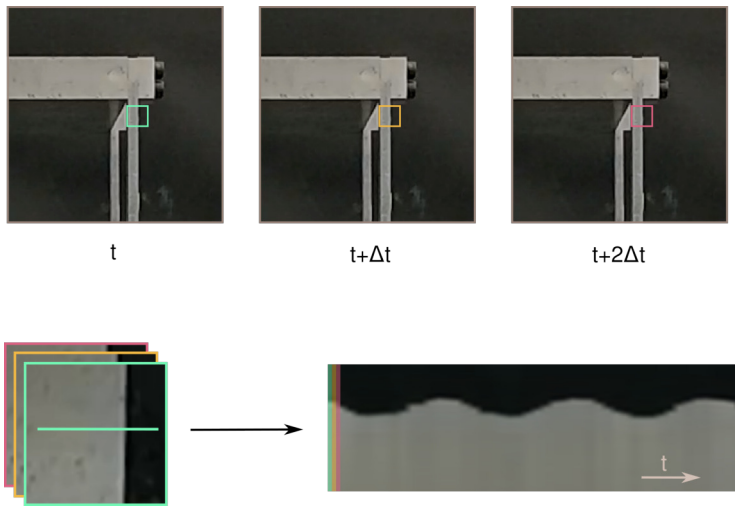


Figure 3: Video sequence transformation to temporal slice [25]

National Laboratory. The second case studies a large-scale reinforced concrete frame structure cast in place [50].

3.2.1. Case 1: Three-story building benchmark

In the first case, the experiments are conducted on the three-story building benchmark, a well-known reference structure made from square aluminium plates ($305 \times 305 \times 25mm$) and prismatic aluminium columns ($177 \times 25 \times 6mm$) attached employing bolted joints forming a four degree-of-freedom system. Each plate is supported by four columns forming a three-story building which is then mounted over linear rails that enable the lateral movement of the structure. Parallel to the rails, an electrodynamic shaker [2] excites the centre-right of the lower floor of the building, providing configurable excitation in intensity and type. Both the shaker and the building are mounted on a rigid steel frame isolated from other sources of vibration. For validation purposes, the three-story building benchmark has been instrumented with piezoelectric accelerometers [3] placed on each floor of the test structure to serve as the ground-truth reference of the controlled damage experiments detailed below. On the other hand, a global shutter industrial camera [4] was placed in front of the structure to capture the full height of the building, acquiring ˜500 frames per second (fps) in 8-bit depth monochrome. The resulting image sequences were processed with both LB-VMM and STB-VMM.

To validate the damage detection capacity and accuracy of STB-VMM, several simulated damage scenarios on a column supporting the first floor are considered. The studied damage scenarios consist in reducing the longitudinal

[2] Brüel & Kjær electrodynamic shaker model 4824
[3] Brüel & Kjær accelerometers type 4519-003
[4] Allied Vision Alvium 1800 U-158c

cross-section of the lower left column of the building by 10%, 20%, 50%, and 100% (removing the column). These section reductions imply a stiffness reduction of the individual column by 20%, 50%, 85%, and 100% respectively. The reduction of the stiffness of the constituent parts of a structure implies changes in its natural frequencies, therefore alerting and highlighting damage. Additionally, a moderate loading scenario was also evaluated by placing a 1.2 kg mass on the top floor of the building in its undamaged configuration.

Since no dynamic operational loading is present on the structure, external excitation is required to reveal the dynamic response. In all of the studied scenarios, two different kinds of excitation limited between 0 and 100Hz have been used: random and periodic sweep. The sweep vibrates sequentially from 0 up to 100Hz and repeats the signal once it is done. This type of excitation generally allows for cleaner dynamic response signals, however, this is a rather unrealistic excitation for larger civil structures subjected to ambient and operational conditions. For this reason, we also used random noise to excite the structure, which contains several frequencies at random and produces slightly more noisy results.

3.2.2. Case 2: Reinforced concrete frame structure

The second case focuses on the data provided by [50], which constructed and tested a $6m$ tall four-story reinforced concrete structure. Each of the stories was $1.5m$ tall by $3m$ wide and $1m$ deep, loaded each with 2700 kg except for the roof, which was loaded with 2580 kg. The test building was placed over a $4 \times 4m$ 25-ton shaker table that excites the structure using recordings of real earthquakes.

Each story of the structure was instrumented with a non-contact high-speed motion capture system [5] with a refresh rate of 128 Hz. This data will be considered as ground-truth for the discussion taking place in the following results section. At the same time, the experiments were recorded in video for post-processing. The camera used for the recording is a consumer-grade device [6] equipped with a wide-angle lens that was placed roughly at a distance of $3m$ in front of the reinforced concrete structure. The recordings were captured at 1920×1080 and 25 fps. Figure 4 shows a frame of an in-progress experiment.

4. Results and discussion

4.1. Case 1: Three-story building benchmark

In this subsection, the results corresponding to the three-story building benchmark and its different configurations are presented and analyzed. Figure 5 shows a comparison between the autospectra obtained using STB-VMM and

[5] NDI Optotrack Certus
[6] Sony HDR-PJ220

Figure 4: Video frame corresponding to the tests performed on the reinforced concrete frame. The left-hand side corresponds to the original unaltered frame, while on the right, an ×20 version of the frame shows the operating deflection shape.

the reference piezoelectric accelerometers. As it was expected, the visual method exhibits a higher noise floor than the accelerometers, however, it only needs a single sensor to obtain these four signals. Moreover, it can be observed that the visual method, based on displacement, is notably more sensitive to lower frequencies. In fact, STB-VMM registers frequencies lower than what our reference piezoelectric accelerometers do. This discrepancy is explained by the fact that the frequency sweep used to excite the structure starts at a very low frequency, displacing the structure and creating a visible peak on the autospectrum. On the other hand, the higher frequencies are better seen using the accelerometers, as they are better suited for this task and generally boast much higher acquisition rates than cameras. When working on civil structures, this might be an acceptable trade-off, as generally, these kinds of structures exhibit frequencies under 15 Hz.

Notice that not all floors of the structure present the same magnitude for their natural frequencies, this is a consequence of their physical position in the length of the building. To be more precise, the magnitude difference observed for each frequency is proportional to the distance between the measured point and the closest node of the modal shape.

After checking that STB-VMM is capable of detecting the correct frequencies in the structure's undamaged conditions, the following results illustrate the sensitivity to damage of the technique. The results presented in figures 6a, 6b, and 6c correspond to the damage scenarios performed on the lower column of the three-story building benchmark, as stated in section 3. The results have been split into three figures to aid in the readability of the results.

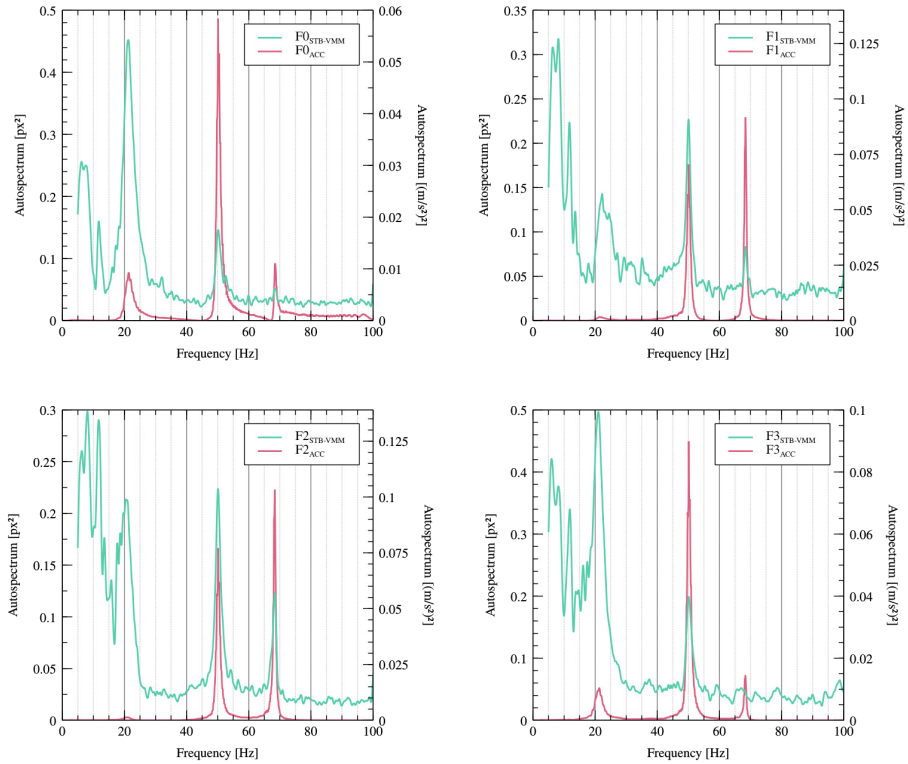


Figure 5: Frequency response functions for each floor of the three-story building benchmark in its undamaged configuration. The structure was excited using a periodic chirp signal ranging from 0 to 100 Hz.

Figure 6a shows the results for the undamaged, the 1.2 kg added mass, and the missing column cases. As expected, the difference between the undamaged state and the missing column case is the most noticeable. On the other hand, the structure loading case produces a more subtle frequency shift still visible in all three methods tested. Figures 6a and 6b present the data obtained using STB-VMM, LB-VMM, and the accelerometers in three separate side-by-side graphs. Comparing the different measurements, the visual methods, which are based on displacement, present a more significant sensitivity for lower frequencies and a noticeably higher noise floor with respect to the contact accelerometers. Nevertheless, the differences between STB-VMM and LB-VMM are smaller. Damage sensitivity is similar for both methods, with LB-VMM presenting a slightly higher noise floor than STB-VMM.

The measurements and magnification for the vision-based graphs have been performed on the exact same coordinates and then post-processed individually to improve signal quality as much as possible. The differences in magnification quality often result in variations in slice image contrast due to blurring, making uniform post-processing

Table 2: Results obtained from periodic excitation

	Accelerometer (Hz)			STB-VMM (Hz)			LB-VMM (Hz)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	21.38	50.12	68.38	21.10	50.33	68.27	21.10	50.36	68.30
Undamaged$_{Repr.}$	21.38	50.12	68.38	21.26	50.11	68.37	21.34	50.15	68.29
Mass	21.25	49.00	68.00	20.12	49.16	68.00	20.47	49.01	67.88
20%	21.38	49.38	67.50	21.36	49.34	67.70	21.02	49.34	67.77
50%	21.38	48.62	67.62	20.86	48.59	67.52	20.78	48.55	67.48
85%	20.62	46.62	66.75	20.42	46.76	66.60	20.40	46.76	66.64
100%	19.88	45.88	62.62	19.80	45.82	62.47	19.72	45.89	62.62

Table 3: Relative error to the reference acceleration data when using periodic excitation

	Error STB-VMM (%)			Error LB-VMM (%)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	1.31%	0.42%	0.16%	1.31%	0.48%	0.12%
Undamaged$_{Repr.}$	0.56%	0.02%	0.01%	0.19%	0.06%	0.13%
Mass	5.32%	0.33%	0.00%	3.67%	0.02%	0.18%
20%	0.09%	0.08%	0.30%	1.68%	0.08%	0.40%
50%	2.43%	0.06%	0.15%	2.81%	0.14%	0.21%
85%	0.97%	0.30%	0.22%	1.07%	0.30%	0.16%
100%	0.40%	0.13%	0.24%	0.80%	0.02%	0.00%

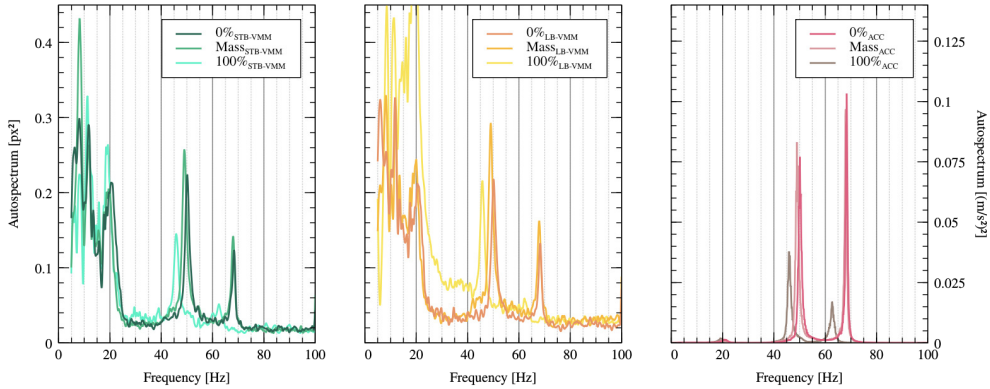
unfair.

In Figure 6b the trend previously observed continues with the three methods being capable of discerning between the presented damage scenarios. This figure presents the different autospectra for the cases where a single column loses 10%, 20%, and 50% of its longitudinal section. This is a challenging test as the resulting spectrums are very close. If the same tests are performed using random noise excitation instead of a periodic sinusoidal signal, the overall conclusions still remain the same, however, a visible increase in noise appears in all methods, as can be observed in Figure 6c.

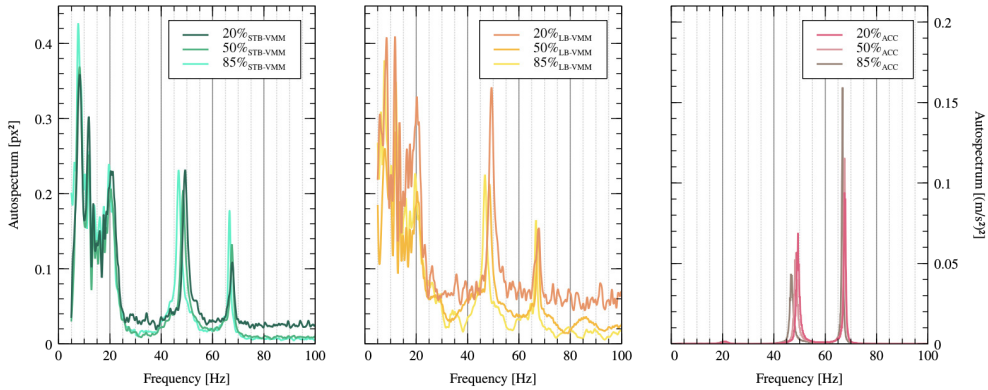
Tables 2-5 gather the numerical results of the experiments performed on the three-story building benchmark as well as the errors of the visual methods with respect to the reference accelerometers. On average, the errors of STB-VMM with respect to the accelerometer data are 0.64% for periodic excitation and 1.03% for random excitation. On the other hand, the same errors for LB-VMM are 0.66% and 1.38% respectively, slightly worse than the results of

Table 4: Results obtained from random excitation

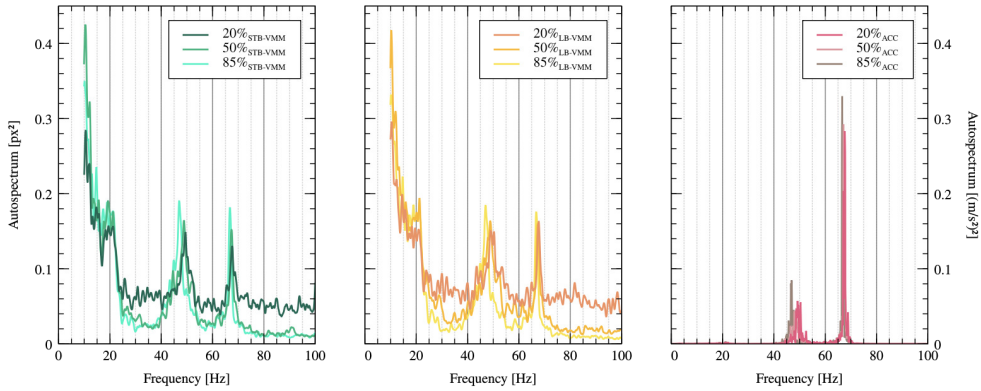
	Accelerometer (Hz)			STB-VMM (Hz)			LB-VMM (Hz)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	22.25	50.38	68.88	22.12	50.35	68.76	21.16	50.25	68.71
Undamaged$_{Repr.}$	22.50	50.38	68.75	22.08	49.66	68.58	22.15	49.98	68.77
Mass	22.25	48.75	67.75	21.16	48.40	67.64	21.16	48.49	67.64
20%	22.25	49.25	67.62	21.35	49.39	67.63	22.18	50.35	67.69
50%	22.25	48.75	67.12	21.23	48.80	67.43	21.47	48.80	67.43
85%	21.12	46.88	66.62	21.01	47.03	66.79	21.50	47.03	66.79
100%	19.75	45.75	62.75	19.73	45.40	62.71	20.90	45.47	62.71



(a) Undamaged, loaded, and missing column damage scenarios performed under periodic excitation ranging from 0 to 100Hz.



(b) Column damage scenarios performed under periodic excitation ranging from 0 to 100Hz.



(c) Column damage scenarios performed under random noise excitation.

Figure 6: Damage sensibility study of the damage scenarios presented in section 3. Each signal corresponds to the second floor of the three-story building benchmark measured under different damage conditions or using different measurement techniques.

Table 5: Relative error to the reference acceleration data when using random excitation

	Error STB-VMM (%)			Error LB-VMM (%)		
	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency	1_{st} order frequency	2_{nd} order frequency	3_{rd} order frequency
Undamaged	0.58%	0.06%	0.17%	4.90%	0.26%	0.25%
Undamaged$_{Repr.}$	1.87%	1.43%	0.25%	1.56%	0.79%	0.03%
Mass	4.90%	0.72%	0.16%	4.90%	0.53%	0.16%
20%	4.04%	0.28%	0.01%	0.31%	2.23%	0.10%
50%	4.58%	0.10%	0.46%	3.51%	0.10%	0.46%
85%	0.52%	0.32%	0.26%	1.80%	0.32%	0.26%
100%	0.10%	0.77%	0.06%	5.82%	0.61%	0.06%

STB-VMM.

4.1.1. Visual noise resilience

To evaluate the effect that visual noise has on VMM measurements, the same video source files used in the previous experiments have been altered by adding white noise. STB-VMM, with its intensive use of image filtering, is expected to perform comparatively better than LB-VMM. Visual noise is a very disruptive element for motion magnification as not only masks sub-pixel motion but destroys key information when the dynamic range is low or noise is sufficiently pronounced. Unfortunately, noise is a general occurrence in all image sensors that gets more pronounced in poorly lit scenes.

Figure 7 showcases the difference in measured frequency response when noise is added to the original recordings. The altered video files have an average Structural Similarity Index Measure (SSIM) [51] of 0.45, which means that the videos have a not insignificant amount of noise while keeping them clearly legible for a human observer. However, as the figure shows, the information loss that occurs is very pronounced, making higher frequencies disappear in the noise and reducing their magnitude.

As was expected, STB-VMM maintains a more stable behaviour than LB-VMM but is still far from the noiseless measurement results. In less extreme situations, STB-VMM is expected to yield better results allowing for the use of more economic image sensors at little or no penalty.

4.1.2. Resolution study

Finally, the effect of reducing the resolution of the source video files has been studied. Like noise, a reduction in resolution deletes information present in the video reducing key features from a few pixels to one or sub-pixel features to variations indistinguishable from noise. This effect is akin to moving the image sensor away from a target subject or reducing the optical zoom. Nevertheless, unlike when adding noise, no spurious results are expected to appear as no new false information is added to the image sequences.

Figure 8 displays the results of reducing the original horizontal resolution by 20%, 50%, and 80%. The videos

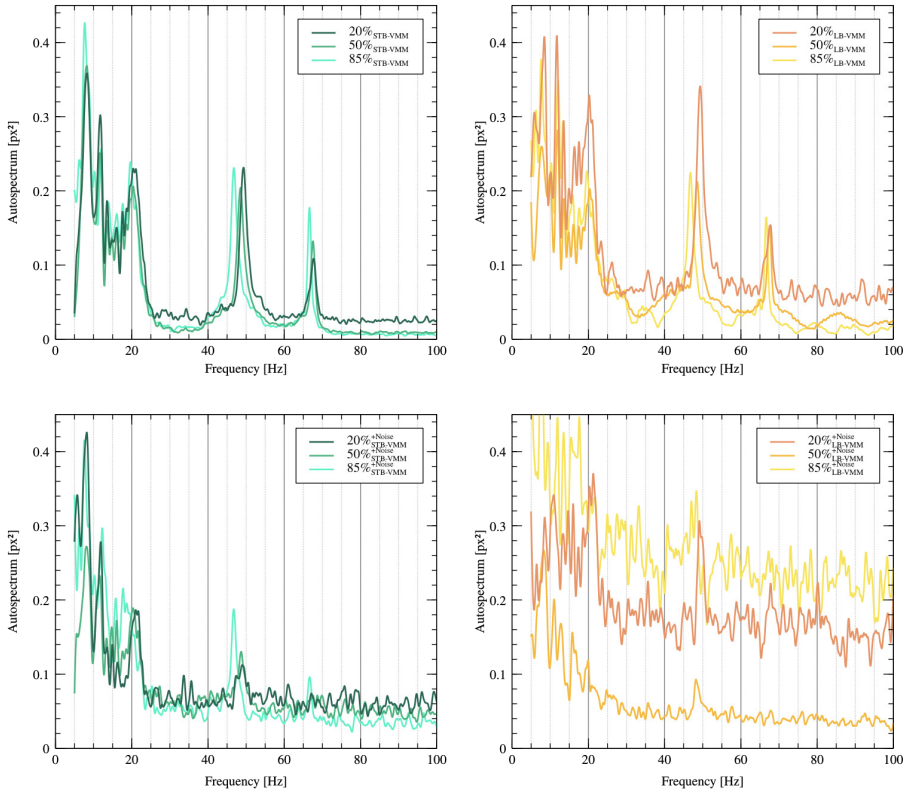


Figure 7: Comparison between the frequency responses obtained from the undamaged video files on top, and the white noise altered ones, at the bottom.

are scaled to maintain the aspect ratio resulting in the following video resolutions: original (1456×384), -20% horizontal resolution (1164×307), -50% horizontal resolution (728×192), and -80% horizontal resolution (291×77). Consequently, in terms of number of pixels, these same sizes contain a 36.1%, 75%, and 96% less pixels than the original.

The results show a clear decrease in magnitude as the resolution decreases, losing the higher frequency peaks progressively to the low-resolution blur. This behaviour makes sense as the higher frequencies manifest as lower amplitude displacements because the excitation power is kept constant in all experiments. When losing pixels to represent movement the shorter displacements evidently end up getting lost beyond what motion magnification can detect.

As shown in Figure 8, differences between STB-VMM and LB-VMM are not particularly significant, unlike when working with noise. In this case, better filtering cannot recover the missing information, and therefore, no significant

improvement is registered.

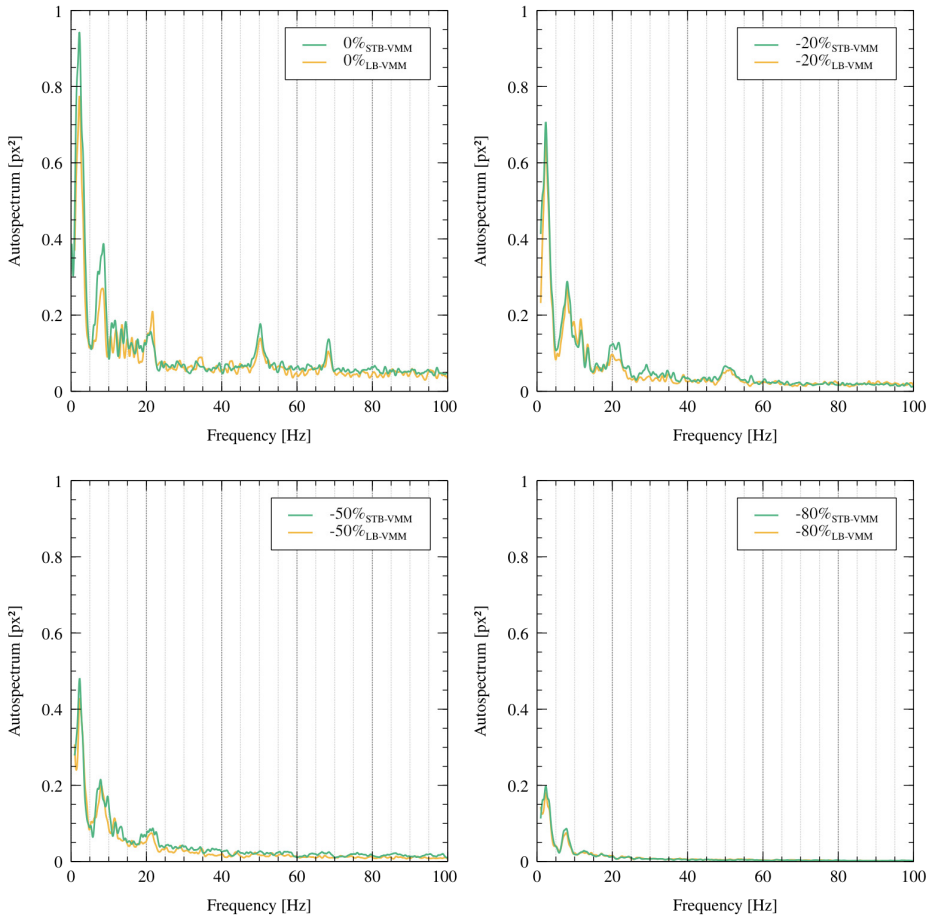


Figure 8: Frequency response graphs of decreasing resolution experiments, from the original sized measure to a reduction of 80% in the horizontal resolution. These measurements correspond to the second floor of the undamaged condition building under periodic excitation.

4.2. Case 2: Reinforce concrete frame structure

The reinforced concrete frame structure offers fewer configuration options and variability than the three-story building benchmark, however, its scale makes it a relevant test for visual-based measurement systems' scalability. Moreover, the videos recorded with a consumer-grade camera by independent third-party researchers improve the credibility of the results.

Figures 9 and 10 show two of the cases shared by Wang et al. [50] and processed using the STB-VMM and

LB-VMM techniques. Cases 1 and 2 show measurements remarkably close to the ground truth data provided, albeit with more noise present. The spectra have been represented from 0 to 2 Hz as the more relevant frequencies are concentrated between 0.6 and 0.9 Hz depending on the floor, which was expected for a structure of its size and rigidity. No notable performance differences appear between the results obtained with STB-VMM and LB-VMM besides some minor differences in noise levels.

Considering the resolution difference between the signal produced by the optical vibrometers at 128 Hz and the camera at 25 Hz, the results are considered very satisfactory. It should be noted that the camera is not in an optimal position for measurement, i.e., directly in front and perpendicular to the measurement plane, nor is it equipped with lenses that do not excessively distort the image. Despite these limitations, the resulting autospectra provides relevant information on the structure's frequencies.

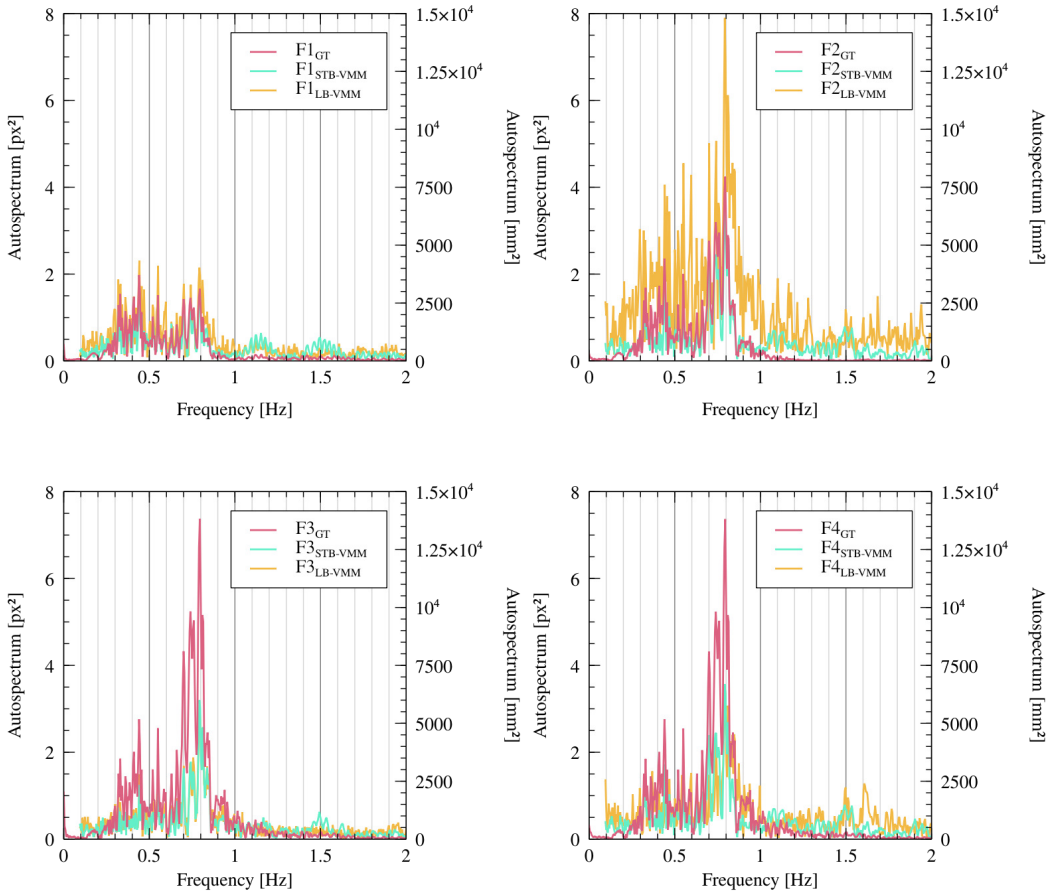


Figure 9: Graphs comparing each floor measurement against case 1's ground truth. [Raw data provided by Wang et al. [50]]

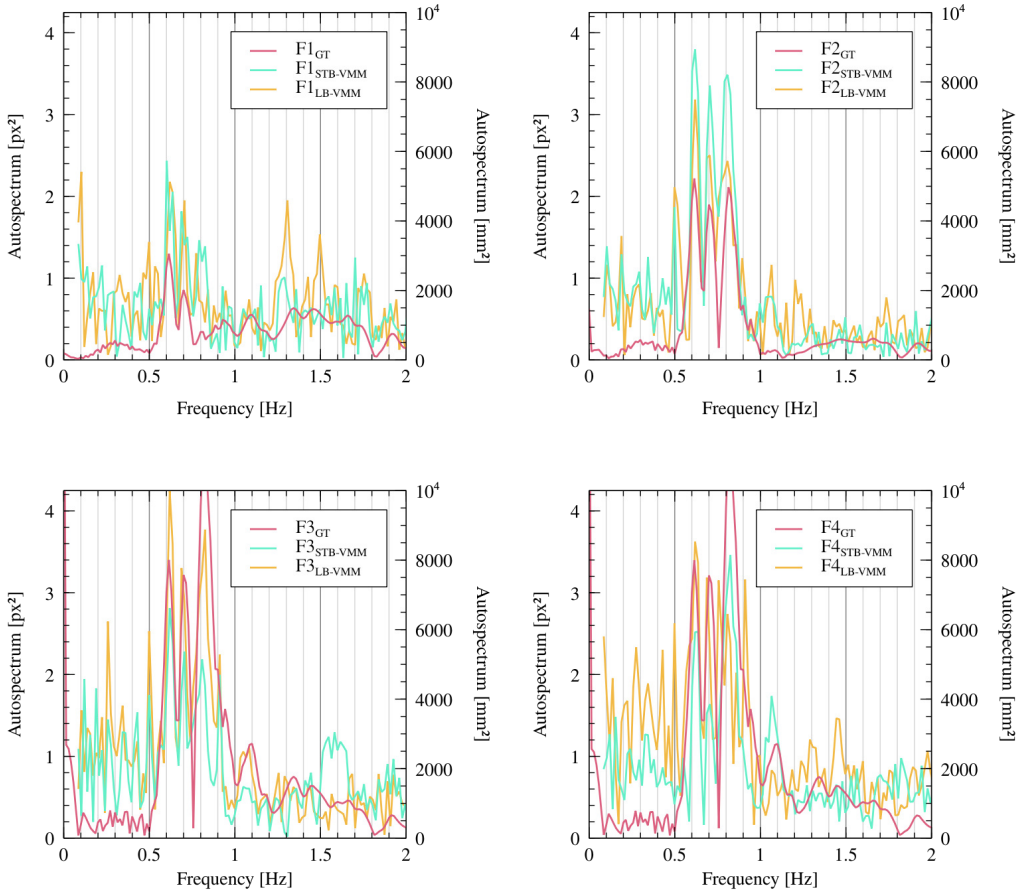


Figure 10: Graphs comparing each floor measurement against case 2's ground truth. [Raw data provided by Wang et al. [50]]

5. Conclusions

This paper validates the use of STB-VMM as a suitable motion magnification framework for SHM applications. Experimental work has shown that the system is capable of identifying a structure's natural frequencies and operating deflection shapes, allowing the algorithm to reliably detect structural damage. Compared to previous research, STB-VMM presents advantages when processing noisy data with respect to LB-VMM, however, in more favourable image conditions, STB-VMM only presents a marginal improvement over LB-VMM. Nevertheless, it is still preferable to use STB-VMM as situations on the field are rarely ideal.

Motion magnification increases the sensitivity of vision-based vibration analysis but comes with limitations, such

as the need for a stationary camera and consistent inter-frame lighting. On the other hand, vision-based techniques such as the ones presented in this work can cut down sensor deployment costs, as a single camera can obtain multiple measurements at once. Of course, the number of useful measurements will depend on the pixel-size of the feature being measured. Therefore, the authors recommend performing a brief feature resolution study on a case-by-case basis in order to obtain suitable results from learning-based motion magnification methods. Feature resolution is a key concept that determines the quality of the output measurements and is affected by many factors such as distance to target, resolution, optical zoom, or sensor bit-depth.

Visual methods present higher noise floors than contact sensors or laser techniques but open up new possibilities for applications and research in structural health monitoring. Newer and cost-effective techniques combined with the transition towards higher digitalization in civil engineering are continuously increasing safety and reducing the costs associated with unplanned maintenance and service disruptions.

Acknowledgements

The authors would like to thank Dr. Chenhui Ren and Prof. Jiazeng Shan from Tongji University for providing the raw data of their reinforced concrete building. Also, the authors would like to gratefully acknowledge the support and funding of the Catalan Agency for Business Competitiveness (ACCIÓ) through the project INNOTEC ISAPREF 2021. Furthermore, the first author would like to acknowledge a Doctoral Scholarship from IQS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could appear to influence the work reported in this paper.

References

[1] C. R. Farrar, K. Worden, An introduction to structural health monitoring, Phil. Trans. R. Soc. A. 365 (1851) (2007) 303–315. doi:10.1098/rsta.2006.1928.

[2] M. Civera, L. Sibille, L. Zanotti Fragonara, R. Ceravolo, A dbscan-based automated operational modal analysis algorithm for bridge monitoring, Meas. 208 (2023) 112451. doi:10.1016/j.measurement.2023.112451.

[3] D. García Cava, L. D. Avendaño-Valencia, A. Movsessian, C. Roberts, D. Tcherniak, On Explicit and Implicit Procedures to Mitigate Environmental and Operational Variabilities in Data-Driven Structural Health Monitoring, Springer International Publishing, 2022, pp. 309–330. doi:10.1007/978-3-030-81716-9_15.

[4] L. D. Avendaño-Valencia, E. N. Chatzi, D. Tcherniak, Gaussian process models for mitigation of operational variability in the structural health monitoring of wind turbines, MSSP 142 (2020) 106686. doi:10.1016/j.ymssp.2020.106686.

[5] J. Font-Moré, G. Reyes-Carmenaty, R. Lado-Roigé, , M. A. Pérez, Performance analysis of vibration-based damage indicators under low-modal information structures, MSSP 190 (2023) 110166. doi:10.1016/j.ymssp.2023.110166.

[6] G. Reyes-Carmenaty, J. Font-Moré, R. Lado-Roigé, M. A. Pérez, Cross-domain transfer learning for vibration-based structural damage classification via convolutional neural networks, Eng Appl Artif Intell(Under review) (2023).

[7] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, D. J. Inman, A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications, MSSP 147 (2021) 107077. arXiv:2004.04373, doi:10.1016/j.ymssp.2020.107077.

[8] D. Tcherniak, L. L. Mølgaard, Active vibration-based structural health monitoring system for wind turbine blade: Demonstration on an operating vestas v27 wind turbine, Struct. Health Monit 16 (5) (2017) 536–550. doi:10.1177/1475921717722725.

[9] O. Avci, M. Gül, N. Catbas, O. Celik, T. Ince, S. Kiranyaz, Real-Time and Web-Based Structural Damage Detection Network for Multiple Structures, 2023, pp. 41–48. doi:10.1007/978-3-031-05449-5_5.

[10] M. Zakaria, E. Karaaslan, N. Catbas, Real-time ai-based bridge inspection using mixed reality platform, in: Structures Congress 2023, 2023, pp. 120–131. doi:10.1061/9780784484777.012.

[11] C. Iacovino, Z. I. Turksezer, P. F. Giordano, M. P. Limongelli, Comparison of bridge inspection policies in terms of data quality, J. Bridge Eng 27 (3) (2022) 04021115. doi:10.1061/(ASCE)BE.1943-5592.0001831.

[12] A. Strauss, S. Kessler, M. P. Limongelli, K. Bergmeister, Existing concrete structures: Life management, testing and structural health monitoring, Struct. Concr 21 (4) (2020) 1212–1212. doi:10.1002/suco.202070042.

[13] Z. H. Warsi, S. M. Irshad, F. Khan, M. A. Shahbaz, M. Junaid, S. U. Amin, Sensors for structural health monitoring: A review, in: 2019 INTELLECT, 2019, pp. 1–6. doi:10.1109/INTELLECT47034.2019.8955453.

[14] F. Ksica, Z. Hadas, J. Hlinka, Application of piezoelectric sensors for structural health monitoring in aerospace, in: 2018 5th IEEE MetroAeroSpace, 2018, pp. 596–601. doi:10.1109/MetroAeroSpace.2018.8453610.

[15] D. Feng, M. Q. Feng, Computer vision for shm of civil infrastructure: From dynamic response measurement to damage detection – a review, Eng. Struct.s 156 (2018) 105–117. doi:10.1016/j.engstruct.2017.11.018.

[16] S. Patsias, W. J. Staszewskiy, Damage detection using optical measurements and wavelets, Struct. Health Monit. 1 (1) (2002) 5–22. doi:10.1177/147592170200100102.

[17] A. Wahbeh, J. Caffrey, S. Masri, A vision-based approach for the direct measurement of displacements in vibrating systems, Smart Mater.s Struct.s 12 (5) (2003) 785–794. doi:10.1088/0964-1726/12/5/016.

[18] J. J. Lee, M. Shinozuka, A vision-based system for remote sensing of bridge displacement, NDT & E Int. 39 (5) (2006) 425–431. doi:10.1016/j.ndteint.2005.12.003.

[19] J. G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W. T. Freeman, O. Buyukozturk, Modal identification of simple structures with high-speed video using motion magnification, J. Sound Vib. 345 (2015) 58–71. doi:10.1016/j.jsv.2015.01.024.

[20] J. G. Chen, N. Wadhwa, F. Durand, W. T. Freeman, O. Buyukozturk, Developments with motion magnification for structural modal identification through camera video, in: J. Caicedo, S. Pakzad (Eds.), Dyn. Civil Struct.s, (2), 2015, pp. 49–57. doi:10.1007/978-3-319-15248-6_5.

[21] J. G. Chen, A. Davis, N. Wadhwa, Video camera-based vibration measurement for condition assessment of civil infrastructure (2015).

[22] M. Lu, Y. Chai, Q. Liu, Observation of tower vibration based on subtle motion magnification, IFAC-PapersOnLine 52 (24) (2019) 346–350. doi:10.1016/j.ifacol.2019.12.434.

[23] J. G. Chen, A. Davis, N. Wadhwa, F. Durand, W. T. Freeman, O. Büyüköztürk, Video camera–based vibration measurement for civil infrastructure applications, J. Infrastructure Syst. 23 (3) (2017) B4016013. doi:10.1061/(ASCE)IS.1943-555X.0000348.

[24] R. Lado-Roigé, J. Font-Moré, M. A. Pérez, Learning-based video motion magnification approach for vibration-based damage detection, Meas. 206 (2023) 112218. doi:10.1016/j.measurement.2022.112218.

[25] R. Lado-Roigé, M. A. Pérez, Vimag: A visual vibration analysis toolbox, JOSS. (Submitted to journal) (2023).

[26] O. Shabi, S. Natan, A. Kolel, A. Mukherjee, O. Tchaicheeyan, H. Wolfenson, N. Kiryati, A. Lesman, Motion magnification analysis of microscopy videos of biological cells, PLOS ONE 15 (11) (2020) 1–18. doi:10.1371/journal.pone.0240127.

[27] A. J. McLeod, J. S. Baxter, S. de Ribaupierre, T. M. Peters, Motion magnification for endoscopic surgery, in: Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 9036, SPIE, 2014, pp. 81–88. doi:/10.1117/12.2043997.

[28] H. Lauridsen, S. Gonzales, D. Hedwig, K. L. Perrin, C. J. Williams, P. H. Wrege, M. F. Bertelsen, M. Pedersen, J. T. Butcher, Extracting physiological information in experimental biology via eulerian video magnification, BMC Biol. 17 (1) (2019) 1–26. doi:10.1186/s12915-019-0716-7.

[29] J. Fei, Z. Xia, P. Yu, F. Xiao, Exposing ai-generated videos with motion magnification, Multimed Tools Appl 80 (20) (2021) 30789–30802. doi:10.1007/s11042-020-09147-3.

[30] N. Wadhwa, J. G. Chen, J. B. Sellon, D. Wei, M. Rubinstein, R. Ghaffari, D. M. Freeman, O. Büyüköztürk, P. Wang, S. Sun, S. H. Kang, K. Bertoldi, F. Durand, W. T. Freeman, Motion microscopy for visualizing and quantifying small motions, Proc. Natl. Acad. Sci 114 (44) (2017) 11639–11644. doi:10.1073/pnas.1703715114.

[31] C. Liu, A. Torralba, W. T. Freeman, F. Durand, E. H. Adelson, Motion magnification, in: ACM SIGGRAPH 2005 Papers, 2005, p. 519–526. doi:10.1145/1186822.1073223.

[32] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. (2012). doi:10.1145/2185520.2185561.

[33] N. Wadhwa, M. Rubinstein, F. Durand, W. T. Freeman, Phase-based video motion processing, ACM Trans. Graph. 32 (4) (2013) 1–10. doi:10.1145/2461912.2461966.

[34] N. Wadhwa, M. Rubinstein, F. Durand, W. Freeman, Riesz pyramids for fast phase-based video magnification, in: 2014 IEEE ICCP, 2014, pp. 1–10. doi:10.1109/ICCPHOT.2014.6831820.

[35] Y. Zhang, S. L. Pintea, J. C. van Gemert, Video acceleration magnification, CVPR (2017). doi:10.48550/arXiv.1704.04186.

[36] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, W. Matusik, Learning-based video motion magnification (2018). doi:10.48550/arXiv.1804.02684.

[37] R. Lado-Roigé, M. A. Pérez, Stb-vmm: Swin transformer based video motion magnification, Knowl.-Based Syst. (2023) 110493doi:10.1016/j.knosys.2023.110493.

[38] N. Wadhwa, H.-Y. Wu, A. Davis, M. Rubinstein, E. Shih, G. J. Mysore, J. G. Chen, O. Buyukozturk, J. V. Guttag, W. T. Freeman, F. Durand, Eulerian video magnification and analysis, Commun. ACM 60 (1) (2016) 87–95. doi:10.1145/3015573.

[39] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context (2015). doi:10.48550/arXiv.1405.0312.

[40] M. Everingham, L. Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338. doi:10.1007/s11263-009-0275-4.

[41] E. Figueiredo, G. Park, J. Figueiras, C. Farrar, K. Worden, Structural health monitoring algorithm comparisons using standard data sets, Tech. Rep. LA-14393, 961604 (2009). doi:10.2172/961604.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows (2021). doi:10.48550/ARXIV.2103.14030.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszko-

reit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2020). doi:10.48550/ARXIV.2010.11929.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2017). doi:10.48550/ARXIV.1706.03762.

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv:1810.04805.

[46] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, J. Wang, Release strategies and the social impacts of language models (2019). arXiv:1908.09203.

[47] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners (2020). arXiv:2005.14165.

[48] OpenAI, Gpt-4 technical report (2023). arXiv:2303.08774.

[49] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better (2021). doi:10.48550/ARXIV.2106.14881.

[50] J. Wang, J. Zhao, Y. Liu, J. Shan, Vision-based displacement and joint rotation tracking of frame structure using feature mix with single consumer-grade camera, Struct Control Health Monit 28 (12) (2021). doi:10.1002/stc.2832.

[51] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

References

1. Chen, H.-P. & Ni, Y.-Q. in *Structural Health Monitoring of Large Civil Engineering Structures* 1–14 (John Wiley & Sons, Ltd, 2018). doi:10.1002/9781119166641.ch1.
2. Wang, H. & Raj, B. *On the Origin of Deep Learning* 2017. arXiv: 1702.07800.
3. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79,** 2554–2558. doi:10.1073/pnas.79.8.2554 (1982).
4. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition* (Pearson Education, 2021).
5. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks* in. **60** (Association for Computing Machinery, New York, NY, USA, 2017), 84–90. doi:10.1145/3065386.
6. OpenAI. *GPT-4 Technical Report* 2023. arXiv: 2303.08774.
7. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* 2019. arXiv: 1810.04805.
8. Xu, H., Durme, B. V. & Murray, K. *BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation* 2021. arXiv: 2109.04588.
9. Bojarski, M. *et al. End to End Learning for Self-Driving Cars* 2016. arXiv: 1604.07316.
10. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65,** 386–408. doi:10.1037/h0042519 (1958).
11. Auer, P., Burgsteiner, H. & Maass, W. A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Netw* **21,** 786–795. doi:10.1016/j.neunet.2007.12.036 (2008).
12. Dubey, S. R., Singh, S. K. & Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **503,** 92–108. doi:10.1016/j.neucom.2022.06.111 (2022).
13. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778. doi:10.1109/cvpr.2016.90.
14. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 4700–4708. doi:10.1109/cvpr.2017.243.

15. Long, J., Shelhamer, E. & Darrell, T. *Fully Convolutional Networks for Semantic Segmentation* 2015. arXiv: 1411.4038.

16. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *International Conference on Medical image computing and computer-assisted intervention* (2015), 234–241. doi:10.1007/978-3-319-24574-4_28.

17. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation* 2018. arXiv: 1807.10165.

18. Cao, H. *et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation* 2021. arXiv: 2105.05537.

19. Wang, L. *et al.* UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **190,** 196–214. doi:10.1016/j.isprsjprs.2022.06.008 (2022).

20. Lecun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. English (US). *Neural Comput.* **1,** 541–551. doi:10.1162/neco.1989.1.4.541. (1989).

21. Lipton, Z. C., Berkowitz, J. & Elkan, C. *A Critical Review of Recurrent Neural Networks for Sequence Learning* 2015. arXiv: 1506.00019.

22. Lin, T., Wang, Y., Liu, X. & Qiu, X. *A Survey of Transformers* 2021. arXiv: 2106.04554.

23. Vaswani, A. *et al. Attention Is All You Need* 2017. arXiv: 1706.03762.

24. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* 2015. arXiv: 1512.03385.

25. Xu, J., Sun, X., Zhang, Z., Zhao, G. & Lin, J. *Understanding and Improving Layer Normalization* 2019. arXiv: 1911.07013.

26. Islam, S. *et al. A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks* 2023. arXiv: 2306.07303.

27. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* 2020. arXiv: 2010.11929.

28. Liu, Z. *et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* 2021. arXiv: 2103.14030.

29. Shallue, C. J. *et al.* Measuring the Effects of Data Parallelism on Neural Network Training. *J. Mach. Learn. Res.* **20,** 1–49. arXiv: 1811.03600 (2019).

30. Choi, D. *et al. On Empirical Comparisons of Optimizers for Deep Learning* 2020. arXiv: 1910.05446.

31. Godbole, V., Dahl, G. E., Gilmer, J., Shallue, C. J. & Nado, Z. *Deep Learning Tuning Playbook* 2023. https://github.com/google-research/tuning_playbook.

32. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5,** 115–133. doi:10.1007/bf02478259 (1943).

33. Hebb, D. O. *The organization of behavior; a neuropsychological theory* doi:10.2307/1418888 (Wiley, Oxford, England, 1949).
34. Bain, A. *Mind and body. The theories of their relation* eng. doi:10.1037/12850-000 (New York : D. Appleton and company, 1873).
35. Minsky, M. & Papert, S. *Perceptrons* doi:10.7551/mitpress/11301.001.0001 (M.I.T. Press, Oxford, England, 1969).
36. Minsky, M. & Papert, S. A. *Perceptrons: An Introduction to Computational Geometry* doi:10.7551/mitpress/11301.001.0001 (The MIT Press, 2017).
37. Kawaguchi, K. A multithreaded software model for backpropagation neural network applications. *ETD Collection for University of Texas, El Paso,* 1–92 (2000).
38. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for boltzmann machines. *Cognitive Sci.* **9,** 147–169. doi:10.1016/s0364-0213(85)80012-4 (1985).
39. Smolensky, P. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* 194–281 (MIT Press, Cambridge, MA, USA, 1986).
40. Jordan, M. I. Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986. https://www.osti.gov/biblio/6910294 (1986).
41. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. en. *Biol. Cybernetics* **36,** 193–202. doi:10.1007/bf00344251 (1980).
42. Salakhutdinov, R. & Hinton, G. E. *Deep Boltzmann Machines* in *International Conference on Artificial Intelligence and Statistics* (2009).
43. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. *Improving neural networks by preventing co-adaptation of feature detectors* 2012. arXiv: 1207.0580.
44. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks* in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Curran Associates Inc., Lake Tahoe, Nevada, 2012), 1097–1105.
45. Russakovsky, O. *et al. ImageNet Large Scale Visual Recognition Challenge* 2014. arXiv: 1409.0575.
46. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* 2015. arXiv: 1409.1556.
47. Szegedy, C. *et al. Going Deeper with Convolutions* 2014. arXiv: 1409.4842.
48. Girshick, R., Donahue, J., Darrell, T. & Malik, J. *Rich feature hierarchies for accurate object detection and semantic segmentation* 2014. arXiv: 1311.2524.
49. Liu, Z., Wu, Z. & Tóth, R. *SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation* 2020. arXiv: 2002.10111.

50. Werbos, P. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting* (Wiley, 1994).

51. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9,** 1735–1780. doi:10.1162/neco.1997.9.8.1735 (1997).

52. Hinton, G. E., Osindero, S. & Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **18,** 1527–1554. doi:10.1162/neco.2006.18.7.1527 (2006).

53. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *CVPR* (2009), 248–255. doi:10.1109/cvpr.2009.5206848.

54. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* 2022. arXiv: 1312.6114.

55. Ioffe, S. & Szegedy, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* 2015. arXiv: 1502.03167.

56. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. *Rethinking the Inception Architecture for Computer Vision* 2015. arXiv: 1512.00567.

57. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning* 2016. arXiv: 1602.07261.

58. Goodfellow, I. J. *et al. Generative Adversarial Networks* 2014. arXiv: 1406.2661.

59. Solaiman, I. *et al. Release Strategies and the Social Impacts of Language Models* 2019. arXiv: 1908.09203.

60. Brown, T. B. *et al. Language Models are Few-Shot Learners* 2020. arXiv: 2005.14165.

61. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* 2018. arXiv: 1608.06993.

62. Howard, A. G. *et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* 2017. arXiv: 1704.04861.

63. Wang, J. *et al. Deep High-Resolution Representation Learning for Visual Recognition* 2020. doi:10.1109/tpami.2020.2983686. arXiv: 1908.07919.

64. Liu, Z. *et al. A ConvNet for the 2020s* 2022. arXiv: 2201.03545.

65. Liang, J. *et al. SwinIR: Image Restoration Using Swin Transformer* 2021. arXiv: 2108.10257.

66. Lado-Roigé, R. & Pérez, M. A. STB-VMM: Swin Transformer based Video Motion Magnification. *Knowl.-Based Syst.,* 110493. doi:10.1016/j.knosys.2023.110493 (2023).

67. Zamir, S. W. *et al. Restormer: Efficient Transformer for High-Resolution Image Restoration* 2022. arXiv: 2111.09881.

68. Touvron, H., Cord, M. & Jegou, H. DeiT III: Revenge of the ViT. doi:10.1007/978-3-031-20053-3_30 (2022).

69. Dang, Y., Hu, Z., Cranmer, M., Eickenberg, M. & Ho, S. *TNT: Vision Transformer for Turbulence Simulations* 2022. arXiv: 2207.04616.

70. Wang, W. *et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions* in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 548–558. doi:10.1109/iccv48922.2021.00061.

71. Touvron, H. *et al. Training data-efficient image transformers and distillation through attention* in *International Conference on Machine Learning* **139** (2021), 10347–10357.

72. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. & Jégou, H. *Going Deeper With Image Transformers* in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 32–42. doi:10.1109/iccv48922.2021.00010.

73. Touvron, H., Cord, M., El-Nouby, A., Verbeek, J. & Jegou, H. Three things everyone should know about Vision Transformers. doi:10.1007/978-3-031-20053-3_29 (2022).

74. Ranftl, R., Bochkovskiy, A. & Koltun, V. *Vision Transformers for Dense Prediction* 2021. arXiv: 2103.13413.

75. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C. & Foresti, G. L. *VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization* in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)* (IEEE, 2021). doi:10.1109/isie45552.2021.9576231.

76. Arnab, A. *et al. ViViT: A Video Vision Transformer* 2021. arXiv: 2103.15691.

77. Dong, X. *et al. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows* 2022. arXiv: 2107.00652.

78. Zhang, Y. *et al. KBNet: Kernel Basis Network for Image Restoration* 2023. arXiv: 2303.02881.

79. Park, N. & Kim, S. *How Do Vision Transformers Work?* 2022. arXiv: 2202.06709.

80. Naseer, M. *et al. Intriguing Properties of Vision Transformers* 2021. arXiv: 2105.10497.

81. Hermann, K. L., Chen, T. & Kornblith, S. *The Origins and Prevalence of Texture Bias in Convolutional Neural Networks* 2020. arXiv: 1911.09071.

82. Pinto, F., Torr, P. H. S. & Dokania, P. K. *An Impartial Take to the CNN vs Transformer Robustness Contest* en. 2022. arXiv: 2207.11347.

83. Bai, Y., Mei, J., Yuille, A. & Xie, C. *Are Transformers More Robust Than CNNs?* en. 2021. arXiv: 2111.05464.

84. Wang, Z., Bai, Y., Zhou, Y. & Xie, C. *Can CNNs Be More Robust Than Transformers?* en. 2023. arXiv: 2206.03452.

85. Rubinstein, M., Wadhwa, N., Durand, F. & Freeman, W. T. Revealing Invisible Changes in the World. *Science* **339,** 519–519 (2013).

86. Verkruysse, W., Svaasand, L. O. & Nelson, J. S. Remote plethysmographic imaging using ambient light. *Opt. Express* **16,** 21434–21445. doi:10.1364/oe.16.021434 (2008).

87. McLeod, A. J., Baxter, J. S., de Ribaupierre, S. & Peters, T. M. *Motion magnification for endoscopic surgery* in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling* **9036** (2014), 81–88. doi:10.1117/12.2043997.

88. Poh, M.-Z., McDuff, D. J. & Picard, R. W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **18,** 10762–10774. doi:10.1364/oe.18.010762 (2010).

89. Mohsen, M. S., Fakhri, A. B., Ahmed, N. M., Mahmood, M. F. & Mohammed, S. L. Video Magnification Techniques: Medical Applications and Comparison of Methods. *IOP Conf. Ser.: Mater. Sci. Eng.* **1105,** 012074. doi:10.1088/1757-899x/1105/1/012074 (2021).

90. Lauridsen, H. *et al.* Extracting physiological information in experimental biology via Eulerian video magnification. *BMC Biol.* **17,** 1–26. doi:10.1186/s12915-019-0716-7 (2019).

91. Civera, M., Zanotti Fragonara, L., Antonaci, P., Anglani, G. & Surace, C. An Experimental Validation of Phase-Based Motion Magnification for Structures with Developing Cracks and Time-Varying Configurations. en. *Shock Vib.* **2021,** 5518163. doi:10.1155/2021/5518163 (2021).

92. Chen, J. G., Wadhwa, N., Durand, F., Freeman, W. T. & Buyukozturk, O. *Developments with Motion Magnification for Structural Modal Identification Through Camera Video* en. in *Dyn. Civil Struct.s, (2)* (eds Caicedo, J. & Pakzad, S.) (2015), 49–57. doi:10.1007/978-3-319-15248-6_5.

93. Chen, J. G. *et al.* Modal identification of simple structures with high-speed video using motion magnification. en. *J. Sound Vib.* **345,** 58–71. doi:10.1016/j.jsv.2015.01.024 (2015).

94. Chen, J. G. *et al.* Video Camera–Based Vibration Measurement for Civil Infrastructure Applications. EN. *J. Infrastructure Syst.* **23,** B4016013. doi:10.1061/(asce)is.1943-555x.0000348 (2017).

95. Chen, J. G., Adams, T. M., Sun, H., Bell, E. S. & Büyüköztürk, O. Camera-Based Vibration Measurement of the World War I Memorial Bridge in Portsmouth, New Hampshire. *J Struct Eng* **144,** 04018207. doi:10.1061/(asce)st.1943-541x.0002203 (2018).

96. Cosco, F., Cuenca, J., Desmet, W., Janssens, K. & Mundo, D. Towards phase-based defect detection: A feasibility study in vibrating panels. *J. Sound Vib.* **537,** 117196. doi:10.1016/j.jsv.2022.117196 (2022).

97. Wadhwa, N. *et al.* Motion microscopy for visualizing and quantifying small motions. *Proc. Natl. Acad. Sci* **114,** 11639–11644. doi:10.1073/pnas.1703715114 (2017).

98. Shabi, O. *et al.* Motion magnification analysis of microscopy videos of biological cells. *PLOS One* **15,** 1–18. doi:10.1371/journal.pone.0240127 (2020).

99. Zhang, J. *et al.* Motion magnification multi-feature relation network for facial microexpression recognition. *Complex Intell. Syst.* **8,** 3363–3376. doi:10.1007/s40747-022-00680-2 (2022).

100. Fei, J., Xia, Z., Yu, P. & Xiao, F. Exposing AI-generated videos with motion magnification. *Multimed. Tools Appl.* **80,** 30789–30802. doi:10.1007/s11042-020-09147-3 (2021).

101. Liu, C., Torralba, A., Freeman, W. T., Durand, F. & Adelson, E. H. *Motion Magnification* in *ACM SIGGRAPH 2005 Papers* (Los Angeles, California, 2005), 519–526. doi:10.1145/1186822.1073223.

102. Zhai, M., Xiang, X., Lv, N. & Kong, X. Optical flow and scene flow estimation: A survey. en. *Pattern Recogn.* **114,** 107861. doi:10.1016/j.patcog.2021.107861 (2021).

103. Wu, H.-Y. *et al.* Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* doi:10.1145/2185520.2185561 (2012).

104. Serres, J. R. & Ruffier, F. Optic flow-based collision-free strategies: From insects to robots. en. *Arthropod Struct. Dev. From Insects to Robots* **46,** 703–717. doi:10.1016/j.asd.2017.06.003 (2017).

105. Zhu, J., Lu, Z. & Zhang, C. A marker-free method for structural dynamic displacement measurement based on optical flow. *Struct. Infrastruct. E.* **18,** 84–96. doi:10.1080/15732479.2020.1835999 (2022).

106. Feng, D. & Feng, M. Q. Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection – A review. en. *Eng. Struct.s* **156,** 105–117. doi:10.1016/j.engstruct.2017.11.018 (2018).

107. He, M., Zhu, C., Huang, Q., Ren, B. & Liu, J. A review of monocular visual odometry. en. *Vis Comput* **36,** 1053–1065. doi:10.1007/s00371-019-01714-6 (2020).

108. Pandey, T., Pena, D., Byrne, J. & Moloney, D. Leveraging Deep Learning for Visual Odometry Using Optical Flow. en. *Sensors* **21,** 1313. doi:10.3390/s21041313 (2021).

109. Wadhwa, N., Rubinstein, M., Durand, F. & Freeman, W. *Riesz pyramids for fast phase-based video magnification* in *2014 IEEE ICCP* (2014), 1–10. doi:10.1109/iccphot.2014.6831820.

110. Zhang, Y., Pintea, S. L. & van Gemert, J. C. *Video Acceleration Magnification* 2017. arXiv: 1704.04186.

111. Wadhwa, N., Rubinstein, M., Durand, F. & Freeman, W. T. Phase-based video motion processing. en. *ACM Trans. Graph.* **32,** 1–10. doi:10.1145/2461912.2461966 (2013).

112. Oh, T.-H. *et al.* Learning-based Video Motion Magnification. en. arXiv: 1804.02684 (2018).

113. Rytter, A. *Vibrational Based Inspection of Civil Engineering Structures* English. PhD thesis (Denmark, 1993).

114. Park, G. & Inman, D. Smart bolts: an example of self-healing structures. *Smart Materials Bulletin* **2001,** 5–8. doi:10.1016/s1471-3918(01)80140-4 (2001).

115. Farrar, C. R., Doebling, S. W. & Nix, D. A. Vibration–based structural damage identification. en. *Philos. trans., Math. phys. eng. sci.* **359** (eds Lieven, N. A. J. & Ewins, D. J.) 131–149. doi:10.1098/rsta.2000.0717 (2001).

116. Farrar, C. R. & Worden, K. An introduction to structural health monitoring. en. *Phil. Trans. R. Soc. A.* **365,** 303–315. doi:10.1098/rsta.2006.1928 (2007).

117. Zhang, C. *et al.* Vibration feature extraction using signal processing techniques for structural health monitoring: A review. *Mech. Syst. Signal Pr.* **177,** 109175. doi:10.1016/j.ymssp.2022.109175. (2023) (2022).

118. Zaparoli Cunha, B., Droz, C., Zine, A.-M., Foulard, S. & Ichchou, M. A review of machine learning methods applied to structural dynamics and vibroacoustic. *Mech. Syst. Signal Pr.* **200,** 110535. doi:10.1016/j.ymssp.2023.110535. (2023) (2023).

119. Doebling, S. W., Farrar, C. R. & Prime, M. B. A summary review of vibration-based damage identification methods. *Shock Vibr. Dig.* **30,** 91–105. doi:10.1177/058310249803000201 (1998).

120. Worden, K., Staszewski, W. J. & Hensman, J. J. Natural computing for mechanical systems research: A tutorial overview. *Mech. Syst. Signal Pr.* **25,** 4–111. doi:10.1016/j.ymssp.2010.07.013 (2011).

121. Farrar, C. R., Sohn, H. & Worden, K. Data normalization : A key for structural health monitoring (2001).

122. Worden, K., Sohn, H. & Farrar, C. Novelty detection in a changing environment: Regression and interpolation approaches. *J. Sound Vib.* **258,** 741–761. doi:10.1006/jsvi.2002.5148 (2002).

123. Peeters, B. & De Roeck, G. One-year monitoring of the Z24-Bridge: environmental effects versus damage events. *Earthquake Eng. Struc.* **30,** 149–171. doi:10.1002/1096-9845(200102)30:2<149::aid-eqe1>3.0.co;2-z (2001).

124. Farrar, C. R. & Worden, K. in *New Trends in Vibration Based Structural Health Monitoring* 1–17 (Springer Vienna, Vienna, 2010). doi:10.1007/978-3-7091-0399-9_1.

125. Fassois, S. D. & Sakellariou, J. S. Time-Series Methods for Fault Detection and Identification in Vibrating Structures. *Philos. trans., Math. phys. eng. sci.* **365,** 411–448 (2007).

126. Friswell, M. I. Damage identification using inverse methods. *Philos. Trans. Royal Soc. A* **365,** 393–410. doi:10.1098/rsta.2006.1930 (2006).

127. Worden, K. & Manson, G. The Application of Machine Learning to Structural Health Monitoring. *Philos. trans., Math. phys. eng. sci.* **365,** 515–537 (2007).

128. Doebling, S. W., Farrar, C. R., Prime, M. B. & Shevitz, D. W. *Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: A literature review* English. Tech. rep. LA-13070-MS (Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 1996). doi:10.2172/249299.

129. Salawu, O. S. Detection of structural damage through changes in frequency: a review. en. *Eng. Struct.* **19,** 718–723. doi:10.1016/s0141-0296(96)00149-6 (1997).

130. Roeck, G. D. The state-of-the-art of damage detection by vibration monitoring: the SIMCES experience. en. *J. Struct. Control* **10,** 127–134. doi:10.1002/stc.20 (2003).

131. Carden, E. P. & Fanning, P. Vibration Based Condition Monitoring: A Review. en. *Struct. Health Monit.* **3,** 355–377. doi:10.1177/1475921704047500 (2004).

132. D.J. Inman C.R. Farrar, V. L. & Steffen, V. en. in *Damage Prognosis* i–xix (John Wiley & Sons, Ltd, 2005). doi:10.1002/0470869097.fmatter.

133. Fritzen, C.-P. Vibration-Based Structural Health Monitoring – Concepts and Applications. *Key Eng. Mat.* **293.** doi:10.4028/www.scientific.net/kem.293-294.3 (2005).

134. Fritzen, C.-P. en. in *Structural Health Monitoring* 45–224 (John Wiley & Sons, Ltd, 2006). doi:10.1002/9780470612071.ch2.

135. García Cava, D. *Data-based vibration structural health monitoring methodology for composite laminated structures* PhD thesis (2016).

136. Font-Moré, J., Reyes-Carmenaty, G., Lado-Roigé, R. & Pérez, M. A. Performance analysis of vibration-based damage indicators under low-modal information structures. *Mech. Syst. Signal Pr.* **190,** 110166. doi:10.1016/j.ymssp.2023.110166 (2023).

137. Font-Moré, J. & Pérez, M. A. *A Performance Metric to Evaluate Frequency-Based Damage Indicators* en. in *European Workshop on Structural Health Monitoring* (Springer International Publishing, Cham, 2023), 485–494. doi:10.1007/978-3-031-07254-3_49.

138. García Cava, D., Avendaño-Valencia, L. D., Movsessian, A., Roberts, C. & Tcherniak, D. en. in *Structural Health Monitoring Based on Data Science Techniques* 309–330 (Springer International Publishing, Cham, 2022). doi:10.1007/978-3-030-81716-9_15.

139. Hu, W.-H. *et al.* Comparison of different statistical approaches for removing environmental/operational effects for massive data continuously collected from footbridges. en. *Struct Control Health Monit* **24,** e1955. doi:10.1002/stc.1955 (2017).

140. Oliveira, G., Magalhães, F., Cunha, Á. & Caetano, E. Vibration-based damage detection in a wind turbine using 1 year of data. en. *Struct Control Health Monit* **25,** e2238. doi:10.1002/stc.2238 (2018).

141. Fassois, S. D. & Kopsaftopoulos, F. P. en. in *New Trends in Structural Health Monitoring* 209–264 (Springer, Vienna, 2013). doi:10.1007/978-3-7091-1390-5_4.

142. Peter Carden, E. & Brownjohn, J. M. W. ARMA modelled time-series classification for structural health monitoring of civil infrastructure. en. *Mech. Syst. Signal Pr.* **22,** 295–314. doi:10.1016/j.ymssp.2007.07.003 (2008).

143. Sohn, H., Czarnecki, J. A. & Farrar, C. R. Structural Health Monitoring Using Statistical Process Control. en. *J. Struct. Eng.* **126,** 1356–1363. doi:10.1061/(asce)0733-9445(2000)126:11(1356) (2000).

144. Basseville, M., Abdelghani, M. & Benveniste, A. Subspace-based fault detection algorithms for vibration monitoring. en. *Automatica J. IFAC* **36,** 101–109. doi:10.1016/s0005-1098(99)00093-x (2000).

145. Mahdavi, S. H. & Xu, C. Time-Domain Structural Damage Identification Using Ensemble Bagged Trees and Evolutionary Optimization Algorithms. *Struct Control Health Monit* **2023,** 6321012. doi:10.1155/2023/6321012 (2023).

146. Zou, Y., Lu, X., Yang, J., Wang, T. & He, X. Structural Damage Identification Based on Transmissibility in Time Domain. *Sensors* **22.** doi:10.3390/s22010393 (2022).

147. Lee, D., Ahn, T.-S. & Kim, H.-S. A metric on the similarity between two frequency response functions. en. *J. Sound Vib.* **436,** 32–45. doi:10.1016/j.jsv.2018.08.051 (2018).

148. Allemang, R. The modal assurance criterion–Twenty years of use and abuse. *Sound Vib.* (2003).

149. Pascual, R., Golinval, J. & Razeto, M. *A frequency domain correlation technique for model correlation and updating* in (1997).

150. Sampaio, R. P. C. & Maia, N. M. M. Strategies for an efficient indicator of structural damage. en. *Mech. Syst. Signal Pr. Special Issue: Inverse Problems* **23,** 1855–1869. doi:10.1016/j.ymssp.2008.07.015 (2009).

151. Zang, C., Friswell, M. & Imregun, M. *Structural Health Monitoring and Damage Assessment Using Measured FRFs from Multiple Sensors , Part II : Decision Making with RBF Network* in (2006).

152. Pérez, M. A. & Serra-López, R. A frequency domain-based correlation approach for structural assessment and damage identification. en. *Mech. Syst. Signal Pr.* **119,** 432–456. doi:10.1016/j.ymssp.2018.09.042 (2019).

153. Göge, D. & Link, M. Assessment of computational model updating procedures with regard to model validation. en. *Aerosp. Sci. Technol.* **7,** 47–61. doi:10.1016/s1270-9638(02)01193-8 (2003).

154. Marinone, T. & Moya, A. *Comparison of FRF Correlation Techniques* en. in *Model Validation and Uncertainty Quantification, Volume 3* (Springer International Publishing, Cham, 2015), 299–309. doi:10.1007/978-3-319-15224-0_32.

155. Sampaio, R., Maia, N., Almeida, R. & Urgueira, A. A simple damage detection indicator using operational deflection shapes. en. *Mech. Syst. Signal Pr.* **72-73,** 629–641. doi:10.1016/j.ymssp.2015.10.023 (2016).

156. Das, S., Saha, P. & Patro, S. K. Vibration-based damage detection techniques used for health monitoring of structures: a review. *J. Civ. Struct. Health Monit.* **6,** 477–507. doi:10.1007/s13349-016-0168-5 (2016).

157. Kawiecki, G. Modal damping measurement for damage detection. en. *Smart Mater. Struct.* **10,** 466. doi:10.1088/0964-1726/10/3/307 (2001).

158. Maia, N. M. M., Silva, J. M. M., Almas, E. A. M. & Sampaio, R. P. C. Damage detection in structures: From mode shape to frequency response function methods. en. *Mech. Syst. Signal Pr.* **17,** 489–498. doi:10.1006/mssp.2002.1506 (2003).

159. Alnefaie, K. Finite element modeling of composite plates with internal delamination. en. *Compos. Struct.* **90,** 21–27. doi:10.1016/j.compstruct.2009.01.004 (2009).

160. Parloo, E., Guillaume, P. & Van overmeire, M. Damage assessment using mode shape sensitivities. en. *Mech. Syst. Signal Pr.* **17,** 499–518. doi:10.1006/mssp.2001.1429 (2003).

161. Messina, A., Williams, E. J. & Contursi, T. Structural damage detection by a sensitivity and statistical-based method. en. *J. Sound Vib.* **216,** 791–808. doi:10.1006/jsvi.1998.1728 (1998).

162. Farrar, C. R. & Doebling, S. W. An overview of modal-based damage identification methods (1997).

163. Carrión, F. J., Quintana, J. A. & Crespo, S. E. Techno-economical and practical considerations for SHM systems. *J. Civ. Struct. Health Monit.* **7,** 207–215. doi:10.1007/s13349-017-0215-x (2017).

164. Sarrafi, A., Mao, Z., Niezrecki, C. & Poozesh, P. Vibration-based damage detection in wind turbine blades using Phase-based Motion Estimation and motion magnification. en. *J. Sound Vib.* **421,** 300–318. doi:10.1016/j.jsv.2018.01.050 (2018).

165. Valente, N. A., Mao, Z., Southwick, M. & Niezrecki, C. *Implementation of Total Variation Applied to Motion Magnification for Structural Dynamic Identification* en. in *Rotating Machinery, Optical Methods & Scanning LDV Methods, Volume 6* (Springer International Publishing, Cham, 2020), 139–144. doi:10.1007/978-3-030-47721-9_17.

166. Fioriti, V., Roselli, I., Tatì, A., Romano, R. & De Canio, G. Motion Magnification Analysis for structural monitoring of ancient constructions. en. *Meas.* **129,** 375–380. doi:10.1016/j.measurement.2018.07.055 (2018).

167. Zimmermann, M., Gülan, U., Harmanci, Y. E., Chatzi, E. & Holzner, M. Structural Health Monitoring through Video Recording. en-US. *J Nondestr Eval* **21** (2016).

168. Holak, K. *A motion magnification application in video-based vibration measurement* en. in *Adv. Mech. Mach. Sci* (ed Uhl, T.) (Springer International Publishing, Cham, 2019), 4135–4144. doi:10.1007/978-3-030-20131-9_412.

169. Cygert, S. & Czyzewski, A. Eulerian motion magnification applied to structural health monitoring of wind turbines. *J. Acoust. Soc. Am.* **144,** 1796. doi:10.1121/1.5067923 (2018).

170. Molina-Viedma, A. J., Felipe-Sesé, L., López-Alba, E. & Díaz, F. A. 3D mode shapes characterisation using phase-based motion magnification in large structures using stereoscopic DIC. en. *Mech. Syst. Signal Pr.* **108,** 140–155. doi:10.1016/j.ymssp.2018.02.006 (2018).

171. Qiu, Q. & Lau, D. Defect detection in FRP-bonded structural system via phase-based motion magnification technique. en. *Struct Control Health Monit* **25,** e2259. doi:10.1002/stc.2259 (2018).

172. Angelosanti, M., Currà, E. & Sabato, A. BIM oriented applications of structural health monitoring based on magnified digital image correlation point-clouds. en. *Autom. Constr.* **148,** 104754. doi:10.1016/j.autcon.2023.104754 (2023).

173. Shang, Z. & Shen, Z. Multi-point vibration measurement and mode magnification of civil structures using video-based motion processing. en. *Autom. Constr.* **93,** 231–240. doi:10.1016/j.autcon.2018.05.025 (2018).

174. Siringoringo, D. M., Wangchuk, S. & Fujino, Y. Noncontact operational modal analysis of light poles by vision-based motion-magnification method. en. *Eng. Struct.s* **244,** 112728. doi:10.1016/j.engstruct.2021.112728 (2021).

175. Sony, S., Laventure, S. & Sadhu, A. A literature review of next-generation smart sensing technology in structural health monitoring. en. *Struct Control Health Monit* **26,** e2321. doi:10.1002/stc.2321 (2019).

176. Chen, W., Yan, B., Liao, J., Luo, L. & Dong, Y. Cable Force Determination Using Phase-Based Video Motion Magnification and Digital Image Correlation. *Int. J. Struct. Stab. Dyn.* **22,** 2250036. doi:10.1142/s0219455422500365 (2022).

177. Sarrafi, A., Poozesh, P., Niezrecki, C. & Mao, Z. *Mode extraction on wind turbine blades via phase-based video motion estimation* in *Smart Materials and Nondestructive Evaluation for Energy Systems 2017* **10171** (SPIE, 2017), 105–116. doi:10.1117/12.2260406.

178. Dorn, C. J. *et al. Automated Extraction of Mode Shapes Using Motion Magnified Video and Blind Source Separation* en. in *Topics in Modal Analysis & Testing, Volume 10* (Springer International Publishing, Cham, 2016), 355–360. doi:10.1007/978-3-319-30249-2_32.

179. Wangchuk, S., Siringoringo, D. M. & Fujino, Y. *Vision-Based Vibration Measurement of Stay-Cables by Video Motion Magnification and Dynamic Mode Decomposition* en. in *Experimental Vibration Analysis for Civil Engineering Structures* (Springer International Publishing, Cham, 2023), 149–162. doi:10.1007/978-3-030-93236-7_14.

180. Choi, A. J. & Han, J.-H. Frequency-based damage detection in cantilever beam using vision-based monitoring system with motion magnification technique. en. *J. Intell. Mater. Syst. Struct.s* **29,** 3923–3936. doi:10.1177/1045389x18799961 (2018).

181. Chen, J. G. *et al. Structural Modal Identification Through High Speed Camera Video: Motion Magnification* en. in *Topics in Modal Anal. I, Volume 7* (ed De Clerck, J.) (Springer International Publishing, Cham, 2014), 191–197. doi:10.1007/978-3-319-04753-9_19.

182. Fioriti, V., Roselli, I., Tati, A., Romano, R. & De Canio, G. *Motion Magnification for Urban Buildings* en. in *Critical Information Infrastructures Security* (Springer International Publishing, Cham, 2018), 253–260. doi:10.1007/978-3-319-99843-5_23.

183. Pérez, E. & Zappa, E. Video Motion Magnification to Improve the Accuracy of Vision-Based Vibration Measurements. *IEEE Trans. Instrum. Meas.* **71,** 1–12. doi:10.1109/tim.2022.3175977 (2022).

184. Liang, Z.-y., Chen, H.-l., Hua, J.-h. & Deng, Y.-c. *A Eulerian Video Magnification Based Structural Damage Identification Method for Scaffold* en. in *Proceedings of the 26th International Symposium on Advancement of Construction Management and Real Estate* (Springer Nature, Singapore, 2022), 1122–1132. doi:10.1007/978-981-19-5256-2_88.

185. Cabo, C. T. d., Valente, N. A. & Mao, Z. *Motion magnification for optical-based structural health monitoring* in *Health Monit. Struct. Biol. Syst. XIV* **11381** (SPIE, 2020), 221–227. doi:10.1117/12.2559266.

186. Yang, Y. *et al.* Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification. en. *Mech. Syst. Signal Pr.* **85,** 567–590. doi:10.1016/j.ymssp.2016.08.041 (2017).

187. Arul Prakash, S. K. *et al.* Detection of System Compromise in Additive Manufacturing Using Video Motion Magnification. *J. Mech. Design* **142.** doi:10.1115/1.4045547 (2019).

188. Śmieja, M., Mamala, J., Prażnowski, K., Ciepliński, T. & Szumilas, Ł. Motion Magnification of Vibration Image in Estimation of Technical Object Condition-Review. en. *Sensors* **21,** 6572. doi:10.3390/s21196572 (2021).

189. Qiu, Q. & Lau, D. *Defect visualization in FRP-bonded concrete by using high speed camera and motion magnification technique* in *Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure 2017* **10169** (SPIE, 2017), 579–587. doi:10.1117/12.2259916.

190. Eitner, M., Miller, B., Sirohi, J. & Tinney, C. Effect of broad-band phase-based motion magnification on modal parameter estimation. en. *Mech. Syst. Signal Pr.* **146,** 106995. doi:10.1016/j.ymssp.2020.106995 (2021).

191. Shang, Z. & Shen, Z. Multi-point Vibration Measurement for Mode Identification of Bridge Structures using Video-based Motion Magnification. arXiv: 1712.06566 (2017).

192. Fioriti, V. *et al.* Motion Magnification Applications for the Protection of Italian Cultural Heritage Assets. *Sensors* **22.** doi:10.3390/s22249988 (2022).

193. Lado-Roigé, R., Font-Moré, J. & Pérez, M. A. Learning-based video motion magnification approach for vibration-based damage detection. *Meas.* **206,** 112218. doi:10.1016/j.measurement.2022.112218 (2023).

194. Gao, S., Zhou, P., Cheng, M.-M. & Yan, S. *Masked Diffusion Transformer is a Strong Image Synthesizer* 2023. arXiv: 2303.14389.

195. Chen, X., Wang, X., Zhou, J., Qiao, Y. & Dong, C. *Activating More Pixels in Image Super-Resolution Transformer* 2023. arXiv: 2205.04437.

196. Zhang, D., Huang, F., Liu, S., Wang, X. & Jin, Z. *SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution* 2023. arXiv: 2208.11247.

197. Cao, J. *et al. Learning Task-Oriented Flows to Mutually Guide Feature Alignment in Synthesized and Real Video Denoising* 2023. arXiv: 2208.11803.

198. Li, W. *et al. MAT: Mask-Aware Transformer for Large Hole Image Inpainting* 2022. arXiv: 2203.15270.

199. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* 2015. arXiv: 1512.03385.

200. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597.

201. Khan, S. *et al.* Transformers in Vision: A Survey. *ACM Comput. Surv.* doi:10.1145/3505244 (2021).

202. Liang, J. *et al. VRT: A Video Restoration Transformer* 2022. arXiv: 2201.12288.

203. Ramachandran, P. *et al. Stand-Alone Self-Attention in Vision Models* 2019. arXiv: 1906.05909.

204. Wu, B. *et al. Visual Transformers: Token-based Image Representation and Processing for Computer Vision* 2020. arXiv: 2006.03677.

205. Li, Y., Zhang, K., Cao, J., Timofte, R. & Van Gool, L. *LocalViT: Bringing Locality to Vision Transformers* 2021. arXiv: 2104.05707.

206. Liu, Y. *et al. Vision Transformers with Hierarchical Attention* 2021. arXiv: 2106.03180.

207. Vaswani, A. *et al. Scaling Local Self-Attention for Parameter Efficient Visual Backbones* 2021. arXiv: 2103.12731.

208. Carion, N. *et al. End-to-End Object Detection with Transformers* 2020. arXiv: 2005.12872.

209. Liu, L. *et al.* Deep Learning for Generic Object Detection: A Survey. en. *Int J Comput Vis* **128,** 261–318. doi:10.1007/s11263-019-01247-4 (2020).

210. Touvron, H. *et al. Training data-efficient image transformers and distillation through attention* 2020. arXiv: 2012.12877.

211. Zheng, S. *et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers* 2020. arXiv: 2012.15840.

212. Cao, H. *et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation* 2021. arXiv: 2105.05537.

213. Liang, D., Chen, X., Xu, W., Zhou, Y. & Bai, X. TransCrowd: weakly-supervised crowd counting with transformers. en. *Sci. China Inf. Sci.* **65,** 160104. doi:10.1007/s11432-021-3445-y (2022).

214. Sun, G. *et al. Boosting Crowd Counting with Transformers* 2021. arXiv: 2105.10926.

215. Chen, H. *et al. Pre-Trained Image Processing Transformer* 2020. arXiv: 2012.00364.

216. Cao, J., Li, Y., Zhang, K. & Van Gool, L. *Video Super-Resolution Transformer* 2021. arXiv: 2106.06847.

217. Wang, Z. *et al. Uformer: A General U-Shaped Transformer for Image Restoration* 2021. arXiv: 2106.03106.

218. Xiao, T. *et al. Early Convolutions Help Transformers See Better* 2021. arXiv: 2106.14881.

219. Lin, T.-Y. *et al.* Microsoft COCO: Common Objects in Context. en. arXiv: 1405.0312 (2015).

220. Everingham, M., Gool, L., Williams, C. K., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **88,** 303–338. doi:10.1007/s11263-009-0275-4 (2010).

221. Keleş, O., Yılmaz, M. A., Tekalp, A. M., Korkmaz, C. & Dogan, Z. *On the Computation of PSNR for a Set of Images or Video* 2021. arXiv: 2104.14868.

222. Wang, L. *A survey on IQA* 2021. arXiv: 2109.00347.

223. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13,** 600–612. doi:10.1109/tip.2003.819861 (2004).

224. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **21,** 4695–4708. doi:10.1109/tip.2012.2214050 (2012).

225. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process Lett.* **20,** 209–212. doi:10.1109/lsp.2012.2227726 (2013).

226. Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S. & Medasani, S. S. *Blind image quality evaluation using perception based features* in *IEEE NCC* (2015), 1–6. doi:10.1109/ncc.2015.7084843.

227. Ke, J., Wang, Q., Wang, Y., Milanfar, P. & Yang, F. MUSIQ: Multi-scale Image Quality Transformer. arXiv: 2108.05997 (2021).

228. *PyTorch Toolbox for Image Quality Assessment* version 0.1.7. 2022. https://pypi.org/project/pyiqa/.

229. Jardine, A. K., Lin, D. & Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Pr.* **20,** 1483–1510. doi:10.1016/j.ymssp.2005.09.012 (2006).

230. Mills, S., of Non-Destructive Testing, B. I. & COMADIT. *Vibration Monitoring & Analysis Handbook* (British Inst. of Non-destructive Testing, 2010).

231. Figueiredo, E., Park, G., Figueiras, J., Farrar, C. & Worden, K. *Structural health monitoring algorithm comparisons using standard data sets* en. Tech. rep. LA-14393, 961604 (2009), LA–14393, 961604. doi:10.2172/961604.

232. Lado-Roigé, R. & Pérez, M. A. ViMag: A Visual Vibration Analysis Toolbox. *Journal of Open Source Software* **8,** 5491. doi:10.21105/joss.05491 (2023).

233. Reyes Carmenaty, G. & Font Moré, J. *PyModal* version 0.0.4. 2020. https://github.com/grcarmenaty/pymodal.

234. Wang, J., Zhao, J., Liu, Y. & Shan, J. Vision-based displacement and joint rotation tracking of frame structure using feature mix with single consumer-grade camera. en. *Struct Control Health Monit* **28.** doi:10.1002/stc.2832 (2021).

235. Gallego, G. *et al.* Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44,** 154–180. doi:10.1109/tpami.2020.3008413 (2022).