1 2 90

UNIVERSIDADE D
COIMBRA

Miguel Mendes Silva

Vision Transformers for face anti-spoofing

Dissertação no âmbito de Mestrado em Engenharia Eletrotécnica e de Computadores, no ramo de Robótica, Controlo e Inteligência Artificial, orientada pelo Professor Doutor Jorge Manuel Moreira de Campos Pereira Batista e apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Julho de 2023

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Vision Transformers For Face Anti-Spoofing

Miguel Mendes Silva

Coimbra, July 2023

Vision Transformers For Face Anti-Spoofing

Supervisor:

Prof. Dr. Jorge Manuel Moreira de Campos Pereira Batista

Jury:

Prof. Dr. Jorge Manuel Miranda Dias

Prof. Dr. Nuno Miguel Mendonça da Silva Gonçalves

Prof. Dr. Jorge Manuel Moreira de Campos Pereira Batista

Dissertation submitted in partial fulfilment for the degree of Master of Science in Electrical and Computer Engineering.

Coimbra, July 2023

Agradecimentos

Antes de mais, gostaria de agradecer ao professor Jorge Batista pela orientação e dedicação com que me ajudou a realizar este trabalho e pela confiança que em mim depositou. Também quero agradecer aos meus colegas de laboratório, Eurico, André e Bruno pela disponibilidade e ajuda prestada ao longo deste último ano. Sem eles, confesso que teria tido muito mais dificuldade em enfrentar certos problemas que encontrei pelo caminho e por essa razão, considero que tiveram um papel importante na realização do meu trabalho.

Quero também agradecer à minha família, por me terem dado a oportunidade de estudar o que gosto e do apoio incondicional que me deram ao longo do meu percurso académico, apesar da exigência. A eles, o meu sincero obrigado.

Por fim, aos colegas de curso e aos amigos de longa data. Quer seja pelo tempo que vivemos juntos ou pelas aventuras partilhadas ao longo dos anos, todos foram importantes para o meu desenvolvimento como pessoa. Somos tão grandes quanto aqueles que nos rodeiam e eles são a prova disso.

Por fim, obrigado por fazerem de mim uma pessoa melhor.

Abstract

Authentication systems based on facial recognition have become increasingly popular in recent years as a convenient approach to verifying individuals. This non-intrusive authentication method analyzes distinct facial properties, compares them, and examines patterns in a person's facial contours. However, the rise of Presentation Attacks (PAs) poses a significant threat to the reliability of this form of authentication, as impostors attempt to bypass the systems by impersonating others using printed photos or 3D masks.

Therefore, to ensure the reliability of facial authentication, it is crucial to develop Face Anti-Spoofing (FAS) algorithms that can effectively defend against all types of spoofing attempts and overcome associated challenges. In addition to the extensively studied Convolutional Neural Networks (CNNs), the emergence of Vision Transformers (ViTs) in other areas of computer vision has sparked interest in utilizing this deep learning architecture in the field of FAS. Furthermore, in addition to RGB data, the incorporation of multi-modal information such as Depth and Infrared, has also shown promising results in detecting more complex attacks.

In this regard, the main objective of this thesis is to explore the use of multi-modal Vision Transformers for the FAS task. Based on existing contributions in the literature, the proposed ViT-based frameworks using multi-modal images will be compared to a CNN-based approach for evaluation and performance comparison. These frameworks will be evaluated at the intra-domain, cross-domain, and zero-shot levels using different Presentation Attack Detection (PAD) datasets. The results aim to demonstrate the effectiveness of attention mechanisms in this context and highlight the benefits of leveraging multi-modal information to distinguish genuine faces from spoofing attempts in FAS applications.

Keyworks: Presentation Attacks, Face Anti-Spoofing, Multi-modal Information, Vision Transformer, Deep Learning

Resumo

Os sistemas de autenticação baseados em reconhecimento facial tornaram-se nos últimos anos cada vez mais populares como uma abordagem conveniente para verificar indivíduos. Este método de autenticação não intrusivo analisa propriedades faciais distintas, compara as mesmas e examina padrões nos contornos faciais de uma pessoa. No entanto, o aumento dos Ataques de Apresentação (PAs) representa uma ameaça significativa para a confiabilidade desta forma de autenticação, uma vez que impostores tentam contornar os sistemas ao fazerem-se passar por outros utilizando fotos impressas ou máscaras 3D.

Portanto, para garantir a confiabilidade da autenticação facial, é crucial desenvolver sistemas de *Anti-Spoofing* Facial (FAS) que permitam a defesa contra todos os tipos de tentativas de falsificação e superar os desafios associados. Para além das Redes Neuronais Convolucionais (CNNs) extensivamente estudadas, a emergência dos *Transformers* em outras áreas de visão por computador despertou interesse em utilizar esta arquitetura no campo de FAS. Por outro lado, para além de informação RGB, a incorporação de informações modais como Profundidade e Infravermelho, também tem mostrado resultados promissores na deteção de ataques mais complexos.

Nesse sentido, o objetivo principal desta tese é explorar o uso de *Vision Transformers* (ViTs) multi-modais para a tarefa de FAS. Baseados em contribuições existentes na literatura, os *frameworks* propostos baseados em ViTs utilizam imagens multi-modais e vão ser comparados a uma abordagem baseada em CNN para avaliação e comparação de desempenho. Estes *frameworks* serão avaliados ao nível de *intra-domain*, *cross-domain* e *zero-shot* usando diferentes *datasets* de Detecção de Ataques de Apresentação (PAD). Os resultados visam demonstrar a eficácia dos mecanismos de atenção nesse contexto e destacar os benefícios de aproveitar informações multi-modais para distinguir faces genuínas de tentativas de falsificação em aplicações de FAS.

Keywords: Ataques de Apresentação, *Anti-Spoofing* Facial, Informação Modal, *Vision Transformer*, *Deep Learning*

“Sou piloto dos meus sonhos e vou voar o mais alto que conseguir."

Contents

List of Acronyms

PA	Presentation Attack
FAS	Face Anti-Spoofing
CNN	Convolutional Neural Network
ViT	Vision Transformer
PAD	Presentation Attack Detection
SVM	Support Vector Machine
PCA	Principle Component Analysis
LSTM	Long Short-Term Memory
OFM	Optical Flow Magnitude
RNN	Recurrent Neural Network
rPPG	Remote PhotoPlethysmoGraphy
SE	Squeeze-and-Excitation
GAP	Global Average pooling
MFE	Modal Feature Erasing
NLP	Natural Language Processing
DeiT	Data-efficient image Transformer
T2T-ViT	Tokens-to-Token Vision Transformer
CeiT	Convolution-enhnanced image Transformer
I2T	Image-to-Tokens

LeFF	Locally-enhanced Feed-Forward
LCA	Layer-wise Class token Attention
BCE	Binary Cross-Entropy
MTSS	Multi-Teacher Single-Student
MAMD	Multi-Level Attention Module with Dropblock
MFAST	Multi-Modal Face Anti-Spoofing Transformer
FM-ViT	Flexible Modal Vision Transformer
CMTB	Cross-Modal Transformer Block
MMA	Multi-headed Mutual-Attention
FA	Fusion Attention
MA-ViT	Modality-Agnostic Vision Transformer
FWT	Feature-Wise Transformation
SSAN	Shuffled Style Assembly Network
MHSA	Multi-head Self-Attention
MLP	Multilayer Perceptron
LOO	Leave-one-out
CE	Cross-Entropy
APCER	Presentation Classification Error Rate
BPCER	Presentation Classification Error Rate
ACER	Average Classification Error Rate
ACC	Accuracy
ROC	Receiver Operation Characteristic
FPR	False Positive Rate
TPR	True Positive Rate

List of Figures

List of Tables

1 Introduction

1.1 Context and Motivation

In recent decades, advances in technology have enabled the development of robust and reliable authentication systems. These systems aim to verify the identity of an entity and thereby confirm its access as legitimate. Of all existing authentication systems, those that require biometric information such as facial recognition, fingerprint or even iris scanning as a form of verification stand out since they are safer, more convenient and effective when compared to other more traditional identification solutions such as password, pin-code or email verification.

Since each person has a series of unique measurable properties on their face, facial recognition is able to uniquely identify, recognize and authenticate a person by comparing and analyzing patterns based on the subject's facial contours in a non-intrusive way. Interest in facial recognition systems and algorithms has been increasing by the business community and according a report made by MordorIntelligence [21], not only this market has been valued at 3.72 billion dollars in 2020, but it is projected to be valued at 11.62 billion by 2026.

In addition, these systems have a wide area of applicability. One of their most prevalent uses is in identity verification and access control scenarios. By comparing an individual's face with a pre-registered image or a database of known identities, these systems can rapidly confirm the person's identity and provide access to personal spaces, including buildings or mobile devices. Moreover, facial recognition systems are also employed in security and surveillance applications. For instance, they can be utilized in airport security to monitor public spaces and identify potential threats by recognizing individuals who have been previously identified as dangerous.

However, the widespread deployment of this technology for security-critical scenarios is still limited and under scrutiny due to its vulnerability to presentation attacks. These attacks are usually carried out by malicious users who make use of digital manipulation and physical means to overcome security systems and gain illegitimate acess to a victim device. Most common type of PAs are print, replay, 3D masks, Mannequin, Glasses, Makeup and Tatto. Figure 1.1 provide some examples of presentation attacks.

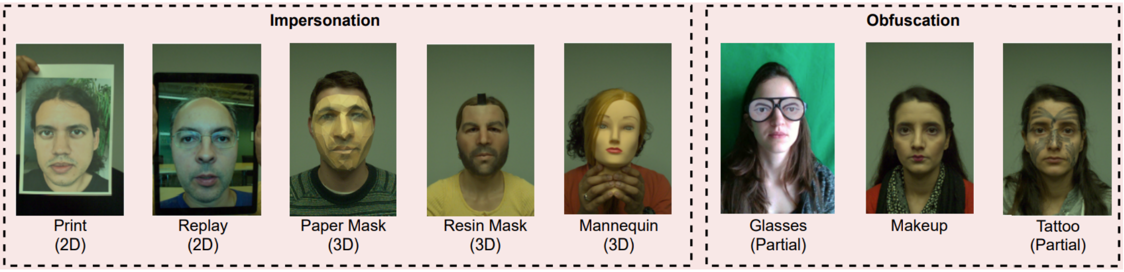


Figure 1.1: Presentation Attacks. Taken from [1].

Based on their typology, PAs can be divided into impersonation and obfuscation attacks. In impersonation attacks, impostors make use of spoof to be recognized as someone else by copying victims facial attributes into spoofing attacks. In obfuscation attacks, impostors use tricks, e.g., wearing glasses, extreme makeup, wig, to avoid being recognized by the system. Futhermore, based on their craft, PAs can be classified into 2D and 3D attacks. 2D attacks primarily revolve around presenting facial attributes using flat or wrapped printed photos, photos with cut-out eyes or mouth, and digital video replays. On the other hand, 3D attacks involve the utilization of printed masks crafted from specialized materials such as paper, resin, or plastic. The diagram illustrated on figure 1.2 summarizes PAs.

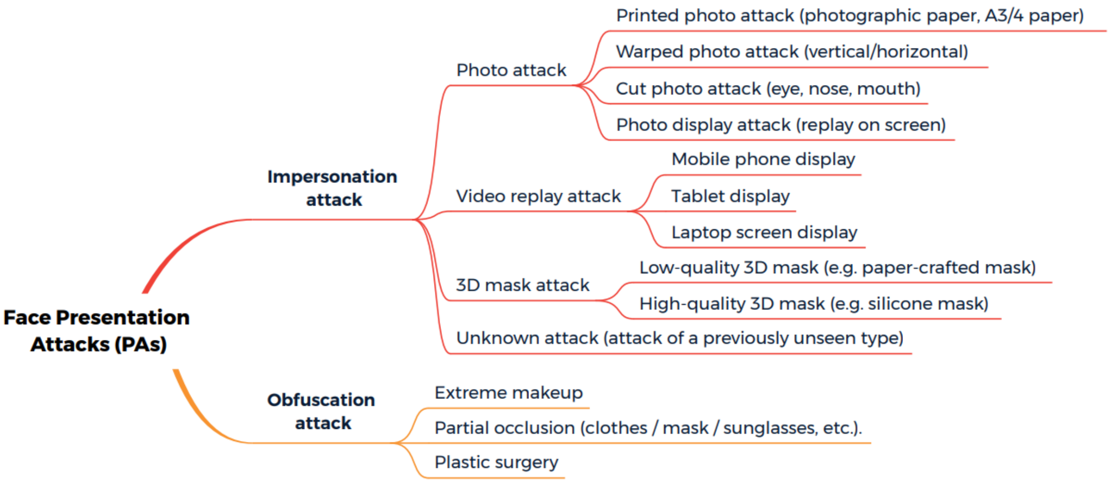

Face Presentation Attacks (PAs)
Impersonation attack
Photo attack
Printed photo attack (photographic paper, A3/4 paper)
Warped photo attack (vertical/horizontal)
Cut photo attack (eye, nose, mouth)
Photo display attack (replay on screen)
Video replay attack
Mobile phone display
Tablet display
Laptop screen display
3D mask attack
Low-quality 3D mask (e.g. paper-crafted mask)
High-quality 3D mask (e.g. silicone mask)
Unknown attack (attack of a previously unseen type)
Obfuscation attack
Extreme makeup
Partial occlusion (clothes / mask / sunglasses, etc.).
Plastic surgery

Figure 1.2: Typology of PAs. Taken from [2].

1.2 Challenges and Breakthroughs

Over the last few years, there has been an increasing focus on the development of FAS algorithms in order to safeguard facial recognition systems against PAs. However, building a robust system that can effectively defend against all types of spoofing attempts is a complex task, mainly due to the high number of factors that need to be considered.

To begin with, apart from traditional 2D attacks, the increasing number of static and dynamic 3D attacks poses a higher threat to PAD systems as they are more realistic in terms of color, texture, and geometry. Additionally, the risk of spoofing attacks has increased not only because it's easier for impostors to find online data about their target, but also because spoofing can be acquired from wider angles, complex scenes, different devices, and materials. These factors, combined with the emergence of spoofing data from multiple domains, present additional challenges for anti-spoofing applications in cross-domain scenarios. For instance, spoofing cues are sensitive to capture conditions, which may drastically vary from one domain to another due to differences in camera devices, sensor types, illumination settings and resolutions, i.e., spoof samples from one domain may be misinterpreted as real samples on another domain and vice-versa.

Moreover, existing PAD datasets are predominantly single-modal, limited in terms of subject diversity, insufficient in representing various types of attacks, and contain relatively small amounts of training data. On one hand, the lack of training data can lead to overfitting issues, resulting in poor generalization to other domains and unseen attack types. On the other hand, datasets restricted to RGB images limit models to effectively learn spoof cues on newly-made more sophisticated and realistic attacks, e.g., 3D masks. To address these challenges, researchers have recently introduced large-scale multi-modal datasets that incorporate multiple modalities, including RGB, Depth, Infrared, and Thermal data. The key ideia behind incorporating multi-modal data is to make it extremely hard for attackers to replicate the properties of a bonafide sample across diferent modalities. For instance, depth data captures information about distances, enabling differentiation between real faces and replay or print attacks that produce flat depth maps. Infrared data, on the other hand, measures the amount of heat radiated from a face and analyzes the differences in appearances between real and spoof faces.

In the light of these problems, multiple deep learning methods have been adopted as framework to distinguish between genuine users and spoofing attacks. From the multitude of solutions, Vision Transformers have recently been studied as an alternative to the extensively used CNNs for the FAS task. Despite showing promising results to date, there are still few works related to the use of attention-based approaches in this field. Notoriously known for its self-attention mechanism, ViT attends to the whole image to capture global dependencies between image patches rather than using convolutional layers and downsampling operations to extract features.

1.3 Objectives

This dissertation intends to develop multi-modal networks utilizing Vision Transformers and compare their performance with a CNN-based approach for FAS task. While CNNs are widely used for classification tasks in computer vision, this thesis aims to investigate the effectiveness of Vision Transformers in detecting spoof attempts and recognizing genuine faces. Building upon existing contributions in the literature, we also aim to evaluate performance under diverse scenarios to assess whether ViT-based architectures can generalize well to unseen domains and attacks and prove how models can benefit from RGB, Depth and Infrared data.

To achieve this, both ViT and CNN-based models will be trained and tested using multiple PAs from the WMCA [22] and CASIA-SURF [7] datasets. The evaluation will cover zero-shot learning, intra-domain, and cross-domain analysis to provide comprehensive insights into the performance of these models in different contexts.

1.4 Document Structure

The outline of this dissertation is the following:

- **Chapter 2**: State-of-the-Art reviews the most relevant works in the literature and provides a theoretical background about the subject.

- **Chapter 3**: Methodology describes the implementation and all approaches used during the course of this work.

- **Chapter 4**: Results and Discussion provide a detailed examination about the experimental results.

- **Chapter 5**: Conclusion and Future Work summarises what was concluded from the results and suggests improvements.

2 State-of-the-Art

This chapter reviews the relevant contributions of the literature concerning the topic of face anti-spoofing and contextualizes the existing architectures and methods that have proven to be usefull in this field. By examining and summarizing the significant findings and advancements in the literature, this chapter offers a comprehensive understanding of the state-of-the-art approaches and techniques employed in face anti-spoofing research.

2.1 Evolution of Face Anti-Spoofing methods

The current literature and much of the previous research predominantly focus on two primary approaches to tackle FAS: handcrafted-based methods and deep learning-based methods.

The first attempts for face PAD were made through liveness detection using handcrafted methods. Liveness cues, such as eye-blinking, blood pulse measure, face and head movement and physiological signals were explored for dynamic representation. Classical handcrafted features (e.g LBP [23] [24], SIFT [25], HOG [26] and DoG [27]) made use of texture, color and motion cues to extract spoofing patterns from various color spaces. However, this type of PAD method shows weak generalization abilities as they are not powerful enough to capture all the possible variations in the acquisition conditions and liveness cues are easily mimicked bv video attacks, making them less reliable. For the context of this thesis, these methods will not be deepened.

Nevertheless, over the last few years, most of the research in this field has shifted towards using deep learning methods such as CNNs and more recently, Transformers. These methods have been heavily proposed for both static and dynamic face PAD due to their strong discriminative feature representation ability. While static methods consider a single frame captured by the sensor, dynamic methods leverage the temporal information obtained from video captures by exploiting motion across multiple video frames. In contrast to handcrafted methods, deep learning methods are able to automatically extract relevant texture features from input data without relying on human expertise to design hand-crafted features. Even though CNNs and Transformers have already shown promising results at detecting spoofing, both methods may suffer from overfitting due to limited amount of training data and weak generalization ability as they are trained to recognize specific types of spoofing attacks. Futhermore, deep learning-based approaches heavily rely on training configurations and fine-tuning of hyperparameters.

2.2 CNN-based methods for Face Anti-Spoofing

Regarding deep learning methods, most of the developed pipelines in the field of FAS have relied on using CNNs as the primary approach for identifying spoofing attempts. In the early stages, most studies relied on end-to-end training of CNNs to learn feature representations from RGB face images and videos, combined with Support Vector Machines (SVMs) to perform classification. Although most common algorithms require labeled data for supervision, they still regard face anti-spoofing as a binary classification problem.

2.2.1 Single cue-based methods

In the context of FAS, CNNs have been widely used for extracting features related to texture and liveness cues. Texture cues focus on the texture properties of the object presented to the system and allow CNNs to learn distinctive texture-based features that are related to local patterns and image details. Liveness cues, on the other hand, refer to dynamic patterns observed in a video sequence such as head movement and facial expressions.

In 2014, Yang *et al.* [3] made the first attempt to detect spoofing attacks using a CNN. The proposed method is illustrated in figure 2.1 and demonstrated superior performance compared to existing handcrafted methods when it came to detecting photo and video replay attacks. The approach utilized an AlexNet to learn texture features and employed a SVM classifier with binary classes. Although the results were promising, the model encountered overfitting issues due to the limited scale and diversity of datasets available at that time.

To alleviate this issue, Li *et al.* [28] proposed fine tuning a ImageNet-pretrained VGG network for the PAD task on the same type of attacks. The approach involved extracting deep partial features from the convolutional layers of the network, and then reducing the dimensionality of these features using Principle Component Analysis (PCA) blocks to prevent

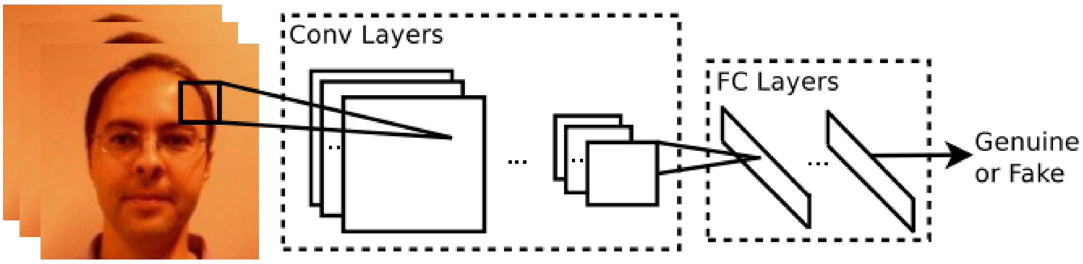


Figure 2.1: First PAD method via CNN. Taken from [3].

overfitting. Once the deep partial features were extracted and reduced, a SVM classifier was trained for classification.

Since CNN architectures cannot extract temporal features themselves, Tu *et al.* [4] investigated using a CNN-LSTM architecture to extract textural features across video frames by focusing on the motion cues. The CNN part, based on a VGG16 network, was used to extract features from each individual frame, while the LSTM network is used to capture the temporal dynamic information across the sequence of features extracted by the CNN. This allowed the model to learn the differences between real human and spoofed faces, based not only on the appeareance of the individual frames but also on the temporal patterns of the face sequence. Figure 2.2 shows the pipeline.

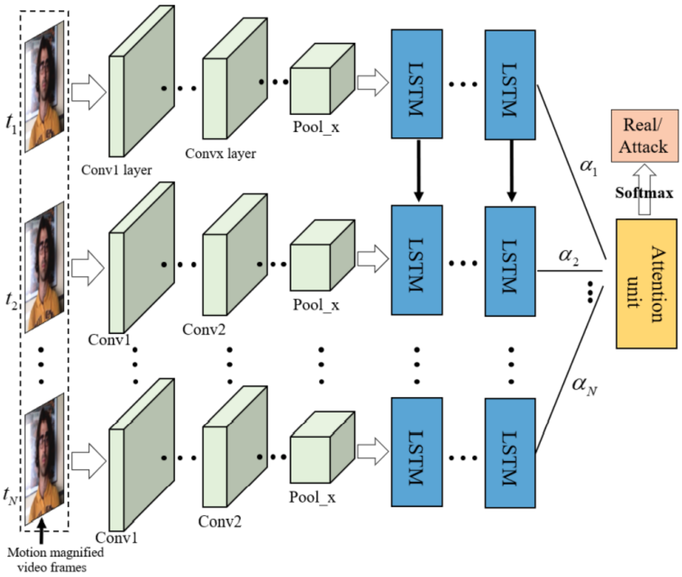


Figure 2.2: CNN-LSTM framework. Taken from [4].

Gan *et al.* [29] also approached FAS as a video sequence classification problem by proposing a 3D CNN to learn spatial and temporal features on multi-frame image level. To preserve the characteristics of temporal dimension between consecutive frames, a 3D convolution was applied on the video sequence. According to the author, unlike common 2D convolutions, also known as stack convolutions, the kernel of the 3D convolution creates more than one feature map by operating in multiple dimensions rather than one dimension on the input image sequence, which is usefull for extracting features of continuous video sequences.

Feng *et al.* [30] explored fusing motion-based cues and image quality-based cues for liveness detection. The proposed CNN has three sub-networks all locally connected with inputs from three different visual cues. The first network uses an image as input and performs face image quality assessment. The second network analyses a face video and extracts motion-based liveness features between face frames with a fixed interval and creates an Optical Flow Magnitude (OFM) map which describes the facial motion pattern. Lastly, the third network calculates an average OFM scene map from the scene video, which is the raw video where the face video was extracted.

Within the same context but with a novel approach, Jourabloo *et al.* [31] introduced a new perspective for solving the face PAD via noise modeling. The proposed CNN framework based on a De-Spoof Net was aimed to estimate the noise of a given spoof image by inversely decomposing it into the live face and spoof noise pattern. For optimization, a new type of loss functions were also designed to encourage the pattern of the spoof images to be ubiquitous and repetitive while aiming for zero noise in the real images. The author's method showed promising results on photo and video replay attacks when compared with other state-of-the-art deep face PAD methods.

2.2.2 Multiple cue-based methods

Alternatively to methods that rely solely on single cues, other approaches in the literature reflect the importance of combining texture, liveness and 3D geometric cues to address the detection of various types of spoofing.

For instance, Liu *et al.* employed 3D-geometric and liveness cues in their work [5]. The proposed approach consists of a CNN-RNN model that utilizes pseudo-depth maps and Remote PhotoPlethysmoGraphy (rPPG) supervision to extract spoofing cues from face videos. The CNN part is responsible for evaluating each frame separately and estimating both depth and feature map of each frame and the RNN part for evaluating the temporal variability across the feature maps by estimating a rPPG signal. The feature maps generated by the CNN part are then fed into a non-rigid registration layer which is responsible for processing the input data for the RNN part. This layer aligns the input data which enables the RNN to compare the feature maps without considering the facial pose or expression and motivates the CNN component to generate zero depth maps either for all frames or for a majority of frames in a given input video sequence. The described method is shown in figure 2.3.

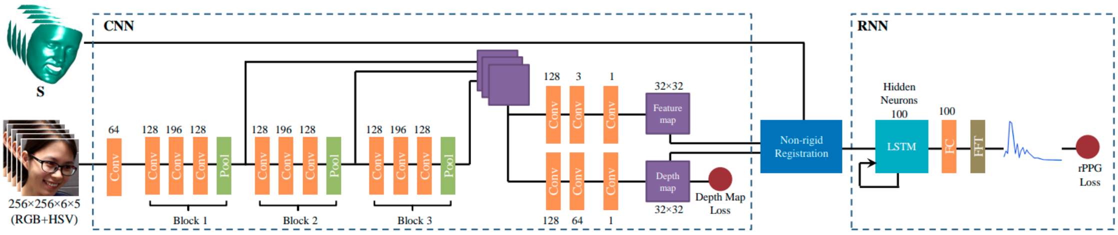


Figure 2.3: CNN-RNN model. Taken from [5].

Atoum *et al.* [6] employed a two-stream CNN that makes use of texture and 3D-geometric cues to differenciate between live and spoof faces. This approach involves extracting local features and generating depth maps from the input image. While the first CNN stream is trained on randomly extracted patches from the input image to learn rich-appearance features, the second stream is used for depth estimation of the full face image. In the end, the outputs from the two streams are combined and used for classification. Figure 2.4 illustrates the proposed approach.

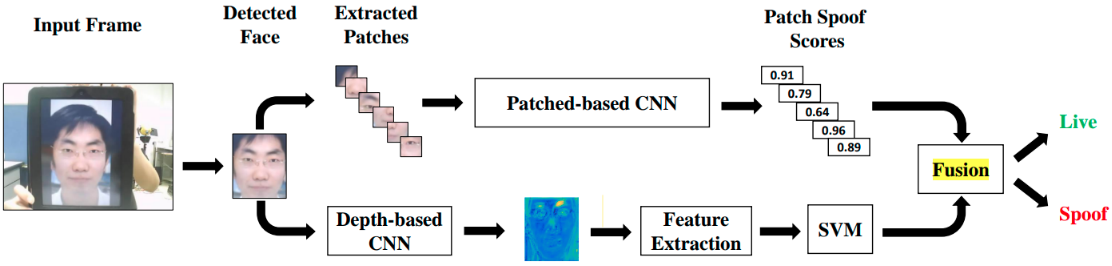


Figure 2.4: Two-stream CNN for FAS. Taken from [6].

2.2.3 Multi-modal based methods

As discussed in the introduction chapter, integrating information from multi-modal data is essential to enhance the robustness of the traditional single-modal PAD algorithms. This is particularly important because certain attacks can only be detected by leveraging information captured from depth and infrared sensors. Despite the fact that late fusion is tipically employed to merge features obtained from each modal input, a significant challenge has been indentifying an optimal approach for fusing information from all modalities without losing any of their intrinsic characteristics.

The creator of the CASIA-SURF dataset, Zhang *et al.* [7] proposed a novel three branch multi-modal network as shown in figure 2.5. Each branch uses ResNet blocks as backbone and takes as input only images from one single modality, making each branch specialized in extracting features from that modality. At deeper levels of each branch, the features are passed through Squeeze-and-Excitation (SE) blocks for feature re-weighting and then concatenated with features from other modalities via Global Average Pooling (GAP) into a single stream.

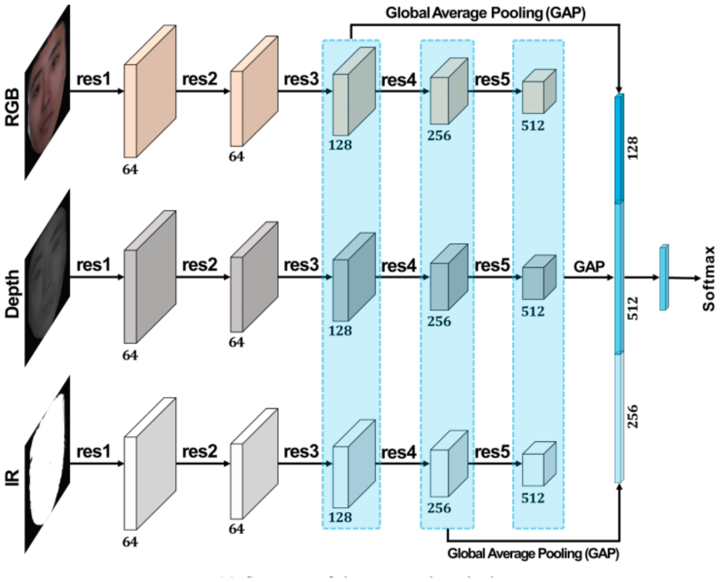


Figure 2.5: Multi-modal CNN. Taken from [7].

A. Parkins *et al.* [8], winner of the Chalearn Face Anti-Spoofing Attack Detection Challenge, adopted a similar approach using a ResNet-based CNN and multi-modal input data. Even though each modality is processed by a unique branch and features are fused via SE and concatenated at deep-level layers, this approach differs from the baseline in terms of feature aggregation. The authors used aggregation blocks to group outputs from multiple layers of the network, with each block using features from the previous one and fusing them with ones from the residual blocks of the main branches, allowing the model to capture multi-layer information. Figure 2.6 illustrates the network and the multi-layer aggregation stategy.

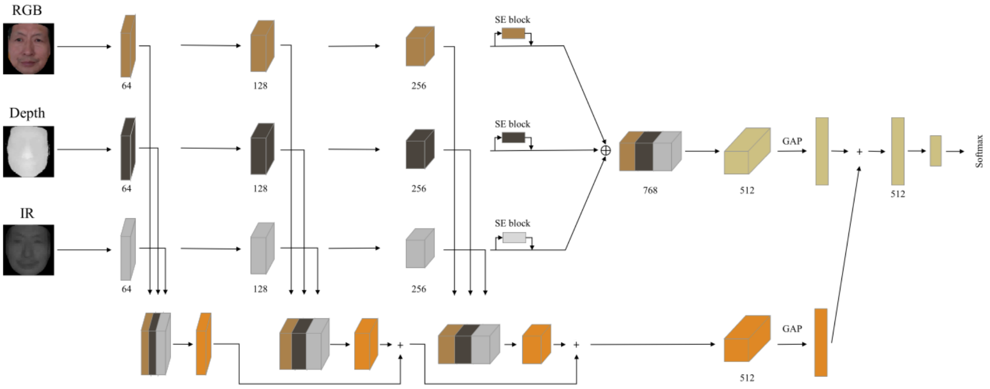


Figure 2.6: Multi-modal CNN with multi-layer feature aggregation. Taken from [8].

The work [9] by T. Shen *et al.* employed a patch-level image approach to extract features using a CNN and a Modal Feature Erasing (MFE) module for multi-modal face anti-spoofing detection, as shown in figure 2.7. The authors trained one single ResNext for each modality to learn rich appearances using randomly selected patches from the images for patch-based feature learning. Then, during training, features from one randomly selected modality were erased to prevent overfitting and better learning.

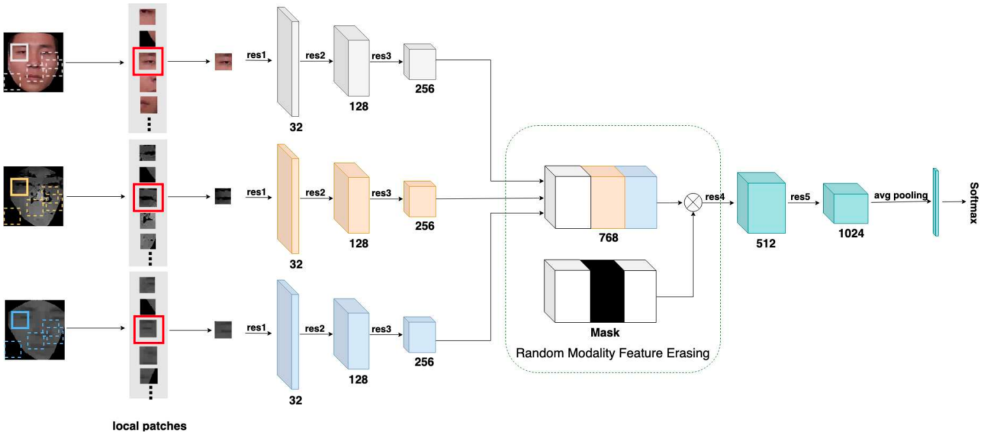


Figure 2.7: Multi-modal CNN with modal feature erasing. Taken from [9].

On the other hand, A. George *et al.* [10] adressed the issue of PAD by proposing a novel multi-modal face detector capable of performing face location and presentation attack detection based on the combined feature representation extracted from multiple modal images, including Grayscale, Depth and Infrared. The authors implemented a RetinaNet with focal loss as an object detector consisting of two subnetworks: one responsible for regressing the bounding box and the other for performing the classification between bonafide and PA. One major difference between the framework illustrared in figure 2.8 and the other works in the literature is that the model does not consider each modality separately. Instead, composite images are used, which result by stacking normalized gray-scale, depth and infrared channels into one multi-channel image.

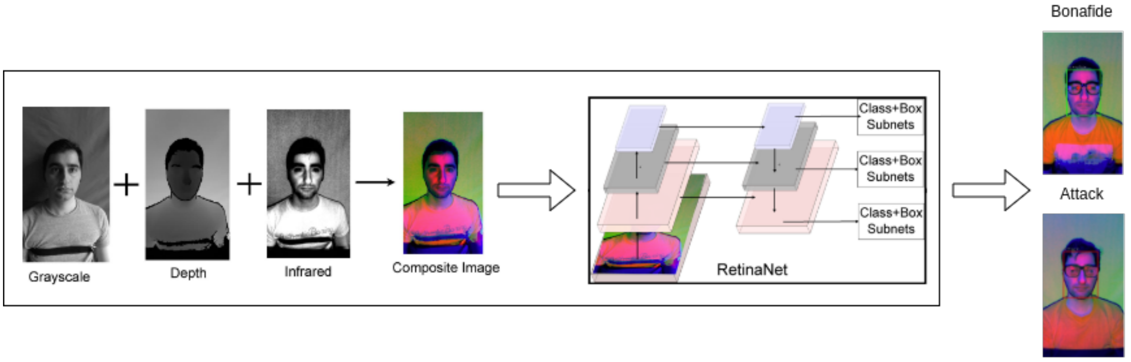


Figure 2.8: Multi-modal face and PA detector. The green boxes indicate bonafide detections and red boxes denote attack detections. Adapted from [10].

2.3 Attention-based methods for Face Anti-Spoofing

While CNNs are generally recognized for their convolutional approach, attention mechanisms have demonstrated their utility in enhancing the performance and interpretability of CNNs and Transformers architectures. From the multitude of attention mechanisms that have been developed in the field of deep learning, each with its own strengths and applications, the most widely used type of attention mechanism in Transformer architectures is self-attention.

Self-attention enables models to focus on relevant information and establish long-range dependencies across the entire input sequence. To accomplish this, a given patch embedding matrix or feature map is processed through three separate linear layers, resulting in query, key, and value matrices (Q, K, V). The query and key matrices are then multiplied element-wise to generate attention scores, forming an attention matrix. This matrix is subsequently normalized using the softmax operator, producing scores that represent the importance of each element relative to others in the input sequence. Finally, the output (Equation 2.1) is obtained as the weighted sum of the values, with each value's weight determined by a compatibility function between the query and the corresponding key.

$$\text{Attention}(Q, V, K) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{2.1}$$

where d denotes the dimensionality of the Q, K and V matrices.

2.3.1 Vision Transformer

Inspired by the success achieved by Transformers [32] in the field of Natural Language Processing (NLP), Dosovitskiy *et al.* [11] has explored using Transformer's original design with minimal changes for image classification. Compared to CNNs, which primarily focus on local features, the proposed Vision Transformer uses self-attention to capture long-distance dependencies between image patches, enabling it to effectively derive global information about a given image.

Specifically, ViT first reshapes an image $X \in \mathbb{R}^{H \times W \times C}$ into a sequence of non-overlapping patches. Then, these patches, also treated as tokens, are linearly embedded into patch embeddings, and 1D positional embeddings are added to encode positional information. Additionaly, an extra learnable embedding, also designated as CLS token, is appended to the sequence of embedded patches, and the resulting sequence is fed to an encoder. The state of the CLS token after training and fine-tuning serves as global feature representation and is used for classification. An overview of the ViT model is depicted in figure 2.9.

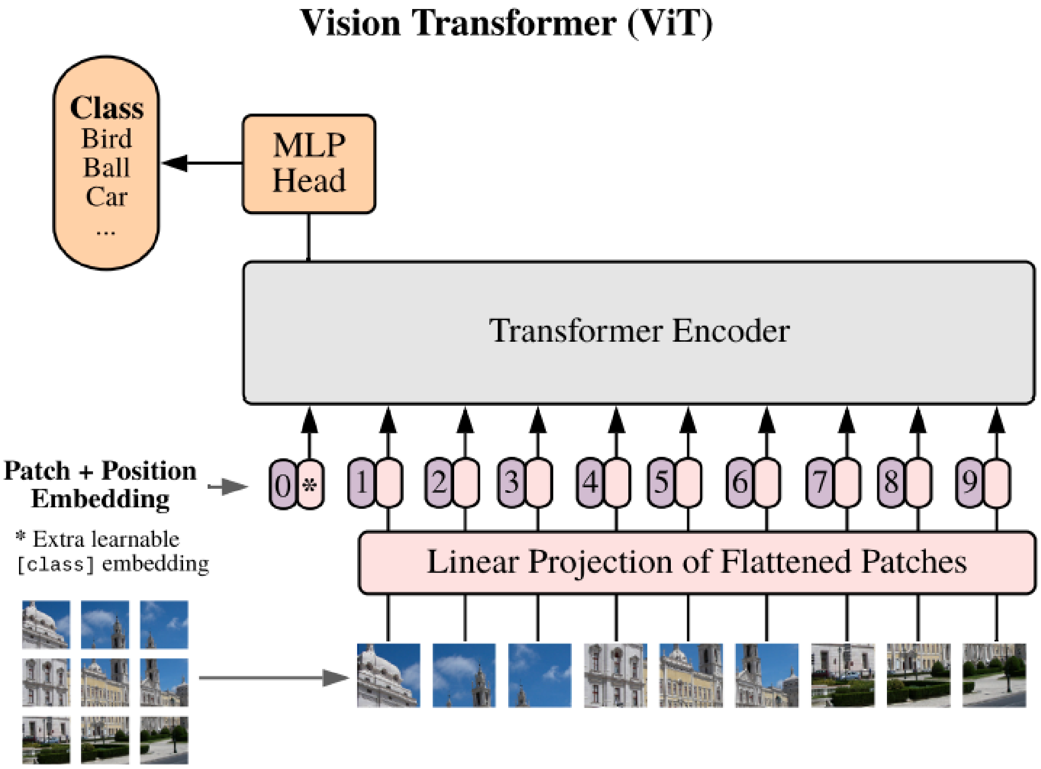

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 2.9: ViT model overview. Taken from [11].

ViT is the first computer vision model to rely exclusively on the Transformer architecture to achieve competitive image classification performance at a large scale. While it may be true that this kind of architecture has shown to be effective for feature extraction, ViT's performance is still below similarly sized convolutional models when trained on small amounts of data. This is due to the fact that Transformers lack some of the inductive biases inherent to CNNs, e.g., translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data. Moreover, the straightforward tokenization of input images by hard split can make ViTs unable to model local structures like edges and lines. Last but not least, they suffer from large model size and require substantial computational resources.

Nevertheless, the authors of ViT found that large-scale training surpasses the limitations of inductive bias and achieves state-of-the-art results on multiple image recognition benchmarks, i.e., ViTs reaches peak performance when pre-trained on large scale datasets (14-300 million images) and fine-tuned on smaller datasets for downstream recognition tasks.

2.3.1.1 Vision Transformer Variants

To adress the issue of large-scale training on huge datasets, Touvron *et al.* [33] proposed a Data-efficient image Transformer (DeiT). The model uses ImageNet as the sole training set and achieves competitive accuracy on the ImageNet benchmark with less data requirements and improved efficiency compared to ViT.

To solve the limitation of simple tokenization in ViT, Tokens-to-Token Vision Transformer (T2T-ViT) [34] incorporates a transformation layer that performs progressive image tokenization by recursively aggregating adjacent tokens into one token.

FocalViT [35], SWIN [36] and TNT [37] are three transformer variants that were designed to capture both global and local dependencies between image patches. FocalViT introduces a novel focal self-attention mechanism which allow each token to attend to its closest surrounding tokens at fine-granularity and to summarized tokens when it goes to farther regions of the image. SWIN constructs a hierarchical representation by first dividing the image into non-overlapping smaller patches and gradually merging them with neighboring patches at deeper transformer levels. Lastly, TNT divides each patch in sub-patches and then uses an inner transformer block to model the relationships between sub-patches and an outer transformer block to capture patch-level intrinsic information.

To enhance feature representation, CrossViT [38] introduces a dual-branch ViT that combines image patches of various sizes. Additionally, it incorporates a cross-attention technique that merges the CLS tokens from one branch with the patch tokens from the other branch, and vice versa. This enables the exchange of information between the branches, facilitating improved information flow and interaction between different parts of the model.

To address the problem of attention collapse in deeper transformers, the DeepVis [39] approach replaces the standard multihead self-attention block with a re-attention module. In ViT models, unlike CNNs, simply stacking more layers does not always grant improved performance as attention maps tend to become more similar in the higher layers, leading to less discriminative feature maps.

2.3.2 Hybrid Vision Transformer

To minimize the drawbacks and limitations of ViTs, some works in the literature have explored incorporating convolutions into transformers to introduce locality and improve the extraction of local features.

Dosovitskiy *et al.* [11] proposed a hybrid architecture that combines a Vision Transformer and a CNN. Rather than splitting the input image into patches, the input sequence to the ViT's encoder is formed by flattening and projecting the feature maps of a CNN to the Transformer dimension. Although the proposed model outperformed ViT for small model sizes, the performance gap disappeared for larger models.

Inspired by [11], Yuan *et al.*[40] proposed a Convolution-enhnanced image Transformer (CeiT) that combines a CNN for low-level feature extraction and a ViT for establishing long-range dependencies. However, CeiT has three main modifications compared to the vanilla ViT used in [11]. Firstly, an Image-to-Tokens (I2T) module is employed to extract patches from generated low-level features. These patches are smaller in size and flattened into a sequence of tokens. The second key change is a Locally-enhnaced Feed-Forward (LeFF) layer that replaces the standard feed-forward network and promotes correlation among neighboring tokens in the spatial dimension. Lastly, a Layer-wise Class token Attention (LCA) is attached at the top of the Transformer to attend over class tokens at different layers.

Even though these two former works only use CNNs to generate features maps which are then fed to a standard ViT, other works in the literature such as [41] and [42], attempted to strategically incorporate convolutional layers directly into ViT's architecture.

2.3.3 ViT-based methods using RGB images for FAS

As previously mentioned at the beggining of this chapter, the majority of research related to FAS has relied on using CNNs. As a result, only a limited number of studies have investigated the application of ViT's capabilities to identify spoofing cues.

The baseline approach in [12] proved that fine-tuning a pre-trained ViT for the PAD task was sufficient to achieve state-of-the-art performance. The authors only replaced the last layer with a fully-connected layer with one output and fine-tuned their model using a Binary Cross-Entropy (BCE) loss. The framework is depicted in figure 2.10. Multiple training stategies were also considered, including partially fine-tuning only a sub-set of layers to reduce overfitting. The proposed model not only achieved excellent results on unseen attack scenarios but it also showed remarkable performance on cross-domain generalization.

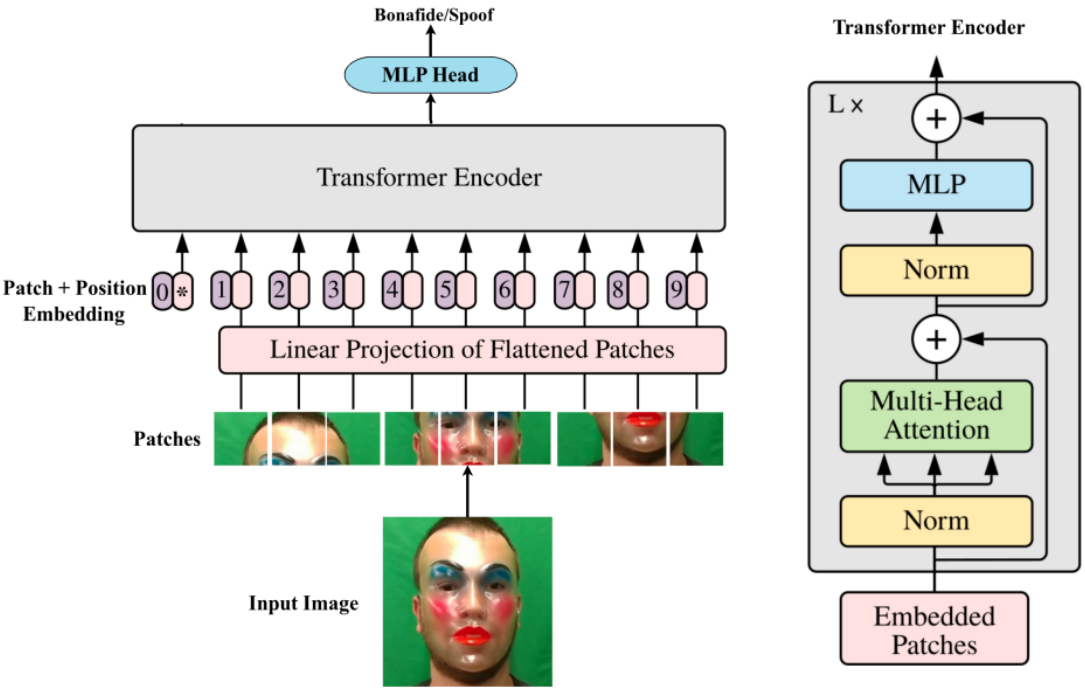


Figure 2.10: ViT framework for PAD task. Taken from [12].

In [13], it was proposed a new Multi-Teacher Single-Student (MTSS) ViT with a multi-level attention design. The proposed model consists of feature extractor ViT followed by a CNN and a Multi-Level Attention Module with Dropblock (MAMD). During training, the goal is to use the Vision Transformer to train a smaller student CNN and improve the student model's performance. Given an input image, the three color channels are converted into YCbCr color space, where Y, Cb and Cr channels are rearranged into a 1-D feature vector, positional encoded and fed into the visual transformer for feature extraction. Then, the MAMD block is used to adress overfitting during training by generating rich attented feature maps from multi level inputs while dropping irrelevant spatial features. These features are consequently converted and resized to match the input size. The overall architecture is shown in figure 2.11.

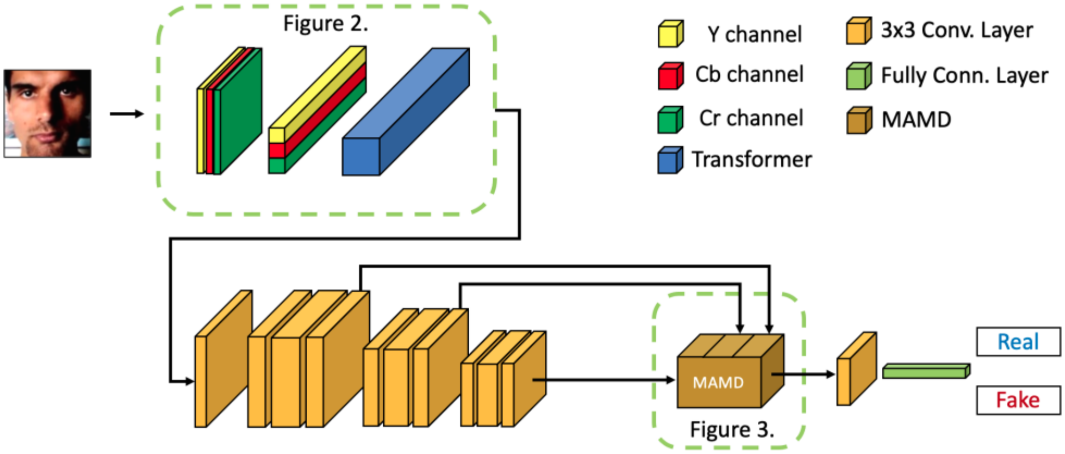


Figure 2.11: MTSS architecture. Taken from [13].

2.3.4 ViT-based methods using multi-modal images for FAS

Samar *et al.* [14] proposed a Multi-Modal Face Anti-Spoofing Transformer (MFAST) that utilizes two branches to independently process RGB and Thermal images. Both positional encodings and CLS tokens are applied to each branch. The resulting features from both branches are fused together and fed to a linear classifier. Figure 2.13 shows the proposed MFAST architecture.

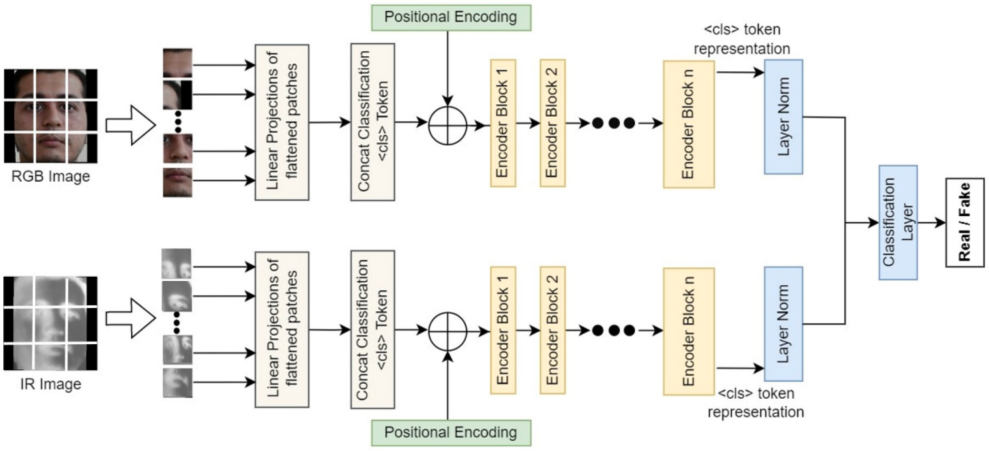

Positional Encoding
<cls> token representation
RGB Image
Linear Projections of flattened patches
Concat Classification <cls> Token
Encoder Block 1
Encoder Block 2
Encoder Block n
Layer Norm
Classification Layer
Real / Fake
IR Image
Linear Projections of flattened patches
Concat Classification <cls> Token
Encoder Block 1
Encoder Block 2
Encoder Block n
Layer Norm
<cls> token representation
Positional Encoding

Figure 2.12: MFAST architecture. Taken from [14].

In a similar approach, Ajian *et al* [15] proposed a flexible modal framework built on a multi-branch ViT. Specifically, the Flexible Modal Vision Transformer (FM-ViT) retains a specific branch for each modality and introduces a novel Cross-Modal Transformer Block (CMTB). Each block incorporates two Multi-headed Mutual-Attention (MMA) layers and two Fusion-Attention (FA) blocks to guide each branch to learn modality-agnostic features. The key innovation behind this approach is that the FA block enforces the CLS tokens of each modal sequence to be used as a query to enchange information with patch tokens of another modal sequence.

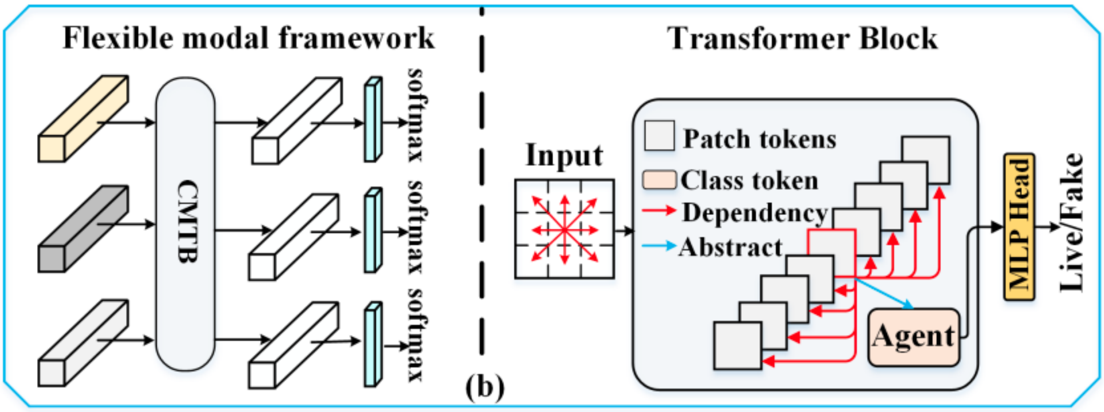


Figure 2.13: FM-ViT pipeline. Taken from [15].

Liu *et al.* [16] studied multiple fusion stategies on his Modality-Agnostic Vision Transformer MA-ViT. Despite the fact that halfway fusion is one of the most commonly used fusion strategies, it suffers from the drawback that if one modality disappears during testing, this method would fail to distinguish between real and spoof. On the other hand, late fusion strategies usually retain a specific branch for each modality to capture different modal information independently and fuses the multi-modal information at the decision level, resulting in large models to store. On the contrary, early fusion is used to reduce computational cost and improve efficiency by projecting multi-modal data into a joint embedding space at input level to capture intra and cross modality interactions within the transformer model. The decribed multi-modal fusion stategies are illustrated on figure 2.14.

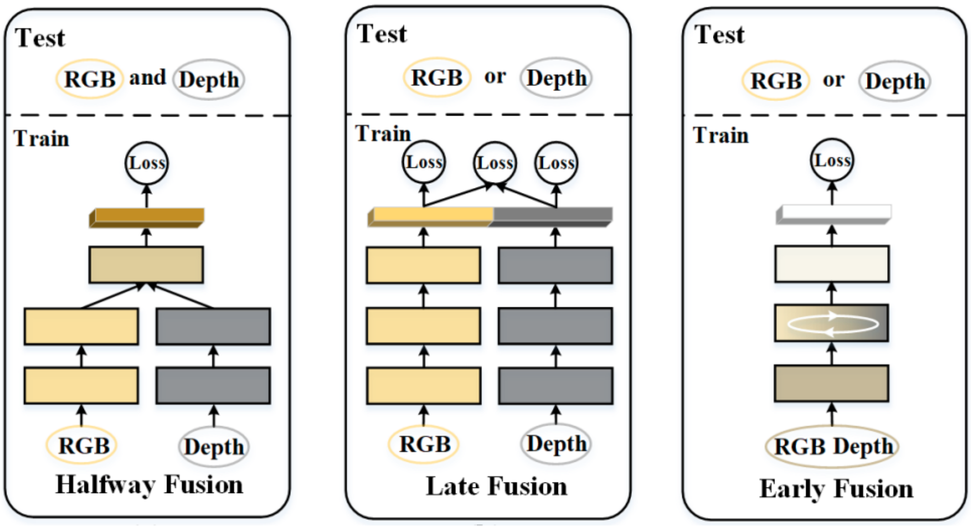


Figure 2.14: Comparison of existing multi-modal fusion strategies. Taken from [16].

2.4 Domain Generalization methods for Face Anti-Spoofing

Despite the notable achievements of previous CNN-based and ViT-based methods for FAS in intra-domain testing, the paradigm shifts for zero-shot learning and cross-domain scenarios where limited data and domain gaps become aditional challenges. This is primarly due to fact that intrinsic image characteristics, e.g., illumination, facial appearance, sensor types, camera quality, may change from domain to domain, leading to feature bias during training and poor generalization towards unseen domains. The disparity between feature spaces of real and spoof faces are often less pronounced within domains than across them. For this reason, it is important for FAS solutions to be trained on one or multiple source domains and adapt to unseen domains or unknown attacks.

The work by [17] proposed a domain invariant MobileViT supervisioned by two losses to learn a domain-invariant latent space. To accomplish this, a concentration loss is employed to encourage real-face embeddings to not be biased to specific domains and converge its embeddings towards the origin. On the contrary, an attack-separatation loss groups multiple attacks and separates their representation from real-faces, i.e., this loss pushes attack embeddings away from the origin. Figure 2.15 overviews the proposed method.

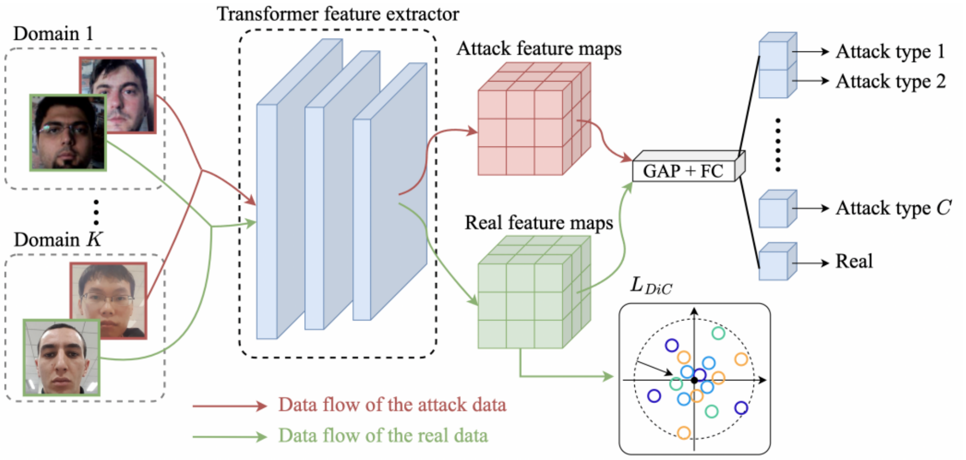


Figure 2.15: Latent space approach. Taken from [17].

Huant *et al.* [18] proposed an adaptive ViT for robust few-shot cross-domain FAS. The key innovation is the introduction of two novel components: ensemble adapters and Feature-Wise Transformation (FWT) layers. Each ensemble adapter first linearly projects the n-dimensional features into a lower dimension m, applies a non-linear activation function GELU and then projects back to n dimensions. The main goal of the adapter is to help adjust feature distribuition of the pre-trained transformer blocks to the face anti-spoofing data, as well as granting training stability and prevent overfitting. Moreover, a cosine similarity loss L_{cos} is applied to the outputs h_i and h_j of the adapter to enforce the learning of diverse features. On the other hand, the FWT layer is only used during training as can be seen as an augmentation technique. This layer aims to increase the diversity of training samples, thus dramatically reducing overfitting. According to the authors, the incorporation of both components have improved the performance of few-shot cross-domain FAS by allowing the model to adapt to new domains, particularly in cases where there is a low volume of training data. Figure 2.16 provides an overview of the framework.

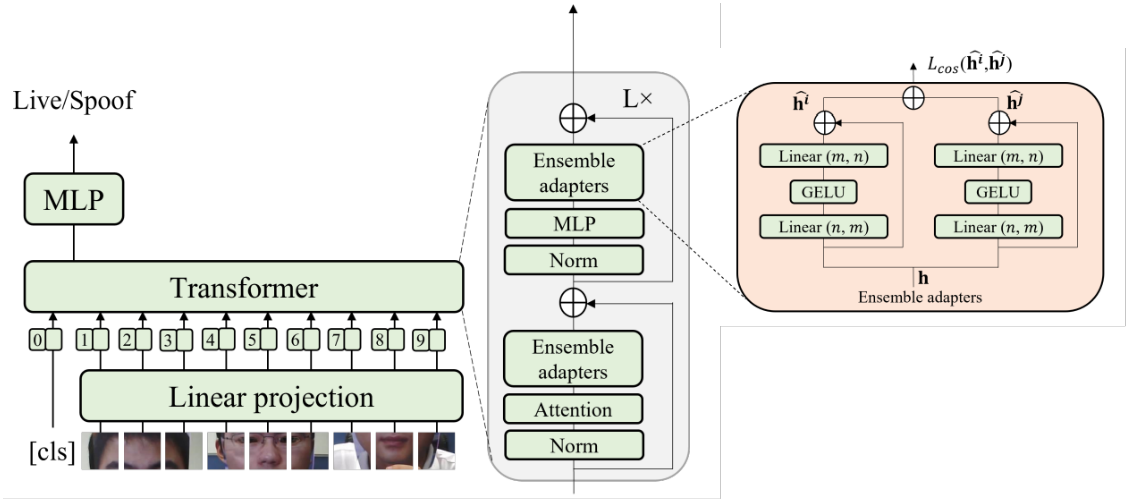


Figure 2.16: Adaptive ViT. Adapted from [18].

2.4.1 Style Augmentation based methods

In the scope of domain generalization, style augmentation involves transferring style between images from different domains, where one image provides visual context, e.g., color, texture, contrast, brightness, and the other provides high-level semantic content. The transfer of style between source domains generates synthetic images that are usefull in several ways for FAS tasks. On one hand, synthetic images can increase the size of training data and diversify inputs, i.e., styled augmentation can generate spoof attacks from live faces, as [43] and [44] demonstrates. On the other hand, they can be used to approximate the semantic space and reduce disparity between source domains, leading to a generalized feature space.

The method proposed in [19] employed a two branch network called Shuffled Style Assembly Network (SSAN) to perform style transfer at feature level across source domains. One of the branches is used to extract style information while the other is used to extract content features, e.g., global semantic features and physical attributes. Rather than performing style transfer image-to-image like in [45], the proposed method uses a cascade of shuffled style assembly layers to reassemble content and style features under a contrastive learning stategy. This stategy emphasizes liveness-related style information and suppresses domain-specific ones by pulling the shuffled-assemly features close or far from an anchor point. The pipeline is shown in figure 2.17.

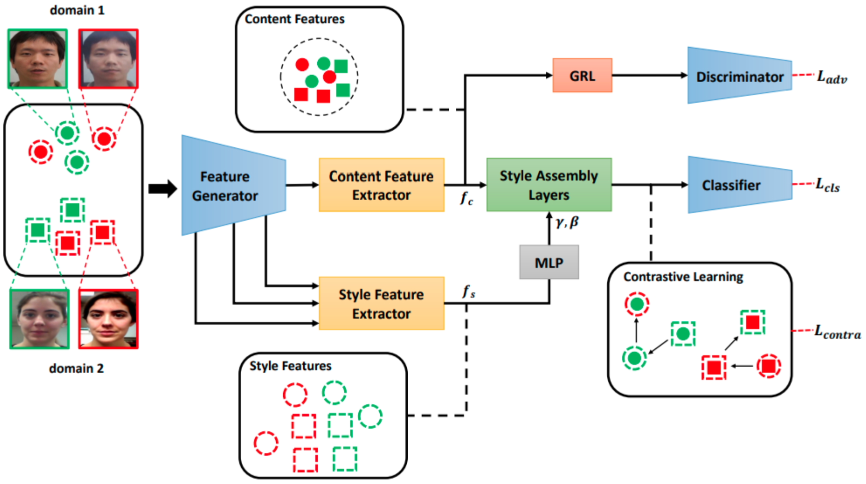


Figure 2.17: SSAN pipeline. Taken from [19].

2.5 Summary

Throughout this chapter, we discussed several methodologies presented in the literature that are based on CNNs and Vision Transformers. These methodologies detail the evolution of this field and have contributed to the development of algorithms aimed at addressing the challenges of FAS.

It is important to note that while FAS research based on CNNs has been extensively explored and will not be the main focus of this dissertation, ViT-based approaches are relatively less utilized in this field, primarily due to their high data requirements. Firstly, although ViT architectures have already shown superior performance compared to CNNs in other computer vision tasks, a pre-trained ViT backbone may not adapt well to the specific facial data used in the task. Secondly, the features extracted from a pre-trained backbone are typically at a high-level, which may not be suitable for detecting subtle low-level information that is crucial in face anti-spoofing.

For this reason, the following chapters will focus on the development of strategies based on these two architectures that incorporate contributions from other relevant works in the literature such as the integration of multi-modal data, methods for merging modalities, and techniques to assist models in cross-domain scenarios, such as style transfer and the utilization of specific loss functions.

3 Methodology

This chapter aims to provide a comprehensive overview of the experimental work developed throughout this thesis.

Initially, sections 3.1, 3.2, and 3.3 delve into the multi-modal baselines (ViT, ResNet, Hybrid) developed in this dissertation. These sections provide a detailed examination of how these baselines were structured and adapted to classify multi-modal data for the task of FAS. The subsequent sections focus on explaining the techniques implemented to enhance the performance of these models, providing insights into the strategies employed to optimize their effectiveness.

Section 3.4 provides a careful explanation of the formulation of the multi-modal fusion strategies. In section 3.5, the style transfer technique is introduced, and the differences between various mixing strategies are explained. Lastly, section 3.6 presents two types of losses. The first is the loss used in all baselines, while the second is a loss employed to enhance performance in cross-domain scenarios.

3.1 Multi-modal ViT Baseline Network

The first goal was to adapt the configuration of the original Vision Transformer [32] to a multi-modal scenario. In the proposed baseline, the backbone parameters are shared across all modalities and the multi-modal learned CLS embeddings (tokens) are fused via concatenation and used for classication. Figure 3.1 overviews the multi-modal ViT pipeline.

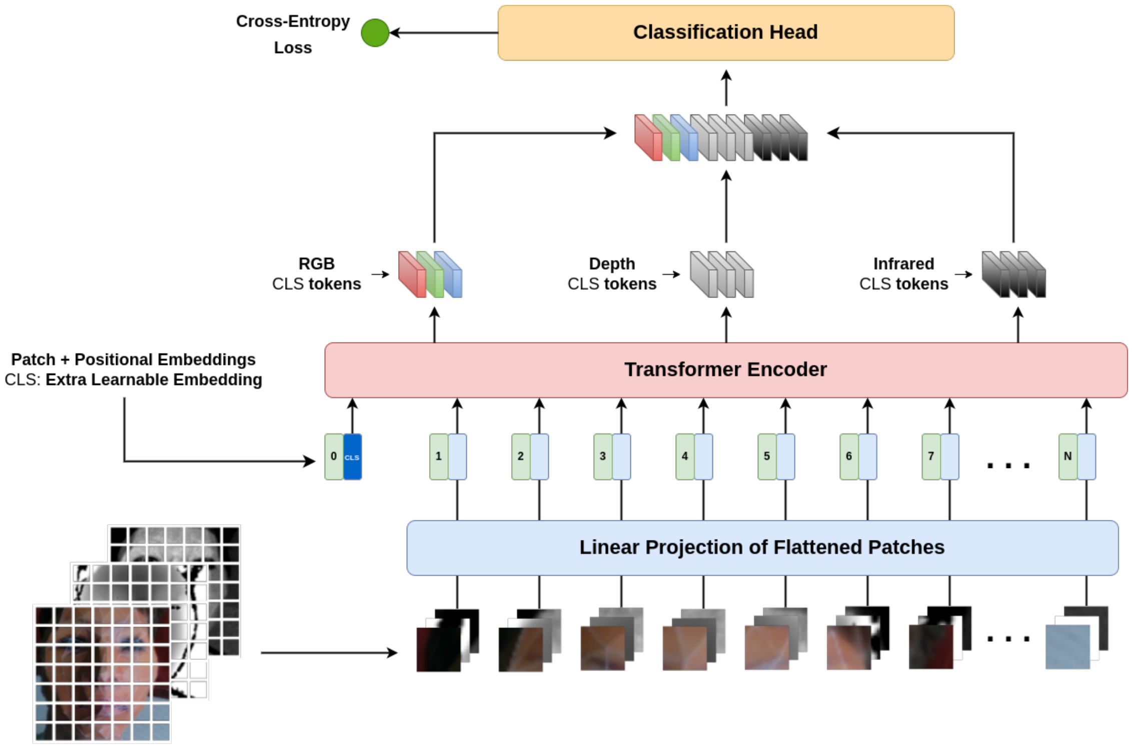


Figure 3.1: Multi-modal ViT pipeline.

3.1.1 Patch Embeddings

The images from different modalities are sequentially fed to the backbone. Since ViTs only process images with 3 channels, each image from the Depth and Infrared modalities had to be stacked up to create an input image with three grayscale channels.

For any given image $X \in \mathbb{R}^{C \times H \times W}$ with height H, width W, and number of channels C, it is first divided into a grid of N non-overlapping patches $x_p^i, i = 1, 2, ..., N$=196. Next, these patches are flattened and tokenized into a sequence of embedded patches, as illustrated in figure 3.2. Each patch x_p^i is linearly projected by a convolutional layer with a $P \times P$ kernel and stride P and mapped to a single 1D vector with dimension D=768. We denoted this projection as $E(x_p^i)$ and D is kept fixed throughout the layers. After the projection of all patches, the sequence of embedded patches, i.e., x_p^1E, x_p^2E,.., x_p^NE, are concatenated to create a $N \times D$ patch embedding matrix.

In addition to the embedded patches, an extra 1D learnable CLS embedding with D-dimension is appended to the patch embedding matrix. This embedding vector interacts with patch tokens at every transformer encoder and serves as global feature representation as it contains global information about the image. Futhermore, a learnable positional embedding P_e is added to each embedded patch vector to encode positional information betwen patches, i.e., this embedding vector is added in matrix form. The resulting sequence of embedding vectors Z_0 with size $N+1 \times D$ serves as input to the encoder and is described by Equation 3.1.

$$Z_0 = \{CLS; x_p^1E; x_p^2E; ...; x_p^NE\} + P_e \quad CLS \in \mathbb{R}^{1\times D}, x_p^iE \in \mathbb{R}^{1\times D}, P_e \in \mathbb{R}^{(N+1)\times D} \tag{3.1}$$

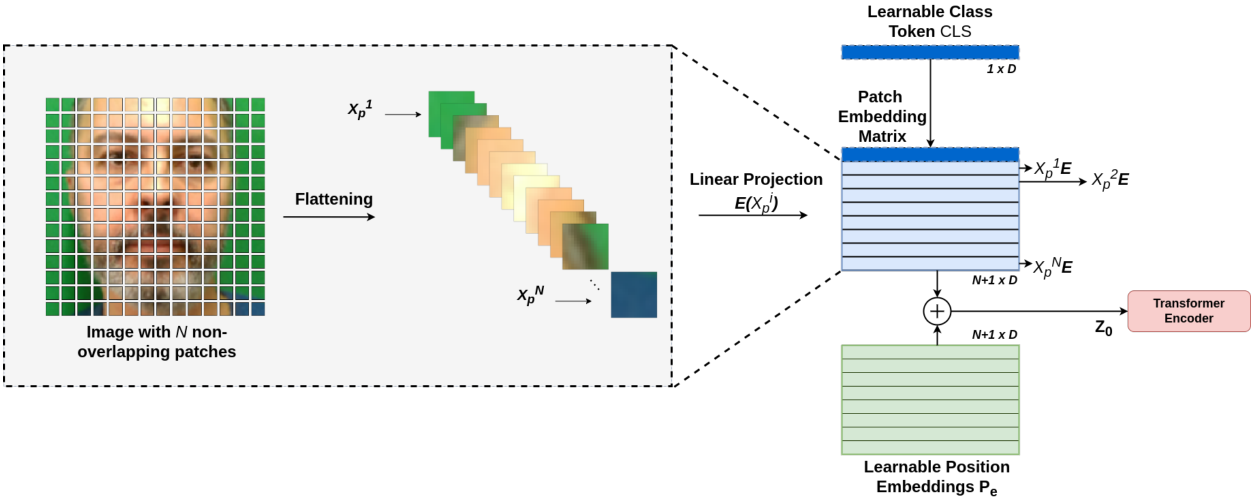


Figure 3.2: Tokenization of image patches.

3.1.2 Positional Encodings

To encode spatial information, positional encodings are generated using sinusoidal functions of varying frequencies and added directly to the embedded patches. The main purpose of these embeddings is to provide the model with information regarding the sequence order of the tokens. Furthermore, these encodings are learnable and implemented as a fixed-dimensional matrix P_e that is updated alongside the model parameters.

$$Pe(pos, i) = \begin{cases} sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}}) & \text{if} \quad \text{i is even} \\ cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}}) & \text{if} \quad \text{i is odd} \end{cases} \tag{3.2}$$

where $pos \in \mathbb{R} : 0 < pos < N$ denotes the token position and $i \in \mathbb{R} : 0 < i < D-1$ represents the current position along the embedding dimension.

3.1.3 Encoder

The Transformer is composed of a series of L=12 encoder blocks. The encoder block is depicted in figure 3.3. Each block comprises alternating Multi-head Self-Attention (MHSA) layers, Multilayer Perceptron (MLP) blocks and normalization blocks. Considering that the sequence of embedded vectors is given by Z_{l-1}, the flow of tokens throughout each encoder block is the following: Z_{l-1} is first normalized, passed through the MHSA block and added to its previous state to produce an output Z_l' as shown in Equation 3.3. The output Z_l' is then normalized and fed to the MLP block for token reprojection. This output is added to its previous state via residual connection to produce Z_l, as shown in Equation 3.4.

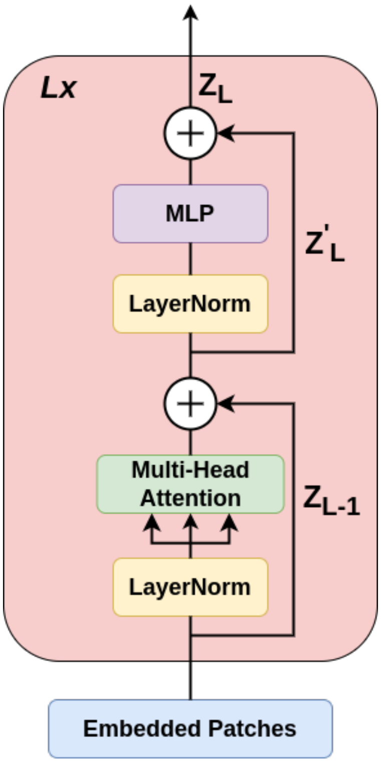


Figure 3.3: Encoder block.

$$Z_l' = MSA(LN(Z_{l-1})) + Z_{l-1},\ Z_l' \in \mathbb{R}^{(N+1)\times D},\ l = 1, ..., L \tag{3.3}$$

$$Z_l = MLP(LN(Z_l')) + Z_l',\ Z_l \in \mathbb{R}^{(N+1)\times D},\ l = 1, ..., L \tag{3.4}$$

The residual connections help the model to mitigate the vanishing gradient problem and allow better flow of information through the layers. In the normalization blocks, the tokens are normalized by a LayerNorm function according to their mean and standard deviation and then are linearly transformed by a set of weights and biases. The MHSA block is used to boost the performance of the vanilla self-attention mechanism lightly described in section 2.3 and is illustrated in figure 3.4.

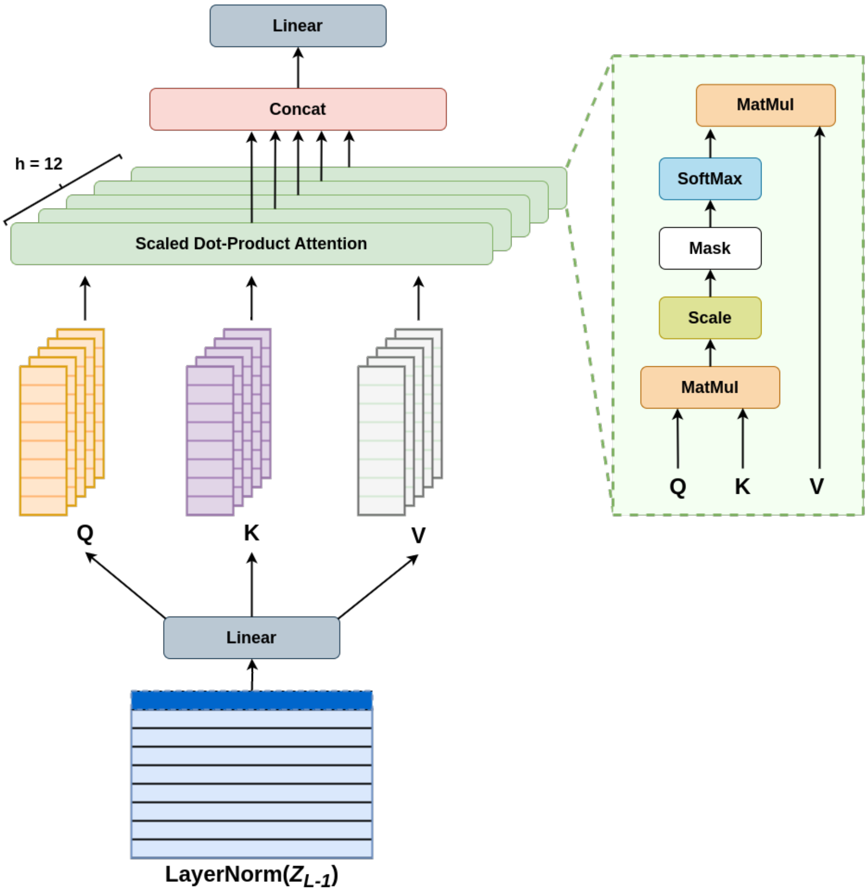


Figure 3.4: Multi-head self-attention block.

At this block, the sequence of embedded vectors Z_{l-1} is first expanded by a fully-connected layer and then broken up into a set of Query, Key and Value matrices (Q, K and V), each one with the same size as Z_{l-1}. Instead of applying a single attention function to the original sets of Q, K and V, each vector is further split into h=12 different paralell heads, i.e., Q, K and V $\in \mathbb{R}^{h \times N \times D/h}$. Each head is responsible for learning a different representation by performing self-attention (equation 3.5) on the smaller split versions of Q, V and K.

$$\text{Attention}(Q, V, K) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{3.5}$$

To accomplish this, the dot product of the query vector with the key vector for each pair of query and key vectors is computed to calculate the attention weights, also known as attention matrix. In pratical terms, this operation represents of much each query matches a given key, i.e., how much each token is paying attention to other tokens in the sequence. The resulting attention scores are then normalized by a softmax operator to scale the values to a probability function and multiplied to the V vector. In the end, the outputs of all attention heads are concatenated to form an embedding vector whose shape is the same as the encoder input Z_{l-1}. Finally, this vector is passed through another linear layer to obtain the final output of the MHSA.

The MLP block is composed by two fully-connected layers with dropouts and GeLU activations. The first fully-connected layer expands the embeddings from D to $4D$ dimensions, while the second layer shrinks them back to D dimensions. For regularization and to prevent overfitting, dropout is applied to the output of each fully-connected layer.

3.2 Multi-modal ResNet Baseline Network

Rather than solely relying on self-attention mechanisms, a multi-modal ResNet baseline was designed to study the effectiveness of a convolutional approach for FAS tasks and to provide a comparation metric with the former ViT based architecture. The proposed pipeline is shown in figure 3.5 and employs three branches to process the different modalities of the input data individually.

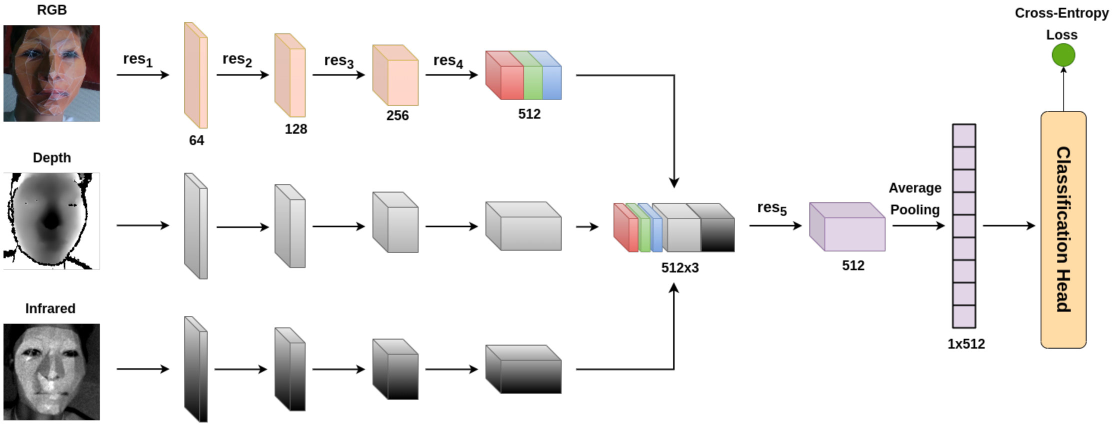


Figure 3.5: Multi-modal ResNet pipeline.

The network can be divided into two main parts. In the first part, each branch of the network uses a full ResNet-18 backbone with res_i blocks, $i = 1, 2, 3, 4$, to extract features from input images. Since each branch deals exclusively with images from one modality, all branches become specialized in extracting features from different modalities, thus enhancing the model ability to capture sensitive modality-related information. In the second part, the deep modal features are fused via concatenation and passed through a fifth ResNet block, denoted by res_5. Unlike the other res blocks, this block is used to decrease the depth of the multi-modal feature map and to withdraw interdependencies between different modalities. The output vector is then averaged using a pooling layer and used for classification.

3.2.1 Residual Connections

Each *res* block is composed of $N=2$ residual blocks and each residual block consists of two 3×3 convolutional layers, followed by batch normalization layers and ReLU activation functions. The first residual block within each *res* block is used to increase the number of channels C of the given feature map $X \in \mathbb{R}^{C \times H \times W}$. To accomplish this, an extra 1×1 convolutional layer is applied to match the shape of the input X with the pre-output $F(X)$ of the residual block. These two feature representations are then added via skip connections before the last ReLU activation to produce an output feature map $F(X)+X$ that has twice the channels and half the feature size of X. This feature map is then fed to the second residual block. An illustration of a *res* block along with its residual blocks is shown in figure 3.6.

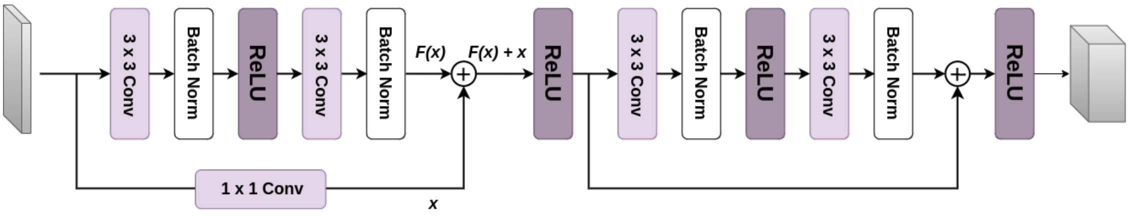


Figure 3.6: Residual Block.

3.3 Multi-modal Hybrid Baseline Network

To combine the strenghts of both Vision Transformers and ResNets, a multi-modal Hybrid baseline was built by merging the properties of the two previous baselines. The proposed Hybrid pipeline is shown in figure 3.7. The model employs a three branch ResNet backbone to extract high-level features from multi-modal images and a ViT backbone to attend to the spatial relationshipts between the extracted features. Rather than splitting the input image into patches and then linearly project each patch, the input sequence fed to the ViT backbone is formed by feature maps extracted from the ResNet backbone.

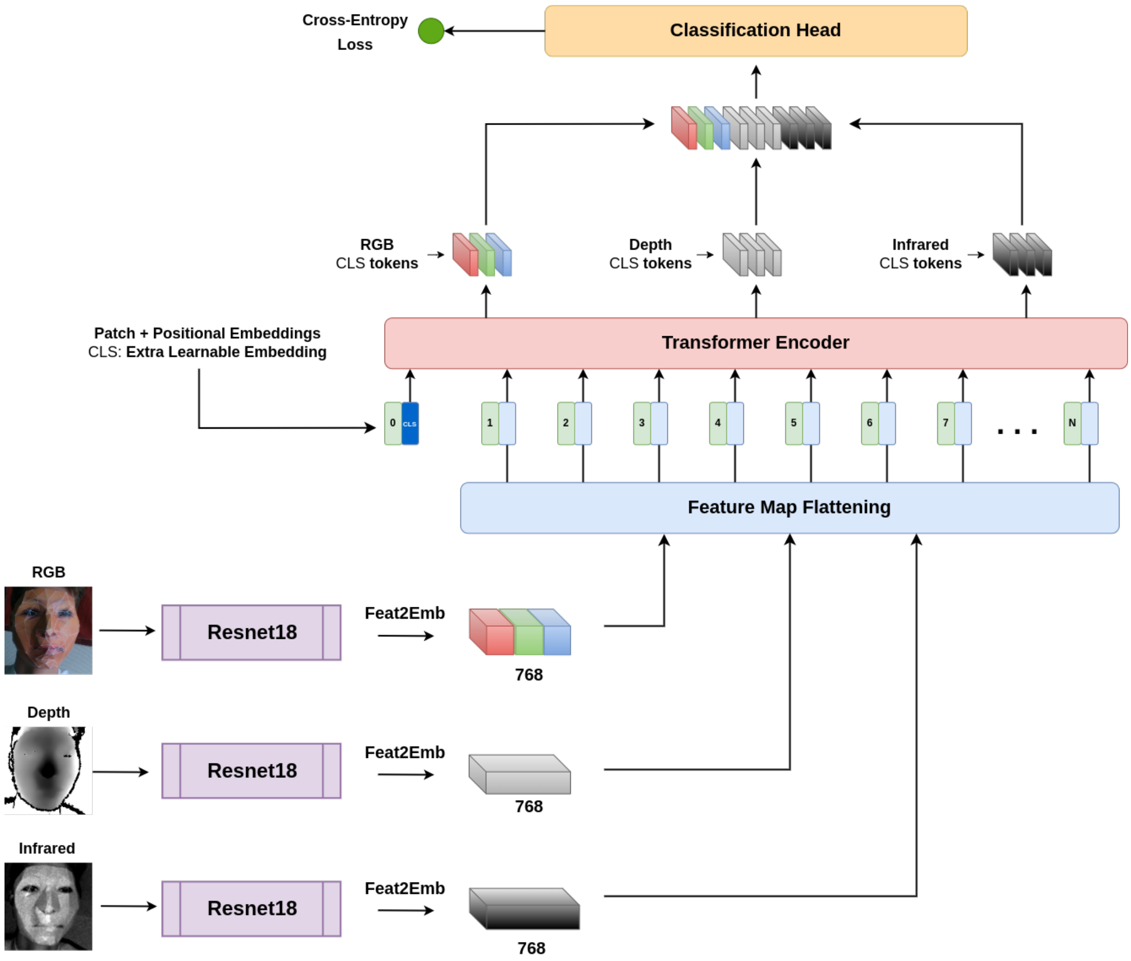


Figure 3.7: Multi-modal Hybrid pipeline.

3.3.1 Hybrid Patch Embeddings

Given an image $X \in \mathbb{R}^{H \times W \times C}$ with height H, width W, and number of channels C, it is first processed by the ResNet backbone for feature extraction. The output of the backbone is a feature map $F_m \in \mathbb{R}^{512 \times 14 \times 14}$ which is then passed through a *Feat2Emb* convolutional layer. This layer serves the puporse to match the shape of the output feature map F_m with the proper dimensions needed to build the patch embedding matrix, i.e., this layer increases the depth of the feature map F_m to dimension D=768 and keeps its spatial size. The final output of the convolutional part of the hybrid model is a feature map $F_m \in \mathbb{R}^{D \times 14 \times 14}$, which is then flattened along dimension D. As a result, each channel of the feature map F^i_m, $i = 1, 2, 3, ...D$ is flatenned by an operation we denoted by $F(F^i_m)$ to produce a 1D vector $F^i_m F$ with size N=196. Analogous to the standard version of ViT, both learnable CLS tokens and positional embedding P_e are added to the sequence of concatenated flattened feature maps. Figure 3.8 provides a visual representation of the tokenization of feature maps.

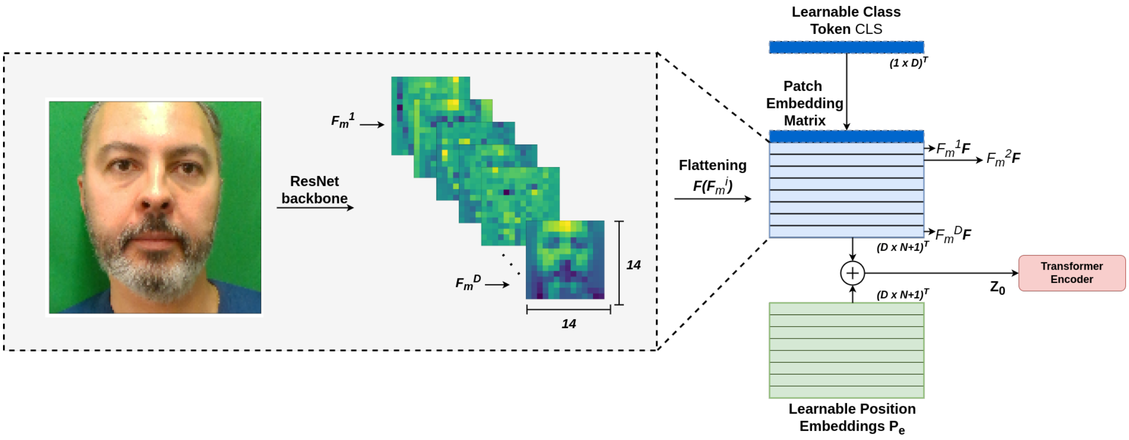


Figure 3.8: Tokenization of feature maps.

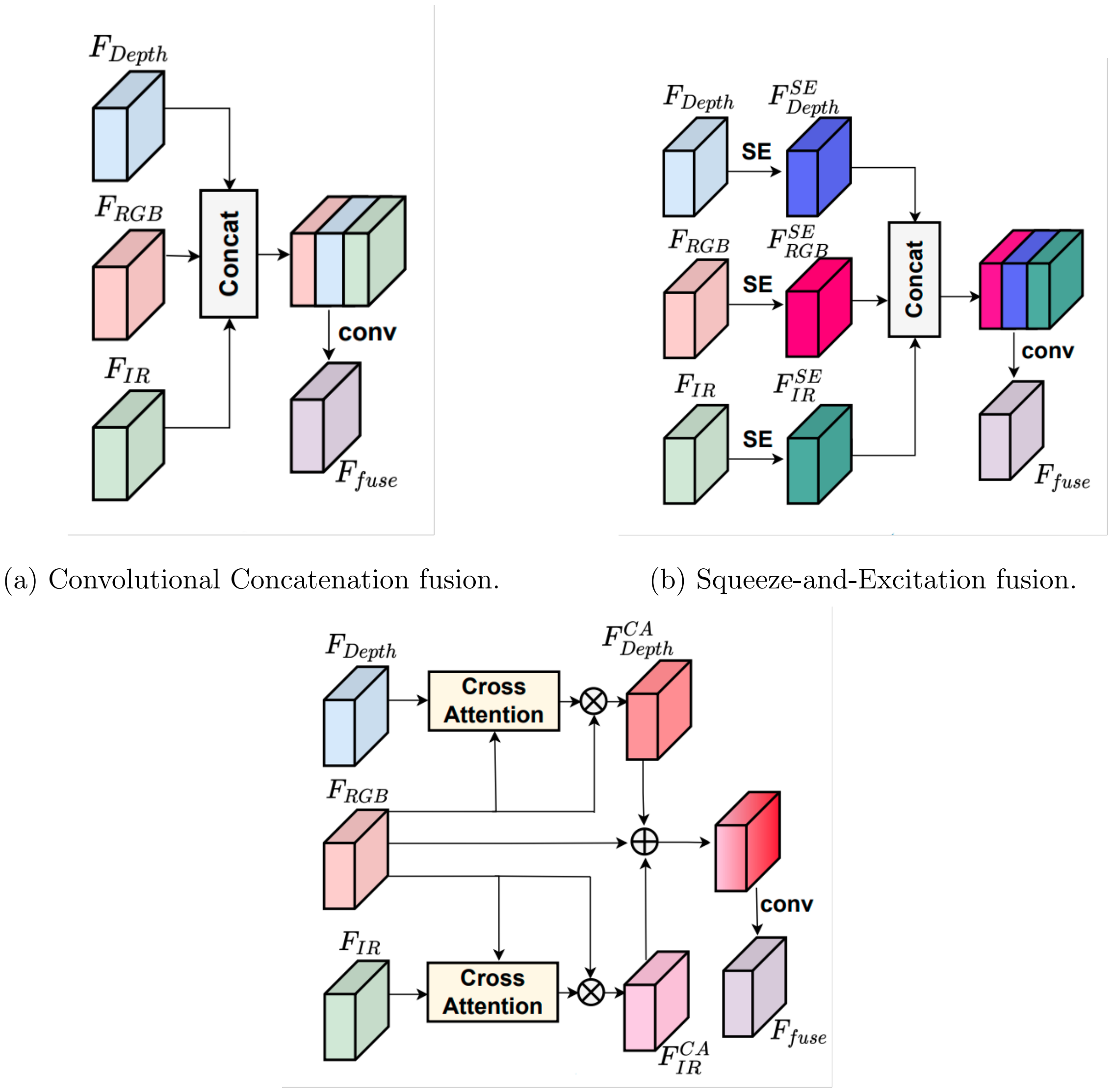
3.4 Multi-modal Fusion Methods

Fusion methods for multi-modal datasets refer to methods used to combine information from different sources of data into a joint representation. In the field of FAS, sophisticated attacks often require information from multiple modalities to be detected. Therefore, several studies in the literature have explored ways of leveraging information from different modal data to increase the robustness of models. Inspired by the work of [20], multiple late fusion strategies were incorporated into baseline models as an alternative to the straightfoward concatenation, which concatenates all features from different modalities without any additional processing.

On one hand, to implement the fusion stategies in ViT and Hybrid baselines, rather than utilizing the class tokens CLS attached to the sequence of vectors of patch tokens, the whole sequence of patch tokens was used as feature map. On the other hand, in the ResNet baseline, the feature map from each branch before fusion was utilized. The framework of each fusion method is shown in figure 3.9.

Convolutional Concatenation

This type of concatenation is illustrated in Figure 3.9a) and differs slightly from the baseline direct concatenation used in the ViT and Hybrid baselines. Instead, the proposed method is very similar to the fusion method used in the ResNet baseline which first concatenates the feature maps (F_{RGB}, F_{Depth}, and F_{IR}) and employs a convolutional layer to mine the dependencies among the modal features. This method is formulated in Equation 3.6.



(a) Convolutional Concatenation fusion.

(b) Squeeze-and-Excitation fusion.

(c) Cross-attention fusion.

Figure 3.9: Multi-modal fusion methods. Taken from [20].

$$F_{fuse} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(\mathrm{Concat}(F_{rgb}, F_{Depth}, F_{IR})))) \tag{3.6}$$

Squeeze-and-excitation fusion

For this type of fusion, a Squeeze-and-Excitation module is utilized in each independent modality branch. This module recalibrates each channel of the feature map to create a more robust representation by enhancing the important features while ignoring the irrelevant ones. The refined features (F_{RGB}, F_{Depth}, and F_{IR}) are then concatenated and aggregated according to the formulation described in Equation 3.7(σ denotes the Sigmoid function). This fusion method is shown in figure 3.9b).

$$\begin{aligned} F_{RGB}^{SE} &= F_{RGB} \cdot \sigma(\mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(\mathrm{AvgPool}(F_{RGB}))))) \\ F_{Depth}^{SE} &= F_{Depth} \cdot \sigma(\mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(\mathrm{AvgPool}(F_{Depth}))))) \\ F_{IR}^{SE} &= F_{IR} \cdot \sigma(\mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(\mathrm{AvgPool}(F_{IR}))))) \\ F_{fuse} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(\mathrm{Concat}(F_{RGB}^{SE}, F_{Depth}^{SE}, F_{IR}^{SE})))) \end{aligned} \tag{3.7}$$

Cross-attention fusion

Rather than fusing features from a heterogeneous space, feature addition in the homogeneous space was also explored. Therefore, relationship maps between F_{RGB} and F_{Depth}/F_{IR} are computed through cross-attention. The resulting normalized modality-interacted maps are then multiplied by F_{RGB} to form cross-attentioned features, namely F_{RGB}^{CA} and F_{IR}^{CA}. Finally, the original RGB features F_{RGB} are added to the cross-attentioned features and fused using an additional convolution. Cross-attention fusion is formulated in Equation 3.8 and illustrated in figure 3.9c).

$$\begin{aligned} \bar{F}_{Depth}^{CA} &= \mathrm{Softmax}(\bar{F}_{Depth}(\bar{F}_{RGB})^{T})\bar{F}_{RGB} \\ \bar{F}_{IR}^{CA} &= \mathrm{Softmax}(\bar{F}_{IR}(\bar{F}_{RGB})^{T})\bar{F}_{RGB} \\ F_{fuse} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(F_{RGB} + F_{Depth}^{CA} + F_{IR}^{CA}))) \end{aligned} \tag{3.8}$$

where F and $\bar{F}$ denote the spatial features and vectorized features, respectively.

3.5 Style Augmentation: Mixstyle

Mixstyle [45] is a type of style augmentation that was designed to regularize the training of a model by mixing style information of source domain features in cross-domain scenarios. Unlike other methods, it does not require the generation of synthetic images of new styles as it is implemented into batch training, i.e., this module performs style transfer by only mixing instance-level feature statistics across images from the same batch.

Considering that an input batch x is composed by images of domains x_1 and x_2, then $x = [x_1^i, x_2^j]$, i.e,. i, j denote the image index position in the respective domain batch. The goal of mixstyle is to generate a reference batch x_{ref} by shuffling image positions between both domains and use it to compute mixed feature statistics using Equations 3.9 and 3.11.

$$\gamma_{mix} = \lambda\sigma(x) + (1 - \lambda)\sigma(x_{ref}) \tag{3.9}$$

$$\beta_{mix} = \lambda\mu(x) + (1 - \lambda)\mu(x_{ref}) \tag{3.10}$$

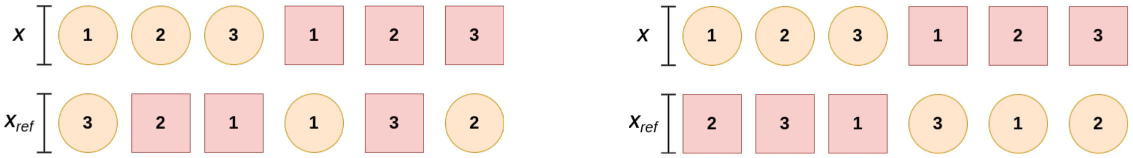
where $\lambda \in \mathbb{R}$ is an intance weight sampled from a Beta distribution, $\lambda \sim Beta(\alpha, \alpha)$ with α=0.1. Then, the mixed feature statistics γ_{mix} and β_{mix} are used to compute the style-normalized x:

$$\text{Mixstyle}(x) = \gamma_{mix}\frac{x - \mu(x)}{\sigma(x)} + \beta_{mix} \tag{3.11}$$

The generation of x_{ref} depends on the shuffling strategy applied to the input batch x. Therefore, two mixing strategies are proposed:

- Random Mix: the order of images in x is randomly shuffled, i.e., x_{ref}=[Shuffle(x)]. In this scenario, images may be styled with images from the same domain.

- Crossbatch Mix: x_{ref} is obtained by swapping domain order and shuffing image positions within each domain. i.e., x_{ref} =[Shuffle(x_2^j),Shuffle(x_1^i)].



(a) Random Mix.

(b) Crossbatch Mix.

Figure 3.10: Types of batch shuffling. Each shape corresponds to a different domain.

Mixstyle is applied on every batch during training and deactivated when the baselines are validating and testing.

3.6 Loss Functions

3.6.1 Cross-Entropy Loss

Given a certain image, it is either classified as real/bonafide or attack/spoof. To measure the difference between the outputs and the true class labels, a cross-entropy loss L_{CE} was implemented across all baseline models. In practical terms, this loss applies a softmax function direcly to the output tensor of the classification head to obtain the predicted class probabilities and computes the negative log-likelihood between the predited probabilities and the true labels.

$$L_{CE} = -\sum_{i=1}^{N} \hat{y}_i \mathrm{log} y_i \tag{3.12}$$

where N is the batch size and $\hat{y}_i$, y_i are the truth label and the predicted probability distribuition of the i^{th} image respectively, i.e., y_i is given by the classification head.

3.6.2 Domain-invariant Concentration Loss

In addition to the cross-entropy loss L_{CE}, this work incorporates a domain-invariant concentration loss, denoted as L_{DiC}, inspired by the work of Liao *et al.* [17]. The purpose of the L_{DiC} loss is to enhance the generalization capability of the baseline models in cross-domain scenarios. By aggregating the features of real face images, this loss facilitates the learning of a domain-invariant representation, which is crucial for achieving robust performance across different domains.

Let $D_1, ..., D_K$ be representation of K datasets, where each dataset specifies one domain. Considering that each domain constains C types of attacks and real images, the previous notation can be further extended to symbolize the set of attacks in domain k, D_k^c, where c is the c-th type of attack with $c \in [1, ..., C]$. In addiction, D_k^{real} indicates the set of real face images within the same domain $k \in [1, ..., K]$. To start with, all real faces from different domains D_k^{real} are treated as a unified positive (non-spoof) class of data, combining them in the following manner:

$$D^R = \bigcup_{k=1}^{K} D_k^{real} \tag{3.13}$$

This unification serves two purposes. Firstly, it aims to provide any given real face image in D_k^{real} with a feature representation that is not biased towards specific domains. This ensures that the learned representation remains invariant to domain changes, enabling the model to generalize well across different domains. Secondly, regardless of the domain of a real face image, the objective is to ensure that its feature embedding is positioned close to the origin of the embedding space. To accomplish this, the loss function L_{DiC} is employed to encourage the learned feature embeddings of real face images in all domains to have smaller norms. The formulation of L_{DiC} is explicitly described in Equation 3.14.

$$L_{DiC} = \frac{1}{N} \sum_{i=1}^{N} 1[x_i \in D^R] \cdot ||f_i|| \tag{3.14}$$

where N denotes de batch size, 1 is the indicator function, '$\cdot$' is the inner product and f_i is the i-th feature embedding of the input image x_i.

Finally, the domain-invariant concentration loss L_{DiC} and the cross-entropy loss L_{CE} are combined to train the baseline backbones in a supervised manner. The total loss L_{total} is defined in Equation 3.15 and is depicted in Figure 3.11.

$$L_{total} = L_{CE} + \lambda L_{DiC} \tag{3.15}$$

where λ is a balance factor, $\lambda \in [0, \infty]$.

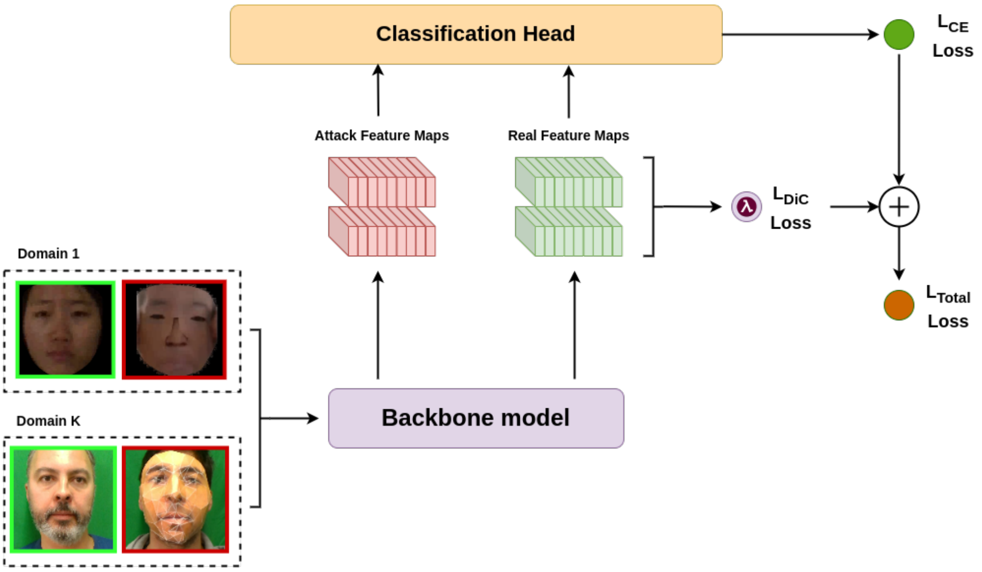


Figure 3.11: Pipeline of the training under L_{DiC} loss.

4 Results and Discussion

This chapter is divided in two main parts. The first part, consisting of sections 4.1, 4.2, 4.3, and 4.4, focuses on technical aspects related to the datasets, implementation details, evaluation metrics and the interpretability of Vision Transformers in the context of FAS. The second part, section 4.5, provides a detailed examination about the experimental results. It starts by covering intra-domain, cross-domain and zero-shot evaluation using all baseline models and then transitions to the different approaches that have been incorporated to enhance the performance of the baselines, specially in the cross-domain scenario.

4.1 Datasets

4.1.1 WMCA

WMCA [22] is a publibly available preprocessed PAD dataset which consists of 1679 short video samples of bonafide representations, 2D and 3D attacks from 72 different individuals. It contains synchronized multi-channel data from several channels including RGB, Depth, Infrared and Thermal. Two sensors, e.g., Intel RealSense SR300 and Seek Thermal Compact PRO, captured the data synchronously during 10 seconds using different resolutions and frame-rate depending on the channel. The color channel was recorded with a resolution of 1920×1080, while the other channels were recorded at 640×480. Furthermore, all data was recorded during several sessions, each with different environmental conditions, such as lighting and background.

The presentations in the database are grouped into two main categories: Bonafide, which is the real representation of individuals and presentation attacks. Due to the wide variety of presentation attacks, they can be further divided into sub-categories. Table 4.1 provides the acquisition conditions of each PA category and figures 4.1 and 4.2 illustrate some examples.

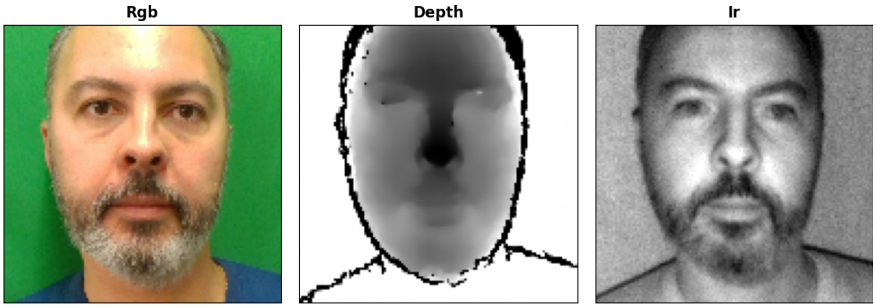


Figure 4.1: RGB, Depth and Infrared bonafide images from the WMCA dataset.

Table 4.1: Acquisition condition for each PA.

Attack type	**Acquisition condition**
Print	Printed face images on A4 paper matte and glossy paper
Replay	Electronic photos and videos recorded on an iPhone 6 and iPad pro 12.9
Fake Head	Several pre-heated manequinn heads
Rigid Mask	Custom made realistic rigid masks and decorative plastic masks
Flexible Mask	Custom made realistic soft silicone masks
Paper Mask	Custom make paper masks based on real identities
Glasses	Different models of disguise glasses with fake eyes and paper glasses

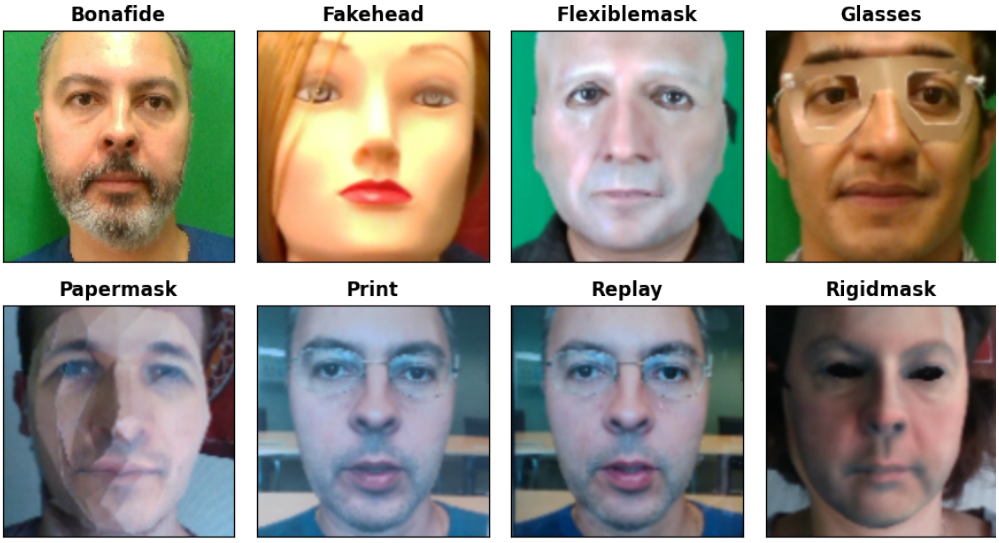


Figure 4.2: Examples of bonafide and presentation attacks.

The total number of 1679 video recordings include bonafide and presentation attacks and are divided into multiple protocols and grouped into three subsets: train, eval and test. Since the video recordings are correlated and consequently are uniformly sampled in the temporal domain, only 50 frames from each video were selected, totalling a sum of 83,950 biometric samples across all subsets.

Table 4.2: WMCA attack distribuition.

Attack type	**Category**	**Videos**
Bonafide	-	347
Print	2D	200
Replay	2D	348
Fake Head	3D	122
Rigid Mask	3D	137
Flexible Mask	3D	379
Paper Mask	3D	71
Glasses	3D	75
Total	-	1679

4.1.1.1 Evaluation Protocols

The split into multiple protocols aims to evaluate the performance of the framework in seen and unseen attack scenarios. The *grandtest* protocol is mainly used for intra-domain performance evaluation and is intended to test the performance of the network in the cases where the attack categories are known a priori. It consists of all attacks distributed in equal proportions across all subsets and the data split is done ensuring almost equal distribution of PA categories and disjoint set of client identifiers in each subset.

The remaining protocols are designed to simulate the Zero-shot/Leave-one-out (LOO) scenarios as they pretend to emulate real-world situations where an unseen attack type is encountered. The main objective is to determine which types of attacks are most easily detectable in the absence of any training data for those specific attacks. This helps verify whether the model is capable of transferring knowledge from previously encountered attack categories to detect the unseen attack. There are a total of seven LOO protocols, each involving the exclusion of a specific attack type during the training and validation phases. During testing, only samples from that particular attack type and bonafide data are used. For instance, in the LOO-2D protocol, all 2D attacks are removed from the training and validation sets, while the test set consists of bonafide samples and 2D attacks exclusively. The same approach is applied to all other protocols. The data distribution for both the *grandtest* and LOO-2D protocols is summarized in Tables 4.3 and 4.4, respectively, providing an overview of how the data is split across the different attack types and genuine samples.

Table 4.3: Data distribution of the *grandtest* protocol.

Attack type	**Train**	**Val**	**Test**
Bonafide	124	108	115
Print	68	66	66
Replay	152	116	80
Fake Head	53	52	17
Rigid Mask	26	62	49
Flexible Mask	110	131	137
Paper Mask	-	-	71
Glasses	31	22	22
Total	564	557	557

Table 4.4: Data distribution of the LOO-2D protocol.

Attack type	**Train**	**Val**	**Test**
Bonafide	124	108	115
Print	-	-	66
Replay	-	-	80
Fake Head	53	52	-
Rigid Mask	26	62	-
Flexible Mask	110	131	-
Total	333	353	262

4.1.2 CASIA-SURF

CASIA-SURF is another large-scale multi-modal PAD dataset that primarily focuses on 2D attacks and includes bonafide samples collected from over 1,000 chinese subjects with a wide range of ages. The data acquisition process involves the use of an Intel RealSense SR300 camera, which simultaneously captures RGB, Depth, and Infrared videos in diverse indoor environments. The resolution varies across the different modalities, with RGB videos recorded at 1280x720 resolution and Depth and Infrared videos recorded at 640x480 resolution. During the video recording sessions, participants were instructed to hold a printed attack image in front of their faces and perform specific actions such as walking towards and away from the camera, turning their heads, and bending the paper. This process aimed to create multiple variations of attack scenarios. For each bonafide sample, there are 6 presentation attacks, each one in the form of a printed flat or curved face image with cut eyes, nose, mouth areas or their combination. Table 4.5 describes CASIA-SURF attacks and figure 4.3 shows some examples of attacks and bonafide samples across all modalities.

Table 4.5: CASIA-SURF attack categories.

Attack type	Attack Description	Regions Cut out
Attack 1	flat printed face	eyes
Attack 2	curved printed face	eyes
Attack 3	flat printed face	eyes and nose
Attack 4	curved printed face photo	eyes and nose
Attack 5	flat printed face photo	eyes, nose and mouth
Attack 6	curved printed face	eyes, nose and mouth

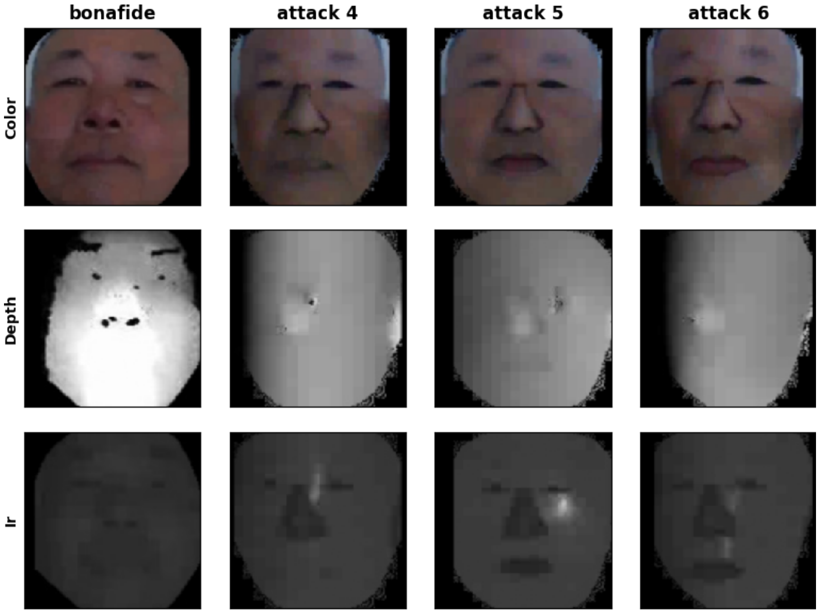


Figure 4.3: Multi-modal bonafide and PAs from the CASIA-SURF dataset. The attacks were taken from the training set.

4.1.2.1 Evaluation Protocol

The owned version of the CASIA-SURF dataset, provided by the laboratory, consists of a total of 29,266 images for training, 9,608 images for validation, and 57,710 images for testing. Following the legacy evaluation stategy introduced by the authors, we have chosen to use live faces and attacks 4, 5, and 6 to train the baselines, while the remaining attacks and live faces are reserved for validation and testing. This selection ensures an equal distribution of flat and curved faces across all subsets. Table 4.6 presents the distribution of samples and the types of attacks included in each subset.

Table 4.6: CASIA-SURF types of attacks and number of samples per subset.

Subset	**Attack**						**Nº Samples**
	1	**2**	**3**	**4**	**5**	**6**	
Train				✓	✓	✓	29,266
Val	✓	✓	✓				9,608
Test	✓	✓	✓				57,710

4.2 Implementation Details

The implementation of this thesis relied on the open-source PyTorch deep learning library, and all the code was written in Python. For the ViT and Hybrid baselines, we adapted the ViT-B base network described in [11] for the multi-modal task. The ViT-based baselines were trained with a batch size of 32, an AdamW optimizer, and a cosine annealing scheduler for 30 epochs.

As for the ResNet baseline, we utilized the pre-existing backbone available in the PyTorch library. This baseline was trained for 50 epochs with an Adam optimizer, a cosine annealing scheduler, and a batch size of 64.

The difference in batch size and epochs between the ViT-based baselines and the ResNet baseline is due to model size and memory constraints, as we were unable to train the ViT-based baselines with the same configuration as the ResNet baseline. Furthermore, the initial weights of all baselines were pre-trained on the ImageNet-1k dataset, and the training was conducted using a Cross-Entropy (CE) loss function. The choice of learning rates will be covered in the upcoming section. Table 4.7 summarizes the training configuration of all baselines.

Table 4.7: Training configuration of the baselines.

Baseline	**pre-trained**	**epochs**	**batch**	**loss**	**optimizer**	**scheduler**
Multi-modal ViT	✓	30	32	CE	AdamW	cosine
Multi-modal Hybrid	✓	30	32	CE	AdamW	cosine
Multi-modal Resnet	✓	50	64	CE	Adam	cosine

To enhance the baseline's performance, multiple data augmentation techniques were applied to all images during training, including random rotation, random cutout, random crop, random grayscale, horizontal flip, and color jitter. These augmentation techniques were applied consistently across all baselines and the images from all modalities were normalized with $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$. However, in the ViT and Hybrid baselines, the images were resized to 224x224, while in the ResNet baseline, a lower resolution of 112x112 was used for resizing the images.

For result reproduction, we used the same seed for every training and testing run and the best model was selected based on the minimum loss observed in the validation set. Finnaly, all baselines were trained using an Nvidia RTX 3090.

4.3 Evaluation Metrics

For the performance evaluation, real/bonafide images are assigned with label 1 and attack images are labeled as 0. Based on these labels, the evaluation metrics are defined as follows:

- TP: Bonafide samples that are correctly predicted as bonafide.
- FP: Spoof samples that are wrongly classified as bonafide.
- TN: Spoof samples that are correctly predicted as spoof.
- FN: Bonafide samples that are wrongly classified as spoof.

These metrics are then used to compute another two sets of metrics that are commonly used among other works in the literature. The first is the standardized ISO/IEC 30107-3 metrics which consists of an Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER) and an Average Classification Error Rate (ACER). These metrics are formulated in the following equations:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$APCER = \frac{FP}{FP + TN} \quad BPCER = \frac{FN}{FN + TN} \quad ACER = \frac{APCER + BPCER}{2} \tag{4.2}$$

where Accuracy (ACC) is the percentage of bonafide and spoof samples correctly identified, APCER is the percentage of spoof samples that are incorrecly classified as bonafide, BPCER is the percentage of bonafide samples that are incorrectly classified as spoof and ACER is the mean of APCER and BPCER.

The second set of metrics is the Receiver Operating Characteristic (ROC) curve which is used in several works in the literature and is used to select a suitable threshold to trade off the False Positive Rate (FPR) and True Positive Rate (TPR) according to the requirement of real applications.

$$FPR = \frac{FP}{FP + TN} \qquad\qquad TPR = \frac{TP}{TP + FN} \tag{4.3}$$

4.4 Vision Transformer Interpretability

In section 3.1.3 we discussed the multi-modal ViT, which comprises a series of encoders. Within each encoder block, there are 12 parallel heads responsible for projecting the embedded patch tokens to multiple sub-spaces using self-attention. As a result, we obtain 12 different projections, where each projection represents a distinct representation of the input image.

In practical terms, each head attends to a specific part of the input image by evaluating the significance of the split version of the query in relation to the split version of the key. These matrices have a shape of 3 × 197 × 64, and each head attends to 197 tokens, where each token has a feature length of 64. In other words, each patch token is projected 768 times via a convolutional layer, resulting in each head attending to 68 projections of each patch. Out of the 197 tokens, the first token represents the CLS token that flows through the Transformer which is appended to the sequence of embeded patches, and the remaining 196 tokens represent the linearly projected patches. Consequently, each head computes attention weights by taking the dot product between the given query and key (equation 3.5). The resulting attention matrix, denoted as $A \in \mathbb{R}^{197\times197}$, provides insights into the importance of different image parts to the Transformer. Figure 4.4 illustrates the computed attention matrices for all attention heads in the first encoder ($L = 1$).

Visualizing the attention weights overlaid on the image offers an intuitive understanding of the regions considered significant by the Transformer. Therefore, it is beneficial to visualize each attention head map separately to comprehend their focus areas. The heatmaps are generated by calculating the mean of each attention matrix along the second dimension and then resize it to a 14 × 14 matrix. Figure 4.5 demonstrates how the attention maps appear throughout the attention heads in the first encoder ($L = 1$).

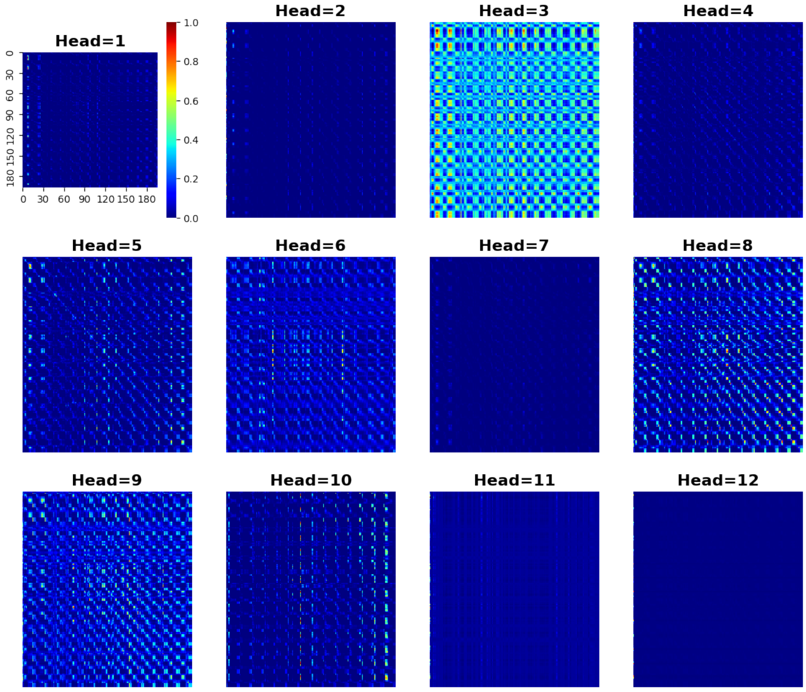


Figure 4.4: Graphical representation of the attention matrix in each paralell head of the first encoder block.

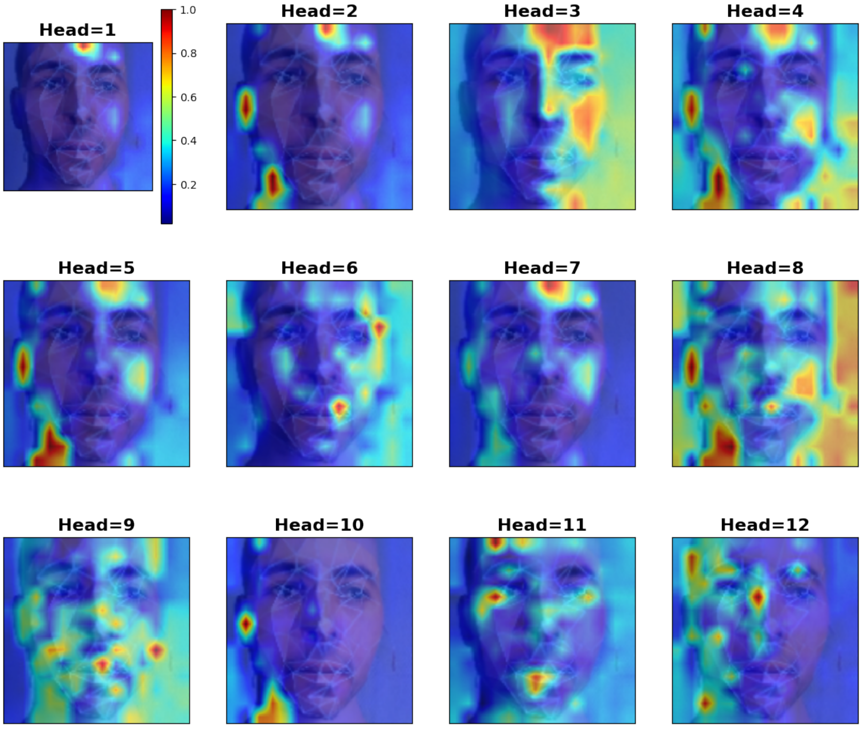


Figure 4.5: Attention maps of all paralell heads in the first encoder block. The red zone refer to the most discriminative location of the image.

Notably, different heads concentrate on different image regions. For example, heads 1, 7 and 10 prioritize local features, while heads 3, 8 and 9 capture information that covers a larger portion of the face, including the background. The outputs of all heads are then concatenated to produce an output whose shape is the same as the input to the encoder. This output is subsequently fed into the next encoder and so on.

To better understand how information flows through all the encoders, the attention weights of all heads are averaged, representing the learned representations at each encoder. These representations are shown in figure 4.6. When analyzing the attention maps at each encoder, we find that the Transformer attends to most of the image at the lowest layers and progressively refines its focus until the last layer, resulting in regions that it considers semantically relevant for classification.

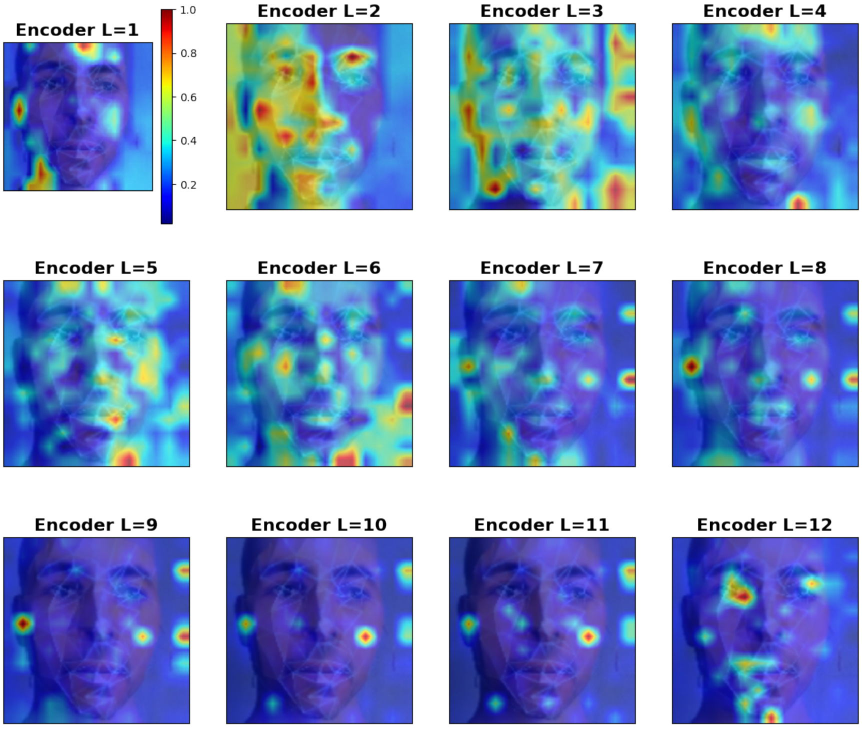


Figure 4.6: Attention maps of all encoders.

4.5 Validation Results

4.5.1 Intra-domain Results

The first stage of evaluation focuses on conducting intra-domain evaluation using the CASIA-SURF dataset and the *grandtest* protocol from the WMCA dataset. This evaluation involves utilizing all the baseline models described in chapter 3 and aims at simulating circumstances where the network has prior knowledge of the attacks it will encounter during the testing phase.

Considering the crucial influence of hyperparameter tuning on deep learning models, an ablation study was performed to determine the optimal learning rate for each baseline. The goal was to identify the learning rate that would yield the best results for the specific dataset. The selection of the best model was based on the accuracy achieved on the test set. The results of this ablation study can be found in appendix A.1, and they are summarized in table 4.8. Until stated otherwise, we will refer to these results as the baseline results. In order to enhance readability, the baselines with the highest Accuracy (ACC) and lowest Average Classification Error Rate (ACER) for the dataset under study are highlighted in cyan. This highlighting allows easier identification of the top-performing baseline in terms of these evaluation metrics.

Table 4.8: Intra-domain evaluation on the CASIA-SURF and the *grandtest* protocol from the WMCA dataset across all baseline models.

Dataset	**Baseline**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	**10^{-3}**	**10^{-4}**
WMCA-*grandtest*	**ViT**	99.73	0.2	1.25	0.64	99.57	99.04	98.54
	ResNet	98.74	1.56	0.10	0.83	98.12	72.87	64.89
	Hybrid	99.34	0.55	1.10	0.82	99.53	84.78	66.54
CASIA-SURF	**ViT**	98.67	1.01	2.07	1.54	97.89	88.18	70.56
	ResNet	99.38	0.43	1.05	0.74	99.59	96.17	89.12
	Hybrid	98.62	0.75	2.85	1.8	97.83	83.29	63.07

Overall, the baselines achieved highly promising results on both datasets. On one hand, the results presented in table 4.8 demonstrate that the ViT-based baselines (ViT and Hybrid) performed better in detecting the various PAs present in the WMCA dataset (99.73% and 99.34% accuracy, 0.64% and 0.82% ACER respectively). On the other hand, the ResNet baseline outperformed the ViT-based baselines on the CASIA-SURF dataset (99.38% accuracy and 0.74% ACER), which primarily consists of 2D attacks such as print and replay attacks. These results indicate that the ViT-based models excel at handling the diverse range of attacks present in the WMCA dataset, while the ResNet baseline is slightly more effective in detecting 2D attacks in a dataset mostly composed of planar attacks.

To further understand the features contributing to the decisions, the activation maps for the ViT and ResNet baselines are shown in figures 4.7 and 4.8.

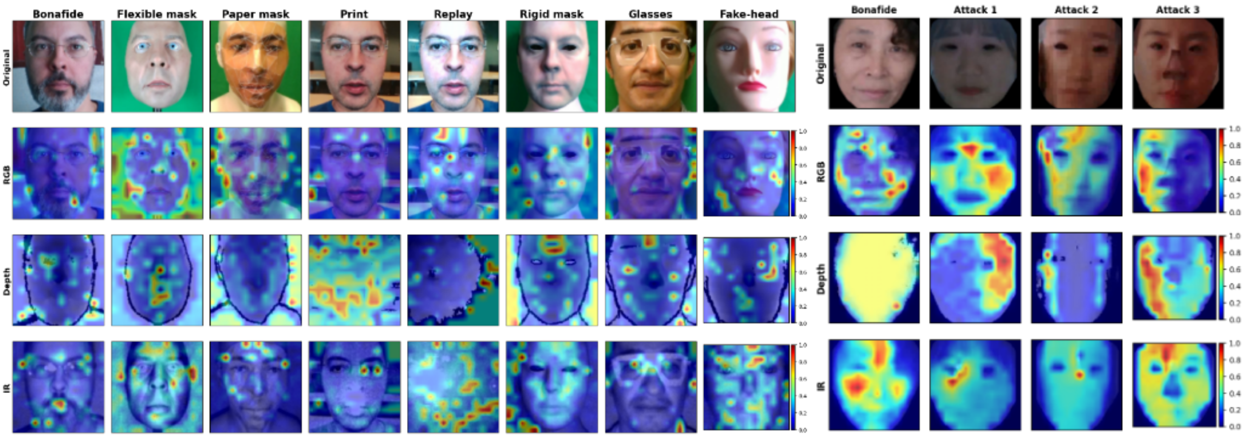


Figure 4.7: Attention maps of the multi-modal ViT baseline for all classes of both datasets (WMCA on the left and CASIA-SURF on the right). The attention maps were built using the attention weights from the last encoder.

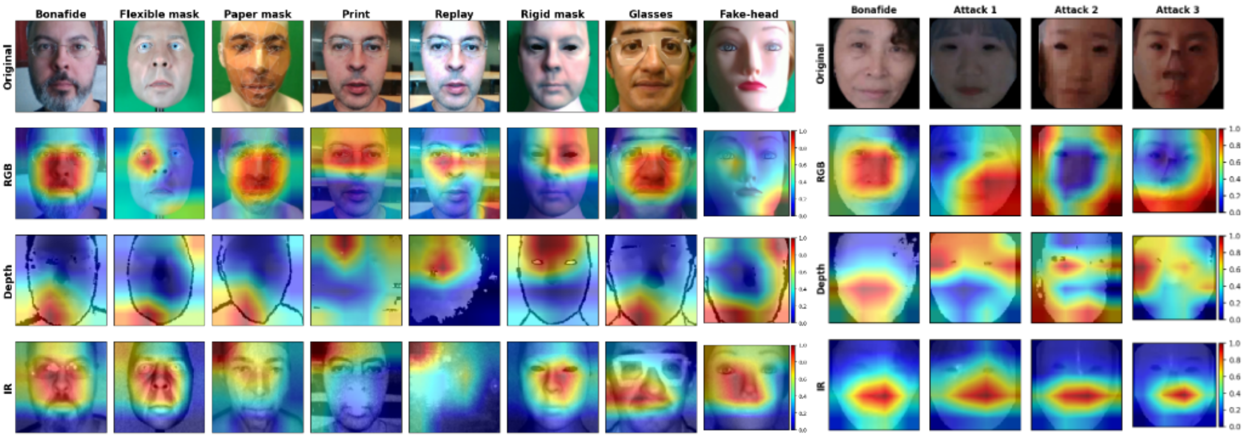


Figure 4.8: Activation maps of the multi-modal ResNet baseline for all classes of both datasets (WMCA on the left and CASIA-SURF on the right). The activations were selected from each individual branch, after the 4^{th} residual block (res_4).

Interestingly, a noticeable distinction exists between the two architectures in terms of their perception of the semantically relevant regions. The attended regions of the ViT baseline are spatially distributed across the image, smaller in scale and more detailed when compared to the heatmaps generated by the ResNet baseline, which suggests that ViTs assign equal relevance to different parts of the image. In contrast, the ResNet baseline demonstrates limitations in weighting different image regions due to its pooling layers and convolutional process. The features extracted from the higher layers of the baseline possess a larger receptive field, enabling them to represent semantic information effectively. However, in terms of resolution, these features exhibit lower detail compared to the activations observed in ViTs.

4.5.2 Cross-domain Results

Following the intra-domain evaluation, the cross-domain evaluation involves training, validating, and testing the models using both the CASIA-SURF and the *grandtest* protocol of the WMCA dataset. The results presented in table 4.9 correspond to the best baseline model selected from a range of learning rates, which are detailed in appendix A.2. This type of evaluation aims to simulate a model that has been trained under different domains, allowing us to assess the generalization capability of the baselines.

Table 4.9: Cross-domain evaluation using the CASIA-SURF and the *grandtest* protocol from the WMCA dataset.

Dataset	**Baseline**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	**10^{-3}**	**10^{-4}**
Cross-Domain	**ViT**	97.42	3.40	0.37	1.89	96.67	81.86	72.45
	ResNet	95.67	4.97	2.62	3.80	89.28	56.18	37.88
	Hybrid	94.49	4.95	7.02	5.99	79.26	51.35	35.42

Although the best performing model is the multi-modal ViT (99.73% accuracy and 1.89% ACER), the performance of all baselines in cross-domain scenario decreased compared to the intra-domain level. This decline in performance can be attributed to the bias introduced by the similarity of images within the same dataset and their differences to the other dataset. When models are evaluated in a cross-domain scenario, they may struggle in correctly classifying bonafide and PAs from the other dataset because the model has learned representations that are specific to one dataset and may not effectively generalize to the feature space of the other dataset.

4.5.3 Zero-shot Learning Results

The last evaluation stage is the zero-shot evaluation, which differs from the previous evaluations where all attack categories are known a priori. In zero-shot learning, models are evaluated in unseen presentation attacks as it indicates their effectiveness in encontering real-world attacks that were not seen during training time. To conduct this evaluation, all baselines were trained and tested on all leave-one-out protocols of the WMCA dataset. For a detailed explanation of these protocols, please revisit section 4.1.1.1. All baselines were tested using the learning rates that were used in the evaluation of the *grandtest* protocol. The results are tabulated in table 4.10.

Initially, the baselines demonstrated mediocre performance in the LOO-2D and LOO-3D protocols, except for the ResNet baseline, which performed significantly better in the LOO-2D protocol (98.43 % ACC and 1.41% ACER) than its counterparts. In practical terms, training a model to detect 2D attacks and then evaluating it on 3D attacks poses a significant challenge for the baselines, as they struggle to extract features that are relevant to the distinctive characteristics of 3D attacks. The same is true for the reverse scenario.

Furthermore, the LOO-Flexible-Mask and LOO-Glasses protocols also yielded poor results across all baselines, indicating a collapse in performance. Among the 3D attacks, the silicone masks utilized in these protocols are the most similar to real faces, which explains the difficulty in detecting these attacks even when leveraging multi-modal information. The infrared range does not capture significant changes in the face, as the masks were pre-heated before the attack acquisition. Moreover, the relief of the masks closely resembles the facial relief of a real person, resulting in very similar depth maps. Attacks involving fake glasses are also difficult to detect, as all baselines struggled to achieve good results (86.85%, 84.76% and 86.06% ACC for the ViT, ResNet and Hybrid respectively). These attacks barely alter the features of a real person, as the fake glasses are easily mistaken for real glasses. Once again, in this situation multi-modal information is of little use, which highlights the difficulty of detecting these 2 types of attacks.

When analysing the performance of three baselines, it is evident that the ViT baseline outperformed the other two baselines in several protocols, including print, replay, paper mask, rigid mask, fake-head, and glasses. This suggests that the attention mechanism employed by ViT is highly effective in detecting various types of attacks. On the contrary, the ResNet baseline demonstrated significantly better performance in the protocols (LOO-2D, LOO-3D and LOO-Flexible mask) where it outperformed the ViT and Hybrid baselines.

Table 4.10: Zero-shot evaluation on the leave-one-out protocols in WMCA dataset.

Protocol	**Baseline**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						$\mathbf{10^{-2}}$	$\mathbf{10^{-3}}$	$\mathbf{10^{-4}}$
2D	**ViT**	74.04	46.41	0.0	23.21	99.74	99.55	99.43
	ResNet	98.43	2.71	0.12	1.42	99.74	99.22	99.04
	Hybrid	89.15	18.84	0.71	9.77	95.67	92.61	87.37
3D	**ViT**	29.35	99.98	0.0	49.99	23.08	6.75	2.35
	ResNet	45.26	77.90	0.0	38.95	34.10	23.04	0.0
	Hybrid	29.96	99.67	0.0	49.83	26.43	7.48	4.43
Print	**ViT**	99.77	0.0	0.37	0.18	100	100	100
	ResNet	99.06	0.0	1.48	0.74	99.91	99.86	99.73
	Hybrid	99.52	0.0	0.75	0.37	100	100	100
Replay	**ViT**	99.81	0.25	0.16	0.20	99.84	99.81	99.76
	ResNet	98.89	1.32	0.96	1.14	98.90	97.27	96.47
	Hybrid	98.73	0.43	1.86	1.14	98.49	97.53	96.45
Paper mask	**ViT**	99.90	0.09	0.1	0.1	100	100	99.97
	ResNet	99.22	0.0	1.25	0.63	99.97	99.90	99.81
	Hybrid	99.32	0.0	1.10	0.55	99.97	99.70	99.51
Flexible mask	**ViT**	58.70	75.97	0.0	37.99	63.72	45.93	32.14
	ResNet	84.03	29.36	0.02	14.69	71.32	63.44	59.22
	Hybrid	65.02	64.25	0.10	32.18	69.51	50.17	26.10
Rigid mask	**ViT**	99.75	0.25	0.24	0.25	99.83	99.76	99.74
	ResNet	98.65	0.0	1.91	0.96	100	99.91	99.90
	Hybrid	98.96	0.38	1.32	0.85	98.82	98.54	97.90
Fake-head	**ViT**	98.98	7.53	0.05	3.79	99.83	99.83	99.83
	ResNet	97.91	0.47	2.90	1.69	97.53	96.12	96.12
	Hybrid	98.82	0.12	1.34	0.73	99.95	98.14	98.14
Glasses	**ViT**	86.85	81.55	0.07	40.81	54.99	50.52	50.05
	ResNet	84.76	93.64	0.24	46.94	20.47	17.55	15.43
	Hybrid	86.06	84.18	0.50	42.34	42.17	33.84	30.94

4.5.3.1 Multi-modal Data vs RGB Spectrum

In order to explore the impact of incorporating Depth and Infrared modalities on the detection capability of attacks, we conducted an experiment using the multi-modal ViT baseline with three modalities (RGB, Depth and Infrared) and compared it to the same baseline using only the RGB spectrum. This evaluation aims to assess the influence of multi-modal data on the performance, rather than focusing on the best performing model.

Table 4.11: Zero-shot evaluation on the leave-one-out protocols in WMCA dataset using only the RGB spectrum on the ViT baseline.

Protocol	**Modalities**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-4}
2D	**RGB+Depth+IR**	74.04	46.41	0.0	23.21	99.74	99.55	99.43
	RGB	43.79	99.53	1.22	50.38	77.81	73.93	71.20
3D	**RGB+Depth+IR**	29.35	99.98	0.0	49.99	23.08	6.75	2.35
	RGB	29.96	99.67	0.0	49.83	11.69	4.12	1.50
Print	**RGB+Depth+IR**	99.77	0.0	0.37	0.18	100	100	100
	RGB	95.26	0.03	7.44	3.74	96.92	94.61	92.10
Replay	**RGB+Depth+IR**	99.81	0.25	0.16	0.20	99.84	99.81	99.76
	RGB	65.28	78.43	4.31	41.37	66.14	62.31	59.97
Paper mask	**RGB+Depth+IR**	99.90	0.09	0.1	0.1	100	100	99.97
	RGB	92.61	4.67	9.04	6.86	79.48	76.26	72.19
Flexible mask	**RGB+Depth+IR**	58.70	75.97	0.0	37.99	63.72	45.93	32.14
	RGB	58.67	75.33	0.83	38.08	51.34	20.89	12.40
Rigid mask	**RGB+Depth+IR**	99.75	0.25	0.24	0.25	99.83	99.76	99.74
	RGB	91.48	17.51	4.77	11.14	61.55	48.61	43.95
Fake-head	**RGB+Depth+IR**	98.98	7.53	0.05	3.79	99.83	99.83	99.83
	RGB	94.97	8.47	4.52	6.50	95.10	88.73	88.73
Glasses	**RGB+Depth+IR**	86.85	81.55	0.07	40.81	54.99	50.52	50.05
	RGB	83.96	71.82	5.37	38.60	14.49	8.35	8.17

Based on the information presented in table 4.11, it can be concluded that the inclusion of multi-modal information is in fact important for the model to achieve better performance in detecting attacks due to the fact that the results worsened in all protocols when using the RGB spectrum alone.

In the case of 2D attacks, it is noticeable that the absence of depth information relative to the face contributes to the degradation of performance. Print and replay attacks, which are relatively similar to a real face, become more difficult to detect in the absence of complementary modal information. As expected, the performance in detecting 3D attacks also deteriorates in all protocols when using only the RGB spectrum, which highlights the importance of including depth and infrared cues that assist in distinguishing between genuine and fake samples in these protocols.

Finally, by analyzing the LOO-Flexible-mask and LOO-Glasses protocols, it is observed that the results when using the three modalities or only RGB are practically the same, which validates the hypothetical conclusion made in the previous section, i.e., the Depth and Infrared spectra have little or no contribution to the detection of these specific attacks.

4.5.4 Fusion Methods Results

The first step to enhance the results of all baselines in both intra-domain and cross-domain scenarios is to incorporate the multi-modal fusion methods described in section 3.4. To accomplish this, the original fusion utilized in each baseline model was substituted with three distinct fusion methods: Concatenation (C), Squeeze-and-excitation (SE), and Cross-Attention (CA). In the ResNet baseline, only the SE and CA were tested. To analyze the individual influence of each fusion method, several tests were carried out and the results are presented in Tables 4.12, 4.13 and 4.14.

Despite the increase in the number of parameters and training time across all baselines due to the introduction of fusion methods, most tests demonstrated improvements in results in comparison to the baseline's.

Among all the fusion methods, SE consistently showed the highest performance improvement, benefiting all baselines in the cross-domain scenario. Particularly, the ViT and ResNet baselines exhibited significant improvements, even in the intra-domain results. For the attention-based baselines (ViT and Hybrid), these improves results validade the use of patch tokens as global feature, specially when they are further recalibrated via squeeze and excitation operations, rather than relying solely on the CLS token to perform classification.

On the contrary, the CA method, while improving the performance of the ViT baseline in the *grandtest* protocol, consistently showed poorer results in the cross-domain scenario in all baselines. This suggests that feature fusion in a homogeneous space may not provide advantages in the context of anti-spoofing.

Table 4.12: Fusion Methods on the Multi-modal ViT Baseline.

Dataset	Fusion			ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
	C	SE	CA					10^{-2}	10^{-3}	10^{-4}
WMCA-*grandtest*	ViT Baseline			99.73	0.2	1.25	0.64	99.57	99.04	98.54
	✓			99.10	0.02	4.28	2.15	99.84	97.39	95.50
		✓		98.08	0.03	9.17	4.60	98.63	93.58	89.23
			✓	99.85	0.04	0.59	0.31	99.86	99.53	99.20
CASIA-SURF	ViT Baseline			98.67	1.01	2.07	1.54	97.89	88.18	70.56
	✓			97.60	0.78	6.13	3.46	95.13	74.62	57.31
		✓		98.92	0.70	1.94	1.32	98.44	93.34	85.46
			✓	98.04	2.59	0.52	1.55	98.97	94.31	68.79
Cross-Domain	ViT Baseline			97.42	3.40	0.37	1.89	96.67	81.86	72.45
	✓			97.37	3.50	0.29	1.90	97.53	70.53	45.15
		✓		98.59	1.79	0.37	1.08	98.49	58.88	33.60
			✓	96.94	3.11	2.92	3.01	92.74	77.70	65.12

Table 4.13: Fusion Methods on the Multi-modal ResNet Baseline.

Dataset	Fusion		ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
	SE	CA					10^{-2}	10^{-3}	10^{-4}
WMCA-*grandtest*	ResNet Baseline		98.74	1.56	0.10	0.83	98.12	72.87	64.89
	✓		99.04	1.12	0.35	0.73	98.85	86.14	74.38
		✓	98.59	1.72	0.23	0.97	97.18	92.97	89.37
CASIA-SURF	ResNet Baseline		99.38	0.43	1.05	0.74	99.59	96.17	89.12
	✓		96.50	3.47	3.55	3.51	87.16	68.23	57.54
		✓	95.62	4.11	4.99	4.55	74.17	21.02	16.22
Cross-Domain	ResNet Baseline		95.67	4.97	2.62	3.80	89.28	56.18	37.88
	✓		96.30	4.98	0.26	2.62	96.66	13.48	6.87
		✓	92.75	7.59	6.35	6.97	56.87	0.0	0.0

Table 4.14: Fusion Methods on the Multi-modal Hybrid Baseline.

Dataset	Fusion			ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
	C	SE	CA					10^{-2}	10^{-3}	10^{-4}
WMCA-*grandtest*	Hybrid Baseline			99.34	0.55	1.10	0.82	99.53	84.78	66.54
	✓			98.91	1.06	1.20	1.13	98.54	40.02	27.84
		✓		98.88	1.22	0.75	0.98	97.53	41.15	28.70
			✓	96.10	4.92	0.02	2.47	95.48	61.48	43.10
CASIA-SURF	Hybrid Baseline			98.62	0.75	2.85	5.89	97.83	82.39	63.07
	✓			96.14	3.28	5.18	4.23	89.27	62.77	27.79
		✓		95.52	3.14	7.56	5.35	82.31	58.82	41.94
			✓	95.03	3.19	9.08	6.14	80.27	53.36	22.22
Cross-Domain	Hybrid Baseline			94.49	4.95	7.02	5.99	79.26	51.35	35.42
	✓			79.32	27.85	1.45	14.65	71.99	48.47	29.19
		✓		96.47	4.20	1.74	2.97	92.88	78.18	65.93
			✓	89.45	12.52	5.27	8.90	41.34	17.64	10.69

4.5.5 Mixstyle Results

To further improve the results of the baselines in cross-domain, we investigated mixing features statistics between the two datasets. The outcomes of these experiments are summarized in Tables 4.15 and 4.16. During training, we applied two mixing strategies, namely random and crossbatch, to all batches to ensure maximum feature mixing using a Beta distribution $Beta(\alpha, \alpha)$ with α=0.1.

In the multi-modal ViT baseline, we placed Mixstyle before the patch projection and after the patch projection, denoted by BP and AP respectively in Table 4.15. While we did not applied Mixstyle within the encoders, the model performs better when Mixstyle is apllied after the patch projection rather than before, despite the degradation of results. Another conclusion is that the random shuffling of features collapsed the performance of the model, indicating that this mixing stategy is not the most suitable for this baseline, as features from one dataset may be stylized with features from the same dataset. In contrast, crossbatch mixing ensures that features from one dataset are only stylized with statistics from one image of the other dataset. When this mixing strategy was employed after the projection of the patches, the model surpassed the baseline version in terms of accuracy, achieving 97.65%

Table 4.15: Domain Generalization of the Multi-modal ViT Baseline using Mixstyle.

Dataset	Mixing	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-4}
Cross-Domain	ViT Baseline	97.42	3.40	0.37	1.89	96.67	81.86	72.45
	Random BP	92.10	10.73	0.33	5.53	90.01	51.09	29.26
	Random AP	93.81	8.08	1.10	4.59	88.17	49.11	27.93
	Random BP+AP	95.83	5.39	0.89	3.14	92.48	66.75	41.35
	Crossbatch BP	91.93	10.92	0.43	5.68	87.64	60.65	24.85
	Crossbatch AP	97.65	2.79	1.16	1.98	95.09	60.66	41.39
	Crossbatch BP+AP	96.07	5.32	0.18	2.75	97.75	57.0	19.35

ACC.

Table 4.16: Domain Generalization of the Multi-modal ResNet Baseline using Mixstyle.

Dataset	Mixing	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-4}
Cross-Domain	ResNet Baseline	95.67	4.97	7.02	5.99	79.26	51.35	35.42
	Random res1	95.30	5.68	2.06	3.87	90.44	75.83	0.0
	Random res12	96.81	3.40	2.65	3.02	86.56	53.07	0.0
	Random res123	96.51	3.06	4.63	3.85	81.15	14.64	4.46
	Crossbatch res1	94.82	6.34	2.09	4.21	84.86	48.49	9.77
	Crossbatch res12	94.90	5.08	5.15	5.12	78.26	47.98	39.20
	Crossbatch res123	94.03	7.69	1.35	4.52	85.61	50.33	44.67

Given that each branch of the standard multi-modal ResNet baseline has four residual blocks denoted by res_{1-4}, we trained different models with MixStyle applied to the first three layers, i.e., according to the original paper [45], placing Mixstyle after the 4^{th} res block breaks the inherent label space. To clarify the notation, res_1 signifies that MixStyle is applied after the first residual block, while res_{12} indicates MixStyle is applied after both the first and second residual blocks, and so on. When analysing table 4.16, the results surpassed the baseline in all experiments when using the random shuffling, which indicates a very different behaviour when compared to the ViT baseline, which outperformed its baseline when using the crossbatch shuffling.

4.5.6 Concentration Loss Results

Another approach to improve cross-domain results involves the utilization of a domain-invariant concentration loss L_{DiC} that concentrates domain-invariant bonafide representations into a non-spoof class of data. One significant advantage of this method is that it does not introduce additional computational complexity to any baseline model.

In contrast to the baseline models, which are only trained with a Cross-Entropy loss L_{CE}, here we add the L_{DiC} loss to the L_{CE} loss during the training of all baselines. Each baseline is trained with different balance factors, denoted as $\lambda = \{0.2, 0.5, 0.7, 0.9\}$ and the L_{DiC} loss is calculated by computing the L1 norm of all feature vectors that represent bonafide images within the same batch of images, as shown in equation 3.14. Tables 4.17, 4.18 and 4.19 show the results.

Table 4.17: Domain Generalization of the Multi-modal ViT Baseline using L_{DiC} loss.

Dataset	**Balance Factor**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	**10^{-3}**	**10^{-4}**
Cross-Domain	**ViT Baseline**	97.42	3.40	0.37	1.89	96.67	81.86	72.45
	λ=0.2	96.36	4.90	0.16	2.53	98.02	15.55	3.24
	λ=0.5	97.10	3.77	0.57	2.17	97.33	79.69	46.47
	λ=0.7	98.44	1.95	0.49	1.22	97.56	60.30	0.05
	λ=0.9	97.82	2.74	0.69	1.71	96.22	33.33	1.63

Table 4.18: Domain Generalization of the Multi-modal ResNet Baseline using L_{DiC} loss.

Dataset	**Balance Factor**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	**10^{-3}**	**10^{-4}**
Cross-Domain	**ResNet Baseline**	95.67	4.97	2.62	3.80	89.28	56.18	37.88
	λ=0.2	94.88	4.60	6.42	5.56	72.78	0.0	0.0
	λ=0.5	95.21	5.56	2.72	4.15	86.80	0.0	0.0
	λ=0.7	93.29	6.97	6.02	6.49	0.0	0.0	0.0
	λ=0.9	95.82	5.24	1.34	3.29	93.18	61.9	32.42

Table 4.19: Domain Generalization of the Multi-modal Hybrid Baseline using L_{DiC} loss.

Dataset	Balance Factor	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-4}
Cross-Domain	**Hybrid Baseline**	94.49	4.95	7.02	5.99	79.26	51.35	35.42
	$\lambda = 0.2$	96.12	4.82	1.34	3.08	92.06	12.85	6.89
	$\lambda = 0.5$	95.76	5.49	0.88	3.19	92.22	49.57	34.97
	$\lambda = 0.7$	93.09	8.91	1.56	5.23	85.37	55.41	28.09
	$\boldsymbol{\lambda = 0.9}$	97.45	1.90	4.28	3.09	92.85	68.72	28.74

From the visualization of the last three tables, all baseline models benefited from concentrating the features of real faces into a single category and pushing them close to the origin, as there is at least one balance factor in each model that boosted the cross-domain results when comparared to its respective baseline.

4.5.7 Unified ViT Pipeline Results

The final stage of evaluation involves integrating all the above-mentioned approaches into a unified pipeline, with the ViT baseline serving as the core. The objective is to determine whether it is feasible to design a network that combines multiple enhancement approaches, each individually boosting the baseline performance, and consequently improving the detection of attacks in cross-domain scenarios, which are the most challenging.

Based on the top-performing results obtained in previous sections, the final pipeline includes the multi-modal ViT baseline backbone combined with a Squeeze-and-Excitation fusion module, a crossbatch AP shuffling strategy for feature styling, and a concentration loss L_{Dic}. The fusion and mixstyle strategies remain fixed, and experiments were conducted using different balance factors λ, previously explicited in section 4.5.6.

Table 4.20: Unified ViT pipeline results on cross-domain scenario.

Dataset	Assembly	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-4}
Cross-Domain	**ViT Baseline**	97.42	3.40	0.37	1.89	96.67	81.86	72.45
	SE + **Crossbatch AP** + $\lambda = 0.2$	95.46	6.21	0.06	3.14	98.00	0.0	0.0
	SE + **Crossbatch AP** + $\lambda = 0.5$	95.58	5.55	0.12	3.48	93.52	0.12	0.12
	SE + **Crossbatch AP** + $\lambda = 0.7$	95.68	5.38	1.45	3.42	94.64	0.0	0.0
	SE + Crossbatch AP + $\boldsymbol{\lambda = 0.9}$	97.96	2.33	1.28	1.80	96.25	0.39	0.04

As table 4.20 shows, the initial three results demonstrate that stacking successful approaches does not necessarily lead to performance enhancements compared to the baseline. In these cases, the model overfitted and produced poor results. However, there was one scenario where the model slightly outperformed the baseline (97.96% Accuracy and 1.80% ACER), which was when a balance factor λ=0.9 was considered. This suggests that even though incorporating a more sophisticated fusion method along with mixstyle and a concentration loss can be effective, it is crucial to note that the model's performance is highly sensitive to parameterization.

The data of table 4.20 and all prior experiments in this thesis utilize transfer learning. All baselines are initialized with pretrained weights and subsequently fine-tuned for a downstream task, where during backpropagation, all model parameters are updated. While Mixstyle and L_{DiC} do not introduce any learnable parameters, the same can't be said about the fusion module SE as it lacks pretrained weights. As a result, one can theorize that the introduction of untrained parameters in the pipeline may distort the pretrained features during training, potentially causing the model to converge to a local minimum. This is because the model consists of a pretrained backbone that contains high-level feature information, combined with a module containing learnable parameters that start training from scratch.

To address this issue, we took insights from the work of Ananya *et al.* [46] and tried to improve the downstream performance of the model by freezing the pretrained backbone parameters for the first 10 epochs. During this period, only the SE module and classification head were trained. This technique aimed to establish a more balanced training process and mitigate any potential distortion that could arise from the introduction of untrained parameters in the early stages of training. The results are presented in table 4.21.

Table 4.21: Unified ViT pipeline results on cross-domain scenario with frozen backbone during the first 10 epochs.

Dataset	**Assembly**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	**10^{-3}**	**10^{-4}**
Cross-Domain	**ViT Baseline**	97.42	3.40	0.37	1.89	96.67	81.86	72.45
	SE + Crossbatch AP + λ= 0.2	99.18	0.56	1.53	1.05	99.24	87.0	22.09
	SE + **Crossbatch AP** + λ= 0.5	98.96	1.03	1.09	1.06	98.86	90.34	71.15
	SE + **Crossbatch AP** + λ= 0.7	97.38	3.42	0.47	1.94	96.78	0.0	0.0
	SE + **Crossbatch AP** + λ= 0.9	98.73	1.49	0.66	1.08	98.78	91	75.66

The results from the last table demonstrate that partially freezing the pretrained part of the network had a significant positive impact on performance, achieving an accuracy of 99.18% and an ACER of 1.05% when considering a balance factor λ of 0.2. Notably, the performance for other balance factors also consistently increased. This indicates that by selectively freezing the pretrained parameters and allowing focused training on specific modules until a certain point, the model's overall performance improved considerably.

5 Conclusion and Future Work

5.1 Conclusion

In conclusion, the primary focus of this dissertation was on developing algorithms for spoof detection, particularly based on Vision Transformers, and addressing the challenges associated with spoofing in authentication systems. Since Transformers have already demonstrated state-of-the-art results in other areas of computer vision, how would they perform as base framework for FAS tasks? To answer this question, three distinct baseline models were developed: one based on a standard Vision Transformer, another based on a CNN, and a Hybrid version that combined characteristics from both. Even though all networks were evaluated using the same protocols, greater emphasis was placed on the ViT-based architecture.

Considering the favorable results in section 4.5 attained by the ViT baseline, one may conclude that this architecture can serve as an alternative to CNNs in the context of FAS. Generally speaking, the self-attention mechanism has proven capable of extracting spoof-specific discriminative features, despite the substantial increase in the number of parameters and training time. In contrast, the performance of the Hybrid baseline, although improved upon the ResNet on certain tasks, consistently fell short of the ViT's performance, indicating that convolutionally generated features do not necessarily assist or enhance the performance of ViTs. Additionally, the study on the fusion methods confirmed that patch tokens can be used on both ViT and Hybrid baselines for classification rather than the usual CLS tokens, as they contain sufficient information about the image. Lastly, incorporating feature mixing strategies in the lower layers of the model and introducing an invariant loss have proven to be effective in developing a suitable pipeline for improving robustness in the cross-domain scenario. This pipeline also showed that a model composed of a dissimilarity of pretrained parameters benefits immensely from adjusting the early stages of training. As section 4.5.7 discussed, when leveraging a pretrained backbone with an untrained module, it is suitable to freeze the pretrained parameters at the start of the training and allow the unitialized parameters to first converge to the backbone rather than training all parameters at once.

By analysing the domain evaluation results in sections 4.5.1 and 4.5.2, the experiments showed that the performance within a single dataset was better than the performance across multiple datasets for all baselines. This is because the capture conditions between WMCA and CASIA-SURF are different, making it more challenging for the models to learn and adapt to the unique characteristics of different datasets simultaneously, ultimately leading to a worsening of results compared to intra-domain level. In zero-shot evaluation, despite the evident collapse in the LOO-3D protocol, there was a notable success across all baselines in detecting unknown attacks during the test phase, which demonstrates the ability of the designed models to transfer knowledge from previously trained attacks to identify unseen attack categories.

Finally, this work also proved that the overall effectiveness of attack detection significantly improved in the presence of multi-modal information. Specifically, the tests conducted on section 4.5.3.1 using the ViT baseline demonstrated that incorporating depth and infrared information greatly enhanced the model's ability to detect 2D and 3D attacks. Naturally, this highlights the necessity to supplement the RGB camera with depth and infrared sensors during face capture to increase security during authentication.

5.2 Future Work

Based on the contributions provided by this thesis, we suggest that future experimental work in this field can be directed towards the following directions:

- Extend the cross-domain evaluation to zero-shot cross-domain. This can be achieved by acquiring a third multi-modal PAD dataset and evaluating the network's generalization ability by training on the first two datasets and testing on the third. On top of that, instead of treating this problem as a binary classification task, it should be extended to a multi-class detection problem, where each class is assigned to a unique PA.

- Alongside the standard ViT used in this work, it would be interesting to delve into the unique properties of some variants and explore their use for spoof detection, as some of them were designed to address the limitations of ViTs. Since these architectures are usually large-scale models, further investigation should focus on scaling these models to low hardware requirements. This is necessary to ensure that they can be used in real-time applications or on devices with limited resources.

- In addition, we believe that the performance can be improved by utilizing other loss functions and exploring different model parameters. It would also be valuable to investigate other data augmentation strategies that are more suitable for FAS.

6 References

[1] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, “Deep learning for face anti-spoofing: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5609–5631, 2022.

[2] Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, “A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices,” *Journal of imaging*, vol. 6, no. 12, p. 139, 2020.

[3] J. Yang, Z. Lei, and S. Z. Li, “Learn convolutional neural network for face anti-spoofing,” *arXiv preprint arXiv:1408.5601*, 2014.

[4] X. Tu, H. Zhang, M. Xie, Y. Luo, Y. Zhang, and Z. Ma, “Enhance the motion cues for face anti-spoofing using cnn-lstm architecture,” *arXiv preprint arXiv:1901.05635*, 2019.

[5] Y. Liu, A. Jourabloo, and X. Liu, “Learning deep models for face anti-spoofing: Binary or auxiliary supervision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 389–398, 2018.

[6] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, “Face anti-spoofing using patch and depth-based cnns,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 319–328, 2017.

[7] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, “Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 182–193, 2020.

[8] A. Parkin and O. Grinchuk, “Recognizing multi-modal face spoofing with face recognition networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[9] T. Shen, Y. Huang, and Z. Tong, "Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1611–1616, 2019.

[10] A. George and S. Marcel, "Can your face detector do anti-spoofing? face presentation attack detection with a multi-channel face detector," *arXiv preprint arXiv:2006.16836*, 2020.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8, IEEE, 2021.

[13] Y.-H. Huang, J.-W. Hsieh, M.-C. Chang, L. Ke, S. Lyu, and A. S. Santra, "Multi-teacher single-student visual transformer with multi-level attention for face spoofing detection," in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, p. 125, BMVA Press, 2021.

[14] A. R. Samar, M. Umer Farooq, T. Tariq, B. Khan, M. O. Beg, and A. Mumtaz, "Multi-modal face anti-spoofing transformer (mfast)," in *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 494–501, 2022.

[15] A. Liu, Z. Tan, Z. Yu, C. Zhao, J. Wan, Y. L. Z. Lei, D. Zhang, S. Z. Li, and G. Guo, "Fm-vit: Flexible modal vision transformers for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, 2023.

[16] A. Liu and Y. Liang, "Ma-vit: Modality-agnostic vision transformers for face anti-spoofing," *arXiv preprint arXiv:2304.07549*, 2023.

[17] C. Liao, W. Chen, H. Liu, Y. Yeh, M. Hu, and C. Chen, "Domain invariant vision transformer learning for face anti-spoofing," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Los Alamitos, CA, USA), pp. 6087–6096, IEEE Computer Society, jan 2023.

[18] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang, "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," in *European Conference on Computer Vision*, pp. 37–54, Springer, 2022.

[19] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4123–4133, 2022.

[20] Z. Yu, A. Liu, C. Zhao, K. H. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6345–6350, 2023.

[21] "Relatório de mercado de reconhecimento facial: Tamanho, participação, crescimento e tendências (2022-27)."

[22] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, 2019.

[23] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[24] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 2636–2640, IEEE, 2015.

[25] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[26] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, 2013.

[27] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision – ECCV 2010* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), (Berlin, Heidelberg), pp. 504–517, Springer Berlin Heidelberg, 2010.

[28] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, 2016.

[29] J. Gan, S. Li, Y. Zhai, and C. Liu, "3d convolutional neural network based on face anti-spoofing," in *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*, pp. 1–5, 2017.

[30] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.

[31] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 290–306, 2018.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, pp. 10347–10357, PMLR, 2021.

[34] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021.

[35] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," *arXiv preprint arXiv:2107.00641*, 2021.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[37] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15908–15919, 2021.

[38] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366, 2021.

[39] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.

[40] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 579–588, 2021.

[41] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

[42] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12175–12185, 2022.

[43] L. T. Menon, A. L. Koerich, A. S. Britto Jr, *et al.*, "Style transfer applied to face liveness detection with user-centered models," *arXiv preprint arXiv:1907.07270*, 2019.

[44] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li, "Improving face anti-spoofing by 3d virtual synthesis," in *2019 International Conference on Biometrics (ICB)*, pp. 1–8, IEEE, 2019.

[45] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.

[46] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

Appendix A

Ablation Study on the baseline models

A.1 Intra-Domain Appendix

Table A.1: Performance of the Multi-Modal ViT baseline on the *grandtest* protocol of the WMCA dataset.

Dataset	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
WMCA *grandtest*	0.1	79.10	0.66	98.33	49.49	2.37	00.42	0.02
	0.01	78.47	08.50	71.36	39.93	4.52	0.59	0.14
	0.001	92.42	6.81	10.52	8.66	52.54	16.57	9.88
	0.0001	97.71	2.61	1.03	1.82	90.19	71.30	50.52
	0.00001	99.18	0.94	0.37	0.65	99.70	89.70	84.61
	0.000001	99.73	0.2	1.25	0.64	99.57	99.04	98.54

Table A.2: Performance of the Multi-Modal ViT baseline on the CASIA-SURF dataset.

Dataset	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
CASIA-SURF	0.1	91.54	7.12	11.55	9.34	62.83	22.80	5.68
	0.01	94.10	4.77	8.51	6.64	76.01	39.17	12.95
	0.001	97.44	1.93	4.0	2.97	92.26	73.81	49.43
	0.0001	98.67	1.01	2.07	1.54	97.89	88.18	70.56
	0.00001	88.87	0.0	15.96	7.98	99.94	96.65	86.81
	0.000001	78.75	30.46	0.0	15.23	99.01	94.98	87.76

Table A.3: Performance of the Multi-modal ResNet baseline on the *grandtest* protocol of the WMCA dataset.

Dataset	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
WMCA-*grandtest*	0.1	96.95	2.92	3.55	3.24	93.83	68.54	60.07
	0.01	96.89	3.40	2.02	2.71	91.37	64.12	58.03
	0.001	98.13	1.85	1.97	1.91	96.45	90.26	86.54
	0.0001	98.74	1.56	0.10	0.83	98.12	72.87	64.89
	0.00001	98.33	1.86	0.90	1.38	96.07	80.83	62.61
	0.000001	97.64	1.54	5.51	3.53	88.52	48.64	36.64

Table A.4: Performance of the Multi-Modal ResNet baseline on the CASIA-SURF dataset.

Dataset	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
CASIA-SURF	0.1	93.99	6.80	4.19	5.50	83.64	67.77	0.00
	0.01	99.38	0.43	1.05	0.74	99.59	96.17	89.12
	0.001	87.78	16.51	2.32	9.41	83.65	64.86	54.19
	0.0001	82.68	24.82	0.02	12.42	89.94	64.27	44.08
	0.00001	90.81	13.14	0.06	6.60	97.87	86.75	57.73
	0.000001	81.17	26.98	13.51	96.22	83.58	63.43	

Table A.5: Performance of the Multi-modal Hybrid baseline on the *grandtest* protocol of the WMCA dataset.

Dataset	LR	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-3}
WMCA-*grandtest*	0.1	90.76	5.57	23.27	14.42	16.96	3.65	1.83
	0.01	87.06	1.49	29.12	56.75	38.16	0.59	0.52
	0.001	98.63	1.32	1.57	1.44	97.25	90.23	86.83
	0.0001	99.12	0.85	1.03	0.94	99.10	50.90	55.54
	0.00001	99.15	0.84	0.89	0.86	99.44	85.97	73.76
	0.000001	99.34	0.55	1.10	0.82	99.53	84.78	66.54

Table A.6: Performance of the Multi-modal Hybrid baseline on the CASIA-SURF dataset.

Dataset	LR	ACC(%)	APCER(%)	BPCER(%)	ACER(%)	TPR@FPR(%)		
						10^{-2}	10^{-3}	10^{-3}
CASIA-SURF	0.1	95.47	2.44	9.33	5.89	83.03	45.09	21.65
	0.01	98.62	0.75	2.85	1.8	97.83	83.29	63.07
	0.001	96.14	3.40	4.90	4.15	87.43	53.76	23.85
	0.0001	70.73	41.92	0.09	21.01	89.17	52.50	18.05
	0.00001	81.78	26.12	0.0001	0.1306	0.9672	0.8449	0.6746
	0.000001	70.37	42.49	0.0	21.24	96.98	80.88	51.83

A.2 Appendix Cross-Domain

Table A.7: Cross-domain performance of the Multi-modal ViT baseline.

Baseline	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
ViT	0.1	87.46	6.75	28.07	17.41	56.00	24.07	6.39
	0.01	91.02	4.23	12.1	8.16	69.59	31.98	18.6
	0.001	96.49	3.47	3.63	3.55	96.49	75.61	60.88
	0.0001	97.42	3.40	0.37	1.89	96.67	81.86	2.45
	0.00001	92.93	9.70	0.01	4.86	99.38	91.79	78.90
	0.000001	95.57	6.08	0.01	3.05	99.64	97.57	87.65

Table A.8: Cross-domain performance of the Multi-modal ResNet baseline.

Baseline	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
ResNet	0.1	**95.67**	4.97	2.62	3.80	89.28	56.18	37.88
	0.01	95.00	6.67	0.54	3.61	91.19	0.0	0.0
	0.001	92.54	10.10	0.39	5.24	95.07	82.55	62.20
	0.0001	91.02	12.30	0.06	6.18	95.17	85.39	74.25
	0.00001	94.62	7.26	0.36	3.81	96.62	91.44	78.17
	0.000001	88.17	16.08	0.44	8.26	93.46	76.64	47.33

Table A.9: Cross-domain performance of the Multi-modal Hybrid baseline.

Baseline	**LR**	**ACC(%)**	**APCER(%)**	**BPCER(%)**	**ACER(%)**	**TPR@FPR(%)**		
						10^{-2}	10^{-3}	10^{-3}
Hybrid	0.1	89.98	9.83	10.52	10.17	54.04	17.43	6.26
	0.01	92.73	6.49	9.37	7.93	62.16	34.25	22.25
	0.001	94.49	4.95	7.02	5.99	79.26	51.35	35.42
	0.0001	75.43	33.73	0.0	16.87	87.51	61.78	41.31
	0.00001	73.53	36.34	0.01	18.17	93.59	80.00	66.67
	-	-	-	-	-	-	-	-