Vision Transformers for Cross-Domain Presentation Attack Detection

Raghupatruni Venkata Harsha Vardhan[1], Tarigoppula V S Sriram[2]
[1]*M.Tech Student, Dept. of CSE, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam*
[2]*Associate Professor, Dept. of CSE, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam*

***Abstract* — Facial biometric systems are essential for modern security and authentication, but they remain highly vulnerable to physical spoofing methods, such as printed photos and screen replays. Because deepfake detection and physical presentation attack detection (PAD) both focus on identifying fabricated faces, recent research has hypothesized that deepfake models might naturally generalize to physical PAD without needing task-specific retraining. This study tests that hypothesis using the Cross-Domain Generalization Framework for Presentation Attack Detection (CDGF-PAD). Within this framework, Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures were evaluated across both in-domain and cross-domain scenarios. ISO/IEC 30107-3 metrics (APCER, BPCER, ACER, and AUC) were used to evaluate the performance of models. These models were trained on a subset of CelebA-Spoof and tested on the CASIA-FASD dataset. Our results show that ViT models trained solely on deepfakes produce near-random accuracy at zero-shot physical PAD. On the other hand, ViTs show significantly greater cross-domain generalization than CNNs when exposed to dataset shifts when explicitly trained on PAD data. In the conclusion, this study shows that physical PAD and deepfake detection are really different problems that require separate training.**

***Index Terms* — Biometric Security, Cross-Domain Generalization, Deepfake Detection, Face Anti-Spoofing, Presentation Attack Detection, Vision Transformers.**

I. INTRODUCTION

Biometric recognition systems are essential for access control, border crossing checks, and financial transactions. But despite the growing use of face recognition techniques, many security threats emerge due to their susceptibility to physical Presentation Attacks (PAD), such as the use of photo prints, video playback, and 3D masks. In addition, rapid development of the deepfake technology has led to the intensification of research in automated PAD detection, where the state-of-the-art results are achieved using Vision Transformer (ViT) models. Although many studies were devoted to the usage of Convolutional Neural Networks (CNN) and different domain generalization techniques for physical face anti-spoofing, more recently the trend in this area switched to transformer models. Still, the following key areas remain unstudied today: There is no zero-shot evaluation of deepfake pre-trained ViTs for PAD. the performance of CNN and ViT models under equal cross-domain generalization conditions has never been evaluated. There is no comparison study for attention maps obtained using deepfakes and PAD training. To fill the existing research gaps, we propose the Cross-Domain Generalization Framework for Presentation Attack Detection (CDGF-PAD).

II. BACKGROUND

Although biometric facial recognition is heavily deployed across critical security sectors, it remains highly vulnerable to physical presentation attacks (PAD) such as printed photographs and video replays. The concurrent rise of deepfake technology has led to advanced detection models, sparking a common but flawed assumption that these deepfake detectors can be seamlessly applied to physical anti-spoofing without task-specific retraining. In reality, the two tasks demand entirely different visual cues: deepfake models focus on digital manipulation artifacts and compression errors, whereas physical PAD requires detecting physical liveness indicators like surface

reflections and texture distortions caused by the spoofing medium. While Convolutional Neural Networks (CNNs) have traditionally dominated PAD research, Vision Transformers (ViTs) are emerging as a powerful alternative due to their global attention mechanisms that better capture liveness representations. Despite this architectural shift, significant research gaps remain; specifically, there are no existing zero-shot evaluations of deepfake-trained ViTs applied to physical PAD, a lack of direct cross-domain performance comparisons between CNNs and ViTs under identical conditions, and no comparative analysis of the visual features prioritized by their respective attention maps.

III. LITERATURE SURVEY

The change in the face anti-spoofing and presentation attack detection pattern has been enormous, from the classic machine learning methods to the current cutting-edge deep learning architectures using the transformers. The initial study of biometric security has been mainly on hand-crafted feature extractors. For example, Chingovska et al. [2] showed that Local Binary Patterns (LBP) works well in extracting the textural information of genuine face and spoofing mediums. However, these traditional methods were not robust enough to meet the needs of diverse real-world presentation attacks. Arashloo et al. [1] provided a comprehensive survey about the evolution of deep learning and its impact on PAD, in which deep-learning models can automatically learn complex and hierarchical feature representations. After this change, Convolutional Neural Networks (CNNs) have come to the fore in physical face anti-spoofing. The researchers' aim was to enhance the performance of CNN to go beyond binary classification. Atoum et al. [9] presented a two-stream CNN architecture, which uses the patch-based features as well as depth map estimations, and Liu et al. [13] highlighted the need to provide auxiliary supervision (e.g. depth and rPPG signals) to help deep models learn genuine liveness cues. To achieve better spatial and temporal representations, to capitalise on this, George and Marcel [10] explored the initial performance of ViTs for zero-shot face anti-spoofing, and Huang et al. [12] suggested adaptive transformer architectures for cross-domain face anti-spoofing with few shots.

The following questions are used to identify the research gap. Although CNN-based methods for domain generalization have been widely researched, and transformer structures have been recently incorporated into various applications, a major challenge remain to be addressed: the intersection of these two domains. There are no empirical studies yet that assess whether the global attention mechanisms in the advanced ViTs trained with only digital deepfakes can be intrinsically generalized to zero-shot physical PAD. Moreover, there is no explicit comparison that looks at how well PAD-supervised ViTs are capable of handling cross-domain data shift compared to conventional CNNs on the same dataset shift and standardized ISO/IEC 30107-3 evaluation metrics. This gap in the literature sets the groundwork and need for the proposed CDGF-PAD framework.

IV. PROBLEM STATEMENT

Despite the critical reliance on facial biometric systems for secure authentication, these frameworks remain highly susceptible to physical presentation attacks (PAD), such as printed photos and screen replays. A prevailing yet untested assumption in current research is that advanced deepfake detection models, particularly Vision Transformers (ViTs), can inherently generalize to detect these physical attacks without task-specific retraining, as both involve identifying artificial faces. However, this assumption conflates two fundamentally distinct challenges: deepfake algorithms target digital manipulation artifacts and compression errors, whereas physical PAD requires the detection of physical liveness cues, such as surface reflections and material textures. Consequently, a critical problem exists in the current literature: it is unknown whether deepfake-supervised ViTs can effectively perform zero-shot physical PAD, nor is there empirical evidence directly comparing the cross-domain generalization capabilities of ViTs against traditional Convolutional Neural Networks (CNNs) under dataset shift conditions. Furthermore, the distinct visual features prioritized by these models remain unanalysed, leaving a significant gap in developing robust, domain-adaptable biometric security systems.

V. PROPOSED SOLUTION

To address these vulnerabilities and research gaps in current face biometric authentication system, this study proposes a comprehensive, framework around the Cross-Domain Generalization Framework for Presentation Attack Detection (CDGF-PAD). This framework is specifically designed to bridge the difference between digital deepfake detection and physical anti-spoofing by systematically evaluating model architectural robustness under real-world dataset shift conditions. The proposed solution operates through a carefully structured, three-phase evaluation pipeline. First, it conflicts the assumption of model transferability by performing a zero-shot evaluation on the model. In this phase, publicly available Vision Transformers (ViTs) that have been exclusively pre-trained on deepfake datasets are deployed directly into physical PAD scenarios without any fine-tuning. This isolates whether models trained to recognize digital compression errors and manipulation artifacts can inherently detect physical liveness cues. Second, the framework conducts a controlled, supervised benchmarking phase to directly compare the generalization capabilities of distinct neural network architectures. A traditional Convolutional Neural Network (ResNet-18) and a transformer-based architecture (ViT-Base-Patch16-224) are trained on a carefully balanced subset of the highly diverse CelebA-Spoof database. By training both models under identical, optimized conditions, the solution ensures a fair comparison of how well localized CNN filters perform against the global attention mechanisms of ViTs when learning physical liveness features. To rigorously test cross-domain adaptability, these supervised models are then evaluated on an entirely unseen external dataset, CASIA-FASD, under both balanced and heavily imbalanced class distributions. All performance outcomes are evaluated using strict, standard ISO/IEC 30107-3 metrics, specifically tracking the APCER, BPCER, ACER, and AUC.

By visualizing the specific facial regions that trigger model decisions, the framework decodes the internal focus of the network. This allows researchers to verify whether a model is incorrectly relying on peripheral background noise and high-frequency textures, or if it is accurately targeting genuine physical liveness attributes such as the nose, cheeks, and periocular regions. Ultimately, this solution not only provides concurrent evidence that deepfake detectors cannot be safely repurposed for physical anti-spoofing, but it also connects a clear architectural pathway by, proving that task-specific, PAD-supervised Vision Transformers offer a highly robust and generalizable defense for modern biometric security systems.

VI. METHODOLOGY

In this work, we propose and introduce the Cross-Domain Generalization Framework for Presentation Attack Detection (CDGF-PAD) to rigorously examine the transferability of deepfake detectors and benchmark the architectural robustness against physical spoofing. The methodology can be divided into several stages, including data preparation, model configuration, and comprehensive evaluation. Characteristic of the framework architecture. Key principles of the framework architecture. The CDGF-PAD pipeline is designed to control performance variables in three sequential evaluation stages: Zero-Shot Transfer Evaluation: Deepfake data is used for pre-training only Vision Transformers (ViTs) to achieve fine-tuning-free identification of Physical Presentation Attacks (PPA). In-Domain PAD Evaluation: Baseline supervised performance with models trained and tested using the same environmental distribution. Cross-Domain PAD Evaluation: Testing the trained models on completely different data sets, and computing their robustness to domain shift in architecture. Students will learn how to select and assemble data sets. Data Set Selection and Composition The experimental design takes advantage of two different facial biometric data sets to impose a strict split between training and cross-domain testing: CelebA-Spoof is a highly-diverse dataset with more than 625,000 images. In order to maintain computational efficiency and strict parity of classes, a balanced subset of 4,000 images has been extracted, consisting of 2,000 "bona fide" (genuine) images and 2,000 "spoof" images. The spoof category includes a variety of physical presentation attacks, such as printed photos, replay screens, and partial face manipulations. This subset was then divided into an in-domain testing set (800 images) and a training set (3,200 images). Cross-Domain Testing Database (CASIA-FASD): In order to ensure the external

evaluation is completely unbiased, CASIA-FASD was exclusively used for the external evaluation process after the training and hyperparameter optimization stages. The framework also employs two testing splits of this dataset: a balanced subset (323 genuine and 323 spoof images) to check the baseline error rates, and an imbalanced subset (10,128 genuine and 55,658 spoof images) to determine the stability of the models in the presence of very unbalanced real-world distribution. Transforming the data to improve its quality. Data Preprocessing and Augmentation Raw images were processed through a detailed and strict preprocessing pipeline to guarantee the uniformity of the inputs across the different architectures. First, it was used spatial face detection and alignment so as to extract the face areas. Images were then cropped and resized to 224×224 pixels in aligned position with bilinear method. Data augmentation techniques such as random horizontal flipping on a probability of 0.5, and random bounding box rotations were used to improve model generalization during the supervised training period. The corrupted images and those whose alignment is off by too many degrees were removed systematically before splitting the data, and exact class balance was checked after the preprocessing. Describe various model architectures and configurations. Four different model setups are investigated and classified into two groups: Unsupervised (Zero-Shot) Models: Two publicly available Vision Transformer architectures, pre-trained only on deepfake detection tasks were tested. The models were fed with the PAD directly without any fine-tuning to validate the hypothesis of cross-task transferability. When explicitly trained to perform physical anti-spoofing, two different PAD model topologies were used: Supervised PAD Models: CNN Baseline: a basic ResNet-18 structure, which is localized and convolution based and extracts spatial features. Transformer Baseline: A model architecture based on ViT-Base-Patch16-224, that is, global self-attention mechanisms. The two supervised architectures were adapted to include the inclusion of binary classifiers in their final layers to provide genuine versus spoof probability scores. Training Protocol and Optimization PyTorch, a deep learning framework, was used for all of the supervised experiments. The Adam optimizer was used to optimize network weights in an efficient way to navigate the high-dimensional loss landscape. Network weights were optimized using the Adam optimizer. Learning hyperparameters have been adjusted for each model topology, due to the differences in architecture. F. Evaluation Metrics Performance was critically evaluated using standardised biometric security measures according to the ISO/IEC 30107-3 guidelines. The classification thresholds were used to extract the following main metrics: APCER (Attack Presentation Classification Error Rate): The percentage of spoof attacks that are wrongly identified as legitimate attacks. Bona Fide Presentation Classification Error Rate (BPCER): Ratio of incorrect rejection of real users. CPER (Correct Public Classification Error Rate): Average of APCER and BPCER, the overall system reliability will be decided by this metric. AUC (Area Under the Curve): It is a measure of the model's overall difference around all possible thresholds obtained from the Receiver Operating Characteristic (ROC) curve.

VII. DUAL MODEL ARCHITECTURE

In this study, two basic and very different architectures of neural networks are employed to thoroughly assess and compare the feature extraction paradigms in the context of physical Presentation Attack Detection (PAD): the Convolutional Neural Network (CNN) and the Vision Transformer (ViT). The models were chosen to compare the effectiveness of local spatial filtering with global self-attention networks in detecting physical liveness cues. The study uses ResNet-18 as the foundation for traditional, convolution-based architectures. The architecture is specially designed to overcome the vanishing gradient problem in deep networks by using a residual learning framework that passes gradients to skip intermediate layers. Feature Extraction: The network takes the 224x224x3 normalized input images, and processes them using a first 7x7 convolutional layer that is followed by max pooling. The main architecture is made up of four successive residual blocks, with two 3x3 convolutional layers, Batch Normalization and ReLU activation functions. ResNet-18 is intrinsically based on local spatial hierarchy, and successfully preserves the high-frequency textural features with short-range influence. In this study, the original 1000-class fully connected layer of ImageNet was removed. The final feature maps were subjected to a global average pooling layer and then passed to a custom

fully connected layer with two outputs representing the probabilities of genuine or spoofing, respectively. ViT-Base-Patch32-14-128. To evaluate the capability of global attention mechanisms in face anti-spoofing, the ViT-Base-Patch16-224 architecture was implemented. The ViT is different from CNNs because it considers an image as a series of unique tokens rather than processing it pixel-by-pixel via sliding windows. A trainable linear projection maps them to a fixed dimensional latent vector size. Positional Encoding and CLS Token: Since there is no intrinsic spatial geometry in transformers, 1D learnable positional embeddings are appended to the patch embeddings to preserve the positional information. Also, a special learnable classification token [class] is at the beginning of the sequence, which represents the whole image. Transformer Encoder: The embedded sequence goes through 12 transformer encoder blocks. The key blocks are mainly Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks with Layer Normalization (LN) and residual connections. The MSA mechanism enables the model to be able to include long-range global dependencies including the structural relationship between a face's nose, cheeks and lighting reflections that is essential for determining 3D physical liveness. The state of the token in the output of the last encoder block is passed to a Multi-Layer Perceptron (MLP) head. The CNN setup was modified to also conduct binary classification (genuine vs. spoof) for the PAD task, just as was done for the terminal layer. Zero-Shot Deepfake Configurations. Two pre-trained deepfake detection models publicly available were used in the zero-shot transferability analysis (Deepfake-A and Deepfake-B). These models are based on the same Vision Transformer (ViT) backbone as mentioned above but are trained using digital manipulation datasets, such as real faces and faces transformed into other images by AI or face-swapping techniques. For the framework pipeline, the pre-processed facial images were fed into these models, with no weight tuning, fine-tuning or domain adaptation done on the physical PAD datasets.

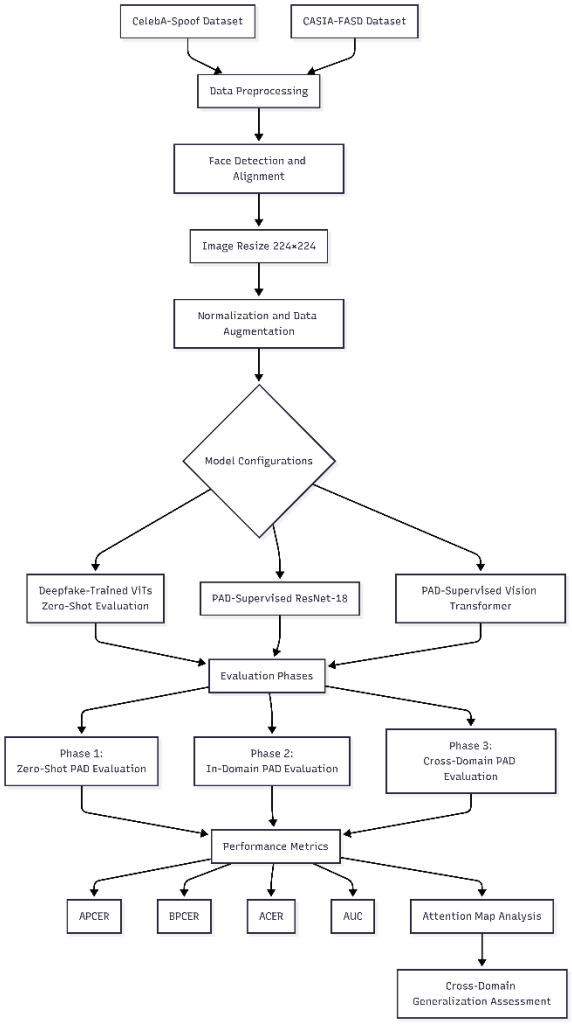

Fig. 1. Dual Model Architecture of the proposed System.

VIII. IMPLEMENTATION

This section provides a detailed description of the procedure for implementing the CDGF-PAD framework, including the software environment, data processing pipeline, and hyperparameter tuning during training and testing of the neural network architectures selected. The implementation of the CDGF-PAD framework will be described, consisting of software environment, data processing pipeline, and hyperparameter tuning during training and testing of the selected neural network architectures.

Data preprocessing was performed and the model was trained and evaluated by the open-source PyTorch deep learning framework. Considering that training computational limit is 4,000 images, it was decided to

take a balanced subset of CelebA-Spoof with 4,000 images for the training phase. This enabled training and evaluation of both Convolutional Neural Network (CNN) and Vision Transformer (ViT) model with the same hardware configurations so that there was no memory issue. Raw images were first subjected to a standard image preprocessing pipeline, to get rid of any background noise in the images and ensure that the neural networks would only be looking at the facial features and not at the background noise: Spatial Face Detection and Alignment: To get the faces into consistent spatial positions they were first detected and aligned using spatial face detection algorithms. Face images were resized to a fixed-size of 224 × 224 pixels to fit the input to the ResNet-18 and ViT architectures, bilinear interpolation was used for resizing. All three colour channels (RGB) were normalised using mean and standard deviation vectors [0.5, 0.5, 0.5] to normalise the pixel intensities. This shifted the pixel values so that they were all in the range of [-1, 1] which helped in speeding up the Gradient Descent and finding a better convergence in the model. For both CNN and ViT supervised models, the Adam optimizer was used to optimize the weights of the network, which helped to stabilize the learning process and optimize the weights. The loss landscape however is different for both transformer and convolutional networks and different learning rates were used for each network architecture: Using a learning rate of 1×10^{-4} and a batch size of 32, trained with ResNet-18. Trained with ResNet-18 on a learning rate of 1×10^{-4} and a batch size of 32. ViT-Base-Patch16-224 (Transformer): The transformer trained with the lower learning rate of 2×10^{-5} and the batch size of 32. A simple trick for getting the model to learn much slower than it normally would and to not get "stuck" too soon in the learning process is to reduce the learning rate. The implementation was done in two different modes of operation: Both pre-trained models Deepfake-A and Deepfake-B were tested in the PyTorch environment (torch.no_grad ()). These splits of CelebA-Spoof were used to fine-tune and backpropagate, and did not provide the classification scores. The baseline models Supervised Execution (Supervised Execution: The ResNet-18 and ViT) were trained in an iterative way on the training set of the CelebA-Spoof. Weights were iteratively updated to convergence, using the loss function, binary cross-entropy. Finally, the models were frozen and then deployed for in and cross domain testing.

IX. RESULTS AND DISCUSSION

A. Quantitative Evaluation

System performance was measured using standardized ISO/IEC 30107-3 metrics: Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER).

Note: High BPCER values indicate an increased and undesirable rejection of genuine users

Table I: Performance Metrics Across Model Configurations

Model	Dataset (Split)	APCER	BPCER	ACER	AUC
ResNet-18 (PAD)	CelebA-Spoof (test)	0.0125	0.0050	0.0087	0.9996
ResNet-18 (PAD)	CASIA-FASD (balanced)	0.5077	0.2910	0.3994	0.6793
ResNet-18 (PAD)	CASIA-FASD (full)	0.0067	0.9353*	0.4710	0.5996
ViT-PAD (CDGF-PAD)	CelebA-Spoof (test)	0.0025	0.0025	0.0025	0.9999
ViT-PAD (CDGF-PAD)	CASIA-FASD (balanced)	0.2198	0.2291	0.2245	0.8284
ViT-PAD (CDGF-PAD)	CASIA-FASD (full)	0.0798	0.6886	0.3842	0.8739**
Deepfake-A	CelebA-Spoof (test)	0.6950	0.2800	0.4875	0.4780
Deepfake-A	CASIA-FASD (balanced)	0.0495	0.9071	0.4783	0.5611
Deepfake-A	CASIA-FASD (full)	0.1096	0.9400	0.5248	—
Deepfake-B	CelebA-Spoof (test)	0.4100	0.6150	0.5125	0.4596
Deepfake-B	CASIA-FASD (balanced)	0.8947	0.0433	0.4690	0.6034
Deepfake-B	CASIA-FASD (full)	0.9870	0.0053	0.4962	—

As demonstrated in Table I, the zero-shot deepfake models produced near-random classification performance in-domain (AUC ~0.45 - 0.47) and exhibited highly unstable decision behaviors under class imbalance. Specifically, Deepfake-A suffered from high bona fide rejection rates (BPCER of 0.9400 on the full split), while Deepfake-B yielded unacceptably high spoof acceptance rates (APCER of 0.9870). Under the highly challenging cross-domain imbalanced split (simulating real-world deployments with bona fide samples and spoof samples), the zero-shot models suffered catastrophic operational failure. Deepfake-A completely biased its decision boundary toward the negative class, yielding a crippling BPCER of 0.9400 (rejecting of genuine users). In Contrast, the Deepfake-B has the issue entirely toward the positive class, resulting in an unmitigated security breach with an APCER of 0.9870 (physical spoof attacks were misclassified as legitimate entries). There is a in difference with the near random classification.

B. ROC Curve Analysis

Receiver Operating Characteristic (ROC) analysis was utilized to map the discriminative capabilities of the models across various decision thresholds.

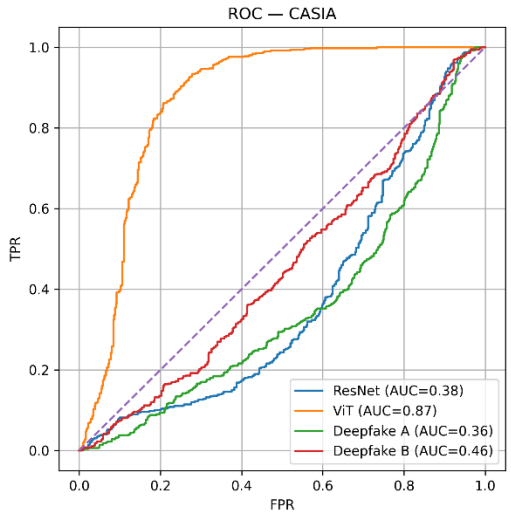

Fig. 2. ROC Curve Analysis for cross-domain evaluation on the CASIA-FASD dataset.

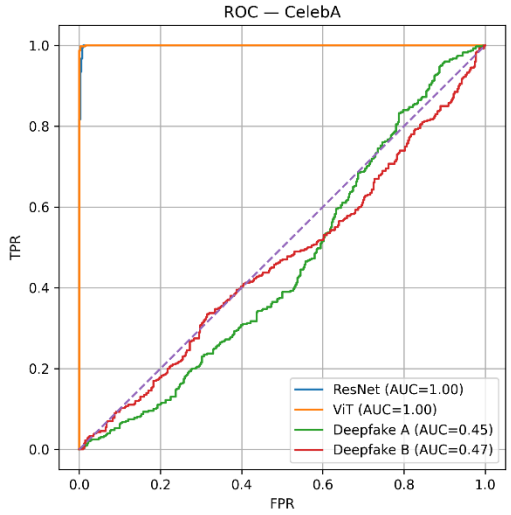

Fig. 3. ROC Curve Analysis for in-domain evaluation on the CelebA-Spoof dataset

The ROC analysis strictly corroborates the quantitative findings: task-specific PAD supervision is mandatory for reliable physical anti-spoofing. Moreover, the ViT-PAD model also showed high discriminative performance on the unseen CASIA-FASD dataset, with an AUC of 0.8284 compared with that of the CNN-based ResNet-18 model which indicates that it is more effective at coping with the domain shift.

C. Attention Map and Feature Focus Analysis

To understand how the internal decision-making processes of the transformer architectures in this hybrid approach attention maps are being used.

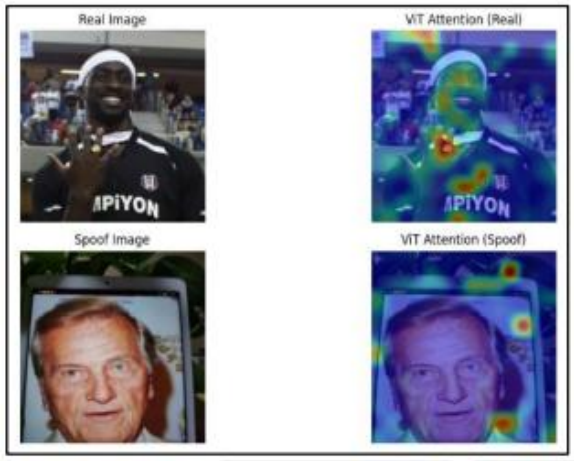

Fig. 4. Qualitative comparison of attention maps. Top: Zero-shot deepfake ViTs focusing on peripheral noise. Bottom: PAD-supervised ViT focusing on physical liveness attributes.

- Deepfake Models: heavily reliable on extracting information from the peripheral regions of the faces, depending on digital textures and compression artifacts.
- PAD-Supervised ViT: Transformer's attention toward physically relevant facial regions specifically the nose, cheeks, and periocular areas. These regions carry critical 3D structural and surface reflection data essential for determining physical liveness.

X. CONCLUSION

In this study, a Cross-Domain Generalization Framework for Presentation Attack Detection (CDGF-PAD) is proposed to comprehensively assess the cross-domain robustness of Vision Transformer (ViT) architectures for facial biometric security. One of the main goals of this research was to empirically validate the commonly held belief that the advanced deepfake detection models learn to generalize to PAD without task-specific retraining. This assumption is shown to be false in the large number of experimental results obtained. When tested on physical datasets, the ViTs that were only trained on digital deepfakes demonstrated very unstable classification performance near random accuracy. This shows that the challenges

of deepfake detection and physical PAD are two different types of biometric challenges, where the detection of the deepfake doesn't equal the detection of physical liveness cues. Therefore, these domains demand a distinct training paradigm for each domain. Moreover, the controlled benchmarking between traditional Convolutional Neural Networks (ResNet-18) and transformer-based architectures showed that the global attention mechanisms seem to be better at dataset shift than the former. Both models achieved high performance in domain, but the CNN's performance was significantly reduced in the cross-domain test. By contrast, the PAD-supervised Vision Transformer outperformed all other models by a large margin in terms of cross domain robustness (AUC of 0.8284 vs. 0.6793 when split balanced) and also had much better stability across skewed class distributions. Overall, this thesis concludes that the deepfake detectors are not readily available to be repurposed for physical anti-spoofing, but explicitly PAD-supervised Vision Transformers can be a highly robust, generalizable and transparent solution to protect modern face biometric systems from real world presentation attacks.

REFERENCES

[1] Arashloo SR, Kittler J, Christmas W. Deep learning for face anti-spoofing: a survey: arXiv. 2017 [cited 2026 May 7]; from https://arxiv.org/abs/1708.05273.

[2] Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG Proceedings of the International Conference of Biometrics Special Interest Group. Piscataway (NJ): IEEE; 2012. p. 1-7.

[3] Dang H, Liu F, Stehouwer J, Liu X, Jain AK. On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway (NJ): IEEE; 2020. p. 5781-90.

[4] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. San Diego (CA): ICLR; 2021.

[5] Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolpashnikova D, Holz T. Leveraging Frequency analysis for deep fake detection. In: International Conference on Machine Learning. New York (NY): PMLR; 2020. p. 3247-58.

[6] Jia Y, Zhang J, Shan S, Chen X. Single-side domain generalization for face antispoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway (NJ): IEEE; 2020. p. 8484-93.

[7] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway (NJ): IEEE; 2019. p. 1-11.

[8] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data- efficient image transformers and distillation through attention. In: International

[9] Atoum Y, Liu Y, Jourabloo A, Liu X. Face anti-spoofing using patch and depth-based CNNs. In: 2017 IEEE International Joint Conference on Biometrics. Piscataway (NJ): IEEE; 2017. p. 319-28.

[10] George A, Marcel S. On the effectiveness of Vision Transformers for zero-shot face anti spoofing. In: 2021 IEEE International Joint Conference on Biometrics. Piscataway (NJ): IEEE; 2021. p. 1-8.