Fusion of vision transformers and convolutional networks for advanced face anti-spoofing

Zhanseri Ikram[1,*,†] and Bauyrzhan Omarov[1,†]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*

Abstract

Face anti-spoofing systems play a crucial role in securing biometric authentication frameworks against presentation attacks. The growing complexity of spoofing techniques demands the development of advanced detection methods that can effectively generalize across various attack forms and environmental conditions. In response to challenges, a new architecture fusing Vision Transformers (ViT), ConvNeXT, and Swin Transformer is proposed for advanced face anti-spoofing. The method combines global contextual modeling with local feature extraction and multi-scale analysis. Detailed evaluations on the OULU-NPU and CASIA-MFSD datasets demonstrate competitive performance across various protocols, with notable improvements in generalization to unseen environmental conditions. Feature space visualizations reveal improved class separability post-fusion, emphasizing the effectiveness of the combined approach. Cross-dataset experiments highlight challenges in domain generalization in bidirectional evaluations between OULU-NPU and CASIA-MFSD. The proposed method advances the state-of-the-art in face anti-spoofing, offering insights into feature fusion strategies and avenues for future research in cross-domain generalization.

Keywords

Face anti-spoofing, machine learning, computer vision, transformers

1. Introduction

Face recognition technologies have rapidly evolutionized and are now essential to various security systems, from personal devices to large-scale surveillance networks. However, the widespread adoption of these technologies has also led to the emergence of face spoofing attacks, where attackers use photos, videos, masks, or other facial representations to deceive recognition systems [1]. The field of face anti-spoofing is a critical component of robust biometric systems and it has undergone significant advancements in recent years, primarily fueled by the rapid evolution of deep learning methodologies.

The fusion of ViT and Convolutional Neural Networks (CNNs) represents a promising route for addressing these challenges. Vision Transformers, introduced by [2], have demonstrated remarkable performance in various computer vision tasks by applying self-attention mechanisms to model long-range dependencies. Conversely, CNNs excel at extracting hierarchical local features and have been the cornerstone of many successful face anti-spoofing approaches [3]. The synergistic integration of these architectures aims to harness their complementary strengths, potentially yielding a more detailed and nuanced representation of facial characteristics pertinent to spoofing detection.

The proposed methodology implies a multi-stream architecture that processes input images through ViT and uses two parallel pathways. In this context ViT offers the potential to capture

DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry, October 16–17, 2024, Almaty, Kazakhstan

* Corresponding author.

† These authors contributed equally.

ikram.zhanseri@outlook.com (Z. Ikram); bauyrzhan313@gmail.com (B. Omarov)

0009-0001-8059-6590 (Z. Ikram); 0000-0002-9312-4429 (B. Omarov)

subtle, global features that may be indicative of spoofing attacks. The proposed approach builds upon recent advancements in face anti-spoofing research, including multi-modal fusion techniques [4], attention mechanisms [5], and domain generalization strategies [6].

The remainder of this paper is organized as follows: Section 2 provides an overview of related works, the advancements and challenges in the domain of face anti-spoofing. In Section 3, we detail the materials and methods used in our study, including the problem statement, the proposed method, evaluation metrics, loss functions, and datasets. Section 4 presents the experimental results, demonstrating the performance of our approach on various benchmarks. Finally, Section 5 offers a discussion of the findings, their implications, and potential directions for future research.

2. Related works

Face anti-spoofing research has made a big growth transitioning from traditional handcrafted feature-based approaches to deep learning-driven methodologies.

Early face anti-spoofing techniques primarily relied on texture analysis to differentiate between genuine and spoofed facial presentations. Local Binary Patterns (LBP) and its variants were extensively employed to capture micro-textural patterns [7]. Subsequent works explored more sophisticated descriptors such as SURF [8] and HOG [9] to boost the discriminative power of extracted features. While these methods demonstrated good results in controlled environments, their performance often degraded under variable lighting conditions and against high-quality spoofing attacks.

The deep learning settled a paradigm shift in face anti-spoofing research. CNNs became the powerful tools for automatically learning hierarchical features from raw input images. [10] proposed a CNN architecture specifically designed for face anti-spoofing, incorporating a pixel-wise supervision strategy to improve localization capabilities. [11] introduced a multi-stream CNN framework that concurrently processed color, depth, and infrared information to bolster spoofing detection accuracy, thus outperforming traditional methods, particularly in scenarios involving diverse spoofing techniques. Recognizing the potential of temporal cues in distinguishing between genuine and spoofed facial presentations, researchers began incorporating motion analysis into anti-spoofing frameworks. [12] proposed a 3D CNN architecture to capture spatio-temporal features from video sequences. Long Short-Term Memory (LSTM) networks were employed by [13] to model the temporal dynamics of facial movements, demonstrating strong robustness against video replay attacks and 3D mask impersonations.

The integration of attention mechanisms into face anti-spoofing models has gained significant interest due to their ability to focus on salient regions. [14] introduced a spatial attention module to emphasize discriminative facial areas for spoofing detection. [15] proposed a channel attention mechanism to adaptively recalibrate feature maps, improving the model's sensitivity to subtle spoofing artifacts.

A continuous challenge in face anti-spoofing lies in the domain shift between training and testing distributions. To address this, several works have explored domain generalization techniques. [16] proposed a multi-adversarial domain generalization framework to learn domain-invariant features. [17] introduced a meta-learning approach to simulate domain shift during training, thereby increase the model's generalization capabilities. The recent success of ViT in various computer vision tasks has sparked interest in their application to face anti-spoofing. [18] adapted the ViT architecture for spoofing detection, demonstrating competitive performance with CNN-based counterparts. [19] proposed a hybrid CNN-Transformer model that applied both local and global feature representations for advanced spoofing detection. Recognizing the limitations of single-modality approaches, researchers have explored the fusion of multiple information sources for robust spoofing detection. [20] proposed a multi-modal framework that combined visible light, infrared, and depth information. [21] introduced a cross-modal fusion strategy that used complementary cues from different sensing modalities, showing that multi-modal approaches can address diverse spoofing scenarios and environmental variations.

3. Materials and methods

The proposed architecture in Figure 1 consists of ViT, ConvNeXT [22], and Swin Transformer [23] to create a robust face anti-spoofing system. ConvNeXT introduces a pure ConvNet approach that incorporates design elements from transformers, achieving performance competitive with state-of-the-art vision transformers while maintaining the efficiency and inductive biases of CNNs. Swin Transformer proposes a hierarchical vision transformer that utilizes shifted windows, enabling efficient modeling of image features at various scales. The methodology uses the strengths of each component to address the complex challenge of distinguishing genuine from spoofed facial presentations.

3.1. Problem statement

Face anti-spoofing systems aim to differentiate between bona fide facial presentations and fraudulent attempts using various spoofing techniques, such as printed photographs, digital displays, or 3D masks. The challenge lies in capturing both fine-grained textural details and global contextual information while maintaining robustness across diverse environmental conditions and attack modalities. Formally, given an input image $I \in \mathbb{R}^{H \times W \times C}$, where $H, W,$ and C represent height, width, and channels respectively, the objective is to learn a function $f \colon \mathbb{R}^{H \times W \times C} \rightarrow \{0,1\}$, where 0 denotes a spoofed presentation and 1 indicates a genuine facial image.

3.2. Proposed method

The proposed architecture comprises three main components. A ViT for global feature extraction, a ConvNeXT module for local feature refinement, and a Swin Transformer for multi-scale feature analysis. The fusion of outputs from ConvNeXT and Swin Transformer yields the final classification decision.

The ViT module partitions the input image *I* into *N* non-overlapping patches, each of size $P \times P$. In our experiment since we use 8x8 patching for 224x224x3 image, the *N* will become 784.

$$z_0 = \left[x_p^1\mathbf{E}, x_p^2\mathbf{E}, x_p^3\mathbf{E}, \dots, x_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \tag{1}$$

$where\ \mathbf{E} \in \mathbb{R}^{(P^2C) \times D}\ and\ \mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$, for embedding D = 768. Let's say f_θ is our ViT model. Then $L_i \in \mathbb{R}^{b \times c_i \times H_i W_i}$ is the output of *i*-th layer, where total number layers is 12 and b is the batch size. Out of 12 layers, we use only last 4 layers, since they represent the higher level information. Each of shape $[batch\_size, 784, 768]$, after concatenation the tensor becomes $[batch\_size, 784, 3072]$.

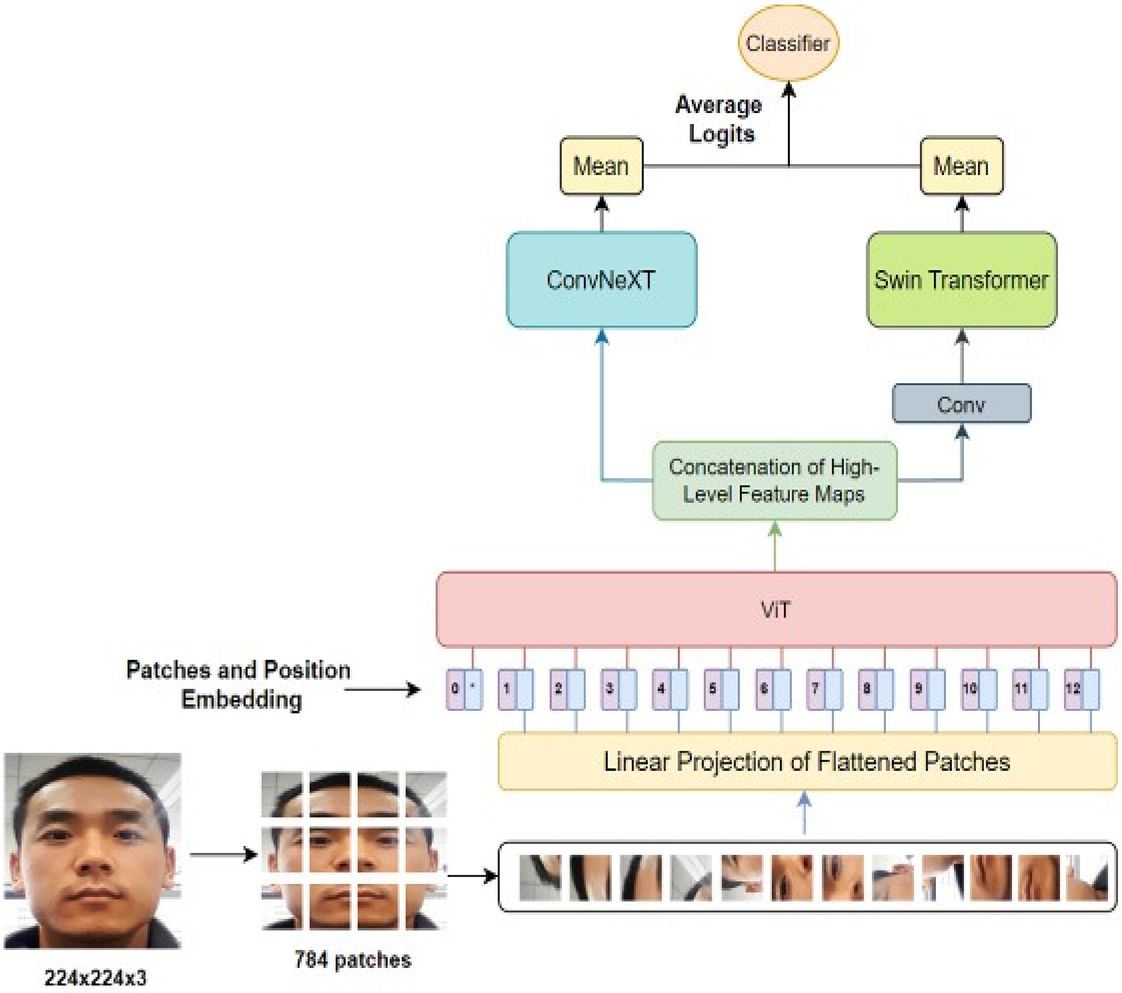

Classifier
Average Logits
Mean
Mean
ConvNeXT
Swin Transformer
Conv
Concatenation of High-Level Feature Maps
ViT
Patches and Position Embedding
0
*
1
2
3
4
5
6
7
8
9
10
11
12
Linear Projection of Flattened Patches
224x224x3
784 patches

Figure 1: The proposed model architecture.

The ConvNeXT module processes the concatenated high-level feature maps from ViT using a series of depthwise separable convolutions and inverted bottleneck layers. For each block *i*.

$$\mathbf{y_i} = DWConv(\mathbf{x_i}) + \mathbf{x_i} \tag{2}$$

$$\mathbf{y_i'} = GELU(\mathbf{y_i}) \tag{3}$$

$$\mathbf{y_i''} = PointwiseConv(\mathbf{y_i'}) \tag{4}$$

where *DWConv* and *PointwiseConv* represent depthwise and pointwise convolutions, respectively, and *GELU* is the Gaussian Error Linear Unit activation function.

The Swin Transformer operates on the same concatenated feature maps, employing shifted window-based self-attention to capture multi-scale contextual information

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = SoftMax\left(\frac{\mathbf{QK}^T}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V} \tag{5}$$

where **Q, K, V** are query, key, and value matrices, *d* is the dimension of queries/keys, and **B** is the relative position bias. Prior to the Swin Transformer, features were passed through a convolutional layer to reduce the number of parameters, thereby facilitating more efficient training. This approach was necessary because the Swin Transformer, unlike ConvNeXT, is comparatively slower in processing.

The outputs from ConvNeXT and Swin Transformer undergo global average pooling:

$$f_{convnext} = GAP(ConvNeXT(z_L)) \tag{6}$$

$$f_{swin} = GAP(SwinTransformer(z_L)) \tag{7}$$

where *GAP* is Global Average Pooling. The final value before feeding to *Sigmoid* function is

$$avg = \frac{f_{convnext} + f_{swin}}{2} \tag{8}$$

so that both have an equal contribution.

3.3. Metrics and loss

Performance metrics used to evaluate the effectiveness of these anti-spoofing systems include APCER, BPCER, and ACER [24].

APCER measures the rate at which the system incorrectly classifies a spoofing attack as a genuine attempt. In other words, it is the proportion of spoofing attempts that are wrongly accepted as legitimate by the system.

$$APCER = \frac{Number\ of\ False\ Accepts}{Total\ Number\ of\ Attack\ Presentations} \tag{9}$$

BPCER measures the rate at which the system incorrectly classifies a genuine attempt as a spoofing attack. metric reflects the system's ability to correctly recognize legitimate users.

$$BPCER = \frac{Number\ of\ False\ Rejects}{Total\ Number\ of\ Genuine\ Presentations} \tag{10}$$

ACER is the average of APCER and BPCER. It is usually used to find a balance between aforementioned metrics:

$$ACER = \frac{APCER + BPCER}{2} \tag{11}$$

For the loss, we used Binary Cross Entropy Loss:

$$BCELoss = -\frac{1}{N}\sum_{i=1}^{N}(y_i * \log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i)) \tag{12}$$

where N - the total number of samples, y_i- true label, $\hat{y}_i$ - predicted probability.

3.4. Dataset

The proposed method was evaluated on two widely recognized datasets in the face anti-spoofing domain: OULU-NPU [25] in Figure 2 and CASIA-MFSD [26] in Figure 3. These datasets provide diverse spoofing scenarios and environmental conditions, enabling detailed assessment of anti-spoofing algorithms.

OULU-NPU combines 4,950 real access and spoofing videos from 55 subjects, captured using six mobile devices with front-facing cameras. The dataset provides four protocols evaluating generalization across unseen environmental conditions, attack types, input sensors, and a combination thereof. Spoofing attacks include print and video-replay using two printers and two display devices. Environmental variations encompass three sessions with different illumination and background settings.

CASIA-MFSD contains 600 video clips of genuine and attack attempts from 50 subjects. The dataset features three image quality categories: low-quality, normal-quality, and high-quality. Spoofing attacks are categorized into three types: warped photo attacks, cut photo attacks, and video attacks. The dataset was collected under varying illumination conditions and with different digital devices, presenting challenges in terms of image quality and attack diversity. The dataset provides seven scenarios from three main protocols with low, high qualities and last is a mix of all train versus mix of all test samples. For this dataset we used only the 3rd protocol, which is a 7th scenario.

For both datasets we took only every 25th frame and 5 frame in total only. Face detection was performed using MTCNN [27] library.

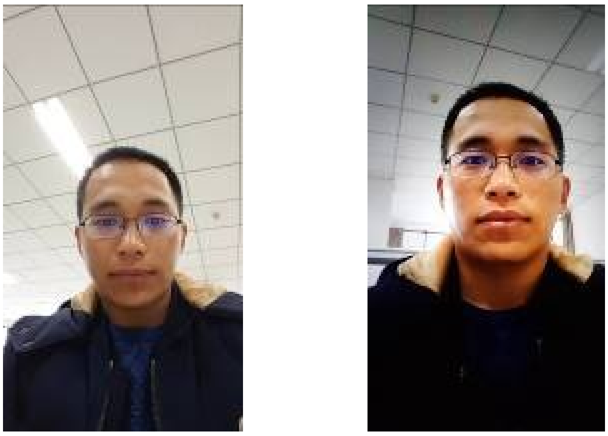


Figure 2: Samples from OULU-NPU. From left to right: “live” face and “replay” attack.

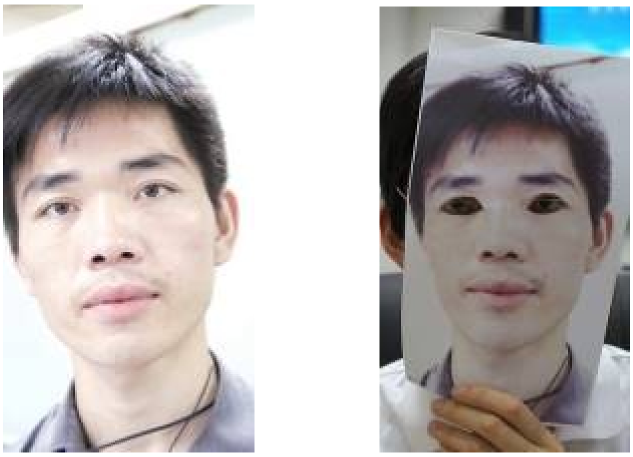


Figure 3: Samples from CASIA-MFSD. From left to right: “live” face and “print” attack.

3.5. Experimental setup

In this research, the experiments were executed using an NVIDIA RTX 4090 GPU with 24GB of VRAM. Batch size is 16, which took about 17GB memory of GPU. The optimizer is Adam with initial learning rate value 0.00001 by reducing using ReduceLROnPlateau scheduler after each 3 epochs without loss decrease for 20 epochs. For the augmentations we used flipping, rotating, random cropping, blurring and changing the brightness.

4. Experiment results

The proposed method went through a different evaluations using the OULU-NPU and CASIA-MFSD datasets, with performance metrics including ACER, APCER and BPCER. The experimental results are presented in Tables 1, 2, and 3, showing the model's performance across various protocols and cross-dataset scenarios.

Table 1
Comparison of state-of-the-are and our methods on OULU-NPU dataset

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	STDN [28]	0.8	1.3	1.1
	CDCN [3]	0.4	1.7	1.0
	DC-CDN [30]	0.5	0.3	0.4
	NAS-FAS [31]	0.4	**0**	**0.2**
	Ours	**0.2**	0.6	0.4
2	STDN [28]	2.3	1.6	1.9
	CDCN [3]	1.5	1.4	1.5
	DC-CDN [30]	**0.7**	1.9	1.3
	NAS-FAS [31]	1.5	**0.8**	**1.2**
	Ours	1.0	2.0	1.5
3	STDN [28]	**1.6±1.6**	4.0±5.4	2.8±3.3
	CDCN [3]	2.4±1.3	2.2±2.0	2.3±1.4
	DC-CDN [30]	2.2±2.8	1.6±2.1	1.9±1.1
	NAS-FAS [31]	2.1±1.3	**1.4±1.1**	1.7±0.6
	Ours	2.0±1.6	1.8±1.4	**1.9±0.8**
4	STDN [28]	**2.3±3.6**	5.2±5.4	3.8±4.2
	CDCN [3]	4.6±4.6	9.2±8.0	6.9±2.9
	DC-CDN [30]	5.4±3.3	2.5±4.2	4.0±3.1
	NAS-FAS [31]	4.2±5.3	**1.7±2.6**	**2.9±2.8**
	Ours	3.0±2.6	4.4±4.8	3.7±2.3

Table 1 delineates the comparative analysis of the proposed method against state-of-the-art approaches on the OULU-NPU dataset across four protocols. In Protocol 1, which evaluates generalization across unseen environmental conditions, the proposed method achieved an APCER of 0.2%, outperforming all baseline methods. The BPCER and ACER values of 0.6% and 0.4% respectively, demonstrate competitive performance, with only NAS-FAS achieving a marginally lower ACER. Protocol 2, assessing generalization to unseen attack types, revealed the proposed method's APCER of 1.0%, ranking second after DC-CDN. While the BPCER of 2.0% was higher than some baselines, the overall ACER of 1.5% remained competitive, matching CDCN and surpassing STDN. For Protocol 3, which examines generalization to unseen input sensors, the proposed method exhibited consistent performance with an APCER of 2.0±1.6%, BPCER of 1.8±1.4%, and ACER of 1.9±0.8%. The results indicate robust generalization capabilities, with the method ranking second in ACER, marginally behind NAS-FAS. Protocol 4, combining all previous challenges, demonstrated the method's resilience in the most demanding scenario. An APCER of 3.0±2.6%, BPCER of 4.4±4.8%, and ACER of 3.7±2.3% were achieved, positioning the proposed approach competitively among top-performing methods.

Table 2
Testing the method on the CASIA-MFSD dataset. The benchmark requires the equal error rates (EER) so that both APCER and BPCER were equal at some threshold

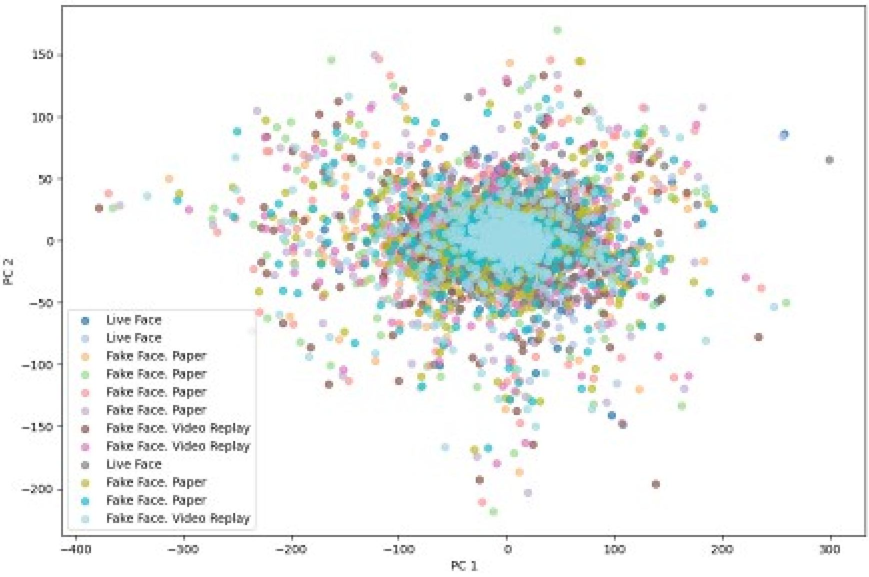
Method	APCER (%)	BPCER (%)	ACER (%)
Fisher Vector [32]	2.80	2.80	2.80
Patch&DepthFusion [33]	2.67	2.67	2.67
3D Synthesis [34]	2.22	2.22	2.22
Ours	**1.68**	**1.68**	**1.68**

Table 2 presents the performance on the CASIA-MFSD dataset. The proposed method achieved an APCER of 1.68%, BPCER of 1.68%, and ACER of 1.68%, indicating high performance in detecting presentation attacks while maintaining a low false rejection rate for genuine presentations.

Table 3
Cross dataset testing of the Proposed method

Trained on	Tested on	APCER (%)	BPCER (%)	ACER (%)
OULU-NPU	CASIA-MFSD	22.5	8.8	15.7
CASIA-MFSD	OULU-NPU	12.3	2.5	7.4

Table 3 shows the cross-dataset generalization capabilities of the proposed method. When trained on OULU-NPU and tested on CASIA-MFSD, the model achieved an APCER of 22.5%, BPCER of 8.8%, and ACER of 15.7%. Conversely, training on CASIA-MFSD and testing on OULU-NPU yielded an APCER of 12.3%, BPCER of 2.5%, and ACER of 7.4%. The disparity in performance between the two cross-dataset scenarios highlights the challenges associated with domain shift and the varying complexity of presentation attacks across datasets.



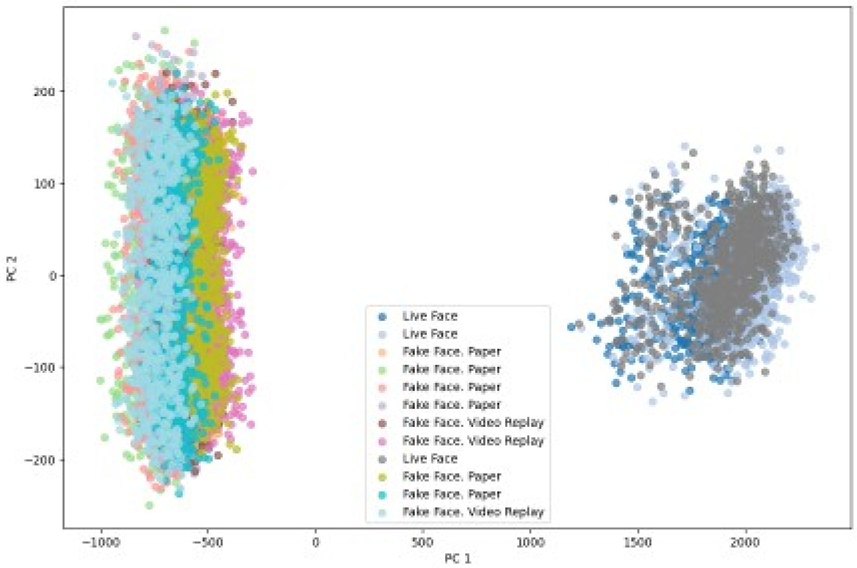


Figure 4: The graph on top shows the PCA of ViT output, while below is the output after ConvNeXT and Swin Transformer blocks fusion is illustrated. The samples' taken from CASIA-MFSD dataset.

Figure 4 presents a visual representation of the feature spaces generated by different components of the proposed architecture. The top graph illustrates the Principal Component Analysis (PCA) of the ViT output, while the bottom graph depicts the PCA after fusion of ConvNeXT and Swin Transformer blocks. Both visualizations use the same samples from the CASIA-MFSD dataset. The ViT output on top graph provides a relatively clustered distribution of features, with considerable overlap between live faces and various types of spoofing attacks. The feature space lacks clear separation between classes, indicating that the ViT alone struggles to distinguish between genuine and fake presentations consistently. In contrast, the fused output of ConvNeXT and Swin Transformer blocks on the bottom graph demonstrates a markedly improved feature distribution. The live face samples form a distinct cluster, which are shown in blue and gray, well-separated from the spoofing attack samples. The fake face presentations, including paper-based and video replay attacks, are grouped together but distinctly apart from the live face cluster.

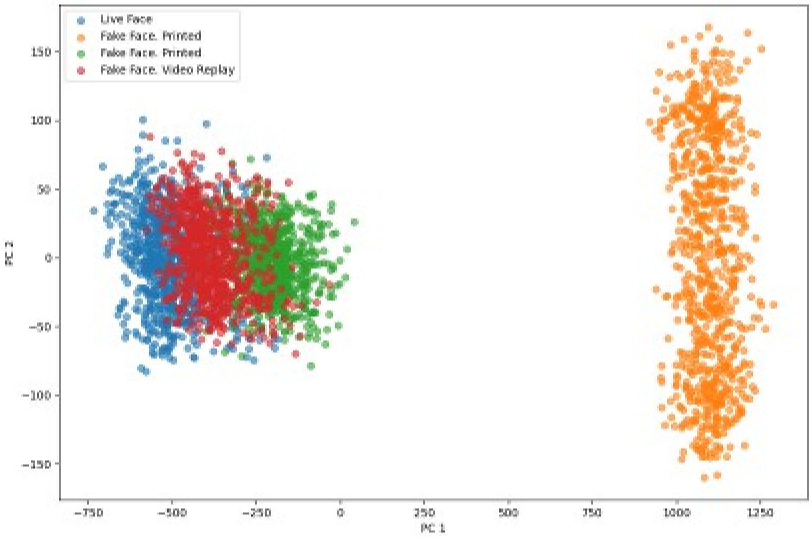


Figure 5: PCA of images on cross-dataset test, showing the domain-shift problem. The model is trained on CASIA-MFSD and tested on OULU-NPU.

Figure 5 illustrates the challenge of domain shift in cross-dataset scenarios. The PCA visualization shows the feature distribution when the model, trained on CASIA-MFSD, is tested on OULU-NPU. The plot reveals four distinct clusters corresponding to live faces and different types of spoofing attacks. Notably, the printed fake faces (orange cluster) are well-separated from other categories, suggesting robust detection of this attack type across datasets. However, the live faces (blue), video replay attacks (red), and another type of printed fake (green) show some overlap, indicating potential challenges in distinguishing these categories in cross-dataset scenarios. The significant shift in feature distribution between the training and testing datasets is evident from the distinct grouping of samples on the right side of the plot.

5. Discussion

The experimental results and feature space visualizations provide valuable insights into the performance and characteristics of the proposed face anti-spoofing method. The fusion of ViT, ConvNeXT, and Swin Transformer demonstrates promising capabilities in addressing the challenges of face presentation attack detection across various scenarios.

On the OULU-NPU dataset, the proposed method shows competitive performance across all four protocols. Notably, in Protocol 1, which evaluates generalization to unseen environmental conditions, the method achieves the lowest APCER among all compared approaches, suggesting that the fusion of global and local feature extraction mechanisms effectively mitigates the impact of varying illumination and background conditions. The method's performance in Protocols 2-4 remains competitive, indicating robustness to unseen attack types and input sensors. The high performance on the CASIA-MFSD dataset further confirms the method's effectiveness in handling diverse spoofing techniques and image qualities. The lowest APCER on this dataset is particularly noteworthy, showing a strong ability to detect all presentation attacks without false acceptances.

The PCA visualizations in Figure 4 provide crucial insights into the feature learning process. The ViT output alone shows limited separation between live faces and spoofing attacks, indicating that global context modeling is insufficient for robust anti-spoofing. However, the fused output of ConvNeXT and Swin Transformer blocks demonstrates a marked improvement in class separability, indicating the importance of combining global contextual information with local textural features and multi-scale analysis for effective spoofing detection. The clear separation between live faces and various attack types in the fused feature space aligns with the strong performance metrics observed on individual datasets. It suggests that the proposed architecture successfully learns discriminative features that generalize well across different spoofing techniques and environmental conditions within a single dataset. The cross-dataset experiments reveal both strengths and limitations of the proposed method. When trained on OULU-NPU and tested on CASIA-MFSD, the model achieves an ACER, which, while not optimal, indicates some degree of generalization. On the other hand, training on CASIA-MFSD and testing on OULU-NPU yields a better ACER, suggesting that the features learned from CASIA-MFSD may be more generalizable. Figure 5 visualizes the domain shift problem inherent [35] in cross-dataset scenarios. The distinct clustering of samples from the test dataset separate from the training dataset distribution highlights the challenge of domain adaptation in face anti-spoofing. The clear separation of printed fake faces in this scenario is encouraging, indicating that certain attack types may be more consistently detectable across domains.

While the proposed method demonstrates strong performance within individual datasets, the cross-dataset results reveal room for improvement in domain generalization. The disparity in cross-dataset performance depending on the training set suggests that the model's generalization capabilities are influenced by the diversity and characteristics of the training data. Future work should focus on addressing the domain shift problem, potentially through techniques such as adversarial domain adaptation or meta-learning approaches [36].

6. Conclusion

In summary, the proposed face anti-spoofing method, which combines Vision Transformer, ConvNeXT, and Swin Transformer, demonstrates high results in detecting presentation attacks across diverse scenarios. Experimental evaluations on OULU-NPU and CASIA-MFSD datasets reveal competitive performance, particularly in generalizing to unseen environmental conditions and attack types. Feature space analysis through PCA visualizations shows the importance of fusing global and local feature representations. The clear separation between genuine and spoofed samples in the fused feature space correlates with the strong performance metrics observed on individual datasets. The proposed method offers a balance between capturing fine-grained textures and modeling long-range dependencies, crucial for robust spoofing detection.

However, cross-dataset experiments expose challenges in domain generalization, with varying performance when transferring between OULU-NPU and CASIA-MFSD. Future work should focus on improving cross-dataset generalization through advanced domain adaptation techniques or meta-learning approaches. Additionally, investigating the individual contributions of each architectural component may lead to further optimizations in the fusion strategy. While the proposed method exhibits strong performance within individual datasets, addressing domain shift remains a critical challenge for real-world deployment of face anti-spoofing systems.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Galbally, J., Marcel, S., & Fierrez, J. (2014). Biometric anti-spoofing methods: A survey in face recognition. IEEE Access, 2, 1530-1552.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

[3] Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., ... & Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[4] Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., ... & Li, S. Z. (2020). A dataset and benchmark for large-scale multi-modal face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[5] Wang, G., Han, H., Shan, S., & Chen, X. (2020). Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[6] Shao, R., Lan, X., Li, J., & Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[7] Määttä, J., Hadid, A., & Pietikäinen, M. (2011). Face spoofing detection from single images using micro-texture analysis. In 2011 International Joint Conference on Biometrics (IJCB).

[8] Pinto, A., Pedrini, H., Schwartz, W. R., & Rocha, A. (2015). Face spoofing detection through visual codebooks of spectral temporal cubes. IEEE Transactions on Image Processing, 24(12), 4726-4740.

[9] Komulainen, J., Hadid, A., & Pietikäinen, M. (2013). Face spoofing detection using dynamic texture. In Asian Conference on Computer Vision.

[10] Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., ... & Wen, F. (2019). Face anti-spoofing: Model matters, so does data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[11] Liu, Y., Jourabloo, A., & Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[12] Li, L., Xia, Z., Jiang, X., Ma, Y., Roli, F., & Feng, X. (2019). 3D Face Mask Presentation Attack Detection Based on Intrinsic Image Analysis. IET Biometrics, 8(5), 342-351.

[13] Xu, Z., Li, S., & Deng, W. (2015). Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR).

[14] Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., ... & Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[15] Yu, Z., Li, X., Niu, X., Shi, J., & Zhao, G. (2020). Face anti-spoofing with human material perception. In European Conference on Computer Vision.

[16] Jia, Y., Zhang, J., Shan, S., & Chen, X. (2020). Single-side domain generalization for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[17] Shao, R., Lan, X., & Yuen, P. C. (2019). Regularized fine-grained meta face anti-spoofing. In Proceedings of the AAAI Conference on Artificial Intelligence.

[18] Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., ... & Zhao, G. (2021). Searching central difference convolutional networks for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[19] Liu, S., Lan, X., & Yuen, P. C. (2021). Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection. In European Conference on Computer Vision.

[20] Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., ... & Li, S. Z. (2019). CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(2), 182-193.

[21] George, A., & Marcel, S. (2021). Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Transactions on Information Forensics and Security, 16, 361-375.

[22] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[23] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[24] ISO/IEC 30107-3:2023. (2023). Information technology — Biometric presentation attack detection — Part 3: Testing and reporting (Edition 2). International Organization for Standardization.

[25] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017).

[26] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., & Li, S. Z. (2012). A face antispoofing database with diverse attacks. In 2012 5th IAPR International Conference on Biometrics (ICB).

[27] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *arXiv*. https://doi.org/10.48550/arXiv.1604.02878.tt.

[28] Liu, Y., Stehouwer, J., & Liu, X. (2020). On disentangling spoof trace for generic face anti-spoofing. In Proceedings of the European Conference on Computer Vision (pp. 406–422).

[29] Yu, Z., et al. (2020). Searching central difference convolutional networks for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5295–5305).

[30] Yu, Z., Qin, Y., Zhao, H., Li, X., & Zhao, G. (2021). Dual-cross central difference network for face anti-spoofing. In Proceedings of the International Joint Conference on Artificial Intelligence (pp. 1281–1287).

[31] Yu, Z., Wan, J., Qin, Y., Li, X., Li, S. Z., & Zhao, G. (2021). NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(9), 3005–3023.

[32] Boulkenafet, Z., Komulainen, J., & Hadid, A. (2017). Face anti-spoofing using speeded-up robust features and Fisher vector encoding. IEEE Signal Processing Letters, 24(2), 141-145. https://doi.org/10.1109/LSP.2017.2654306.

[33] Atoum, Y., Liu, Y., Jourabloo, A., & Liu, X. (2017). Face anti-spoofing using patch and depth-based CNNs. In 2017 IEEE International Joint Conference on Biometrics (IJCB) (pp. 319-328). IEEE. https://doi.org/10.1109/BTAS.2017.8272724.

[34] Guo, J., Zhu, X., Xiao, J., Lei, Z., Wan, G., & Li, S. (2019). Improving Face Anti-Spoofing by 3D Virtual Synthesis. In 2019 International Conference on Biometrics (ICB) (pp. 1-8). https://doi.org/10.1109/ICB45273.2019.8987385.

[35] Sun, Y., Liu, Y., Liu, X., Li, Y., & Chu, W.-S. (2023). Rethinking domain generalization for face anti-spoofing: Separability and alignment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Available at arXiv:2303.13662 [cs.CV]. https://doi.org/10.48550/arXiv.2303.13662.

[36] Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., & Kot, A. C. (2022). Learning Meta Pattern for Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security, 17, 1201-1213. https://doi.org/10.1109/TIFS.2022.3158551.