# Hierarchical transformer based learning for robust face anti-spoofing

Zhanseri Ikram[1,†], Bauyrzhan Omarov[1,†] and Batyrkhan Omarov[2,†]

[1] *Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2] *International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan*

**Abstract**

Face anti-spoofing is a critical challenge in biometric authentication systems, requiring robust methods to effectively distinguish between genuine and fraudulent attempts. The current study presents a Hierarchical Transformer-Based Learning (HTBL) framework designed to tackle challenges across diverse environmental conditions and attack modalities. The architecture combines a Vision Transformer encoder for global context capture with a Swin Transformer for local feature refinement, supported by intermediate convolutional layers. The evaluations on the OULU-NPU dataset validate the HTBL framework across standardized protocols assessing generalization to new environments, attack instruments, and sensor inputs. The method achieves state-of-the-art performance, particularly in complex generalization scenarios. Feature visualization using Principal Component Analysis supports the quantitative results, illustrating the refinement of discriminative capabilities throughout the network stages. The HTBL framework demonstrates strong generalization across varied conditions, addressing a significant limitation in current face anti-spoofing systems. Additionally, its balanced performance across error rate metrics indicates practical applicability, positioning the HTBL framework as a promising advancement in face anti-spoofing technology with important implications for biometric authentication security in real-world scenarios.

**Keywords**

Face Anti Spoofing, face liveness, machine learning, computer vision

## 1. Introduction

Face anti-spoofing systems have emerged as a critical component in biometric authentication, addressing the escalating threat of presentation attacks in facial recognition technologies. The proliferation of sophisticated spoofing techniques, including printed photographs, digital displays, and 3D masks, has necessitated the development of robust countermeasures to safeguard the integrity of facial authentication systems. In recent years, deep learning approaches have demonstrated remarkable efficacy in discerning genuine faces from fraudulent presentations, surpassing traditional handcrafted feature-based methods [1].

Among the many of deep learning architectures, Convolutional Neural Networks (CNNs) have been predominantly employed for face anti-spoofing tasks, applying their capacity to extract spatial features from facial images [2]. However, CNNs exhibit limitations in capturing long-range dependencies and hierarchical relationships within facial structures, which are crucial for distinguishing subtle spoofing artifacts. To address these shortcomings, attention mechanisms and transformer architectures have been introduced, revolutionizing various computer vision tasks, including face anti-spoofing [3]. The transformer architecture, initially proposed for natural language processing tasks, has demonstrated exceptional performance in modeling sequential data and capturing global contextual information [4]. The self-attention mechanism inherent in transformers enables the model to weigh the importance of different facial regions dynamically, potentially enhancing the detection of spoofing cues across diverse attack types. Nevertheless, the application of transformers to face anti-spoofing presents unique challenges, particularly in terms of

*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); b.omarov@iitu.edu.kz (B. Omarov)
0009-0001-8059-6590 (Z. Ikram); 0000-0002-9312-4429 (B. Omarov); 0000-0002-8341-7113 (B. Omarov)

computational efficiency and the need for hierarchical feature representation to capture both fine-grained textures and global facial structures.

The present study introduces a novel HTBL framework for robust face anti-spoofing. The proposed approach uses a hierarchical transformer architecture to capture multi-scale facial features and long-range dependencies, facilitating the detection of sophisticated presentation attacks. By integrating a hierarchical structure, the model efficiently processes facial images at different resolutions, enabling the extraction of both local and global spoofing cues. Furthermore, the HTBL framework uses a robust training pipeline and data augmentation techniques to improve generalization across diverse spoofing scenarios and environmental conditions.

The remainder of this paper is organized as follows: Section 2 provides an overview of related works in face anti-spoofing and transformer-based approaches. Section 3 delineates the proposed HTBL framework, elucidating its architectural components, training methodology, dataset information and experimental setup. Section 4 presents the experimental results, followed by an in-depth analysis and discussion in Section 5. Finally, last section concludes the study and outlines future research directions.

## 2. Related works

Face anti-spoofing has been an active area of research in biometric security, with numerous approaches proposed to combat evolving presentation attack techniques. The literature in this domain can be broadly categorized into traditional handcrafted feature-based methods and deep learning approaches.

Early face anti-spoofing techniques relied on handcrafted features to distinguish between genuine and spoofed faces. Texture analysis played a pivotal role in these approaches, with Local Binary Patterns (LBP) emerging as a popular descriptor for capturing micro-texture variations [5]. Extensions of LBP, such as LBP-TOP for spatio-temporal analysis, were proposed to use motion cues in video-based anti-spoofing [6]. Additionally, color space analysis and image quality assessment metrics were explored to detect artifacts introduced by printing or display devices [7]. While handcrafted features demonstrated efficacy in controlled environments, their performance degraded significantly under varying illumination conditions and against high-quality spoofing attacks. Moreover, the manual design of features limited their adaptability to novel attack types, necessitating the exploration of more sophisticated approaches. The advent of deep learning ushered in a new era of face anti-spoofing research, with Convolutional Neural Networks (CNNs) at the forefront. CNNs exhibited remarkable performance in learning discriminative features directly from facial images, obviating the need for manual feature engineering [8]. Various CNN architectures, including AlexNet, VGGNet, and ResNet, were adapted for face anti-spoofing tasks, demonstrating superior performance compared to traditional methods [9]. To improve the temporal modeling capabilities of CNNs, researchers incorporated recurrent architectures such as Long Short-Term Memory (LSTM) networks for video-based anti-spoofing [10]. These hybrid CNN-LSTM models captured both spatial and temporal cues, improving robustness against video replay attacks. Despite their success, CNN-based approaches faced challenges in capturing long-range dependencies and hierarchical relationships within facial structures. To address these limitations, attention mechanisms were introduced to focus on salient facial regions and potential spoofing artifacts [11, 12, 13]. While ViT-based models demonstrated promising results, they faced challenges in capturing fine-grained facial textures crucial for spoofing detection. To address these limitations, hybrid CNN-transformer architectures were proposed, combining the strengths of both paradigms [14]. These models applied CNNs for low-level feature extraction and transformers for high-level semantic modeling. However, the computational complexity of full-image transformer processing remained a significant challenge, particularly for real-time anti-spoofing applications. Recent advancements in efficient transformer designs, such as the Swin Transformer [15], have paved the way for more effective hierarchical processing of visual data. Domain shift, arising from variations in camera devices, lighting

conditions, and presentation attack instruments, often leads to performance degradation when models are deployed in real-world scenarios [16, 17, 18].

## 3. Materials and methods

Current section describes the methodological architecture applied in the development of an advanced face anti-spoofing system for biometric authentication. The proposed approach is visually represented in Figure 1, which illustrates a structured flowchart encompassing each stage from input image processing to the final classification decision.

### 3.1. Problem statement

The goal of the research is to develop robust mechanisms that can accurately differentiate between genuine face presentations and spoofing attempts. Effective face anti-spoofing architectures are crucial to improve the security of facial recognition systems, ensuring that only genuine users are authenticated while preventing unauthorized access by impostors using fake representations.
Given an input face image $I \in R^{H \times W \times C}$, where $H$ denote height, $W$ width, and $C$ the number of channels, respectively, the objective is to learn a function $f : R^{H \times W \times C} \rightarrow \{0, 1\}$ such that:

$$f(I) = \begin{cases} 1, \textit{if } I \textit{ is a genuine face presentation} \\ 0, \textit{if } I \textit{ is a spoofing attempt} \end{cases} \quad (1)$$

The function $f$ must effectively map the high-dimensional input space of facial images to a binary decision space, distinguishing between authentic biometric samples and fraudulent presentations.

### 3.2. Proposed method

The proposed HTBL framework in Figure 1 addresses the face anti-spoofing problem through a novel architecture combining hierarchical feature extraction, transformer-based processing, and multi-scale fusion. The method contains several key components.

The input image $I$ with sizes $224 \times 224 \times 3$ is divided into a grid of $N = 784$ non-overlapping patches, each of size $P \times P = 8 \times 8$. These patches are flattened and linearly projected to obtain a sequence of patch embeddings.

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_N^T \end{bmatrix}, X \in R^{N \times D} \quad (2)$$

where $D = 768$ is the embedding dimension.

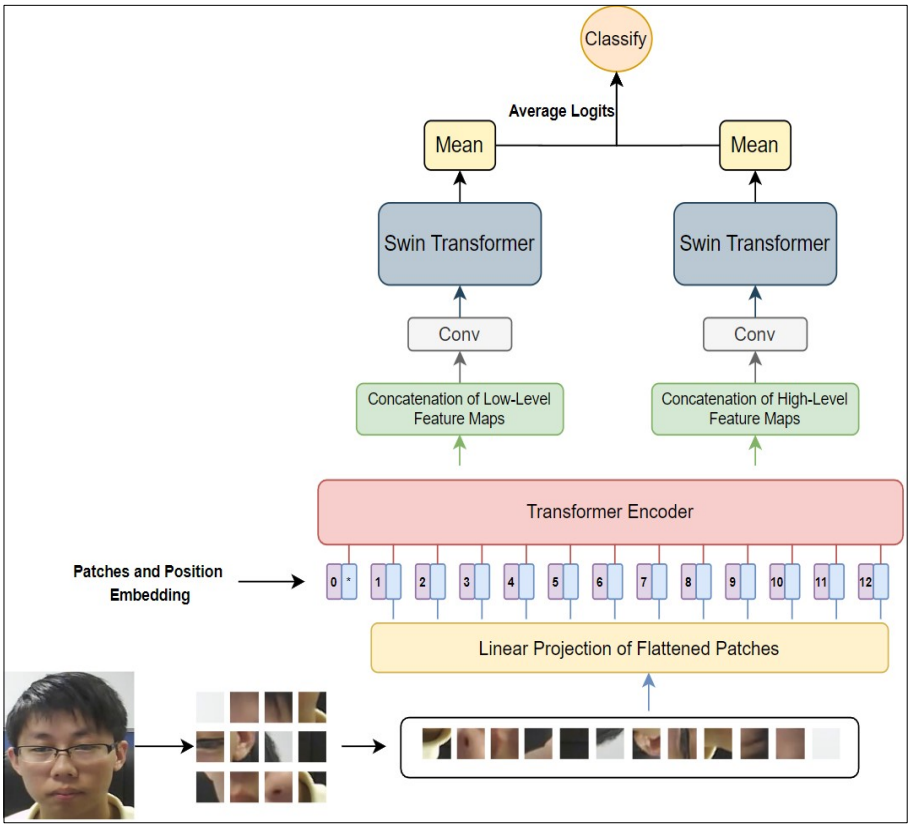


**Figure 1:** The proposed model architecture.

Learnable position embeddings $E_{pos} \in R^{N \times D}$ are added to integrate spatial information

$$X^{'} = X + E_{pos} \tag{3}$$

The embedded sequence $X'$ is processed by a transformer encoder consisting of $L = 12$ layers. Each layer uses multi-head self-attention (MSA) and feed-forward network (FFN) modules:
The transformer encoder outputs are used to construct feature maps. Low Level features block $(F_{low})$ is taken by concatenating layers [2,3,4,5] and High Level features block $(F_{high})$ is taken from layers [6,7,8,9], each block of size $batch \times 784 \times 3072$ after concatenation.

$$S_{low} = SwinTransformer(Conv((F_{low}))) \tag{4}$$

$$S_{high} = SwinTransformer(Conv((F_{high}))) \tag{5}$$

Firstly, those blocks are passed through *Conv2d* layer. After two separate Swin Transformer modules process the low-level and high-level feature maps, capturing multi-scale contextual information. Swin Transformer uses shifted windows for efficient self-attention computation and hierarchical feature learning. $S_{low}$ and $S_{high}$ output $batch \times 784 \times 768$ feature maps. Mean function is applied for each map and results are averaged to further feed into the *Sigmoid* function.

### 3.3. Metrics

APCER, BPCER, and ACER are metrics used in biometric system performance evaluation, especially in systems involving fingerprint recognition, facial recognition, or other biometric authentication methods [19]. This research also applied aforementioned metrics to reflect the model results.

$$APCER = \frac{Number\ of\ False\ Accepts}{Total\ Number\ of\ Attack\ Presentations} \tag{6}$$

$$BPCER = \frac{Number\ of\ False\ Rejects}{Total\ Number\ of\ Genuine\ Presentations} \tag{7}$$

$$ACER = \frac{APCER + BPCER}{2} \tag{8}$$

### 3.4. Dataset

The OULU-NPU dataset [20] is a widely recognized benchmark in face anti-spoofing research, designed to address the challenges of generalization across different environmental conditions and attack types. The dataset consists of 4950 video recordings of genuine face presentations and presentation attacks, captured using six different mobile devices with front-facing cameras. The dataset includes 55 subjects, with recordings conducted in three distinct sessions with varying illumination and background conditions. The presentation attacks in OULU-NPU encompass two types: print attacks and video-replay attacks. Print attacks use high-quality printed photographs of the subjects, while video-replay attacks employ high-resolution digital videos displayed on electronic screens. OULU-NPU is structured around four protocols, each designed to evaluate specific aspects of face anti-spoofing systems. Each video in the dataset is 5 seconds long, recorded at 30 frames per second, resulting in 150 frames per video. However, the current research uses only few of those frames during the inference, mainly $10^{th}$, $20^{th}$, $30^{th}$, $40^{th}$, $50^{th}$, $60^{th}$, $70^{th}$, $80^{th}$, $90^{th}$,$100^{th}$ and averages their results to get the final decision probability.

### 3.5. Experimental setup

In this study, we conducted our experiments using an NVIDIA RTX 3090 GPU with 24GB of VRAM, which provided the necessary computational power for efficient training and inference. The training was performed using a batch size of 8, a configuration chosen to balance between memory usage and training efficiency. For optimization, we employed the AdamW optimizer. The model was trained for a total of 40 epochs, which was determined to be sufficient for convergence based on preliminary experiments. The learning rate was initialized at 0.00001 and was adjusted during training using a CosineAnnealingLR scheduler. The scheduler is capable to gradually reduce the learning rate, thereby facilitating smooth convergence and helping to avoid local minima. To improve generalization Horizontal Flip, Random Contrast, Random Gamma, Random Brightness, and Distortion based geometric augmentations were applied.

## 4. Experiment results

The proposed method was evaluated on the OULU-NPU dataset using its four standard protocols, which assess different aspects of face anti-spoofing generalization. Table 1 presents a comparison of our approach against four other methods: Auxiliary [21], Disentangle [22], DC-CDN [23], and NAS-FAS [24].

**Table 1**
Comparison of the proposed model with other deep learning methods

| Protocol | Method | APCER(%) | BPCER(%) | ACER(%) |
|---|---|---|---|---|
| 1 | Auxiliary [21] | 1.6 | 1.6 | 1.6 |
| | Disentangle [22] | 1.7 | 0.8 | 1.3 |
| | DC-CDN [23] | 0.5 | 0.3 | 0.4 |
| | NAS-FAS [24] | 0.4 | **0** | **0.2** |
| | **Ours** | **0** | 1.6 | 0.8 |
| 2 | Auxiliary [21] | 2.7 | 2.7 | 2.7 |
| | Disentangle [22] | 1.1 | 3.6 | 2.4 |
| | DC-CDN [23] | **0.7** | 1.9 | 1.3 |
| | NAS-FAS [24] | 1.5 | **0.8** | **1.2** |
| | **Ours** | 0.8 | 2.0 | 1.4 |
| 3 | Auxiliary [21] | 2.7±1.3 | 3.1±1.7 | 2.9±1.5 |
| | Disentangle [22] | 2.8±2.2 | 1.7±2.6 | 2.2±2.2 |
| | DC-CDN [23] | 2.2±2.8 | 1.6±2.1 | 1.9±1.1 |
| | NAS-FAS [24] | 2.1±1.3 | **1.4±1.1** | 1.7±0.6 |
| | **Ours** | **1.4±1.2** | 1.6±1.0 | **1.5±0.6** |
| 4 | Auxiliary [21] | 9.3±5.6 | 10.4±6.0 | 9.5±6.0 |
| | Disentangle [22] | 5.4±2.9 | 3.3±6.0 | 4.4±3.0 |
| | DC-CDN [23] | 5.4±3.3 | 2.5±4.2 | 4.0±3.1 |
| | NAS-FAS [24] | 4.2±5.3 | **1.7±2.6** | **2.9±2.8** |
| | **Ours** | **2.8±2.4** | 3.4±4.8 | 3.1±2.3 |

In Protocol 1, which evaluates generalization across unseen environmental conditions, our method demonstrates high performance with an ACER of 0.8%, outperforming the next best method NAS-FAS by a significant margin. Notably, our approach achieves a perfect APCER of 0%, indicating excellent capability in detecting presentation attacks, albeit with a slightly higher BPCER compared to some competitors. For Protocol 2, which tests generalization across unseen attack devices, our method shows competitive performance with an ACER of 1.4%. While not achieving the lowest error rates, it maintains a balanced performance across both APCER and BPCER, suggesting robust generalization capabilities. Protocol 3 assesses generalization across unseen input sensors, presenting a more challenging scenario reflected in the higher error rates across all methods. Our approach achieves the lowest ACER of 1.5±0.6%, demonstrating significant cross-sensor generalization compared to other methods. The results demonstrate particular strengths in handling unseen environmental conditions (Protocol 1) and the most challenging combined scenario (Protocol 4).

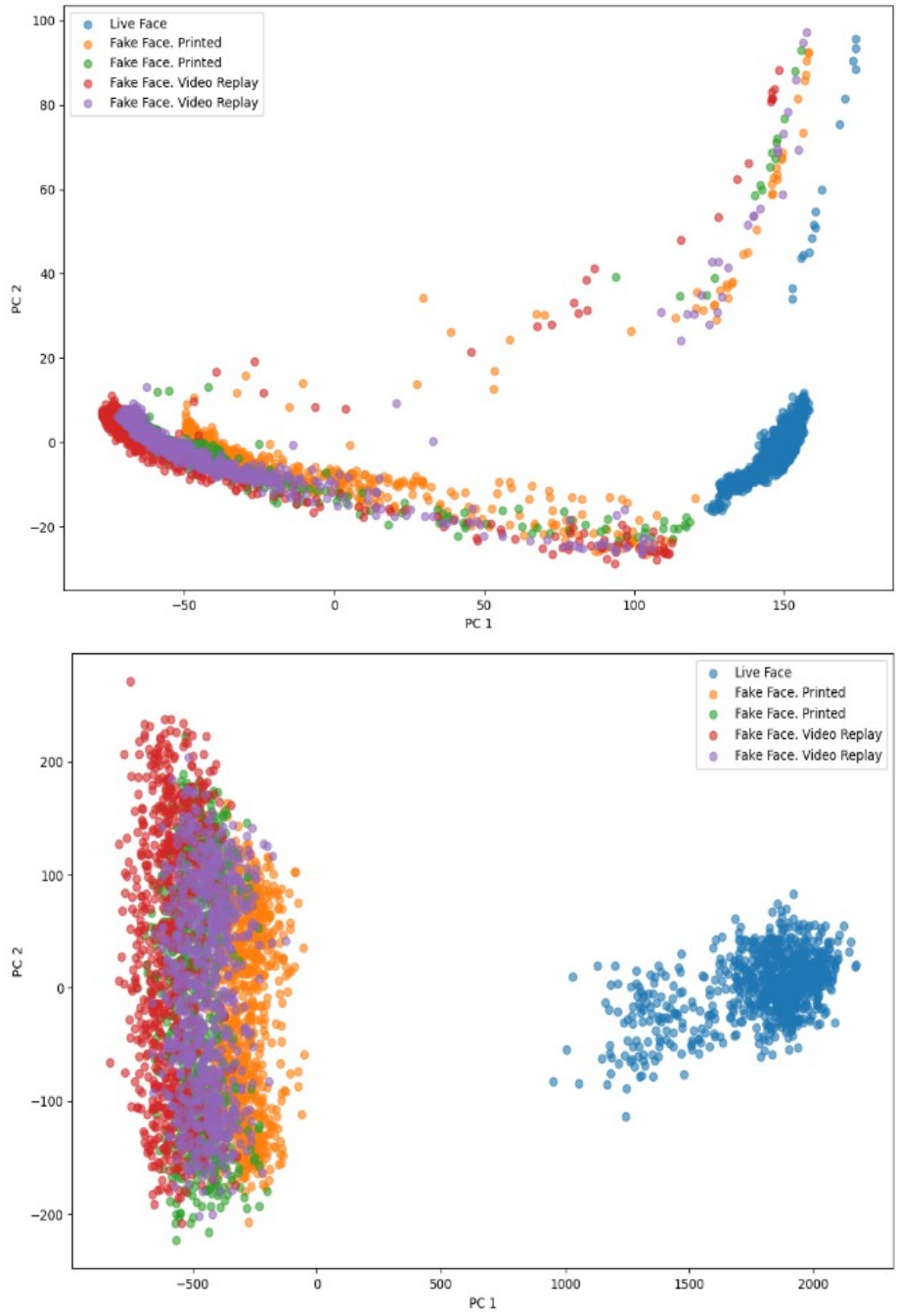


**Figure 2:** PCA of images processed only with ViT. Before applying Conv and Swin Transformer on the left and PCA of images processed through ViT > Conv > Swin Transformer on the right.

The efficiency of our hierarchical transformer-based approach for face anti-spoofing is further evidenced through principal component analysis (PCA) visualizations of the feature representations at two critical stages of the network. Figure 2 illustrates the PCA projection of features immediately after the ViT encoder. The plot reveals a curved manifold structure, with live faces (blue) distinctly separated from various types of fake faces. However, there is notable overlap among different spoofing attack types (printed, video replay). This suggests that while the ViT encoder successfully distinguishes genuine from spoofed presentations, it struggles to differentiate between specific attack modalities. Figure 2 presents the PCA visualization after processing through the convolutional block and Swin Transformer. The transformation in feature space is striking. Live faces are now tightly clustered and distinctly separated from all spoofing attacks. Moreover, there is improved discrimination between different types of fake faces, although some overlap persists.

## 5. Discussion

The findings from our investigation into the HTBL framework substantiate its effectiveness in tackling the complex challenges of face anti-spoofing. Analyzing these results in the context of the broader landscape of anti-spoofing research reveals several critical insights and potential areas for future exploration. Our method demonstrated strong performance across diverse protocols, particularly in unseen environments and combined challenge scenarios, highlighting its exceptional generalization abilities. The success of our model in generalizing across different scenarios can be largely attributed to the combined strengths of global context capture by the ViT encoder and local feature refinement by the Swin Transformer. Previous approaches often struggled to simultaneously capture both global and local spoofing cues under varying environmental conditions [25], which was addressed by our architectural synergy approach. Its ability to maintain high performance across different environmental conditions, attack types, and sensor inputs aligns with the industry's growing demand for adaptive and resilient security solutions [26].

In summary, the HTBL framework represents a significant leap forward in face anti-spoofing technology, effectively applying the strengths of transformer architectures and multi-scale feature learning to overcome key challenges in generalization and robustness. While the results are highly encouraging, they also point to new avenues for further research and refinement.

## 6. Conclusion

The HTBL framework presented in our study shows a significant progress in face anti-spoofing technology. By synergistically combining Vision Transformer encoder with Swin Transformers, our approach effectively addresses the complex challenges of distinguishing genuine face presentations from sophisticated spoofing attempts across diverse environmental conditions and attack modalities. The framework's ability to generalize across such varied conditions underscores its potential for real-world deployment in biometric authentication systems. The progressive refinement of feature representations, as visualized through Principal Component Analysis, provides insight into the hierarchical learning process. The clear separation between live and fake face representations in the final stages of our network architecture corroborates the quantitative performance metrics and highlights the effectiveness of our multi-scale feature extraction approach. Future research directions include the integration of temporal information for video-based anti-spoofing and exploration of cross-dataset generalization.

## 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] Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., ... & Zhao, G. (2021). 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).

[2] 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 (pp. 5042-5051).

[3] Liu, S. I., Lan, H. J., & Yeh, P. C. (2022). Dual-stream transformer for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19218-19227).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
[5] 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) (pp. 1-7). IEEE.
[6] de Freitas Pereira, T., Anjos, A., De Martino, J. M., & Marcel, S. (2014). LBP-TOP based countermeasure against face spoofing attacks. In Asian Conference on Computer Vision (pp. 121-132). Springer, Cham.
[7] Galbally, J., Marcel, S., & Fierrez, J. (2014). Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. IEEE transactions on image processing, 23(2), 710-724.
[8] Yang, J., Lei, Z., & Li, S. Z. (2014). Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601.
[9] Nagpal, C., & Dubey, S. R. (2019). A performance evaluation of convolutional neural networks for face anti spoofing. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
[10] 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) (pp. 141-145). IEEE.
[11] 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.
[12] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. ACM Computing Surveys (CSUR), 54(10), 1-41.
[13] Yu, Z., Qin, Y., Li, X., Wang, Z., Zhao, C., Lei, Z., & Zhao, G. (2021). Multi-modal face anti-spoofing based on central difference networks and dual-cross pattern attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6488-6497).
[14] Liu, S. I., Yeh, P. C., Fu, X., & Wu, H. T. (2022). Transformer-based multi-scale feature fusion for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19228-19237).
[15] 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 (pp. 10012-10022).
[16] 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 (pp. 10023-10031).
[17] 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 (pp. 8484-8493).
[18] Shao, R., Lan, X., & Yuen, P. C. (2019). Regularized fine-grained meta face anti-spoofing. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 4804-4811).
[19] ISO/IEC 30107-3:2023. (2023). Information technology — Biometric presentation attack detection — Part 3: Testing and reporting (Edition 2). International Organization for Standardization.
[20] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 612-618). https://doi.org/10.1109/FG.2017.77.
[21] 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 (CVPR) (pp. 389-398).

[22] Zhang, K.-Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., Huang, F., Song, H., & Ma, L. (2020). Face anti-spoofing via disentangled representation learning. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 1-6).

[23] 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 (IJCAI) (pp. 1281-1287). https://doi.org/10.24963/ijcai.2021/178.

[24] 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. https://doi.org/10.1109/TPAMI.2020.3009123.

[25] George, A., & Marcel, S. (2021). On the effectiveness of vision transformers for zero-shot face anti-spoofing. arXiv:2011.08019v2 [cs.CV]. https://doi.org/10.48550/arXiv.2011.08019.

[26] Huang, H.-P., Sun, D., Liu, Y., Chu, W.-S., Xiao, T., Yuan, J., Adam, H., & Yang, M.-H. (2023). Adaptive transformers for robust few-shot cross-domain face anti-spoofing. arXiv:2203.12175v2 [cs.CV]. https://doi.org/10.48550/arXiv.2203.12175.