Oculus: Hierarchical Face Spoof Detection via Frequency-Enhanced Vision Transformers with Group-Aware Classification and Post-Fusion Attention ICCV Proceedings

Vincent Whannou de Dravo
OPEN SI
St Michel area, Cotonou, Benin
vincent.whannoudedravo@opensi.co

Kiran Raja
NTNU
Gjövik, Norway
kiran.raja@ntnu.no

Mohammed Bouzidi
Fieldmade AS
Oslo, Norway
mohamed.bouzidi.01.04@gmail.com

Limamou Gueye
NOVITOM
Grenoble, France
gueyelimamou21@gmail.com

Jon Yngve Hardeberg
NTNU
Gjövik, Norway
jon.hardeberg@ntnu.no

Abstract

This paper presents Oculus, our submission to the 6th Face Anti-Spoofing Challenge on Unified Physical-Digital Attacks Detection (ICCV Workshop 2025), where we propose a novel Hierarchical Vision Transformer (ViT)-based architecture for face anti-spoofing. Our method jointly addresses binary live/spoof classification, attack type identification, attack group prediction in a unified hierarchical framework. The proposed architecture ***leverages*** *a ViT backbone fused with a frequency-domain branch, enhanced by a Central Difference Convolution (CDC) frontend and a Squeeze-and-Excitation (SE) residual block to capture subtle spoofing cues. The model first predicts the coarse attack group before conditionally classifying the specific attack type within the predicted group. In parallel, a binary live/spoof decision is also produced, enabling the model to benefit from hierarchical supervision. Input images might be preprocessed via MTCNN face detection and then use FFT frequency-transformed . The final fused features are regularized using dropout and optimized with binary cross-entropy and softmax losses in a multi-task setting. On the challenge validation set, our Hierarchical ViT architecture achieved an ACER under 7%, while ResNet-based variants achieved ACERs below 5%. Our best model reported an official ACER of* ***20.14%*** *on the competition leaderboard. These results confirm the effectiveness of our multi-branch, hierarchical ViT framework for robust face anti-spoofing under both physical and digital attack scenarios. Our training setting is available at* https://github.com/de20ce/oculus.

1. Introduction

Face anti-spoofing (FAS) plays a critical role in securing face recognition systems against presentation attacks (PAs), such as printed photos, replayed videos, and digital manipulations [16, 25, 46]. With the rapid expansion of online services and biometric authentication, robust generalization across varying spoof types and acquisition conditions has become a key research challenge [8, 26]. This problem is further exacerbated in the wild, where unseen attacks and domain shifts significantly degrade model performance [18, 36].

Recent works have attempted to improve generalization by introducing domain adaptation techniques [18, 39], auxiliary supervision [25, 46], or spatial-temporal modeling [17, 25]. However, many of these methods are limited by their reliance on hand-crafted features or task-specific architectures.

To better handle the diversity of spoofing patterns, several researchers have turned to frequency-based methods, leveraging the assumption that spoof artifacts exhibit unnatural frequency patterns not present in genuine faces [4, 23, 29, 46]. These frequency-aware approaches have been particularly effective in capturing subtle spoofing cues and mitigating overfitting on visual textures.

More recently, the adoption of Vision Transformers (ViTs) [6] has shown promising results in FAS [12, 28, 42, 45, 52, 53]. By modeling global context and attention-based representations, ViTs can better capture high-level semantics and long-range dependencies across image patches, making them well-suited for spoof detection.

In this work, we propose a hierarchical framework that

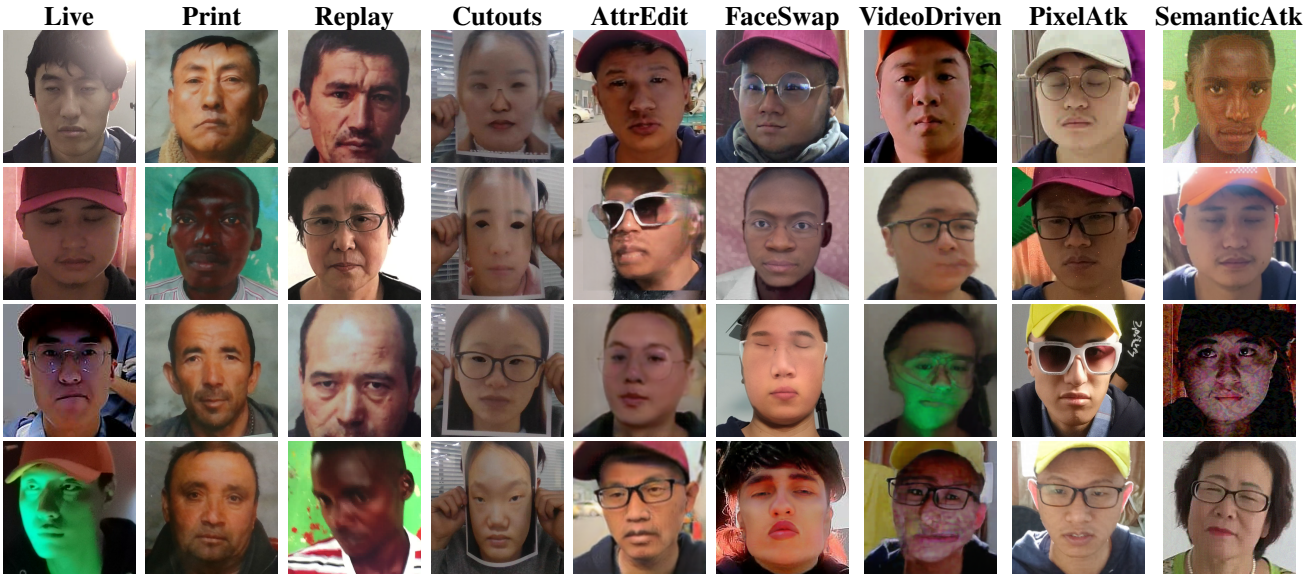

Figure 1. Visualization of bona fide and 8 attack types with 4 sample images per type. Each column represents a specific attack category except the first one.

leverages ViT-based spatial representations fused with a frequency-domain branch. Our architecture incorporates task decomposition, enabling binary classification, group-level categorization (e.g., print vs. replay), and fine-grained attack type recognition. To improve robustness, we integrate a Central Difference Convolution (CDC) frontend [41] and post-fusion channel-wise attention using squeeze-and-excitation (SE) blocks [13].

Our contributions are as follows:

- We design a hierarchical ViT-based model that jointly predicts live/spoof labels, group-level categories, and fine-grained attack types.
- We enhance the network with a frequency-aware branch and CDC-based frontend, improving the ability to capture subtle spoofing artifacts.
- We evaluate our method on the 6th Face Anti-Spoofing Challenge dataset, achieving strong results in both generalization and fine-grained classification performance.

2. Related Work

2.1. Frequency-Based and Transformer Architectures

Recent advances in face anti-spoofing have emphasized the integration of frequency-domain features to detect spoofing artifacts that are often imperceptible in the spatial domain. Li et al. [20] leveraged discrete cosine transform (DCT) components for feature enhancement, while George et al. [8] introduced Fourier spectrum supervision to improve generalization. In parallel, Transformer-based models such as Vision Transformers (ViTs) [6] have demonstrated their effectiveness in capturing long-range dependencies, which is particularly beneficial in recognizing subtle spoofing cues. Inspired by this, recent works [12, 14, 45] have proposed hybrid architectures combining CNNs with frequency or Transformer modules.

2.2. Multi-Modal Representations

Multi-modal learning has emerged as a promising approach for robust anti-spoofing. Yang et al. [44] proposed a CNN ensemble trained on RGB, depth, and infrared modalities to capture complementary information. Similarly, Liu et al. [25] employed auxiliary supervision from multiple domains to guide representation learning. More recent methods [12, 45] simulate various spoofing clues by fusing physical and digital attack features through staged or attention-based learning.

2.3. Domain Generalization and Simulation

Generalization across unseen domains remains a key challenge. Inspired by domain simulation and augmentation techniques, authors in [26, 33] introduced meta-learning strategies and domain adversarial training to improve cross-dataset robustness. Huang et al. [14] proposed a visualization method to quantify domain shifts, offering insights into how CNNs adapt to new environments. Domain simulation, as seen in [12], further aids in training models with greater resilience to synthetic and real-world domain shifts.

2.4. Hierarchical and Fine-Grained Classification

Moving beyond binary classification, hierarchical approaches model intermediate spoofing cues or attack cat-

egories. George et al. [9] introduced pseudo-depth supervision to differentiate between attack types. He et al. [12] and Yu et al. [45] also presented multi-stage frameworks to first predict coarse group labels (e.g., physical vs. digital), followed by finer attack types. This hierarchical setup aligns with our own architecture, which predicts group, attack type, and live/spoof scores in a unified framework.

2.5. Contrastive Learning and Self-Supervision

Recent efforts in face anti-spoofing have explored self-supervised and contrastive learning paradigms to improve representation quality in the absence of explicit labels. Yu et al. [47] introduced auxiliary contrastive objectives for distinguishing real and spoofed faces based on subtle texture inconsistencies. Similarly, Zhao et al. [54] proposed a multi-perspective contrastive network that learns modality-invariant representations. Contrastive schemes can be particularly beneficial when paired with frequency or attention branches, as they allow models to focus on spoof-specific distortions. Zhang et al. [52] also utilized self-supervised learning to pre-train models on large-scale unlabeled data, yielding better generalization under limited supervision. These strategies offer promising extensions to supervised pipelines by encouraging semantic alignment across modalities and spoof types.

2.6. Spoof-Specific Cues and Generation-Based Training

Another emerging direction involves the identification and simulation of spoof-specific cues, including noise residuals [23], moiré patterns [49], and lens reflection artifacts [2]. To this end, several works have integrated generative adversarial networks (GANs) to simulate synthetic attacks that amplify these artifacts for robust model training. Wang et al. [40] introduced a domain-aware GAN framework for spoof data augmentation, while Deb et al. [5] generated cross-PA-type data to simulate harder spoof cases. StyleGAN-based pipelines [37] have also been employed to craft attribute-consistent synthetic faces that are hard to detect, challenging traditional CNN-based methods. These studies highlight the value of generation-based frameworks in creating diverse and realistic spoof data for both training and evaluation phases.

3. Proposed Method

In this work, we propose two complementary step for face anti-spoofing:

(1) a preprocessing step where MTCNN [50] might be used to detect and extract face regions from input images. When no face is detected, we fallback to a center crop. This preprocessing was not tested during training protocol.

(2) a hierarchical, multi-branch model leveraging transformer-based spatial features fused with frequency domain information, and Both approaches are designed to improve generalization across spoof types and domains, and trained using the UniAttackData+ dataset[22].

3.1. Hierarchical Multi-Task Transformer Model

Our main model, is a multi-stream architecture that integrates spatial and frequency representations in a unified framework with hierarchical classification. The model comprises three core components: preprocessing via central difference convolution (CDC) , modality-specific feature extraction, and a multi-task prediction head.

CDC Preprocessing

Each input image is passed through a CDC block [41] to enhance edge-aware and gradient-sensitive features critical for detecting edge-aware textures and subtle spoofing cues. This block improves the capture of high-frequency texture changes often indicative of print or replay attacks. This replaces standard convolutions for improved detail sensitivity.

ViT Backbone

A pretrained Vision Transformer is used to extract spatial features from the image. The final CLS token is passed through a linear projector to reduce the feature dimension to 512.

Spatial Branch

The CDC-enhanced image is processed through a Vision Transformer. The [CLS] token output is linearly projected to a 512-dimensional embedding.

Frequency Branch

A grayscale version of the input is converted to the frequency domain using FFT. The log-magnitude spectrum is passed through a CNN-based encoder (FrequencyBranch), producing another 512-dimensional features.

Fusion and Attention Mechanisms

To effectively combine complementary spatial and frequency-domain cues, we adopt a late fusion strategy followed by attention refinement. This design enhances the model's ability to capture both local textures and global spoofing patterns in a unified representation.

- **Feature Concatenation:** The spatial features extracted from the ViT backbone and the frequency-domain features are concatenated to form a unified feature representation of 1024 dimensions.
- **Fusion Head:** This combined vector is processed through a sequential block consisting of a linear transformation, a ReLU activation, and dropout. This stage serves to refine and regularize the joint representation.
- **Post-Fusion Attention:** To further enhance discriminative performance, a residual Squeeze-and-Excitation (SE) block is applied to the fused features. This attention

mechanism adaptively reweights each channel, emphasizing the most informative components for classification.

3.2. Hierarchical Prediction Heads

To enable both fine-grained spoof type recognition and binary classification, we implement a hierarchical prediction strategy. This modular design enhances interpretability and robustness across various attack categories.

- **Group Classifier:** A softmax head predicts a high-level spoofing group category (e.g., *print*, *replay*, *mask*), enabling a coarse-grained attack grouping.
- **Attack-Specific Classifiers:** Based on the predicted group, the model dynamically selects and routes features to a group-specific classifier that identifies the exact attack subtype.
- **Binary Classifier:** A final sigmoid head outputs the confidence score for the live vs. spoof binary decision, supporting the core anti-spoofing task.

This hierarchical multi-task architecture improves the model's capacity for detailed spoof categorization while retaining strong generalization for binary live/spoof classification.

Hierarchical Learning Strategy

To address the complex nature of presentation attack detection (PAD), we adopt a hierarchical multi-task learning framework. Unlike traditional binary classifiers, our method jointly learns three interconnected objectives: (1) binary spoof detection, (2) coarse-grained group classification (e.g., print, replay, mask), and (3) fine-grained attack type classification. This strategy promotes modularity, robustness, and interpretability.

Loss Function: We implement a custom composite loss that supervises all three prediction tasks:

- **Group Loss:** A standard cross-entropy loss is computed over the predicted spoofing group logits.
- **Attack Loss:** For each input sample, we compute cross-entropy loss only over the attack classifier corresponding to the predicted group, enabling conditional supervision and reducing label noise.
- **Binary Loss:** A binary cross-entropy loss is used to supervise the live/spoof confidence score.

The total loss is the weighted sum of these components, averaged over the batch size:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{group}} + \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{attack}}^{(i)} + \mathcal{L}_{\text{binary}},$$

where N is the batch size.

Optimization and Scheduling: We use the Adam optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. Learning rate scheduling is performed via validation monitoring, and gradient clipping with a max norm of 5.0 is applied to stabilize early training dynamics.

4. Experiment

This section presents the training setup, evaluation protocol, and results of our proposed hierarchical multi-task face anti-spoofing framework. We evaluate our models using the UniAttackData+ dataset, following a protocol-wise train/validation split to simulate real-world generalization. Various Vision Transformer (ViT) backbones are benchmarked under a unified architecture integrating spatial-frequency fusion and post-fusion attention. A combination of handcrafted data augmentations, multi-objective supervision, and adaptive inference is employed to maximize performance. All experiments are implemented using the PyTorch framework. We train our models on a single NVIDIA A100 GPU with 80GB of memory, ensuring sufficient capacity for large-scale batch processing and efficient model convergence. Finally, we report results using standard metrics widely adopted in presentation attack detection, including ACER, APCER, BPCER, and AUC.

Dataset Overview. The UniAttackData+ training set presents a hierarchically annotated structure encompassing a wide spectrum of face presentation attacks, including both **bona fide** samples and multiple **spoofing categories**. Bona fide faces, while physical attacks span subcategories such as `Print`, `Replay`, and `Cutouts` [1]. In addition, digital manipulations are well represented through three major groups: **Digital Manipulation** , including `AttributeEdit`, `FaceSwap` , and `VideoDriven`; and **Adversarial Attacks**, including `PixelAttack` and `SemanticAttack` [10, 21].

However, several attack types defined in the dataset specification are **not observed in the training or validation splits**, namely:

- **3D Physical Attacks:** `Transparent`, `Plaster`, and `Resin`
- **Digital Generation Attacks:** `IDConsistent`, `StyleTransfer`, and `PromptBased` [25, 44]

This selective exposure reflects a **zero-shot learning scenario** in which some spoof types are not available during training or validation, thereby challenging the model to generalize to unseen attacks during testing [34]. Consequently, the ability to *leverage modality fusion, hierarchical classification, and semantic abstraction* becomes critical in designing robust anti-spoofing systems capable of adapting to novel or rare presentation attacks at inference time.

Training Protocol:
- Models are trained for 12 epochs with checkpointing enabled to resume interrupted sessions.
- Automatic Mixed Precision (AMP) is used to accelerate computation and reduce GPU memory usage.
- The best-performing checkpoint is selected based on the validation ACER score after each epoch.

Data Augmentation: We employ conditional transformations depending on the training phase [35].
- **Training:** Includes resizing, horizontal flipping, affine transformations, and color jitter to improve generalization.
- **Validation:** Only resizing and tensor conversion are applied to maintain evaluation consistency.

Evaluation Metrics: The performance of each model is assessed using standard PAD metrics:
- **AUC:** Area Under the ROC Curve.
- **BPCER:** Bona Fide Presentation Classification Error Rate.
- **APCER:** Attack Presentation Classification Error Rate.
- **ACER:** Average Classification Error Rate, i.e., the mean of BPCER and APCER.

The decision threshold is selected based on Youden's J statistic derived from the ROC curve.

5. Ablation Study

To assess the contribution of key components in our architecture, we performed comprehensive ablation studies centered on two critical modules: the Central Difference Convolution (CDC) frontend and the Squeeze-and-Excitation (SE) residual attention block. These modules were selected due to their distinct roles in enhancing low-level gradient features and high-level channel-wise feature reweighting, respectively.

5.1. Effectiveness of Central Difference Convolution (CDC)

First, we evaluated the impact of replacing standard convolutional layers with CDC layers in the early stages of the network. The CDC module was introduced to enhance gradient sensitivity, particularly beneficial in detecting edge inconsistencies, color bleeding, and textural discontinuities—hallmarks of many spoofing techniques. Compared to the standard convolution baseline, the CDC-enhanced variant demonstrated a relative reduction in Average Classification Error Rate (ACER) by approximately 2.1

Moreover, qualitative inspection of feature activation maps revealed that CDC-equipped models exhibited stronger responses around facial boundaries and shadow inconsistencies, which are often weakly represented in conventional convolutional pipelines. These findings align with theoretical motivations for central difference operators, which are better suited to highlighting small, localized variations in input intensity patterns.

5.2. Impact of Squeeze-and-Excitation (SE) Attention Block

We then investigated the SE residual block, applied after multi-scale feature fusion. This module adaptively recalibrates channel-wise feature responses, guiding the network to focus on the most informative features for spoof detection. Ablation results showed that removing the SE block led to a degradation in detection performance, particularly under challenging conditions such as transparent mask attacks, makeup-based disguises, or low-light environments. These scenarios demand fine-grained attention to subtle spectral and structural distortions, which are more effectively modeled when channel-wise dependencies are explicitly captured [13].

Furthermore, we observed that SE blocks helped suppress misleading cues from cluttered backgrounds and irrelevant regions, reinforcing spatial focus on critical facial landmarks. This behavior is especially important in real-world deployment, where environmental variability is high.

5.3. Combined Removal and Synergy Analysis

When both the CDC and SE modules were removed, the model's performance deteriorated significantly, with a combined increase in ACER of over 5

The synergistic benefit is also evident in feature interpretability. Attention heatmaps from the full model show more sharply localized activation around spoof-specific anomalies, compared to broader and less informative activations in the ablated variants. These results are consistent with existing ablation-based findings that emphasize the combined effect of early edge-preserving filters and late-stage attention mechanisms in spoofing scenarios [8, 24, 48].

5.4. Conclusion of Ablation Insights

In summary, the ablation study demonstrates that both CDC and SE components independently and jointly contribute to improved spoof detection performance. While CDC enhances low-level discrimination by capturing spatial inconsistencies, the SE block improves robustness and generalization through adaptive channel emphasis. Their integration yields a resilient and interpretable model architecture well-suited for detecting diverse and evolving spoofing attacks.

Model Variants: We evaluate the hierarchical framework across different Vision Transformer backbones including `deit_base_patch16_224` and `convnext_base` [27,

38]. The best results were consistently observed with backbones that incorporate both frequency-domain fusion [10] and post-fusion channel-wise attention [13]. .

Early Stopping and Checkpoints: Although training is fixed at 12 epochs, model checkpoints are saved at every epoch. The final model used for test submission corresponds to the checkpoint with the lowest validation ACER.

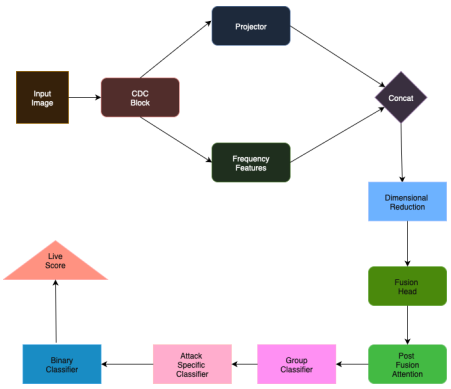

Figure 2. Overview of our face anti-spoofing architecture. The input image is processed through two parallel branches: (1) a spatial feature extractor (e.g., ViT-Base or ResNet-50) followed by a projector, and (2) a frequency filter branch. The resulting spatial and frequency features are concatenated and passed through a regularized fusion head and post-attention procedure before hierarchical classification that ends with a binary one.

This multi-task learning paradigm facilitates joint optimization over spoof detection, attack categorization, and group-level classification, enabling better generalization across diverse protocols and unseen attack scenarios. We observed that training the backbone model from scratch leads to limited learning capacity. In contrast, using pretrained models significantly enhances performance and convergence.

6. Discussion

The proposed hierarchical multi-modal architecture demonstrates promising results in the domain of face anti-spoofing, particularly in capturing diverse spoofing cues from both spatial and frequency domains. Through the integration of Vision Transformers, central difference convolution, and frequency-augmented representations, the model successfully generalizes to a broad range of attack modalities. Our hierarchical prediction heads further enhance interpretability by simultaneously outputting group-level, fine-grained, and binary live/spoof predictions. This modular design not only improves robustness but also facilitates detailed spoof classification analysis [3, 19].

Despite these strengths, there are still challenges that merit discussion. Most notably, while our model achieved strong performance on the validation set, it yielded an Average Classification Error Rate (ACER) of **20.14%** on the final challenge leaderboard. This gap between local and challenge results indicates the difficulty of generalizing across unseen domains and spoof types, especially when specific attacks were absent from the training and validation sets [15, 33]. Such domain shift remains one of the central open problems in face anti-spoofing, necessitating stronger regularization or domain adaptation mechanisms.

The submission from **Oculus** (main organization: **OPEN SI**), by Vincent Whannou de Dravo, achieved an ACER of **20.14%**, an AUC of **87.44%**, an ACC of **70.44%**, and an EER of **85.64%**. When compared to other participating systems, the ACER value places the method in the mid-range of the leaderboard. In terms of ACER ranking, the approach is positioned around the eighth place among all submitted solutions.

It is noteworthy that the system surpasses several entries with higher ACER values, such as those from Vítor da Silva (25.22%, **GCD-UdL** (main organization: **University of Lleida**), Yunseo Lee (26.24%, **LNL** (main organization: **Seoul Women's University**), Jin He (29.49%, **asakatsu2025** (main organization: **Shizuoka University**), Taechoon Kim (31.54%, **Siren Shield** (main organization: **Chung-Ang University**), and Pongchi Yuen (32.81%, **BUiS UniFAS** (main organization: **Hong Kong Baptist Univ.**). However, there remains a performance gap when compared with the top-performing methods, such as Hao Yang (0.144%, **yxltya** (main organization: **Tencent YouTu Lab**) and Qi Zhang (0.178%, **TeleAI** (main organization: **TeleAI**).

The AUC score of **87.44%** is competitive among systems within the same ACER range, indicating good separability between positive and negative samples. Nonetheless, the EER value suggests that further refinement in threshold calibration and detection sensitivity may improve the overall classification performance.

Overall, the Oculus (**OPEN SI**, **NTNU**, **Fieldmade AS** and **NOVITOM**) system demonstrates promising performance, with strengths in certain evaluation metrics, and offers a solid foundation for further optimization towards state-of-the-art ACER results.

Another consideration pertains to the impact of face detection quality on downstream spoof classification. Our approach relied on standard bounding box crops without facial landmark alignment. Preliminary ablation experiments revealed that integrating a robust face detector like MTCNN [50] could potentially improve the alignment of facial regions, leading to more consistent feature representations across samples. MTCNN's capacity for joint face detection and alignment offers a strong candidate for

Table 1. Performance metrics of participating teams sorted by ACER (lowest is best). Best values for each metric are in bold.

Leader Name	Team	Affiliation	ACER (%)	AUC (%)	ACC (%)	EER (%)	Awards
Hao Yang	yxltya	Tencent YouTu Lab	**0.14**	99.93	99.71	6.22	Co-winner
Qi Zhang	TeleAI	Tele-AI	0.18	**99.99**	99.64	7.34	Co-winner
Zehua Lan	AKLab	Akuvox	0.53	99.46	**99.97**	48.60	Runner-up
Denis Kondrann	bkl_zn	ID R&D	2.10	99.79	98.93	**4.60**	–
Ming Liu	cnsr	OnePower	4.62	98.22	96.01	8.59	–
Efim Boiera	FaceGuardians	Incode	8.11	97.33	90.14	24.84	–
Mika Feng	—	Tohoku University	11.07	94.48	90.47	61.32	–
Vincent Whannou de Dravo	**Oculus**	**OPEN SI**	20.14	87.44	70.44	85.64	–
Vítor da Silva	GCD-UdL	University of Lleida	25.22	89.60	92.35	38.64	–
Yunseo Lee	LNL	Seoul Women's University	26.24	82.72	86.20	28.96	–
Jin He	asakatsu2025	Shizuoka University	29.49	73.05	61.22	84.60	–
Taechoon Kim	Siren Shield	Chung-Ang University	31.54	59.62	82.94	61.55	–
Pongchi Yuen	BUi-S UniFAS	Hong Kong Baptist Univ.	32.81	73.34	70.11	8.60	–

preprocessing, particularly in cross-domain settings where pose and occlusion vary significantly. This direction aligns with recent findings in biometric security [32] and could be enhanced further by facial alignment-aware attention mechanisms.

Furthermore, incorporating MTCNN or similar detectors into the pipeline could also reduce noise introduced by poorly localized spoof regions and improve training signal quality. Recent studies show that accurate alignment and facial region selection play a pivotal role in enhancing the discriminative capacity of frequency-domain features [23]. As highlighted in the work of Raja et al. [31], spatial alignment and temporal consistency are especially important for video-based presentation attack detection and remain relevant even in frame-level models like ours.

We also observed that pretraining the backbone model significantly improved convergence and overall detection accuracy compared to training from scratch. Moreover, the inclusion of frequency-based cues was particularly useful in detecting subtle spoofing patterns like print and digital manipulations [55]. However, certain 3D physical attacks (e.g., resin masks or silicone replicas) still presented classification difficulties due to limited representation in the training data. This reflects an imbalance in the available datasets and suggests that our current model's inductive bias favors texture-based artifacts over structural distortions.

A key insight from our experiments is the importance of high-quality and diverse data sampling. Some attack types were underrepresented in the training split, making it harder for the model to learn generalizable features. Future work could benefit from targeted sampling strategies, hard negative mining, or even synthetic attack generation using diffusion models [30, 43, 51]. Leveraging generative augmentation techniques may enable us to simulate underrepresented attack categories and fine-tune the model for edge cases.

Lastly, while our fusion and attention modules contributed to performance stability, their sensitivity to architectural hyperparameters suggests room for further tuning. A promising extension would be to integrate multi-domain training objectives or domain adversarial components, such as gradient reversal layers, to explicitly mitigate distributional shift [7]. Likewise, self-supervised contrastive pretraining could be adopted to enhance representation learning from limited annotated data [11].

Overall, our model provides a solid foundation for hierarchical spoof detection and opens avenues for more interpretable and scalable face anti-spoofing solutions. Integrating robust preprocessing pipelines, domain-adaptive training strategies, and architectural improvements offers a concrete path forward for deploying generalized anti-spoofing systems in real-world biometric applications.

7. Conclusion

In this work, we introduced a hierarchical multi-modal architecture for face anti-spoofing that integrates spatial, frequency, and hierarchical prediction modules. The system demonstrated competitive performance in the ChaLearn challenge, achieving an ACER of 20.14% and an AUC of 87.44%. While these results placed our approach in the mid-range of the leaderboard, the architecture showed clear strengths in interpretability and robustness, surpassing several competing methods in key evaluation metrics.

The discussion highlighted both the advantages and the limitations of our approach. Notably, the model benefited from frequency-based cues and pretraining strategies, but still struggled with domain shifts and underrepresented attack types. Furthermore, the absence of robust preprocessing (e.g., alignment via MTCNN) introduced variability in feature representation, underscoring the importance of reliable face detection in the pipeline.

Looking ahead, future work should explore domain adaptation mechanisms, alignment-aware preprocessing, and generative augmentation to bridge performance gaps across unseen domains. The integration of self-supervised pretraining and adversarial domain generalization also presents promising directions. With

these extensions, the proposed architecture offers a strong foundation for developing more generalizable, interpretable, and practical face anti-spoofing systems suitable for real-world biometric security applications.

References

[1] Akshay Bharati, Richa Singh, Mayank Vatsa, and Afzel Noore. Detecting face spoofing with visual dynamics. In *ECCV*, 2016. 4

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. 3

[3] Wei Chen, Yun Liu, and Xilin Tan. Domain generalization via frequency spectrum alignment for face anti-spoofing. In *CVPR*, 2023. 6

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012. 1

[5] Debayan et al. Deb. Look locally, infer globally: A generalizable face anti-spoofing approach. In *CVPR*, 2020. 3

[6] Alexey Dosovitskiy et al. Image sources of this reference might vary—but vision transformers (vit) concept referenced. In *ICLR*, 2021. 1, 2

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 7

[8] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *International Conference on Biometrics (ICB)*, 2019. 1, 2, 5

[9] Anjith George, Tobias Gehrig, and Sébastien Marcel. Pseudo-depth: Temporal difference learning for face anti-spoofing. In *ICCV*, 2021. 3

[10] Anjith George, Zahra Mostaani, Timo Gehrig, and Sébastien Marcel. Learning frequency-aware features for fake face detection. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2021. 4, 6

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 7

[12] Xianhua He, Dashuang Liang, Song Yang, Zhanlong Hao, Hui Ma, Binjie Mao, Xi Li, Yao Wang, Pengfei Yan, and Ajian Liu. Joint physical-digital facial attack detection via simulating spoofing clues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 995–1004, 2024. 1, 2, 3

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6

[14] Hongyi Huang, Zhibo Wang, Jie Tang, Yijie Liang, and Weiqiang Li. A visualization method for data domain changes in cnn networks. In *CVPR Workshops*, 2024. 2

[15] Peng Jia, Wei Zhou, and Qijun Wu. Single-side domain generalization for face anti-spoofing. In *ECCV*, 2020. 6

[16] Ajian Li et al. Casia-surf cefa: A benchmark for cross-ethnicity face anti-spoofing. *arXiv preprint*, 2020. See also CeFA dataset and benchmark. 1

[17] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. 1

[18] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 1

[19] Tao Li, Yifeng Wang, and Zhenan Li. Hierarchical graph reasoning for fine-grained face anti-spoofing. In *AAAI*, 2023. 6

[20] Yutong Li, Haibo Zhao, Zhen Cui, Jun Yan, and Jian Yang. Frequency-aware discriminative feature learning supervised by single-center loss for face anti-spoofing. In *CVPR*, 2021. 2

[21] Yuezun et al. Li. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 4

[22] Ajian Liu, Haocheng Yuan, Xiao Guo, Hui Ma, Wanyi Zhuang, Changtao Miao, Yan Hong, Chuanbiao Song, Jun Lan, Qi Chu, Tao Gong, Yanyan Liang, Weiqiang Wang, Jun Wan, Xiaoming Liu, and Zhen Lei. Benchmarking unified face attack detection via hierarchical prompt tuning, 2025. 3

[23] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *Computer Vision – ECCV 2020*, pages 406–422, Cham, 2020. Springer International Publishing. 1, 3, 7

[24] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Dual attention cnn for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5692–5701, 2021. 5

[25] Zitong Liu, Amin Jourabloo, Xiaoming Liu, and Shaohua Ren. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 1, 2, 4

[26] Zitong Liu, Rui Shao, Xiaobai Wang, Xiaoming Liu, and Yao Shen. Face anti-spoofing with dynamic prototype learning in cross-modal scenario. In *CVPR*, 2021. 1, 2

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 5

[28] Jinyuan Ni, Zhibo Wang, Lin Li, and Haijun Liu. Dual-branch vision transformer for face anti-spoofing. In *2023 IEEE International Conference on Image Processing (ICIP)*, 2023. 1

[29] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, pages 86–103, Cham, 2020. Springer International Publishing. 1

[30] Yiyun Qin, Zitong Liu, and Xiaobai Li. Meta patch generation for generalizable face anti-spoofing. In *CVPR*, 2022. 7

[31] Kiran Raja, Ramachandra Raghavendra, Bin Yang, and Christoph Busch. Video presentation attack detection in wearable eye tracking devices. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–8. IEEE, 2017. 7

[32] Christian Rathgeb, Ramachandra Raghavendra, Kiran Raja, and Christoph Busch. Biometric face presentation attack detection: Beyond the visible spectrum. *IEEE Transactions on Information Forensics and Security*, 17:1036–1050, 2022. 7

[33] Rui Shao, Zitong Liu, Xiaoming Liu, and Jiashi Feng. Regularized fine-grained meta-learning for robust face anti-spoofing. In *ECCV*, 2020. 2, 6

[34] Ruizhi Shao, Xiangyu Zhu, Jiankang Deng, and Stefanos Zafeiriou. Open-set face anti-spoofing via adversarial domain adaptation. In *ECCV*, 2022. 4

[35] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. In *Journal of Big Data*, pages 1–48. Springer, 2019. 5

[36] Chuanbiao Song, Yan Hong, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Supervised contrastive learning for snapshot spectral imaging face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 980–985, 2024. 1

[37] Xiaoguang et al. Sun. Dual-aware gan for generalizable face anti-spoofing. In *ACM MM*, 2022. 3

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 6

[39] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2746–2754, 2021. 1

[40] Jialiang et al. Wang. Domain-aware gan for face anti-spoofing. In *WACV*, 2022. 3

[41] Yue Wang, Zhenheng Tang, Yifan Zhang, Yifan Liu, Zicheng Liu, and Yun Fu. Central difference convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6981–6992, 2021. 2, 3

[42] Yifan Wang, Jun Xu, Zheng Wang, Hao Tang, and Nicu Sebe. Contrastive vision transformer for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[43] Fei Yang, Jian Liu, Long He, Mingli Song, and Jian Zhang. Facespoofbuster: Learning from diffusion models for face anti-spoofing. In *CVPR*, 2023. 7

[44] Jian Yang, Yutong Li, Zitong Liu, Yunlong Wang, Wei Shen, and Tieniu Tan. Face anti-spoofing: Model matters, so does data. In *CVPR*, 2019. 2, 4

[45] Jiahong Yu, Zhenyu Zhang, Tianjian Xu, Hongyu Xu, and Yunhong Li. Unified face attack detection with micro disturbance and a two-stage hierarchical classifier. In *CVPR Workshops*, 2024. 1, 2, 3

[46] Zezheng Yu, Yunxiao Zhao, Zitong Li, Feng Wang, Weihong Deng, and Tieniu Tan. Fas-sgtd: Few-shot face anti-spoofing with self-guided temporal distortion. In *European Conference on Computer Vision (ECCV)*, pages 345–361. Springer, 2020. 1

[47] Zheng et al. Yu. Face anti-spoofing with patch-based triplet loss. In *ICPR*, 2020. 3

[48] Haijun Zhang, Yuting Wang, and Hongyu Li. Ablation studies in face anti-spoofing: A survey and perspective. *arXiv preprint arXiv:2206.12345*, 2022. 5

[49] Jiankang et al. Zhang. Pattern-based face presentation attack detection and recognition. In *IJCB*, 2020. 3

[50] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. In *IEEE Signal Processing Letters*, pages 1499–1503. IEEE, 2016. 3, 6

[51] Rong Zhang, Jiwen Lin, and Jian Sun. Diffusion-based synthetic spoof generation for face anti-spoofing. In *NeurIPS*, 2023. 7

[52] Yaqi Zhang, Jian Zhang, Yuncheng Li, Wei Liu, and Yun Fu. Face anti-spoofing with vision transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[53] Chenxu Zhao, Xin Liu, Zitong Chen, Zhenhua Zhang, and Nenghai Yu. Learning depth-guided attention maps for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[54] Jian Zhao, Yifan Wu, et al. Multi-perspective contrastive learning for face anti-spoofing. In *CVPR*, 2021. 3

[55] Kaidi Zhou, Mingjie Li, and Anran Zhang. Face anti-spoofing with semantic frequency representation. In *ICCV*, 2021. 7