UniAttackData+: Unified Physical-Digital Attack Detection+ Challenge

Hui Ma[1*], Ajian Liu[2*], Yongze Li[1], Chuanbiao Song[3], Xiao Guo[4], Changtao Miao[5], Wanyi Zhuang[5], Junze Zheng[1], Shunxin Chen[6], Yan Hong[3], Jiabao Guo[7], Jiankang Deng[8], Jun Lan[3], Weiqiang Wang[3], Tao Gong[5], Qi Chu[5], Sergio Escalera[9], Hugo Jair Escalante[10], Xiaoming Liu[4], Zhen Lei[2], Isabelle Guyon[11], Yanyan Liang[1†], Jun Wan[2 †]
[1]M.U.S.T., China; [2]CASIA, China; [3]Ant Group, China; [4]MSU, USA;
[5]USTC, China; [6]NJUPT, China; [7]HFUT, China; [8]ICL, UK; [9]UB, Spain;
[10]INAOE, Mexico; [11]UPSaclay, France;
[1]{3220006153,2009853gii30008,3230006102}@student.must.edu.mo,
[1]yyliang@must.edu.mo, [2]{ajian.liu,zhen.lei,jun.wan}@ia.ac.cn

Abstract

Unified physical-digital face attack detection aims to develop a universal model capable of simultaneously detecting both digital and physical spoofing attacks. However, existing training datasets generally lack comprehensive coverage of diverse attack types, which limits the generalization ability of detection models and hinders their effectiveness in real-world applications. To address this limitation, we released a significantly expanded dataset, UniAttackData+, at the 6th Face Anti-Spoofing Workshop @ICCV 2025. The dataset includes 2,875 participants from three major demographic groups—Africa, East Asia, and Central Asia—and contains 18,250 real videos collected under diverse lighting conditions, backgrounds, and acquisition devices. For each participant, 54 types of attacks were simulated, comprising 14 physical and 40 digital attack variants, resulting in a total of 679,097 high-quality forged videos. Based on this dataset, we organized a Unified Attack Detection Challenge, which attracted 137 participating teams, with 12 teams advancing to the final round. The final rankings were determined based on results re-verified by the organizers. This paper reviews the challenge by introducing the dataset construction, protocol definitions, evaluation metrics, and competition results, and analyzes the top-performing algorithmic solutions, while also outlining future directions for unified physical-digital attack detection. Challenge Website: https://sites.google.com/view/face-anti-spoofing-challenge/welcome/challengeiccv2025?authuser=0

1. Introduction

Face Anti-Spoofing (FAS) is an important component to ensure the security and reliability of biometric authentication systems [19, 25]. The current mainstream attack methods can be divided into two categories: physical forgery attacks [7, 22, 39, 40] and digital forgery attacks [14, 15, 17]. Physical attacks mainly include printing attacks, replay attacks, and 3D mask attacks, where attackers forge faces through physical media to deceive the recognition system. Such attacks usually introduce visible artifacts, such as color distortion, moiré, and reflected light spots, which provide certain detection clues. In contrast, digital forgery attacks are more covert. Common forms include identity manipulation, adversarial samples, and synthetic images based on generative adversarial networks. They often silently tamper with facial images at the pixel level, making them difficult to detect by the human eye or traditional algorithms. Although there have been many advances in the research of physical attack detection and digital forgery detection, most current methods still treat the two as independent tasks and lack a unified modeling and evaluation framework. This fragmented research paradigm limits the generalization ability of the detection model and is not conducive to unified defense against diverse attacks in actual deployment.

Unified physical-digital face attack detection aims to develop a universal model capable of simultaneously detecting both digital and physical spoofing attacks [4, 6, 8, 11, 26, 37]. JFSFDB [37] proposes the first joint FAS and forgery detection benchmark, which combines visual appearance and rPPG physiological signals for authentication through a dual-branch physiological network and a weighted fusion strategy. UniAttackDetection [8] proposes a unified attack detection framework based on visual language models, which effectively learns unified and specific

*Equal contributions.
†Co-Corresponding Author.

Team	Affiliation
Facevengers	Tencent YouTu Lab & Shanghai Jiao Tong University
TeleAI	Tele-AI
AKLab	akuvox
bklzhn	ID R&D
CMSR	OnePower
FaceGuardians	incode
Tohoku Aoki Lab	Tohoku University
GCD-UdL	University of Lleida
LNL	Seoul Women's University
asakatsu2025	Shizuoka University
Siren Shield	Chung-Ang University
BU-S UniFAS	Hong Kong Baptist University

Table 1. List of qualified teams and affiliations for this challenge.

knowledge features through a teacher-student hint module, a unified knowledge mining module, and a sample-level hint interaction module. [11] significantly improves the model's detection ability for "unseen" attack types through data augmentation technology of simulated physical attack clues and digital attack clues. This method won the first place in the "Unified Physical-Digital Face Attack Detection" task of the 5th CVPR 2024 FAS Challenge, proving its effectiveness and generalization ability in cross-attack type detection. The MoAE-CR [4] framework effectively utilizes category information to improve attack detection performance by introducing a hybrid attack expert module at the feature level and a disaggregation module and a clustering distillation module at the loss level. Based on the idea of hierarchical cue adjustment, HiPTune [26] adaptively selects and integrates classification criteria from different semantic spaces by constructing a visual cue tree and a dynamic cue interaction module to improve the robustness of the model against various attacks.

However, existing training datasets generally lack comprehensive coverage of diverse attack types, which limits the generalization ability of detection models and hinders their effectiveness in real-world applications. To address this limitation, we released a significantly expanded dataset, UniAttackData+, at the 6th FAS Workshop @ICCV 2025. The dataset includes 2,875 participants from three major demographic groups—Africa, East Asia, and Central Asia—and contains 18,250 real videos collected under diverse lighting conditions, backgrounds, and acquisition devices. For each participant, 54 types of attacks were simulated, comprising 14 physical and 40 digital attack variants, resulting in a total of 679,097 high-quality forged videos. As shown in Tab.1, we organized a Unified Attack Detection Challenge, which attracted 137 participating teams, with 12 teams advancing to the final round.

The contributions of this paper are summarized as:

- We organized this challenge based on the UniAttackData+ dataset, demonstrating its effectiveness as a valuable resource for advancing research in Unified Physical-Digital Attack Detection.
- We conduct a systematic analysis of the technical solutions submitted by the participating teams, and, drawing from the insights gained through this competition, propose potential avenues for future research in the field of unified face attack detection.

2. Related Work

Face Anti-spoofing (FAS) datasets for Challenges. The **Print-Attack** [1] dataset is specifically constructed to investigate the vulnerability of 2D face recognition (FR) systems to printed photo attacks. It contains 400 video samples, including 200 genuine access attempts and 200 print attacks, involving 50 subjects. The dataset also includes a standardized evaluation protocol and an example motion analysis algorithm, demonstrating how the correlation between facial motion and the surrounding scene can be leveraged for attack detection. The **Replay-Attack** [5] dataset consists of 1,300 video clips featuring various attack types across 50 subjects, all recorded under both controlled and challenging illumination conditions. It simulates real-world scenarios in which FR systems may be deceived through the replay of high-resolution facial images or videos using different media: printed photos (print attacks), iPhone screens (mobile attacks), and iPad screens (high-definition attacks). The dataset is divided into three mutually exclusive identity subsets for training, development, and testing, providing a standardized framework for training and evaluating spoof detection algorithms under diverse attack conditions. The **OULU-NPU** [2] dataset contains 5,940 high-resolution videos from 55 subjects, covering three shooting environments, six smartphone models, and multiple demonstration attack methods. To evaluate the robustness of the algorithm in dealing with unknown attacks and environmental changes, OULU-NPU designed four well-defined evaluation protocols, each of which introduces at least one "unseen condition" in the test set, such as new devices, new backgrounds, or new attack types, so as to more realistically reflect the generalization ability of the method in practical applications. The **CASIA-SURF** [40] dataset contains 21,000 multimodal videos from 1,000 subjects, each of which contains RGB, depth, and infrared information, aiming to promote the development of multimodal FAS research. The research team also proposed a novel multimodal multiscale feature fusion method, which strategically weights multiscale features to enhance the performance of channels with higher information content while suppressing modal features with more noise at a specific scale. The **CASIA-SURF CeFA** [21] is a dataset for studying racial bias in FAS, covering data from 1,607 individuals from

three races and three modalities, and including both 2D and 3D attack types. A notable feature of this dataset is that it clearly annotates racial information, providing reliable support for systematic research on racial bias issues. The **CASIA-SURF HiFiMask** [22] aims to make up for the limitations of existing 3D mask attack detection benchmarks. It contains more than 54,600 video data, collected from 75 individuals wearing 225 high-fidelity masks, using seven different types of sensors. This dataset is committed to bridging the gap between academic research and actual security needs of FR systems, with higher authenticity and challenges. Based on HiFiMask, the authors further proposed a novel Contrastive Context-aware Learning framework, which fully exploits the contextual information in the data pairs between real faces and high-fidelity disguised faces to enhance the supervisory signal in the adversarial attack detection task. The **CASIA-SURF SuHiFiMask** [7] is a large-scale dataset for FAS research, which aims to improve the security of FR systems in remote surveillance scenarios. The dataset covers attack samples of 101 individuals of different ages in 40 different surveillance environments, systematically reflecting the complex conditions in real surveillance scenarios. SuHiFiMask is characterized by its high degree of restoration of common surveillance challenges (such as low-resolution images, environmental noise, etc.). **UniAttackData** [8] data realizes the unified modeling and evaluation of physical and digital attacks for the first time, covering 1,800 subjects. Each subject experienced 2 physical attacks and 12 digital attacks. The attack methods used were all advanced methods in the past three years, which fully reflects the real threats and challenges faced by the current FAS system. However, existing training datasets generally lack comprehensive coverage of diverse attack types, which limits the generalization ability of detection models and hinders their effectiveness in real-world applications. To address this limitation, we released a significantly expanded dataset, UniAttackData+ [26], at the 6th FAS Workshop @ICCV 2025. The dataset includes 2,875 participants from three major demographic groups—Africa, East Asia, and Central Asia—and contains 18,250 real videos collected under diverse lighting conditions, backgrounds, and acquisition devices. For each participant, 54 types of spoofing attacks were simulated, comprising 14 physical and 40 digital attack variants, resulting in a total of 679,097 high-quality forged videos.

Evolution of Face Anti-Spoofing. Early single-modal methods [33, 36] typically utilized CNN to extract image features and used binary classification to identify fake faces. Domain Generalization FAS [3, 9, 10, 12, 13, 16, 18, 27, 29–32, 34, 42–46] is committed to not only performing well on training data domain, but also being effective on unseen domain. Although domain generalization compensates for the disadvantages of single-modal methods in unknown domains to a certain extent, its ability to handle diverse and complex data is still limited, and it cannot fully utilize the complementary information between different modalities. Multi-modal methods [28, 35] have proven to be effective in alleviating the above problems with the motivation of indistinguishable fake faces may exhibit quite different properties under the different spectrums. Despite the success, they require consistent modal inputs during the testing phase as during the training phase, making it impossible to deploy relevant algorithms on a large scale in practical scenarios. Flexible Modal FAS [20, 23, 24, 38, 41] aims to improve the model's adaptability to any given deployed modality. Unified physical-digital face attack detection aims to develop a universal model capable of simultaneously detecting both digital and physical spoofing attacks [4, 6, 8, 11, 26, 37]. JFSFDB [37] proposes the first joint FAS and forgery detection benchmark, which combines visual appearance and rPPG physiological signals for authentication through a dual-branch physiological network and a weighted fusion strategy. UniAttackDetection [8] proposes a unified attack detection framework based on visual language models, which effectively learns unified and specific knowledge features through a teacher-student hint module, a unified knowledge mining module, and a sample-level hint interaction module. [11] significantly improves the model's detection ability for "unseen" attack types through data augmentation technology of simulated physical attack clues and digital attack clues. This method won the first place in the "Unified Physical-Digital Face Attack Detection" task of the 5th CVPR 2024 FAS Challenge, proving its effectiveness and generalization ability in cross-attack type detection. The MoAE-CR [4] framework effectively utilizes category information to improve attack detection performance by introducing a hybrid attack expert module at the feature level and a disaggregation module and a clustering distillation module at the loss level. Based on the idea of hierarchical cue adjustment, HiPTune [26] adaptively selects and integrates classification criteria from different semantic spaces by constructing a visual cue tree and a dynamic cue interaction module to improve the robustness of the model against various attacks.

3. Challenge Overview

Challenge Protocol & Evaluation Metrics. As shown in Tab.2, the UniAttackData+ dataset comprises 2,875 identities from three major demographic groups—Africa, East Asia, and Central Asia—and includes a total of 697,347 video segments (each with at least 25 frames), collected under diverse lighting conditions, backgrounds, and acquisition devices. Among them, 18,250 are live face videos, while 679,097 are fake videos generated through 54 types of attacks per subject, including 14 physical and 40 digital variants (55,950 physical and 623,147 digital attack

Dataset	Attack Type (each ID)	# Datasets / Data	# ID	Physical Attacks		Digital Attacks	
				Dataset Name	No.	# Categories	No.
GrandFake	Incomplete	6 sets: 789,412 (I) (Live: 341738, Fake: 447674)	96,817	SiW-M	128,112 (I)	Adv (6)	116,641 (I)
						DeepFake (6)	202,921 (I)
JFSFDB	Incomplete	9 sets: 27,172 (V) (Live: 5,650, Fake: 21,522)	356	SiW	3,173 (V)	DeepFake (4)	13,752 (V)
				3DMAD	85 (V)		
				HKBU	588 (V)		
				MSU	210 (V)		
				3DMask	864 (V)		
				ROSE	2,850 (V)		
UniAttackData	Complete	1 set: 28,706 (V) (Live: 1,800, Fake: 26,906)	1,800	CASIA-SURF CeFA	5,400 (V)	Adv (6)	10,706 (V)
						DeepFake (6)	10,800 (V)
UniAttackData+	**Complete**	**3 set: 697,347 (V) (Live: 18,250, Fake: 679,097)**	**2,875**	**CASIA-SURF**	**6,000 (V)**	**Adv (16)**	**266,576 (V)**
				CeFA	**9,000 (V)**	**DeepFake (17)**	**242,859 (V)**
				HiFiMask	**40,950 (V)**	**Generation (7)**	**113,712 (V)**

Table 2. UniAttackData+ surpasses all other datasets in scale and diversity. "V" and "I" indicate video- and image-based counts, respectively. The numbers following each attack type in the "# Catagories" row represent the number of algorithms included in that category.

videos). We adopt Protocol 3 from the dataset paper [26] as the evaluation standard for the competition. Under this protocol, models are typically trained on relatively simple attack types and tested on more advanced and complex ones, aiming to assess their generalization ability from simple to sophisticated attacks. The dataset used in this study adopts a unified three-level coding scheme to annotate all samples, aiming to distinguish between live faces and various types of presentation attacks. Live samples are uniformly labeled as $\mathbf{0\_x\_x}$, representing live faces collected under natural conditions. Attack samples are categorized into two main types: Physical Attacks ($\mathbf{1\_x\_x}$) and Digital Attacks ($\mathbf{2\_x\_x}$). Physical Attacks include 2D attacks ($\mathbf{1\_0\_x}$), such as Print ($\mathbf{1\_0\_0}$), Replay ($\mathbf{1\_0\_1}$), and Cutouts ($\mathbf{1\_0\_2}$); as well as 3D attacks ($\mathbf{1\_1\_x}$), including Transparent Mask ($\mathbf{1\_1\_0}$), Plaster Mask ($\mathbf{1\_1\_1}$), and Resin Mask ($\mathbf{1\_1\_2}$). Digital Attacks are further divided into three subcategories: Digital Manipulation ($\mathbf{2\_0\_x}$), including Attribute Edit ($\mathbf{2\_0\_0}$), Face Swap ($\mathbf{2\_0\_1}$), and Video-Driven ($\mathbf{2\_0\_2}$); Digital Adversarial ($\mathbf{2\_1\_x}$), including Pixel-Level ($\mathbf{2\_1\_0}$) and Semantic-Level ($\mathbf{2\_1\_1}$) attacks; and Digital Generation ($\mathbf{2\_2\_x}$), including ID Consistent ($\mathbf{2\_2\_0}$), Style ($\mathbf{2\_2\_1}$), and Prompt-based ($\mathbf{2\_2\_2}$) generation attacks.

In evaluating the performance for this challenge, we adopted the ISO/IEC 30107-3 metrics, which include the Attack Presentation Classification Error Rate (APCER), the Normal/Bona Presentation Classification Error Rate (NPCER/BPCER), and the Average Classification Error Rate (ACER). These metrics quantify the detection system's accuracy in identifying live and fake face presentations. The determination of ACER for the test set relies on the Equal Error Rate (EER) established during the development phase. Additionally, we used the Area Under Curve (AUC) metric as an additional performance measure, assessing model discrimination between fake and real samples across thresholds. Rankings were primarily based on the ACER, with AUC as a secondary criterion, to thoroughly evaluate each algorithm's efficacy against spoofing attacks.

Challenge Process and Timeline. The challenge was held on the CodaLab platform, including two phases as follows: Development Phase (May 20, 2025 – June 13, 2025): During this phase, participants were provided with a labeled training set for model development and an unlabeled validation set for performance evaluation. Participants were allowed to submit predictions on the validation set multiple times and receive real-time feedback via the leaderboard, enabling iterative optimization of their methods. Final Phase (June 13, 2025 – June 28, 2025): The organizers released the ground-truth labels of the validation set and provided an unlabeled test set. Participating teams were required to submit predictions on the test set, using models trained exclusively on the training set. It is important to emphasize that the use of the test data for training purposes was strictly prohibited. The final submission on the CodaLab platform was considered as the official entry. The final rankings were determined based on the performance of the submitted code on the test set, as evaluated by the organizers. To ensure reproducibility, winning teams were required to publicly release their code under the appropriate license and submit a detailed technical report describing their solution in order to be eligible for the prizes.

4. Competition solutions

We focus on summarizing and analyzing the algorithmic strategies of the top 10 ranked teams.

4.1. Facevengers. To address the challenge of detecting both physical and digital face attacks in a generalized setting, the Facevengers team proposes a unified semantic and texture feature analysis framework. Their solution centers around the complementary strengths of CLIP for semantic features and Variational Autoencoders (VAE) for modeling generalized texture representations. To enhance the model's generalization ability on synthetic attacks, VAE is employed to construct paired reconstructed images, helping the model learn texture patterns characteristic of generative

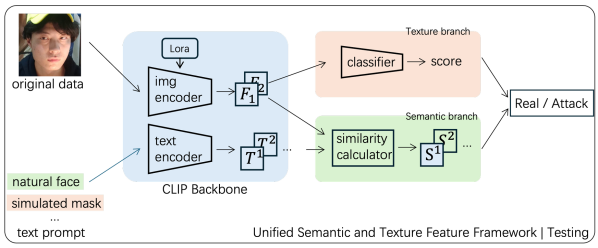

Figure 1. Overall framework of the Facevengers team, which integrates a VAE-based texture representation module with a LoRA-enhanced CLIP backbone to unify semantic and texture features for generalized face attack detection.

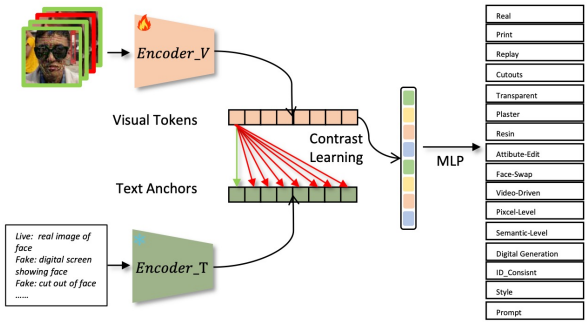

Figure 2. The TeleAI team leverages contrastive learning to pull image tokens closer to their corresponding text anchors while pushing them away from anchors of other spoof types — especially unseen types. This fully exploits the strong vision-language alignment capability of CLIP to enhance generalization.

forgeries. In parallel, semantic representations extracted via CLIP are leveraged to improve robustness against semantically obvious 3D physical attacks such as those involving masks. To better fuse semantic and texture features, LoRA is integrated into the backbone (ViT-Large), allowing efficient fine-tuning without disrupting CLIP's original semantic encoding ability (as illustrated in Fig. 1). During training, a combined dataset including the ChaLearn training set, MSCOCO 2017, and VAE-reconstructed images is used. Without relying on external tools or facial segmentation modules, this method exhibits strong generalization and unbiased detection performance, even when exposed to previously unseen attacks.

4.2. TeleAI. To tackle the challenges posed by the diverse nature of attack cues across physical and digital categories, the TeleAI team proposes a multimodal contrastive learning centered around semantic anchor modeling. Their method establishes robust associations between visual features and descriptive textual concepts, enabling the model to distinguish between various known and unseen attack types through semantically grounded representations (as shown in Fig. 2). For each attack category, this team manually designs five descriptive prompts which are encoded using the CLIP text encoder, forming class-specific semantic anchors. Face images are processed by a visual encoder fine-tuned via Low-Rank Adaptation (LoRA), which maintains efficient adaptation with reduced computational cost. The core training strategy employs a contrastive loss to promote alignment between visual features and their corresponding text anchors, while simultaneously distancing these features from other spoof types. Notably, to improve model robustness and generalization, TeleAI employes several targeted data augmentation strategies. A novel H.264 compression augmentation is applied to mimic real-world compression artifacts. In addition, the team uses advanced AIGC-based generative models (PulID, ConsistentID, InstantID) to synthesize diverse digital attack examples, as well as 3D-rendered resin mask attacks via the Infinity framework to simulate complex physical spoofing scenarios. By incorporating semantic anchors, multimodal alignment, and generation-based augmentation, this framework achieved generalization among unified face attack detection task.

4.3. AKLab. To address the domain gap between validation and test sets, the AKLab team proposes a lightweight semi-supervised framework that leverages local feature geometry and neighborhood diffusion. Their method, titled KNN-based Neighborhood Diffusion, avoids relying on extensive labeled test data and instead builds upon a reliable validation-based anchor mechanism to identify live faces within the test set. This team first trains a ResNet34 backbone model using all available training and validation data (excluding live faces from the validation set). During data preprocessing, Short-Cut Refinement Face Detector (SCRFD) is used for face detection and alignment (excluding the test set), and a 5-crop augmentation is applied to the "Live" and physical attack samples to improve data diversity. For training, only standard augmentations such as random cropping and horizontal flipping are adopted. The model is trained with a 5-class classification head, covering multiple spoof subtypes. After training, the model is used to extract feature embeddings for test samples and a small set of 13 known live samples from the validation set. These embeddings are then projected into 2D using t-SNE, where the live validation faces naturally forms a distinct cluster (as illustrated in Fig. 3). To enhance robustness, AKLab applies a KNN-based outlier removal step to these validation samples and computes a refined KNN centroid. Using this center as a reference, the 100 nearest test feature points are identified and labeled as live faces. To determine optimal feature space clustering, the team introduces a novel InAndR metric, which considers three factors: 1) the relative compactness of the KNN cluster, 2) the inclusion ratio of the validation samples within the cluster, and 3) the inter-center distance normalized by cluster spread. This strategy

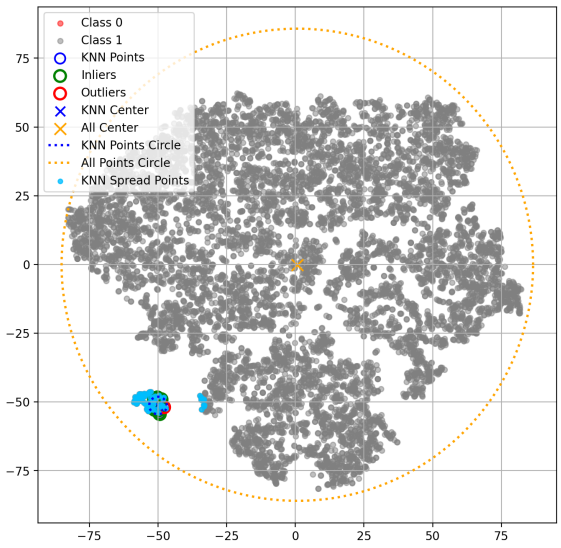

Figure 3. A distinct cluster emerges in the semi-supervised framework (AKLab), well-separated from the remaining data and corresponding to validation live faces.

allows the team to tune their detection threshold based on geometric properties rather than direct supervision.

4.4. bklzhn. The bklzhn team introduces several notable innovations to improve generalization in unified face attack detection. Central to their approach is a pairwise training strategy based on cosine similarity between live and spoof samples, ensuring semantically meaningful supervision through real data rather than synthetic approximations. By filtering live-attack pairs using a tunable threshold, their method emphasizes high-quality relational supervision, avoiding noisy or misleading examples. To enhance robustness without overfitting to spoof-specific artifacts, they employ an asymmetric augmentation strategy that applies diverse transformations only to live samples while keeping spoof images unmodified. This prevents the model from learning augmentation-induced artifacts as spoof cues. Furthermore, their model benefits from dual-task supervision, combining focal loss for spoof detection with supervised contrastive loss for discriminative feature embedding. To improve regularization, CutMix is incorporated during training, helping the model generalize better under occlusion or distribution shifts.

4.5. CMSR. The CMSR team introduces a novel conceptual perspective for understanding physical face attacks. Instead of treating physical attacks (e.g., print, replay) as entirely fake samples, they propose modeling them as a superposition of "real face information" and "physical forgery traces". This fundamental insight shifts the goal from binary classification to recovering the intrinsic authenticity embedded within spoofed images. Building on this framework, CMSR develops targeted image processing

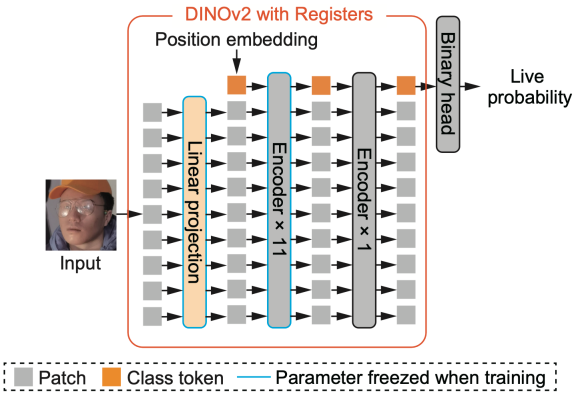

Figure 4. Tohoku Aoki Lab uses pre-trained DINOv2 with registers and a binary classification head.

techniques such as bilateral filtering to selectively suppress forgery artifacts (moiré patterns, reflections, and color distortions) while preserving and enhancing live facial structures. These refined physical attack samples are repurposed as high-quality synthetic augmentations for real faces, enabling the model to learn more discriminative live cues under challenging conditions. This forgery decomposition–based augmentation strategy is implemented without relying on external data and significantly improves generalization and live-face recall. Their work presents a promising direction for semantic reinterpretation and reuse of attack data, bridging the gap between physical realism and machine-level representation learning.

4.6. FaceGuardians. FaceGuardians proposes ARTEMIS: Artefact-centric Robust Training with Engineered Masks and Identity Suppression, a framework that explicitly focuses on spoof artifact learning rather than identity representation. The backbone of ARTEMIS is a frozen DINOv2 (ViT/L), from which high-level embeddings are extracted and optimized using a combination of hypersphere loss, binary cross-entropy loss, and triplet loss. These losses push live samples onto a unit-radius sphere while forcing spoofs away, thereby achieving identity-agnostic feature separation. To eliminate the "identity → spoof" shortcut often exploited in spoof detection, ARTEMIS injects synthetically generated spoofs for every live image in the training set. These include over 5,000 high-fidelity 3D-mask renders based on 68-landmark alignment, as well as identity-agnostic adversarial examples generated using tools like PhotoMaker and ConsistentID. These synthetic spoofs preserve visual diversity while breaking the identity correlation. The framework follows a two-stage pipeline. First, a large teacher model is trained on full-resolution images using the aforementioned multi-loss scheme. Then, knowledge is distilled into a compact ViT-Base student model through feature-matching and distillation losses. This yields a lightweight model with 86M parameters that retains the

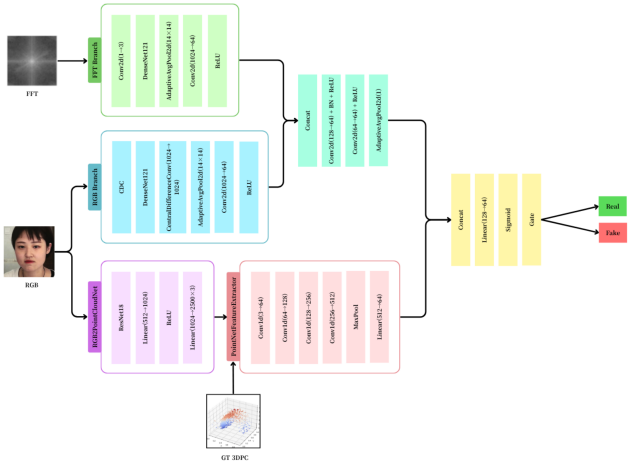

Figure 5. The RFPCNet (RGB-Frequency-PointCloud Network) is a novel face anti-spoofing solution that integrates spatial, frequency, and geometric cues through a three-branch architecture.

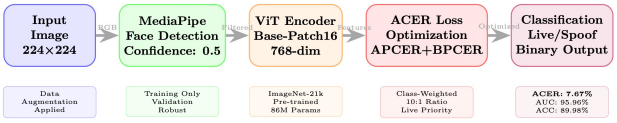

Figure 6. The RFPCNet (RGB-Frequency-PointCloud Network) is a novel face anti-spoofing solution that integrates spatial, frequency, and geometric cues through a three-branch architecture.

teacher's generalization ability.

4.7. Tohoku Aoki Lab. Tohoku Aoki Lab introduces a simple yet effective framework built upon the self-supervised vision transformer model DINOv2 with registers. Their approach leverages pre-trained DINOv2 weights and integrates a lightweight binary classification head, while freezing most of the encoder parameters to preserve general visual features. As shown in Fig. 4, the architecture appends a linear projection and classification head to the register-enhanced DINOv2 backbone. To enhance learning under class imbalance and domain shift, the team adopts Focal Loss with a gamma value of 2 and equal class weights, emphasizing harder samples. Although no additional data or multi-modal input is used, the team highlights the potential extensibility of register-enhanced DINOv2 to broader ViT-based face attack detection tasks, laying groundwork for future work in low-data or domain-shifted scenarios.

4.8. GCD-UdL. To address the significant class imbalance between live and attack samples, the GCD-UdL team proposes a novel Iterative binary training strategy that prioritizes high-frequency fraud classes in the early learning phases. This method is specifically designed to improve generalization under imbalanced training conditions and to enhance the model's robustness against diverse attack types. Instead of directly training on all spoof classes simultaneously, the team decomposes the attack category into multiple binary classification tasks, where the model is initially trained to distinguish live samples from the most frequent attack class. Once this phase converges or reaches 15 epochs, a new attack class is introduced, and training continues iteratively. This reverse-frequency schedule continues until all attack types are included, concluding with a final consolidation phase that fine-tunes the model using all attack classes. This strategy avoids the need for full dataset rebalancing and integrates well with transfer learning setups. By iteratively exposing the model to attack classes in a structured manner, the method effectively reduces class bias and prevents catastrophic forgetting.

4.9. LNL. To effectively capture diverse spoofing cues across spatial, frequency, and geometric domains, team LNL proposes a novel multi-branch architecture named RFPCNet. The model integrates three complementary modalities (RGB textures, frequency-domain signals, and 3D geometric structures) to enhance robustness against both physical and digital attacks. As shown in Fig. 5, RFPCNet includes an RGB branch with Central Difference Convolution (CDC) modules embedded before and after a DenseNet-121 backbone, improving sensitivity to micro-texture distortions and edge artifacts. A second branch processes FFT-based magnitude maps extracted from grayscale images using another DenseNet-121 to detect periodic attack patterns. The third branch, based on a ResNet-18 encoder, predicts 3D point clouds from RGB inputs, which are then supervised using Chamfer loss against offline-generated pseudo ground-truth 3DPCs obtained via 3DDFA_V2. These predicted 3D point clouds are further encoded by a PointNet to extract geometric depth features. To adaptively integrate the multi-modal cues, a gated fusion module is introduced. This module dynamically weights the importance of pooled RGB-frequency features and 3D point cloud features based on learned confidence scores. The network is trained in a multi-task setting, combining spoof classification and 3D reconstruction with a loss function composed of Chamfer distance and weighted cross-entropy. In particular, live samples are assigned triple weight to address class imbalance, and a balanced sampling strategy ensures proportional representation of different attack types. While the design introduces additional computational cost due to 3D point cloud generation and multi-branch inference, the method demonstrates strong generalization across attack modalities, highlighting the advantage of combining spatial, frequency, and geometric representations in a unified framework.

4.10. asakatsu2025. To address the severe class imbalance and performance trade-off in face attack detection, team asakatsu2025 proposes a ViT-based framework with an ACER-optimized loss function that directly targets the Average Classification Error Rate as primary evaluation

Team	Contact	ACER(%)↓	AUC(%)↑	ACC(%)↑	EER(%)↓
Facevengers	Ke-Yue Zhang (zkyezhang@tencent.com)	0.144	99.93	99.71	6.22
TeleAI	Qi Zhang (zhangq139@chinatelecom.cn)	0.178	99.99	99.64	7.34
AKLab	Zehua Lan (zehua.lan@akuvox.com)	0.53	99.46	99.97	48.60
bklzhn	Denis Kondranin (denis.kondranin@idrnd.net)	2.10	97.90	98.93	4.60
CMSR	Ming Liu (liuming2@cmsr.chinamobile.com)	4.62	98.22	96.01	8.59
FaceGuardians	Efim Boieru (efim.boieru@incode.com)	8.11	97.33	90.14	24.84
Tohoku Aoki Lab	Mika Feng (mika@aoki.ecei.tohoku.ac.jp)	11.07	94.80	90.47	61.32
GCD-UdL	Vítor da Silva (vitor.dasilva@udl.cat)	25.22	89.60	92.35	38.64
LNL	Yunseo Lee (yunseo528@swu.ac.kr)	26.24	82.72	86.20	28.96
asakatsu2025	Jin Jie (ohki@sec.inf.shizuoka.ac.jp)	29.49	73.05	61.22	84.60
Siren Shield	Taehoon-Kim (kimth@vilab.cau.ac.kr)	31.54	69.52	82.94	61.55
BU-S UniFAS	Pongchi Yuen (csyangji@comp.hkbu.edu.hk)	32.81	73.34	70.11	8.60

Table 3. Final team results on ACER, AUC, ACC, and EER. ↓ / ↑ indicate that smaller/larger scores correspond to better performance.

metric. Their approach integrates a series of innovations in loss design, data sampling, and augmentation tailored specifically for the face attack detection task. As shown in Fig. 6, the pipeline adopts a Vision Transformer (ViT-Base-Patch16-224) pre-trained on ImageNet-21k. Before training, input images are filtered using MediaPipe face detection to remove samples without valid faces, enhancing training quality. In the data augmentation stage, the model incorporates FAS-specific transformations, such as color distortion, reflection simulation, hand trembling, and low-resolution degradation, applied with various probabilities. The central contribution lies in their ACER-optimized loss, which combines traditional cross-entropy loss with a hinge-based formulation that explicitly optimizes APCER and BPCER:

$$\mathcal{L}_{\text{Combined}} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{ACER}\mathcal{L}_{\text{HingeACER}}, \quad (1)$$

$$\mathcal{L}_{\text{Hinge }ACER} = \lambda_{APCER}\mathcal{L}_{APCER} + \lambda_{BPCER}\mathcal{L}_{BPCER}. \quad (2)$$

Here, the APCER loss focuses on penalizing false positives, while the BPCER loss prioritizes preserving live sample predictions, both regulated by tunable margin-based hinge terms. The team sets $\lambda_{APCER} = 0.2$ and $\lambda_{BPCER} = 0.8$ to emphasize live sample preservation, aligning with real-world deployment needs. To mitigate class imbalance, the method uses 10:1 class-weighted training in favor of live samples and implements a WeightedRandomSampler to ensure equal representation across batches. They also employ cosine annealing with warmup and BPCER-based early stopping to stabilize training and avoid overfitting.

5. Conclusion

In the ICCV 2025 Challenge on Unified Physical-Digital Face Attack Detection, the Facevengers team secured first place with an outstanding ACER of 0.144%, significantly outperforming the runner-up TeleAI (0.178%) and demonstrating strong generalization capabilities. Their success lies in the integration of CLIP-based semantic features with VAE-based texture modeling, combined with efficient LoRA fine-tuning, enabling robust performance against unseen attack types. TeleAI followed closely, achieving the best AUC (99.99%) through a multimodal contrastive learning approach and a semantic anchor mechanism, validating the effectiveness of its text-vision alignment strategy. Notably, AKLab achieved an ACER of 0.53% under limited resources by employing a semi-supervised KNN-based neighborhood diffusion method, highlighting the potential of geometric features in label-scarce scenarios. In contrast, teams relying on traditional CNNs or single-modality inputs (e.g., LNL, asakatsu2025) generally exhibited ACER scores above 20%, primarily due to their limited capacity to capture cross-attack commonalities. Overall, the combination of semantic-texture pre-trained models (e.g., CLIP) and lightweight fine-tuning techniques has emerged as the most effective paradigm for unified detection, while data augmentation and cross-modal alignment remain key to enhancing robustness. Looking ahead, unified detection tasks based on multimodal large models (e.g., Qianwen) are expected to advance toward stronger cross-modal understanding and generalization. By integrating visual, textual, and auditory cues, such models can better capture complex attack patterns and adapt to unseen scenarios. Techniques like large-scale pretraining, prompt learning, and contrastive learning with semantic anchors will further enhance robustness and interpretability. We hope UniAttackData+ and the insights from this challenge will inspire future research on more generalizable face attack detection systems.

6. Acknowledgments

We would like to express our sincere thanks to Facevengers, TeleAI, AKLab, bklzhn, CMSR, FaceGuardians, Tohoku Aoki Lab, GCD-UdL, LNL, asakatsu2025, Siren Shield, and BU-S UniFAS teams for code submissions and contributions to algorithm implementation and performance optimization.

References

[1] André Anjos and Sébastien Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011. 2

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. 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)*, pages 612–618. IEEE, 2017. 2

[3] Rizhao Cai, Zitong Yu, Chenqi Kong, Haoliang Li, Changsheng Chen, Yongjian Hu, and Alex C Kot. S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *IEEE Transactions on Information Forensics and Security*, 2024. 3

[4] Shunxin Chen, Ajian Liu, Junze Zheng, Jun Wan, Kailai Peng, Sergio Escalera, and Zhen Lei. Mixture-of-attack-experts with class regularization for unified physical-digital face attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2195–2203, 2025. 1, 2, 3

[5] 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. IEEE, 2012. 2

[6] Debayan Deb, Xiaoming Liu, and Anil K Jain. Unified detection of digital and physical face attacks. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 1, 3

[7] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 1, 3

[8] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, and Zhen Lei. Unified physical-digital face attack detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 749–757. International Joint Conferences on Artificial Intelligence Organization, 2024. Main Track. 1, 3

[9] Jiabao Guo, Ajian Liu, Yunfeng Diao, Jin Zhang, Hui Ma, Bo Zhao, Richang Hong, and Meng Wang. Domain generalization for face anti-spoofing via content-aware composite prompt engineering. *arXiv preprint arXiv:2504.04470*, 2025. 3

[10] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. *International Journal of Computer Vision*, pages 1–22, 2024. 3

[11] 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*, pages 995–1004, 2024. 1, 2, 3

[12] Yan He, Fei Peng, Rizhao Cai, Zitong Yu, Min Long, and Kwok-Yan Lam. Category-conditional gradient alignment for domain adaptive face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2024. 3

[13] Chengyang Hu, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Rethinking generalizable face anti-spoofing via hierarchical prototype-guided distribution refinement in hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1032–1041, 2024. 3

[14] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020. 1

[15] Liming Jiang, Wayne Wu, Chen Qian, and Chen Change Loy. Deepfakes detection: The deeperforensics dataset and challenge. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 303–329. Springer International Publishing Cham, 2022. 1

[16] Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 188–199, 2024. 3

[17] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 1

[18] Yongze Li, Ning Li, Ajian Liu, Hui Ma, Liying Yang, Xihong Chen, Zhiyao Liang, Yanyan Liang, Jun Wan, and Zhen Lei. Fa^3-clip: Frequency-aware cues fusion and attack-agnostic prompt learning for unified face attack detection. *arXiv preprint arXiv:2504.00454*, 2025. 3

[19] Ajian Liu. Ca-moeit: Generalizable face anti-spoofing via dual cross-attention and semi-fixed mixture-of-expert. *International Journal of Computer Vision*, pages 1–14, 2024. 1

[20] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186, 2022. 3

[21] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 2

[22] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 1, 3

[23] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, S. Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face

anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023. 3

[24] Ajian Liu, Hui Ma, Junze Zheng, Haocheng Yuan, Xiaoyuan Yu, Yanyan Liang, Sergio Escalera, Jun Wan, and Zhen Lei. Fm-clip: Flexible modal clip for face anti-spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8228–8237, 2024. 3

[25] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[26] Ajian Liu, Haocheng Yuan, Xiao Guo, Hui Ma, Wanyi Zhuang, Changtao Miao, Yan Hong, Chuanbiao Song, Jun Lan, Qi Chu, et al. Benchmarking unified face attack detection via hierarchical prompt tuning. *arXiv preprint arXiv:2505.13327*, 2025. 1, 2, 3, 4

[27] Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[28] Zohreh Mostaani, Anjith George, Guillaume Heusch, David Geissbuhler, and Sebastien Marcel. The high-quality wide multi-channel attack (hq-wmca) database, 2020. 3

[29] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *ICCV*, 2023. 3

[30] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24563–24574, 2023.

[31] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *TIFS*, 2020.

[32] Xudong Wang, Ke-Yue Zhang, Taiping Yao, Qianyu Zhou, Shouhong Ding, Pingyang Dai, and Rongrong Ji. Tf-fas: twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In *European Conference on Computer Vision*, pages 148–168. Springer, 2024. 3

[33] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, 2020. 3

[34] Jingyi Yang, Zitong Yu, Xiuming Ni, Jia He, and Hui Li. Graph guided video vision transformer for face anti-spoofing. *arXiv preprint arXiv:2408.07675*, 2024. 3

[35] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *CVPRW*, pages 650–651, 2020. 3

[36] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, 2020. 3

[37] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *arXiv:2208.05401*, 2022. 1, 3

[38] Zitong Yu, Rizhao Cai, Yawen Cui, Ajian Liu, and Changsheng Chen. Visual prompt flexible-modal face anti-spoofing, 2023. 3

[39] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. 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*, pages 919–928, 2019. 1

[40] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 1, 2

[41] Guanghao Zheng, Yuchen Liu, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Towards unified representation of invariant-specific features in missing modality face anti-spoofing. In *ECCV*, 2024. 3

[42] Tianyi Zheng, Bo Li, Shuang Wu, Ben Wan, Guodong Mu, Shice Liu, Shouhong Ding, and Jia Wang. Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19:4058–4069, 2024. 3

[43] Tianyi Zheng, Bo Li, Shuang Wu, Ben Wan, Guodong Mu, Shice Liu, Shouhong Ding, and Jia Wang. Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. *IEEE transactions on information forensics and security(TIFS)*, 2024.

[44] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6009–6018, 2022.

[45] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023.

[46] Hang Zou, Hui Zhang, Yuan Zhang, Hui Ma, Dexin Zhao, Qi Zhang, and Qi Li. Multi-angle consistent generative nerf with additive angular margin momentum contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 930–939, 2024. 3