Benchmarking Vision Foundation Models for Domain-Generalizable Face Anti-Spoofing

Anonymous CVPR submission

Paper ID 24

Abstract

Face Anti-Spoofing (FAS) remains challenging due to the requirement for robust domain generalization across unseen environments. While recent trends leverage Vision-Language Models (VLMs) for semantic supervision, these multimodal approaches often demand prohibitive computational resources and exhibit high inference latency. Furthermore, their efficacy is inherently limited by the quality of the underlying visual features. This paper revisits the potential of vision-only foundation models to establish a highly efficient and robust baseline for FAS. We conduct a systematic benchmarking of 15 pre-trained models, such as supervised CNNs, supervised ViTs, and self-supervised ViTs, under severe cross-domain scenarios including the MICO and Limited Source Domains (LSD) protocols. Our comprehensive analysis reveals that self-supervised vision models, particularly DINOv2 with Registers, significantly suppress attention artifacts and capture critical, fine-grained spoofing cues. Combined with Face Anti-Spoofing Data Augmentation (FAS-Aug), Patch-wise Data Augmentation (PDA) and Attention-weighted Patch Loss (APL), our proposed vision-only baseline achieves state-of-the-art performance in the MICO protocol. This baseline outperforms existing methods under the data-constrained LSD protocol while maintaining superior computational efficiency. This work provides a definitive vision-only baseline for FAS, demonstrating that optimized self-supervised vision transformers can serve as a backbone for both vision-only and future multimodal FAS systems.

1. Introduction

Face recognition technology, which identifies individuals based on unique facial biometrics, has become a cornerstone of modern authentication systems due to its inherent advantages, including low implementation cost, contactless operation, and high user convenience [23]. While face recognition systems are engineered for robustness against common environmental variations, such as changes in head pose, illumination, and image blur, this resilience is paradoxically exploited by malicious actors through spoofing attacks. Specifically, the ease with which presentation materials (e.g., printed photos or replayed videos) can be used to impersonate a registered user poses a critical security threat to face recognition systems. These attempts are known as Presentation Attacks (PAs), and they are becoming increasingly prevalent since high-quality face images are accessible online, making PAs a realistic and easily executed method to bypass security protocols [35].

To effectively counter PAs, the core challenge for Face Anti-Spoofing (FAS) is to detect minute and fine-grained discrepancies between a genuine live face and a spoofed presentation. This requires extracting subtle but highly discriminative features that reveal the artifacts of the presentation medium, such as the distinctive texture of paper, the moiré patterns from a display device, or the lack of authentic 3D information such as depth and light reflection. Consequently, numerous deep learning methods utilizing Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) [12] have been proposed to learn these inherent spoofing cues [7, 13, 14, 17, 22, 48, 51–54, 58]. While these methods often achieve high accuracy in controlled settings (i.e., intra-dataset evaluation), their primary limitation emerges when applied to real-world scenarios. Their performance degrades significantly when encountering unseen attack types, new acquisition devices, or novel environmental conditions. This is known as domain gap in FAS.

This domain gap, which leads to degraded detection accuracy against unseen attacks, primarily stems from two technical factors. The first is the limited capability of the model structure to resist domain-specific biases acquired during training, and the second is the deficiency of the feature extractor in producing robust domain-invariant features across diverse real-world scenarios. So far, CNN-based FAS methods, such as CDCN [54], NAS-FAS [53], and PatchNet [48], have focused on extracting fine-grained local features inherent to spoofing attacks. Notably, PatchNet reformulated FAS as a fine-grained patch-type recognition

CVPR #24

problem and improved generalization by learning highly discriminative features from local capture characteristics. However, many of these CNN-based methods suffer from an inherent lack of generalization capability in their initial parameters against unknown attacks and environmental changes, primarily because they do not leverage large-scale pre-trained models as their backbones. Recently, FAS methods utilizing ViT [7, 14, 51, 52] have been proposed, demonstrating superiority over CNNs in their ability to capture both local and global features integrally. Nevertheless, most pre-trained ViTs employed by these methods rely on supervised learning with limited image datasets, resulting in limitations in their capacity to extract robust and domain-invariant generic features. For instance, while Segment Anything Model (SAM) [19] used by Chen et al. [7] is ViT-based, it is specifically designed for segmentation, not for extracting the discriminative liveness cues required by the FAS task. Therefore, we consider that leveraging a vision foundation model, which is pre-trained on massive amounts of data in a self-supervised manner and possesses high versatility and robustness, is crucial for the feature extractor to achieve domain generalization in FAS.

Recently, several approaches have emerged that leverage Vision-Language Models (VLMs) to enhance generalization through semantically rich textual supervision [27, 45, 56]. However, these VLM-based multimodal techniques typically require massive computational resources (e.g., large-scale LLMs) and suffer from slow inference speeds, making them difficult to deploy in real-world, resource-constrained authentication systems. Furthermore, their performance remains inherently constrained by the quality and domain-invariance of the underlying visual features. Therefore, instead of simply adopting computationally expensive VLMs, achieving highly efficient and robust domain generalization in FAS requires a systematic evaluation of vision-only foundation models to establish a solid baseline.

To explore the potential of vision-only models, Feng et al. demonstrated the utility of intermediate ViT features for FAS [14]. They also demonstrated that DINOv2 with Registers [11] effectively suppresses attention artifacts and captures fine-grained spoofing cues [13]. However, their evaluations were limited to intra-dataset protocols (e.g., SiW [30], OULU-NPU [4]), which fail to measure the robustness against real-world domain shifts. In this paper, we extend this research direction by conducting a comprehensive analysis of various vision foundation models specifically under severe cross-domain scenarios (e.g., the MICO protocol and Limited Source Domains). Specifically, we comprehensively evaluate 15 different visual feature extractors, encompassing supervised CNNs, supervised ViTs, and self-supervised ViTs. Through our extensive benchmarking, we demonstrate that DINOv2 with Registers [11], combined with Face Anti-Spoofing Data Augmentation (FAS-Aug) [5], Patch-wise Data Augmentation (PDA) [52], and Attention-weighted Patch Loss (APL) [52] serves as the most effective and robust feature extractor. By establishing this strong vision-only baseline, we show that an optimized self-supervised vision transformer can achieve highly competitive domain generalization comparable to existing methods in the MICO protocol. Furthermore, our baseline achieves higher accuracy than existing methods under the Limited Source Domains (LSD) protocol, demonstrating the effectiveness of DINOv2 in data-constrained scenarios. Consequently, our comprehensive analysis provides a practical and generalizable vision-only baseline for FAS, offering an effective visual feature extractor that can serve as a strong backbone for both vision-only methods and advanced multimodal approaches.

2. Related Work

This section gives an overview of research in FAS, ranging from traditional to advanced foundation model-based approaches.

2.1. Early FAS Methods

FAS is essential for securing face recognition systems against PAs such as printed photos and 3D masks [35]. While early methods relied on handcrafted features [3, 9, 20, 38, 39], the field quickly shifted to deep learning. CNN-based approaches, such as PatchNet [48] and NAS-FAS [53], achieve high accuracy in controlled, intra-domain settings by learning discriminative local representations or utilizing auxiliary cues [30, 50, 54]. However, these methods often suffer from severe performance degradation when exposed to unseen datasets or novel attack types.

2.2. Domain Generalization for FAS

To address the cross-domain challenge, various Domain Generalization (DG) strategies have been proposed to learn domain-invariant representations. These methods employ meta-learning (e.g., RFMeta [43], MADDG [42]), feature disentanglement (e.g., DR-MD-Net [49], MFAE [58]), or domain alignment techniques [8, 21, 25, 28, 29, 31, 46, 48, 59]. Recent approaches also leverage physics-based data synthesis [5] to simulate diverse attack variations. Despite these advancements, achieving true domain invariance against real-world shifts (e.g., unseen capture devices, lighting conditions, and novel spoof materials) remains a highly challenging and open problem.

2.3. Vision Foundation Models in FAS

Recently, the pursuit of robust representations has driven FAS research toward large-scale foundation models. Several approaches leverage Vision-Language Models (VLMs), such as CLIP [40], or employ multimodal architectures (e.g., FLIP [45], I-FAS [56], InstructFLIP [26]) to enhance

CVPR #24

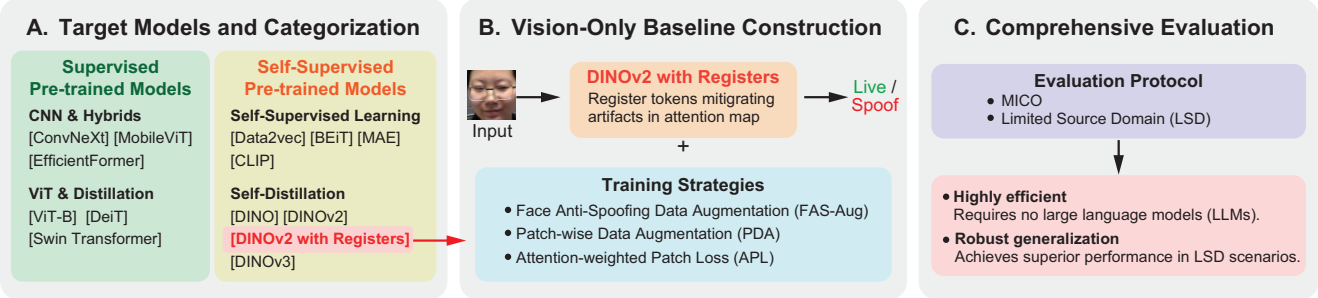

Figure 1. Overview of our comprehensive benchmarking framework and the proposed vision-only baseline. (A) Target Models and Categorization: We systematically categorize and evaluate 15 diverse vision foundation models across different pre-training paradigms (Supervised CNNs, Supervised ViTs, and Self-Supervised & Multi-modal ViTs). (B) Vision-Only Baseline Construction: Based on the benchmark insights, we construct a highly efficient baseline utilizing DINOv2 with Registers to mitigate attention artifacts, further enhanced by robust training strategies including FAS-Aug, PDA, and APL. (C) Comprehensive Evaluation: Extensive evaluations under the standard MICO and Limited Source Domain (LSD) protocols demonstrate that our vision-only approach achieves superior domain generalization.

generalization through semantic text guidance. However, these VLM-based methods typically demand massive computational resources and suffer from high inference latency, limiting their practical deployment in real-world authentication systems. In contrast, self-supervised vision-only foundation models like DINO [6] and DINOv2 [37] provide a highly efficient and robust visual foundation. They are effective at capturing fine-grained, localized spatial features essential for detecting minute spoofing artifacts without relying on semantic labels. Specifically, DINOv2 with Registers [11] introduces auxiliary tokens to suppress attention artifacts. This stabilization has proven crucial for FAS, allowing the model to accurately capture fine-grained spoofing cues [13]. While these vision models show significant promise, a comprehensive benchmarking of their domain generalization capabilities in FAS is still lacking. In this study, we revisit vision-only foundation models for FAS and establish a strong baseline through comprehensive benchmarking.

3. Benchmarking Framework and Baseline

The primary goal of this section is to establish a systematic benchmarking framework to identify the most effective vision foundation models for FAS. We first categorize a set of pre-trained models used for feature extraction. Subsequently, we describe the construction of our baseline model, which integrates the most promising backbone with FAS-Aug [5], PDA [52] and APL [52].

3.1. Categorization of Vision Foundation Models

To systematically evaluate how different architecture designs and pre-training strategies contribute to robustness against spoofing attacks, we classify the selected models along two primary axes: architecture designs and pre-training strategies.

3.1.1. Architecture Design

We evaluate models across a spectrum of architectures, ranging from modernized convolutions to attention mechanisms, to understand their baseline feature extraction capabilities.

CNNs and Hybrids: ConvNeXt [33] serves as a modernized CNN designed to achieve performance competitive with Vision Transformers (ViTs) by incorporating recent architectural advancements. To address the trade-off between representational power and efficiency, we include MobileViT [36], a lightweight hybrid model that integrates the inductive biases of CNNs with the global modeling capabilities of Transformers, and EfficientFormer [24], which optimizes Transformer structures for high-speed inference.

Vision Transformers: We evaluate ViT-B [12], the standard ViT backbone that captures global context by processing images as a sequence of patches. Additionally, we include the Swin Transformer [32], which achieves multi-scale representations through a hierarchical structure and shifted window-based self-attention.

3.1.2. Pre-training Strategies

Beyond network architecture, the method of pre-training significantly impacts a model's ability to extract robust, domain-invariant features critical for cross-domain FAS.

Supervised Learning and Distillation: Several models in our benchmark rely on traditional supervised learning. To improve data efficiency in this setting, we evaluate DeiT [47], which utilizes a teacher-student distillation strategy to train ViTs effectively without relying on massive, fully-annotated datasets.

Self-Supervised Learning (SSL): SSL models have demonstrated exceptional transferability across diverse

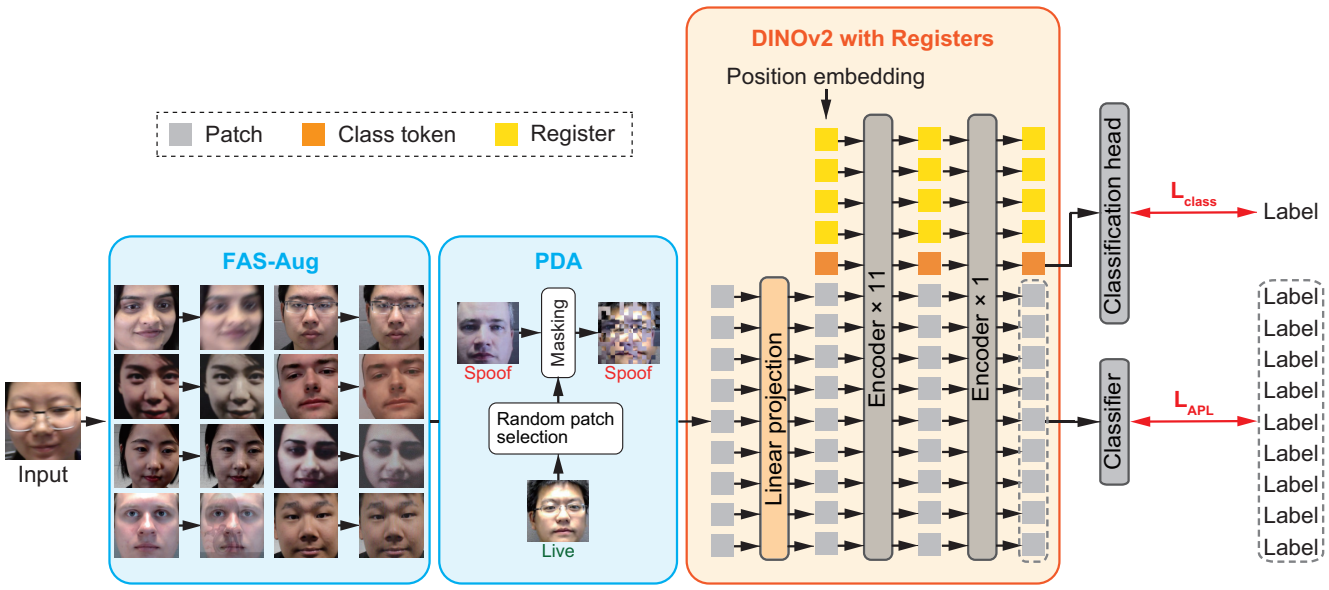

Figure 2. Overview of our proposed vision-only baseline. During training, input images are augmented by Face Anti-Spoofing Data Augmentation (FAS-Aug) and Patch-wise Data Augmentation (PDA) to simulate diverse spoofing artifacts and enforce robust local representations. The processed images are then fed into the DINOv2 with Registers backbone. Finally, the network is optimized using a dual-level supervision strategy: a global classification loss ($\mathcal{L}_{class}$) derived from the class token, and a local Attention-Weighted Patch Loss ($\mathcal{L}_{APL}$) applied to the patch tokens.

downstream tasks by learning without explicit labels. We include Masked Image Modeling (MIM) approaches such as BEiT [2] and MAE [16]. While BEiT treats images as discrete tokens to learn semantic representations by predicting masked patches, MAE focuses on reconstructing original pixel values from highly masked inputs. We also evaluate Data2vec [1], which learns the essential structure of input data by predicting latent representations of augmented views, and CLIP [40], a vision-language foundation model that gains semantically aligned visual representations through contrastive learning on massive image-text pairs.

Self-Distillation (DINO Family): Central to our analysis is the DINO family, which relies on a self-distillation paradigm. Starting from the original DINO [6], we evaluate DINOv2 [37] and its refined variant, DINOv2 with Registers [11]. The latter specifically addresses feature map artifacts by introducing dedicated register tokens, yielding smoother representations that are critical for capturing fine-grained spoofing cues. Finally, we include DINOv3 [44], which scales the training strategies of the DINOv2 framework to further enhance adaptability to complex visual tasks.

3.2. Baseline Method

This section describes the proposed baseline method, designed to enhance FAS generalization against unseen attacks. Recognizing that the rich visual representations learned through massive-scale self-supervised pre-training significantly improve detection performance, we adopt the vision foundation model DINOv2 [37] as our core architecture. Specifically, we utilize DINOv2 with Registers [11] to mitigate attention perturbations and focus on fine-grained visual cues. Moreover, robustness across diverse spoofing scenarios is achieved by combining FAS-Aug [5] and PDA [52] with a dual-level loss function incorporating Attention-Weighted Patch Loss (APL) [52]. The overall pipeline is illustrated in Fig. 2.

3.2.1. Model Architecture

The baseline model initiates by fine-tuning the DINOv2 ViT-B/14 with Registers [11] without freezing any layers. A 224×224 pixel input image is divided into 14×14 non-overlapping patches and processed through the 12-layer Transformer encoder. The incorporation of register tokens, defined as learnable tokens appended to the input sequence, is critical for achieving the fine-grained feature extraction required for FAS. Prior studies demonstrate that standard ViTs suffer from an "attention spike" phenomenon, where anomalously high attention concentrates on irrelevant background areas due to the model repurposing patch tokens for global information storage [11]. Register tokens mitigate this noise, effectively stabilizing the attention mechanism and yielding smoother, more interpretable attention maps. This enables the model to accurately capture the minute visual cues essential for spoof detection. A classification head is appended to the final layer. During inference, the out-

CVPR #24

put probability of the class token from this head, P_{Live}, is computed. The image is classified as "Live" if P_{Live} exceeds a predetermined threshold, and "Spoof" otherwise. This threshold is determined as the Equal Error Rate (EER) point on the training set, where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR) [52].

3.2.2. Data Augmentation

To maximize the model's robustness and generalization capability across domains, we integrate two complementary augmentation techniques applied exclusively during the training phase. First, FAS-Aug [5] is employed to simulate various forms of degradation and attack artifacts. This strategy encompasses eight augmentation types, including photography noise (e.g., hand trembling, low resolution), print attack artifacts (e.g., color distortion), and display attack artifacts (e.g., moiré patterns). Photography noise is applied without altering the label, while simulations of print and display attacks result in the label being reassigned to "Spoof." During training, one of these eight augmentations is randomly applied to each input image based on its original specification. Second, we utilize PDA [52], a patch-level augmentation method tailored for Vision Transformers. Specifically, we employ the Live Patch Mask component, which randomly replaces certain patches in a spoofed image with corresponding patches from a live image with a probability of $P = 0.5$. The model is then trained to classify the substituted live patches as "Live," the remaining spoofed patches as "Spoof," and the overall image as "Spoof." This patch-level mixing increases task difficulty, compelling the model to learn highly discriminative and robust local representations.

3.2.3. Loss Function

During network training, the baseline method employs a dual-level loss function to facilitate robust learning at both the global (image) and local (patch) levels:

$$\mathcal{L}_{total} = \mathcal{L}_{class} + \mathcal{L}_{APL}. \tag{1}$$

The global loss, $\mathcal{L}_{class}$, corresponds to the standard binary classification loss computed from the class token output. The local loss, APL [52] ($\mathcal{L}_{APL}$), is computed for each patch using the attention map of the class token from the 12th encoder block as a weighting factor. This patch-wise supervision enhances the model's ability to detect localized spoofing artifacts. Furthermore, for the binary classification within $\mathcal{L}_{APL}$, the L2-constrained Softmax loss [41] is adopted to ensure balanced feature learning between the "Live" and "Spoof" classes, thereby improving classification stability. The L2-constrained Softmax loss enforces all feature vectors to have a fixed L2 norm α and is defined by:

$$\text{L2Softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^{\top}f(\boldsymbol{x}_i)+b_{y_i}}}{\sum_{j=1}^{C}e^{W_j^{\top}f(\boldsymbol{x}_i)+b_j}}, \tag{2}$$

$$\text{s.t. } \|f(\boldsymbol{x}_i)\|_2 = \alpha, \quad \forall i = 1, 2, \ldots, N,$$

where $\boldsymbol{x}_i$ is the input image, N is the mini-batch size, W is the weight matrix of the fully-connected layer, $f(\boldsymbol{x}_i)$ is the extracted feature vector, C is the number of classes, α is a hyperparameter, W_{y_i} is the column of W corresponding to the ground-truth label y_i, and b_{y_i} is the bias term.

4. Experiments and Discussion

This section describes experiments to demonstrate the effectiveness of the baseline method in detecting face spoofing attacks.

4.1. Evaluation Protocols and Metrics

We conduct cross-dataset evaluations to measure the generalization capability of the FAS methods to unseen domains. We follow the widely adopted MICO protocol involving four benchmark datasets: MSU MFSD [10] (M), IDIAP Replay Attack [9] (I), CASIA-FASD [57] (C), and OULU-NPU [4] (O). This protocol employs a leave-one-out strategy, where the model is trained on three datasets and tested on the remaining unseen dataset. To further evaluate robustness under data-constrained conditions, we also conduct a Limited Source Domain (LSD) evaluation, using only two datasets for training. Performance is evaluated by the Half Total Error Rate (HTER) and the Area Under the ROC Curve (AUC).

4.2. Datasets and Preprocessing

The datasets used for evaluation are the four aforementioned datasets for cross-dataset testing. All video frames are preprocessed to extract the face region, as the input requires single images focused on the face. For all four datasets used in the cross-dataset evaluation, Dlib [55] is utilized for face detection. All extracted face regions are resized to 224×224 pixels, and pixel values are normalized to have zero mean and unit variance per channel. The FAS methods used in the experiments method assumes a single image input rather than a video sequence. Therefore, frames are sampled from the videos. For the cross-dataset evaluation, a total of 5 frames are sampled at a fixed interval.

4.3. Implementation Details

The proposed baseline is fine-tuned using the AdamW [34] optimizer, with training running for 200 epochs and utilizing early stopping based on a patience of 20 epochs. The mini-batch size is set to 32. In the proposed baseline, DINOv2 ViT-B/14 with Registers [11] pre-trained model is fine-tuned using the training data specified by the evaluation protocol. Training inputs are augmented with FAS-Aug [5] ($P = 0.2$), PDA [52] ($P = 0.2$), and standard augmentations (Random Horizontal Flip, Random Rotation, and Random Brightness). The learning rates are set as 5×10^{-5} for

Table 1. Comparison of various vision foundation models as feature extractors for FAS. Evaluation metrics are reported in AUC [%]↑. Models are categorized into: Supervised CNNs, Supervised ViTs, and Self-Supervised & Multi-modal ViTs. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	CIO→M ↑	OMI→C ↑	OCM→I ↑	ICM→O ↑	Avg. ↑
ConvNeXt [33]	95.95	95.78	84.87	84.26	90.22
ViT-B [12]	91.81	96.82	**95.30**	91.96	93.97
EfficientFormer [24]	93.67	93.55	81.61	84.55	88.35
MobileViT [36]	96.73	94.99	82.92	90.80	91.36
DeiT-tiny [47]	90.10	69.40	68.09	78.63	76.56
DeiT-base [47]	88.82	83.21	77.36	75.17	81.14
Swin Transformer [32]	97.20	<u>98.78</u>	92.44	93.42	95.46
Data2vec [1]	88.82	95.13	77.36	93.47	88.70
BEiT [2]	97.25	95.80	82.24	88.89	91.05
MAE [16]	94.41	90.00	84.06	91.84	90.08
CLIP [40]	**98.52**	94.95	92.20	95.95	95.41
DINO [6]	97.01	97.00	91.12	94.80	94.98
DINOv2 [37]	96.90	98.02	91.05	<u>96.23</u>	<u>95.55</u>
DINOv2 with Registers [11]	<u>97.47</u>	98.30	<u>93.83</u>	95.38	**96.25**
DINOv3 [44]	95.42	**98.80**	89.21	**97.13**	95.14

the classification head and 5×10^{-6} for the DINOv2 encoder.

4.4. Experiments and Discussion

We conduct four sets of experiments to validate our baseline: (i) Comparison of vision foundation models to identify the most effective feature extractor; (ii) Cross-dataset evaluation under the MICO protocol; (iii) LSD evaluation; and (iv) Analysis of computational efficiency.

(i) Comparison of Vision Foundation Models: In this experiment, we evaluate the effectiveness of various pre-trained backbones to determine the optimal feature extractor for FAS. To ensure a fair comparison of the inherent representation power of backbones, FAS-Aug, PDA and $\mathcal{L}_{APL}$ were not applied in this specific set of experiments. We compare models categorized by their pre-training paradigms as defined in Sect. 3.1:

- Supervised Models: ConvNeXt [33] (CNN), ViT-B [12], EfficientFormer [24], MobileViT [36], DeiT [47], and Swin Transformer [32].
- Self-Supervised & Multi-modal Models: Data2vec [1], BEiT [2], MAE [16], CLIP [40], DINO [6], DINOv2 [37], DINOv2 with Registers, and DINOv3 [44].

For architectures without a CLS token (ConvNeXt, MobileViT, Swin Transformer), the average pooling of all patch tokens from the final layer is fed into the classification head. Table 1 summarizes the results under the MICO protocol. The results indicate that self-supervised models generally outperform supervised counterparts. This result suggests that self-supervised pre-training yields features with higher transferability and is more effective at capturing discriminative cues for spoofing detection across different datasets. Among the self-supervised models, DINOv2 with Registers demonstrates the most robust and consistent domain generalization. While DINOv3 achieves the highest accuracy in specific scenarios (OMI→C and ICM→O), its performance suffers a significant drop in the challenging OCM→I protocol. In contrast, DINOv2 with Registers maintains high accuracy across all protocols without severe degradation. As indicated by the overall average AUC, DINOv2 with Registers achieves the highest mean performance (96.25%), effectively outperforming both DINOv3 (95.14%) and CLIP (95.41%). This consistency confirms its ability to extract robust, generalized features essential for reliable face anti-spoofing, thereby justifying its selection as the core architecture for our baseline.

(ii) Cross-dataset Evaluation: We perform cross-dataset evaluation using the MICO protocol to validate the domain generalization capability of the baseline method, with results summarized in Table 2. The methods are grouped into conventional vision-only approaches (first group), the baseline itself (second group), and VLM-based SOTA methods (third group). Compared with conventional vision-only methods, ours shows overall competitive performance across most transfer settings, though not uniformly the best. In CIO→M, it achieves an AUC comparable to several strong image-only approaches and close to the top-

CVPR #24

Table 2. Experimental results for MICO [%]. The best and second-best results among vision-only methods are highlighted in **bold** and underline, respectively. VLM-based multimodal methods are listed at the bottom for reference.

Method	CIO→M		OMI→C		OCM→I		ICM→O		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
DRDG [29]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75	15.67	92.04
ANRL [28]	10.83	96.75	17.83	89.26	16.03	91.04	15.67	91.90	15.09	92.24
SSDG-R [18]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.29	95.31
SSAN-R [48]	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.82	96.46
PatchNet [48]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.91	95.95
TransFAS [51]	7.08	96.69	9.81	96.13	10.12	95.53	15.53	91.10	10.64	94.86
DiVT-M [25]	**2.86**	**99.14**	8.67	96.92	**3.71**	**99.29**	13.06	94.04	7.08	97.35
SA-FAS [46]	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	7.83	96.42
IADG [59]	5.41	98.19	8.70	96.44	10.62	94.50	8.86	97.14	8.40	96.57
GAC-FAS [21]	5.00	97.56	8.20	95.16	4.29	98.87	8.60	97.16	**6.52**	97.19
Li [22]	12.92	94.33	9.26	96.98	10.87	95.46	15.13	91.43	12.05	94.55
DiffFAS-V [15]	**2.86**	98.41	10.11	96.32	6.36	97.89	8.11	97.27	6.86	97.47
Baseline	8.86	96.95	**4.49**	**98.92**	9.81	96.70	**7.35**	**98.07**	7.63	**97.66**
FLIP [45]	4.95	98.11	0.54	99.98	4.25	99.07	2.31	99.63	3.01	99.20
CFPL [27]	1.43	99.28	2.56	99.10	5.43	98.41	2.50	99.42	2.98	99.05
I-FAS [56]	0.32	99.88	0.04	99.99	3.22	98.48	1.74	99.66	1.33	99.50

performing methods in this group. On OMI→C, it exceeds the reported results of conventional methods under this protocol. For OCM→I, however, our baseline has slightly lower performance than several recent approaches. This suggests that while DINOv2 with Registers provides solid cross-domain generalization, there remains room for improvement under this particular domain shift. On ICM→O, it also exceeds the reported results of conventional methods under this protocol. Overall, these results indicate that this method offers stable and competitive performance across different cross-dataset scenarios, without consistently dominating every setting. When compared to VLM-based SOTA methods (FLIP, CFPL, and I-FAS), it remains competitive despite its simpler design. While these large-scale VLM approaches achieve the highest AUC scores (all above 99% in several settings), the baseline method narrows the gap substantially, for example, 98.92% vs. 99.99% on OMI→C and 98.07% vs. 99.66% on ICM→O, without relying on extremely large multimodal backbones. Beyond direct performance comparisons, these state-of-the-art VLM-based methods primarily employ the CLIP image encoder for visual feature extraction. As demonstrated in our benchmarking results (Table 1), self-supervised models like DINOv2 with Registers capture more robust and fine-grained visual cues than CLIP, even without relying on semantic alignment. Therefore, our established vision-only baseline is not only an effective independent approach but also has the potential to serve as a superior visual backbone for these multimodal frameworks. Replacing their existing image encoders with our optimized feature extractor could further enhance the performance of future VLM-based FAS systems. Overall, the results indicate that this baseline provides a strong domain-invariant feature representation. Despite not leveraging massive VLM architectures, it achieves highly competitive cross-dataset performance, demonstrating that carefully designed vision-only models can approach the performance of large VLM-based systems while maintaining architectural simplicity and efficiency.

(iii) Limited Source Domain: We further evaluate the robustness under resource constraints using the limited source domain setting of the MICO protocol. This challenging configuration uses only two source datasets (MSU-MFSD and Idiap Replay-Attack) for training, testing the model's ability to generalize from severely restricted domain diversity. Table 3 shows the experimental results. While prior vision-only methods exhibit a severe performance drop under this setting, our approach maintains exceptional accuracy. We achieve state-of-the-art results among vision-only methods, surpassing all previous works (e.g., HTER 8.29% and AUC 97.10% on MI→C). This superiority is obtained despite using a significantly light architecture. This result strongly validates the effectiveness of our vision-only method in capturing domain-invariant cues and adapting to unseen conditions, a key property for resource-constrained deployment.

Table 3. Experimental results for limited source domains [%]. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	MI→C		MI→O	
	HTER↓	AUC↑	HTER↓	AUC↑
DRDG [29]	31.28	71.50	33.35	69.14
ANRL [28]	31.06	72.12	30.73	74.10
SSDG-M [18]	31.89	71.29	36.01	66.88
SSAN-R [48]	30.00	76.20	29.44	76.62
DiVT-M [25]	20.11	86.71	23.61	85.73
IADG [59]	24.07	85.13	18.47	90.49
GAC-FAS [21]	16.91	88.12	17.88	89.67
FAS-Aug [5]	16.89	90.06	15.10	92.69
DiffFAS-V [15]	15.06	92.83	16.19	92.62
Baseline	**8.29**	**97.10**	**12.11**	**95.36**

Table 4. Parameter count for different methods. The lowest parameter counts are highlighted in **bold**.

Method	Trainable Params [M]	Total Params [M]
CFPL [27]	94	157
FLIP [45]	170	170
I-FAS [56]	104	3,100
Baseline	**87**	**87**

(iv) Parameter Count and Computational Complexity: A comparison of model sizes across recent state-of-the-art FAS methods as seen in Table 4 clearly illustrates the substantial parameter efficiency of our DINOv2-based approach. I-FAS [56] relies on a large-scale multimodal architecture that integrates a CLIP ViT-L/14 image encoder (approximately 304M parameters) with the OPT-2.7B language model (approximately 2.7B parameters), resulting in a total model size of roughly 3.1B parameters. Although only 104M parameters (the GAC module) are trainable during optimization, the full architecture must still be stored and executed during inference, leading to considerable computational and memory overhead. CFPL [27] adopts a more compact multimodal design built upon CLIP ViT-B (approximately 86M parameters) and its 63M-parameter text encoder, augmented with two lightweight Q-Formers (approximately 3.5M parameters each). This results in a total model size of approximately 157M parameters, with 94M trainable parameters during training. Similarly, FLIP [45] employs the same CLIP ViT-B backbone and text encoder, reaching approximately 150M-170M parameters during training. While FLIP reduces inference complexity by retaining only the 86M-parameter image encoder after precomputing text embeddings, its training process still depends on a multimodal framework. In contrast, our DINOv2 with Registers method adopts a purely vision-based architecture without any language model or auxiliary text encoder. Using DINOv2 ViT-B/14 with Registers (approximately 86M parameters), our approach achieves competitive cross-domain generalization performance while maintaining a total parameter count of only 87M. This corresponds to less than 3% of the parameters required by I-FAS and roughly 55% of CFPL's full architecture, while remaining comparable to FLIP's inference-time model size. Importantly, our results demonstrate that similar generalization performance can be achieved without the substantial parameter overhead inherent to vision-language models. By relying solely on self-supervised visual representations, our framework avoids the architectural complexity, memory footprint, and computational burden associated with multimodal large language models, highlighting that effective FAS generalization does not necessitate a VLM-based design.

5. Conclusion

In this paper, we revisited the potential of vision-only foundation models for FAS to address the computational limitations of recent VLMs. Through a comprehensive benchmarking of 15 diverse pre-trained vision models under severe cross-domain scenarios, we demonstrated that self-supervised vision models possess superior domain generalization capabilities. Based on these insights, we established a robust and highly efficient vision-only baseline utilizing DINOv2 with Registers combined with FAS-Aug, PDA, and APL. Extensive experiments demonstrated that our baseline not only achieves competitive performance under the standard MICO protocol but also establishes a new state-of-the-art in data-constrained LSD scenarios. The results also indicated that an optimized self-supervised vision transformer provides a highly practical standalone solution for resource-constrained applications, while also offering a potent visual backbone to further advance future multimodal FAS systems.

References

[1] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *Proc. Int'l Conf. Machine Learning*, pages 1298–1312, 2022. 4, 6

[2] H. Bao, L. Dong, S. Piao, and F. Wei. BEiT: BERT pre-training of image transformers. *Proc. Int'l Conf. Learning Representations*, pages 1–13, 2022. 4, 6

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2016. 2

[4] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid.

OULU-NPU: A mobile face presentation attack database with real-world variations. *IEEE Int'l Conf. Automatic Face Gesture Recog.*, pages 612–618, 2017. 2, 5

[5] R. Cai, C. Soh, Z. Yu, H. Li, W. Yang, and A. Kot. Towards data-centric face anti-spoofing: Improving cross-domain generalization via physics-based data synthesis. *Int. J. Comput. Vis.*, pages 1–22, 2024. 2, 3, 4, 5, 8

[6] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, and P. Bojanowski. Emerging properties in self-supervised vision transformers. *Int. Conf. Comput. Vis.*, pages 9650–9660, 2021. 3, 4, 6

[7] X. Chen, Y. Jia, and Y. Wu. Fine-grained annotation for face anti-spoofing. *CoRR*, abs/2310.08142, 2023. 1, 2

[8] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. Jin. Generalizable representation learning for mixture domain face anti-spoofing. *AAAI*, pages 1132–1139, 2021. 2

[9] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. *Int. Conf. Biometrics Special Interest Group*, pages 1–7, 2012. 2, 5

[10] A. K. Jain D. Wen and H. Han. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, pages 746–761, 2015. 5

[11] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformer needs register. *Int. Conf. Learn. Represent.*, 2024. 2, 3, 4, 5, 6

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent.*, 2021. 1, 3, 6

[13] M. Feng, Gallin-Martel P. A., K. Ito, and T. Aoki. Optimizing DINOv2 with registers for face anti-spoofing. *Int. Conf. Comput. Vis. Worksh.*, pages 3256–3262, 2025. 1, 2, 3

[14] M. Feng, K. Ito, T. Aoki, T. Ohki, and M. Nishigaki. Leveraging intermediate features of vision transformer for face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 3464–3472, 2025. 1, 2

[15] X. Ge, Yu Z. Liu, X. and, J. Shi, C. Qi, J. Li, and H. Kälviäinen. DiffFAS: Face anti-spoofing via generative diffusion models. *Eur. Conf. Comput. Vis.*, pages 144–161, 2024. 7, 8

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 16000–16009, 2022. 4, 6

[17] X. He, D. Liang, S. Yang, Z. Hao, H. Ma, M. Binjie, X. Li, Y. Wang, P. Yan, and A. Liu. Joint physical-digital facial attack detection via simulating spoofing clues. *IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 995–1004, 2024. 1

[18] Y. Jia, J. Zhang, S. Shan, and X. Chen. Single-side domain generalization for face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 8484–8493, 2020. 7, 8

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, and R. Girshick. Segment anything. *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 2

[20] J. Komulainen, A. Hadid, and M. Pietikäinen. Context based face anti-spoofing. *Int. Conf. Biometrics: Theory, Applications and Systems*, pages 1–8, 2013. 2

[21] B. Le and S. Woo. Gradient alignment for cross-domain face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 188–189, 2024. 2, 7, 8

[22] D. Li, G. Chen, X. Wu, Z. Yu, and M. Tan. Face anti-spoofing with cross-stage relation enhancement and spoof material perception. *Neural Networks*, 175:106275, 2024. 1, 7

[23] S. Li and A. Jain. *Handbook of Face Recognition*. Springer, 2011. 1

[24] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren. EfficientFormer: Vision transformers at MobileNet speed. *Advances in Neural Information Processing Systems*, pages 12934–12949, 2022. 3, 6

[25] C. Liao, W. Chen, H. Liu, Y. Yeh, M. Hu, and C. Chen. Domain invariant vision transformer learning for face anti-spoofing. *IEEE/CVF Winter Conf. Applications of Comput. Vis.*, pages 6087–6096, 2023. 2, 7, 8

[26] K. Lin, Y. Tseng, K. Huang, J. Wu, and W. Cheng. InstructFLIP: Exploring unified vision-language model for face anti-spoofing. *ACM Int. Conf. Multimedia*, pages 2987 –2996, 2025. 2

[27] A. Liu, S. Xue, J. Gan, J. Wan, Liang Y., J. Deng, S. Escalera, and Z. Lei. CFPL-FAS: Class free prompt learning for generalizable face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 222–232, 2024. 2, 7, 8

[28] S. Liu, K. Zhang, T. Yao, M. Bi, S. Ding, J. Li, M. Huang, and L. Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. *ACM Int. Conf. Multimedia*, pages 1469–1477, 2021. 2, 7, 8

[29] S. Liu, K. Zhang, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, Y. Xie, and L. Ma. Dual reweighting domain generalization for face presentation attack detection. *IJCAI*, pages 867–873, 2021. 2, 7, 8

[30] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 389–398, 2018. 2

[31] Y. Liu, Y. Chen, W. Dai, M. Gou, C. Huang, and H. Xiong. Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. *Eur. Conf. Comput. Vis.*, pages 511–528, 2022. 2

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. *Proc. IEEE/CVF Int'l Conf. Computer Vision*, pages 10012–10022, 2021. 3, 6

[33] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11976–11986, 2022. 3, 6

[34] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. Int. Conf. Learn. Represent., 2019. 5

[35] S. Marcel, J. Fierrez, and N. Evans. *Handbook of Biometric Anti-Spoofing*. Springer, 2023. 1, 2

[36] S. Mehta and M. Rastegari. MobileVit: Light-weight, general-purpose, and mobile-friendly vision transformer. *Proc. Int'l Conf. Learning Representations*, pages 1–13, 2022. 3, 6

[37] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *Trans. Machine Learning Research*, 2024. 3, 4, 6

[38] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Inf. Forensics Secur.*, 11(10):2268–2283, 2016. 2

[39] T. F. Pereira, A. Anjos, J. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? *Int. Conf. Biometrics*, pages 1–8, 2013. 2

[40] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *Proc. Int'l Conf. Machine Learning*, pages 8748–8763, 2021. 2, 4, 6

[41] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *CoRR*, abs/1703.09507, 2017. 5

[42] R. Shao, X. Lan, J. Li, and P. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 10023–10031, 2019. 2

[43] R. Shao, X. Lan, and P. C. Yuen. Regularized fine-grained meta face anti-spoofing. *AAAI*, 34(7):11974–11981, 2020. 2

[44] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. DINOv3. *CoRR*, abs/2508.10104:1–52, 2025. 4, 6

[45] K. Srivatsan, M. Naseer, and K. Nandakumar. FLIP: Cross-domain face anti-spoofing with language guidance. *Int. Conf. Comput. Vis.*, pages 19685–19696, 2023. 2, 7, 8

[46] Y. Sun, Y. Liu, X. Liu, Y. Li, and W. Chu. Rethinking domain generalization for face anti-spoofing: separability and alignment. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 24563–24574, 2023. 2, 7

[47] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. *Proc. Int'l Conf. Machine Learning*, pages 10347–10357, 2021. 3, 6

[48] C. Wang, Y. Lu, S. Yang, and S. Lai. PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 20281–20290, 2022. 1, 2, 7, 8

[49] G. Wang, H. Han, S. Shan, and X. Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6677–6686, 2020. 2

[50] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, G. Qi, J. Wan, and Z. Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *CoRR*, abs/1811.05118:1–15, 2018. 2

[51] Z. Wang, Q. Wang, W. Deng, and G. Guo. Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Trans. Biom. Behav. Identity Sci.*, 4(3):439–450, 2022. 1, 2, 7

[52] K. Watanabe, K. Ito, and T. Aoki. Spoofing attack detection in face recognition system using vision transformer with patch-wise data augmentation. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, pages 1561–1565, 2022. 2, 3, 4, 5

[53] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao. NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(9):3005–3023, 2020. 1, 2

[54] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao. Searching central difference convolutional networks for face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 5295–5305, 2020. 1, 2

[55] D. Zhang, J. Li, and Z. Shan. Implementation of dlib deep learning face recognition technology. *Int. Conf. Robots & Intelligent System*, pages 88–91, 2020. 5

[56] G. Zhang, K. Wang, H. Yue, A. Liu, G. Zhang, K. Yao, E. Ding, and J. Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. *AAAI*, pages 9896–9904, 2025. 2, 7, 8

[57] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. *Int. Conf. Biometrics*, pages 26–31, 2012. 5

[58] T. Zheng, B. Li, S. Wu, B. Wan, G. Mu, S. Liu, S. Ding, and J. Wang. MFAE: Masked frequency autoencoders for domain generalization face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, pages 4058–4069, 2024. 1, 2

[59] Q. Zhou, K. Zhang, T. Yao, X. Lu, R. Yi, S. Ding, and L. Ma. Instance-aware domain generalization for face anti-spoofing. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 20453–20463, 2023. 2, 7, 8