This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Advancing Cross-Domain Generalizability in Face Anti-Spoofing: Insights, Design, and Metrics

Hyojin Kim[1], Jiyoon Lee[1,2], Yonghyun Jeong[1], Haneol Jang[3], YoungJoon Yoo[1,4†]
[1]Naver Cloud,[2]Korea University, [3]Hanbat National University, [4]Chung-Ang University
hyojin.kimm@navercorp.com, jiyoonlee@korea.ac.kr, yonghyun.jeong@navercorp.com,
hejang@hanbat.ac.kr, yjyoo3312@cau.ac.kr

Abstract

This paper presents a novel perspective for enhancing anti-spoofing performance in zero-shot data domain generalization. Unlike traditional image classification tasks, face anti-spoofing datasets display unique generalization characteristics, necessitating novel zero-shot data domain generalization. One step forward to the previous frame-wise spoofing prediction, we introduce a nuanced metric calculation that aggregates frame-level probabilities for a video-wise prediction, to tackle the gap between the reported frame-wise accuracy and instability in real-world use-case. This approach enables the quantification of bias and variance in model predictions, offering a more refined analysis of model generalization. Our investigation reveals that simply scaling up the backbone of models does not inherently improve the mentioned instability, leading us to propose an ensembled backbone method from a Bayesian perspective. The probabilistically ensembled backbone both improves model robustness measured from the proposed metric and spoofing accuracy, and also leverages the advantages of measuring uncertainty, allowing for enhanced sampling during training that contributes to model generalization across new datasets. We evaluate the proposed method from the benchmark OMIC dataset and also the public CelebA-Spoof and SiW-Mv2. Our final model outperforms existing state-of-the-art methods across the datasets, showcasing advancements in Bias, Variance, HTER, and AUC metrics.

1. Introduction

In response to the widespread adoption of deep learning in face recognition, detecting spoofing attacks like printed or video-displayed faces, also called presentation attacks, to the recognition system has become paramount for security. Although sophisticatedly spoofed images often challenge human detection capabilities, the advances of recent face anti-spoofing (FAS) studies [26, 27, 37, 39, 42] demonstrate that deep classification networks can effectively differentiate between authentic and counterfeit facial images. Nonetheless, developing a broadly generalized spoofing detection model, capable of accommodating diverse conditions, is still a challenging problem due to the inherent complications in spoofing datasets.

Specifically, each spoofing dataset possesses distinct attributes, stemming from varying data acquisition environments like camera capture conditions and backgrounds, introducing notable dataset biases. Given that visual cues for spoofing predominantly reside in the nuanced high-frequency image domain [2], these biases significantly impede the extraction of reliable spoofing detection cues, thereby compromising the generalizability of FAS models. Recent Face Anti-Spoofing (FAS) research [13, 16, 24, 25, 27, 32–34] employs evaluation metrics to showcase model generalizability over four benchmark anti-spoofing datasets, one step further to the cross dataset-domain adaptation (DA) [12, 15, 40, 46] for FAS problem. The domain generalization (DG) performance is evaluated by training on three of these datasets and subsequently evaluating the model's performance on the remaining one. While the evaluation is a standard measure of FAS DG performance, real-world application of existing per-frame FAS methods to video inputs often yields unstable predictions across consecutive frames.

Our analysis begins with the observations as:

- Previous frame-wise FAS models successful for the existing FAS measurement fail to robustly capture the spoofness of the subsequent frames sharing similar semantic features.
- Scaling up of the model size cannot effectively enhance the overall FAS performance including the robustness issue our first observations. Such observations underscore the necessity of developing a novel metric to gauge model generalizability and a proper method to design a FAS model, and second, effective model design to improve

†Corresponding author.

FAS performance beyond scaling up the model size.

In this paper, we present a novel perspective on a paradigm for evaluating and designing FAS models. First, we introduce a novel evaluation metric, variance, and bias, focusing on temporal coherence and noise resilience for robustness assessment. Our frame-level analysis validates these metrics, providing quantitative evidence supporting the instability of prior FAS methods. Second, we highlight the efficacy of ensemble methodologies in enhancing spoofing detection performance for both the previous and the newly introduced evaluation protocols. Through extensive empirical evaluations, we demonstrate that the ensemble strategy applying Monte-Carlo (MC) dropout [8] is a pivotal design option for making a FAS model that enhances precise spoofing detection and superior generalizability. Last, through empirical analyses to determine the optimal backbone for MC-dropout-based ensembling, we introduce ECLIPS, a FAS model based on the CLIP vis encoder [22]. Our proposed ECLIPS achieves state-of-the-art FAS detection performance while maintaining frame-level prediction consistency. It is worth noting that, instead of processing consecutive video frames, which incurs high computational costs, we demonstrate the efficacy of the MC-dropout-based ensembling method in improving frame-level FAS prediction consistency.

In demonstrating the effectiveness of our proposed approach, we conduct comprehensive qualitative and quantitative experiments across four benchmark FAS datasets: OULU [1] (**O**), CASIA [45] (**C**), Idiap [4] (**I**), MSU-MFSD [35] (**M**), and additionally to the CelebA-Spoof [43] and SiW-Mv2 [20] datasets. The experimental results demonstrate that our ensemble-based FAS model ECLIPS outperforms current state-of-the-art FAS models by a notable margin including recent FAS methods leveraging auxiliary multi-modality both for FAS accuracy and for the frame-level prediction robustness validated through the proposed protocol.

In summary, we summarize the contributions of the proposed method as follows:

- We provide observations for real-world FAS applications for video input, with a focus on frame-level prediction robustness and model scalability, while introducing the remaining challenges in current FAS research.
- We introduce temporal and noise-aware robustness metrics, by calculating variance and bias, demonstrating their potential as reliable indicators for ensuring the generalizability of FAS models.
- Through empirical analyses, we highlight the proposed ensemble approach as a potent design strategy for FAS models. Consequently, our proposed FAS model ECLIPS reports state-of-the-art performance, excelling in FAS generalization capabilities.

2. Related Work

Domain Generalization for FAS To deal with the well-known dataset disparities inherent to the FAS, many FAS studies have been proposed, emphasizing cross-dataset domain (cross-domain) evaluations via domain adaptation (DA) [12, 15, 40, 46], domain generalization (DG) [13, 16, 24, 25, 27, 32–34], or both [26]. While the DA paradigm allows for few-shot adaptation to the target domain, DG offers a more rigorous assessment by gauging zero-shot generalization across datasets. Given real-world scenarios where DA might be infeasible, the DG approach gains prominence due to its straightforward applicability. After the initiative DG approach [24] proposing Mult-adversarial domain generalization, many follow-up studies have been proposed by meta-laerning [25], style-shuffling [33], new losses [16, 27, 34] based on dataset analyses, and multi-modaility [26]. These studies typically evaluate the generalization ability of their methods by using four representative benchmark datasets: OULU [1], CASIA [45], Idiap [4], and MSU-MFSD [35], leaving one dataset for evaluation from the model trained by the other three datasets. Advancing beyond the previous work, we demonstrate that even with consistent improvements in conventional evaluation frameworks, the reported outcomes might not assuredly reflect detection proficiency across diverse datasets, as evidenced by our newly introduced FAS dataset.

Designing Principles for FAS Followed by the improvement of deep classification architecture [7, 11], we have observed consistent enhancement in the detection capability of the presentation attack (PA) samples by using the classification architecture as their backbone. Most previous FAS methods [16, 25, 27, 29, 33, 34, 36] have configured the PA detection model as a form of classification network using ImageNet [6] pre-trained ResNet [11]-like architecture. Instead of scaling up the backbone, FAS methods have focused on the data domain-specific improvements, such as analyzing the effect of input configuration [29, 36] of feature space manipulation [27, 34], based on the fixed ResNet backbone. A notable size and performance scale-up of the architecture is followed by the use of vision transformer (ViT) [7] pre-trained by large-scale image and text multi-modal dataset [23] as CLIP visual encoder. Followed by the initiative attempts [9, 16] using ImageNet-pre-trained ViT for FAS, adoption of ViT pre-trained as CLIP [22] visual encoder reported impressive performance enhancement [26].

In this paper, we investigate the effect of backbone scale-up in various aspects including depth, channel, and pre-trained dataset on DG scenario accompanied by the novel robustness evaluation metrics and show the efficacy of the ensembling-based method. Beyond previous attempts [14, 21] using the ensembling method for frame-

Sample No.	Frame-wise						Video-wise		
GT Label = 1	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Avg. Prob.	Prob.	Bias	Variance
ResNet18 [11]	0.97(1)	0.33(0)	0.52(1)	0.17(0)	0.98(1)	0.59(1)	0.79(1)	0.044	0.40
ViT-B/16 [7]	0(0)	0.55(1)	0.98(1)	0(0)	0.17(0)	0.4(0)	0.34(0)	0.435	0.31
FLIP [26]	0.93(1)	0.98(1)	0.96(1)	0.54(1)	0.82(1)	1.0(1)	0.84(1)	0.025	0.14
ECLIPS	0.99(1)	0.98(1)	0.97(1)	0.88(1)	0.99(1)	1.0(1)	0.96(1)	0.001	0.05

Table 1. Quantitative analysis of prediction variability in frame-by-frame images from commonly used backbone models [7, 11]. The numerical values within frame-wise indicate the prediction probabilities for each frame, where the numbers in parentheses represent a decision of spoofing (0) or live (1). In the video-wise results, the average probability (Prob.) of all frames is presented, with the overall video prediction denoted within the parentheses. Additionally, the results include calculations of bias and variance.

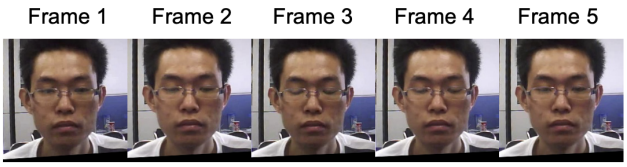

Figure 1. Visualization of images containing a sequence of five frames extracted from the video within the CASIA [45]. The images demonstrate a range of facial expressions and pose variations.

skipping [21] and integrating data-domain specific models [14], the ensemble approach as a pivotal mechanism for theoretically addressing DG in the FAS, underpinned by foundational FAS insights and corroborative analyses.

3. Observations

In this section, we initiate our discussion of designing a generalizable FAS model grounded in observations from FAS evaluations concerning *temporal robustness* and the *scalability* of the prediction. To quantitatively measure the temporal robustness, we introduce the terms Variance and Bias.

3.1. Temporal Robustness

As shown in Figure 1, the provided image shows a sequence of five frames from a video, labeled as Frame 1 through Frame 5. Each frame features the same individual, capturing subtle changes in facial expression and pose, which are critical for evaluating the consistency and robustness of face anti-spoofing models across different moments in time. This sequence demonstrates how frame-wise analysis can reveal variations in the model's prediction accuracy, thus emphasizing the importance of a model's ability to maintain stable performance throughout a video sequence. Table 1 provides a quantitative analysis corresponding to the images shown in Figure 1. The figure illustrates sequence moments within the same video, while Table 1 evaluates the performance of various models at each respective moment.

The computational expressions for these methodologies are outlined as follows: Frame-wise probability refers to the probability score assigned to an individual frame, indicating whether the frame is likely live or a spoof. Frame-wise prediction is expressed verbally as the average of outcomes where each outcome is 1 if the probability score of the frame exceeds a certain threshold and 0 otherwise. This translates to a binary classification where scores above the threshold indicate a live frame and scores below indicate a spoof. Video-wise probability is explained as the average of the probability scores of all frames within a video being compared against a threshold to decide the overall nature of the video. Video-wise Prediction involves making a binary decision, which is determined as live (1) if the average probability score of all frames exceeds the threshold and as spoof (0) if it does not.

Concerning bias and variance within these metrics, they are defined as follows:

- **Bias** $B(\cdot)$ is quantified by the mean squared error (MSE) between the actual labels and the video-wise probability scores. This measures how accurately the video-wise predictions reflect the true outcomes.

$$B(Y_i) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2. \tag{1}$$

  In this formula, $\mathcal{N}$ represents the total number of videos, Y_i is the actual label for the ith video (live or spoof), and $\hat{Y}_i$ is the predicted video-wise probability for the ith video.

- **Variance** is calculated as the average of the standard deviations of the frame-wise probability scores across all videos. This standard deviation reflects the consistency of frame-wise probability scores within a video, providing insights into the model's reliability across different frames.

$$\begin{aligned} V(P_i) &= \frac{1}{N}\sum_{i=1}^{N}\sigma^2(P_i), \\ \sigma^2(P_i) &= \frac{1}{M_i}\sum_{j=1}^{M_i}(P_{ij} - \bar{P}_i)^2. \end{aligned} \tag{2}$$

  The Variance $V(\cdot)$ is computed as the mean of the stan-

Method (%)	FAS	Backbone Architecture	MBParams	GFLOPs	Size or Latency	OCI→M		OMI→C		OCM→I		ICM→O	
						HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)
ResNet [11]		ResNet18	42.63	2.38	2.51	10.24	96.15	14.72	92.76	17.00	91.72	15.03	92.34
		ResNet34	81.19	4.80	3.89	**8.57**	**96.22**	**11.93**	**94.51**	18.00	91.13	**13.11**	93.31
		ResNet50	89.68	5.40	7.06	10.24	93.97	13.94	93.97	**12.95**	**94.46**	13.70	**93.70**
EfficientNet-V2 [28]		Tiny	47.71	2.50	19.19	11.67	94.24	16.16	92.64	20.45	83.42	16.09	91.84
		Small	83.88	3.84	20.47	**8.81**	**95.86**	**12.71**	**93.12**	22.50	86.46	**15.91**	**92.00**
		Medium	193.70	8.08	30.87	12.86	94.99	22.50	78.87	**14.83**	**93.48**	17.95	90.16
ViT [7]		Tiny/16	20.93	1.41	5.40	11.43	95.68	15.94	91.16	21.90	**88.94**	**16.07**	90.82
		Small/16	82.36	5.54	5.45	**8.57**	96.08	14.72	92.77	**20.00**	85.32	14.44	**91.61**
		Base/16	326.72	22.00	6.31	10.24	**96.10**	**13.94**	**93.91**	21.00	78.32	17.99	89.38
SAFAS [27]	✓	ResNet18	42.63	2.38	2.51	10.24	**95.97**	**12.04**	**94.28**	**8.90**	96.96	11.40	95.08
	✓	ResNet34	81.19	4.80	3.89	**8.33**	95.94	12.04	94.28	9.05	**97.79**	**10.24**	**95.99**
	✓	ResNet50	89.68	5.40	7.06	11.19	95.53	13.38	94.19	10.50	96.31	11.25	95.56
PatchNet [29]	✓	ResNet18	42.63	2.38	7.06	**10.24**	93.9	16.05	90.05	22.00	88.12	**17.22**	88.53
	✓	ResNet34	81.19	4.80	3.89	11.67	95.13	15.94	89.33	**14.05**	**93.48**	17.80	**88.90**
	✓	ResNet50	89.68	5.40	7.06	11.43	**95.85**	**13.38**	**93.25**	19.05	90.03	18.10	88.26

Table 2. Comparative experiment of FAS methodologies for **HTER** (Half Total Error Rate) and **AUC** (Area Under the Curve) with respect to model scale variation. The overall quantitative results show that scaling-up of the model size does not yield substantial improvements in FAS applications.

dard deviations (denoted by σ) of frame-wise probability scores across all videos. M_i denotes the total number of frames in the ith video, P_{ij} represents the probability score of the jth frame in the i video, and $\bar{P}_i$) is the mean of the frame-wise probability scores for the ith video.

This distinction between frame-wise and video-wise methodologies not only highlights their unique analytical perspectives but also underlines the complexities involved in assessing the reliability and precision of video-based live detection systems. Through the evaluation of bias and variance, a deeper understanding of the model's performance is achieved, allowing for a balance between accuracy and consistency across various videos and frames.

Table 1 quantitatively illustrates the model's performance in terms of bias and variance across individual frames within a single video, highlighting the disparity in frame-wise predictions. For each model, The Frame-wise section in Table 1 displays the probability scores (with predicted labels in parentheses) for five representative frames and the average probability of prediction (Avg. Prob.) across these frames. The Video-wise section further quantifies the overall prediction probability (Prob.), bias, and variance, offering insights into the model's stability and generalization capability. Notably, the our proposed model demonstrates remarkably low variance in frame-wise predictions, signifying consistent performance across the video's duration, which is crucial for reliable face anti-spoofing in dynamic scenarios.

3.2. Scalability

Although it is generally observed that larger models tend to exhibit higher performance, there are scenarios where this does not hold true. Given the datasets commonly used in the FAS task (OULU [1], CASIA [45], Idiap [4], MSU-MFSD [35]), characterized by its relatively small scale, it becomes imperative to validate performance across various backbone scales. This necessity stems from the nuanced understanding that while larger models have the potential to achieve higher accuracies, their effectiveness can be constrained by factors such as overfitting, increased computational requirements, and the curse of dimensionality, especially in the context of limited data availability. Therefore, we conducted experiments to validate the performance in terms of the scalability of various models. As seen in Table 2 illustrates the performance variations according to the scale of models well-utilized in the FAS problem, including ResNet [11], EfficientNet-V2 [28], and ViT [7], in conjunction with anti-spoofing applications of contrastive learning [27] and patch-based learning [29] methods. Through these results, we confirm that irrespective of the type of model and learning method, the impact of the model's scale on training performance is negligible.

4. ECLIPS

Guided by the overall insights, we propose a temporal and noise-aware robustness framework called ECLIPS for cross-domain FAS. The proposed framework is comprised of two main components: a base learner module and a decision fusion module, as illustrated in Figure 2. We describe each component in the following sections.

4.1. Base Learner Module

Compared to datasets used in other vision tasks, commonly used datasets for face anti-spoofing, such as OULU [1], CASIA [45], Idiap [4], and MSU-MFSD [35], have a notably smaller scale. This limited data scope can cause models to overfit, which in turn reduces their ability to generalize across different domains. Simply increasing the scale of the model is not an effective alternative when data is scarce.

To address this issue, we employ ensemble techniques that leverage multiple base learners to capture diverse do-

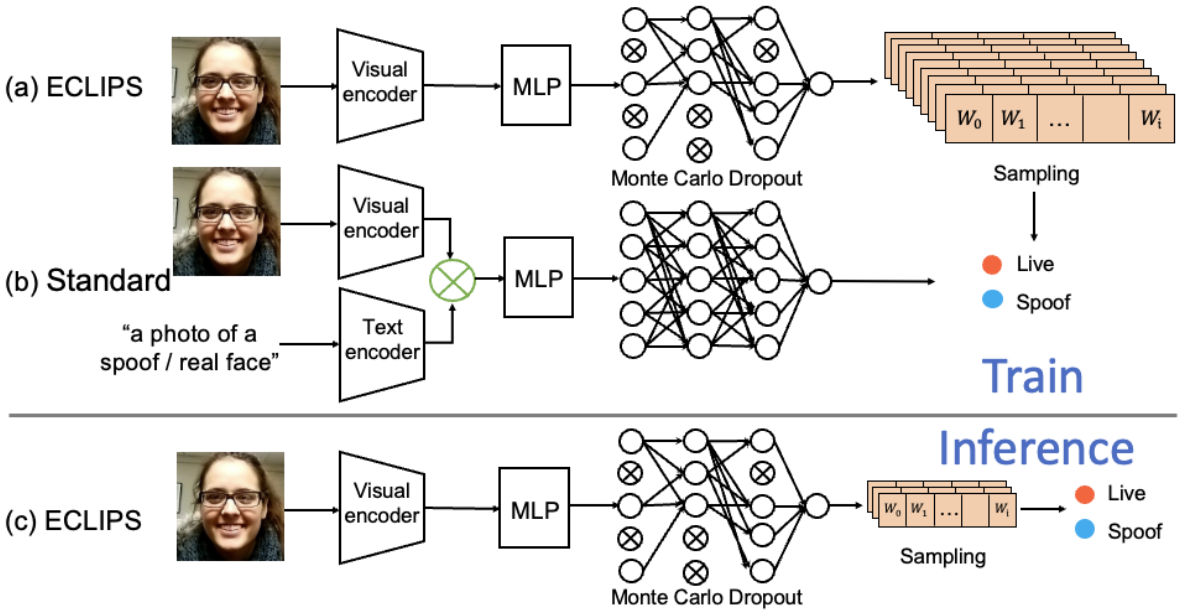

Figure 2. Architecture of the proposed FAS Model ECLIPS. (a) The ECLIPS model for training utilizes only a Visual Encoder. (b) The Standard variant is a dual-stream model that integrates textual information with visual features. (c) The ECLIPS, at inference, only utilizes a Visual Encoder.

main characteristics. Our goal is to assess the effectiveness of the ensemble approach by comparing models with the same backbone architecture as base learners, irrespective of the backbone model used. Ultimately, we determine that the CLIP [22] is highly effective for use as a base learner. In the proposed fusion learning, we adopt the CLIP as our backbone base learner for tackling the FAS challenge.

Despite the constraints posed by the limited FAS training data, capturing its inherent diversity during the CLIP [22] training process is crucial for robustness against emerging attack types. We adopt diverse pre-training strategies for each base learner at the data and model levels to address these novel attacks effectively. In the final base learner, we use the MC dropout [8] technique at the model level to improve the generalization ability.

4.2. Decision Fusion Module

Ensemble learning trains multiple base learners and aggregates their outputs using specific rules. While many ensemble models focus on the architecture and use averaging to predict outputs, this simplistic averaging often leads to suboptimal performance, lacking adaptability to data and sensitivity to biases in base learners. Errors, especially from overfitting in deep learning architectures, can further exacerbate this issue.

To address these challenges, we adopt an approach where we learn weights for each model, constructing a decision fusion module to aggregate probability values from each model for the final prediction. By learning these weights, the fusion module reflects the contributions of each base learner, promoting diversity while assigning higher weights to more important models.

Figure 2 illustrates the architecture of the ECLIPS model, seamlessly integrating a CLIP Visual Encoder and a Multilayer Perceptron (MLP) with Monte Carlo Dropout for uncertainty estimation. During training, the CLIP Visual Encoder processes image inputs, followed by applying dropout with a probability of 0.5 to the extracted features. These dropout-applied features are then used for generating predictions through sampling, iterating this process 10 times to capture uncertainty and combat overfitting. During inference, the process is streamlined for efficiency without compromising robustness, with only 3 samplings performed, ensuring accurate predictions in real-time scenarios.

4.3. Implementation Details

In the preprocessing phase, the Multi-Task Cascaded Convolutional Networks (MTCNN) [41] was used to identify facial bounding boxes. These boxes were then expanded by a padding of 0.5 to crop the faces. After this step, we used a total of $M = 32$ frames each frame from the real and fake videos into the model. In our architecture, we incorporated a feature embedding layer with a dropout layer, setting the dropout rate at 0.5, and utilized a CLIP Visual encoder based on the ViT-B/16 as the backbone.

To establish the final logit for a classifier layer, we sample the logit values and average these samples. For the

Method (%)	OCI→M		OMI→C		OCM→I		ICM→O		Average	
	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)
MADDG (CVPR' 19) [24]	17.69	88.06	24.50	87.51	22.19	84.99	27.98	80.02	23.09	85.89
MDDR (CVPR' 20) [30]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47	20.64	86.43
NAS-FAS (TPAMI' 20) [38]	16.85	90.42	15.21	92.64	11.63	96.98	13.16	94.18	14.21	93.80
RFMeta (AAAI'20)SAFAS [25]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16	16.97	90.69
D^2AM (AAAI'21) [3]	12.70	95.66	20.98	85.58	15.43	91.22	15.27	90.87	16.09	90.58
DRDG(IJCAI '21) [18]	12.43	95.81	19.05	88.79	15356	91.79	15.63	91.91	15.66	91.82
self-DA (AAAI' 21) [31]	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30	19.65	87.15
ANRL (ACM MM' 21) [17]	10.83	96.75	17.85	89.26	16.03	91.04	15.67	91.90	15.09	92.23
FGHV (AAAI' 21) [19]	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55	12.87	93.51
SSDG-R (CVPR' 20) [13]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.28	95.06
SSAN-R (CVPR' 22) [33]	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.80	96.21
PatchNet (CVPR' 22) [29]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.90	95.19
GDA (ECCV' 22) [46]	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60	11.45	94.65
SA-FAS (CVPR' 23) [27]	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	7.83	96.17
DIVT-M (WACV' 23) [16]	2.86	99.14	8.67	96.62	3.71	99.29	13.06	94.04	7.07	97.27
ViT (ECCV' 22) [12]	1.58	99.68	5.70	98.91	9.25	97.15	7.47	98.42	6.00	98.04
FLIP-MCL (ICCV' 23) w/ViT-B/16 [26]	4.95	98.11	0.54	99.98	4.25	99.07	2.31	99.63	3.01	99.19
FLIP-MCL (ICCV' 23) w/ViT-L/14 [26]	4.52	98.40	0.11	99.99	3.45	99.50	1.81	99.63	2.47	99.13
ECLIPS (w/Monte Carlo Dropout)	1.43	99.87	0.78	99.18	1.60	99.16	3.21	99.35	**1.76**	**99.39**

Table 3. Quantitative FAS evaluation result below HTER and AUC metrics. Across transfer learning protocols OMIC, it achieved the lowest average HTER and the highest AUC scores. This suggests a superior efficacy of the OMIC protocol in enhancing the robustness and reliability of FAS systems against spoofing attempts, underscoring its potential for integration into advanced biometric authentication frameworks. Denoted in blue, confirming its status as the new state-of-the-art in the field.

training setup, an input size of 224x224 and a batch size of 16 over 100 epochs for the learning. For optimization purposes, the ADAM optimizer is employed, configured with a learning rate of $1e^{-6}$ and a weight decay to $1e^{-6}$.

5. Experiments

5.1. Experimental Setup

Dataset and Domain Generalization Protocols We evaluated the proposed FAS model using both the public and newly presented FAS datasets. First, followed by the previous domain generalization (DG) setting, we use for the public datasets OULU [1] (**O**), CASIA [45] (**C**), Idiap [4] (**I**), and MSU-MFSD [35] (**M**) for the evaluation. In the DG scenario, the training and test set consists of different sets of datasets such as the training dataset pairs of **O**, **C**, and **I** datasets and the test dataset as M, abbreviated as (**OCI→M**). In addition to the previous setting, we use public FAS datasets: CelebA-Spoof [44] (**A**) and SiW-Mv2 [10] (**S**), to show the tendency the existing DG evaluation, as (**OCIM→A**). From the newly introduced setting, we measure the credibility of the previous DG evaluation by analyzing the correspondence between the two test datasets, CelebA-Spoof and SiW-Mv2. From this setting, we aim to show both the superiority of the proposed FAS method and also the DG representation credibility of the proposed robustness-aware evaluation metric.

Evaluation Metrics First, following the conventional settings of [26, 27], we evaluate the DG performance using metrics that have been previously established: Half Total Error Rate (HTER), Area Under the Receiver Operating Characteristic Curve (AUC), and True Positive Rate (TPR) at a fixed False Positive Rate (FPR). Additionally, we assess the robustness of the FAS method by introducing measures of Variance and Bias in Section 3.1. These metrics are designed to quantify robustness in terms of temporal stability and sensitivity to noise.

5.2. Cross-Domain Performance

In this section, we compare the performance of models using the OCIM benchmark, designed to assess traditional cross-domain performance. The OCIM benchmark experiments, conducted with a leave-one-out protocol, include four transfer scenarios denoted as OCI→M, OMI→C, OCM→I, and ICM→O. In our comprehensive study, which aims to enhance FAS methodologies, we evaluate our model's performance not only through traditional metrics like HTER and AUC but also by proposing the inclusion of novel generalization metrics such as bias and variance.

Traditional Evaluation Table 3 presents a refined evaluation of FAS models on the OCIM dataset, demonstrating superior HTER and AUC metrics. Leveraging the FLIP-MCL [26] as a baseline, we have developed a model that not only achieves greater computational efficiency but also improves performance and generalization. Furthermore, by incorporating innovative fusion techniques such as Monte Carlo Dropout, our model gains additional advantages in robustness and reliability, as evidenced in the reported results.

Method (%)	OCI→M		OMI→C		OCM→I		ICM→O		Average	
	Bias (↓)	Variance (↓)	Bias (↓)	Variance (↓)	Bias (↓)	Variance (↓)	Bias (↓)	Variance (↓)	Bias (↓)	Variance (↓)
ResNet18 [11]	0.085	0.081	0.080	0.128	0.094	0.055	0.258	0.127	0.129	0.097
XceptionNet [5]	0.088	0.085	0.076	0.141	0.109	0.114	0.295	0.149	0.142	0.122
EfficientNet-V2/Tiny [28]	0.078	0.093	0.075	0.129	0.085	0.104	0.319	0.146	0.139	0.118
ViT-B/16 [7]	0.084	0.072	0.075	0.126	0.075	0.080	0.131	0.135	0.091	0.103
SA-FAS (CVPR' 23) [27]	0.071	0.085	0.159	0.131	0.064	0.107	0.231	0.122	0.131	0.111
FLIP-MCL (ICCV' 23) [26]	0.025	0.064	0.029	0.077	0.054	0.121	0.152	0.143	0.065	0.101
ECLIPS	0.021	0.075	0.028	0.098	0.039	0.076	0.051	0.054	0.034	0.075

Table 4. Comparative Evaluation of ECLIPS against state-of-the-art models across four transfer scenarios on bias and variance metrics. This table showcases the superior performance of our ECLIPS model, demonstrating the lowest average bias and variance among all tested models, thereby establishing a new benchmark in the field of FAS.

Method (%)	GFLOPs	OCIM→A		OCIM→S			
		HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	Bias (↓)	Variance (↑)
ViT-B/16 [7]	22.00	38.64	66.09	44.82	58.12	0.44	0.04
ViT-B/16 (w/Sampling)	66.00	20.46	83.82	29.52	70.23	0.31	0.16
ViT-B/16 (w/Monte Carlo Dropout)	66.00	17.76	89.52	28.92	79.80	0.20	0.16
CLIP-V [22]	11.27	13.05	92.46	30.27	77.51	0.24	0.15
CLIP-V (w/Sampling)	33.81	8.53	95.57	27.57	80.38	0.23	0.15
FLIP-MCL (ICCV' 23) [26]	103.57	10.73	94.86	28.54	79.49	0.26	0.08
ECLIPS	33.81	8.47	96.16	26.70	81.29	0.18	0.14

Table 5. Evaluation of domain generalization for models trained on OCIM to new datasets. The CelebA-Spoof and SiW-Mv2 datasets encompass a vast array of spoof types, surpassing the scope of the original OCIM dataset. Notably, CelebA-Spoof comprises data without accompanying videos, thus excluding Variance and Bias considerations.

Importantly, our research highlights the considerable benefits of using an ensemble of models over a single model architecture. This ensemble approach enhances the representational capacity of our system and increases its resilience against various spoofing attacks, thereby improving its generalization across unseen data domains. Ultimately, our model achieves state-of-the-art performance with fewer parameters compared to the existing FLIP-MCL Large model. Our experimental findings underline the advantages of employing a constellation of smaller backbones over a single larger one to boost model generalization, aligning with the paradigm shift suggested by our initial experiments.

Temporal Robustness Evaluation This chapter provides comparison results for variance and bias, designed to consider the model's confidence and temporal robustness. Table 4 presents a rigorous evaluation of our proposed ECLIPS model in four transfer scenarios, following the same methodology as described in Table 3. The results demonstrate our model's exemplary performance in terms of Bias and Variance. Notably, ECLIPS achieves the lowest average Bias and Varance, outperforming other methods, including several state-of-the-art (SOTA) models. These findings underscore the advanced capacity of ECLIPS to provide consistent and reliable predictions while effectively managing uncertainties, cementing its position as the new benchmark in the FAS domain.

5.3. Real-World Use-Case Performance

The CelebA-Spoof dataset, constructed on the foundation of the CelebA dataset, is specifically designed for large-scale face recognition and anti-spoofing research, incorporating a spectrum of spoofing attacks. It encompasses a variety of imaging conditions, including lighting, pose, expression, and the presence or absence of makeup, making it a robust resource for evaluating the generalizability of models. SiW-Mv2, addresses face spoofing in real-world scenarios, incorporating various attack types and conditions indoors and outdoors. It plays a pivotal role in testing the resilience of deep learning-based anti-spoofing systems against the intricacies of real-world variability. Leveraging models trained on the OCIM protocol, testing on the public CelebA-Spoof dataset, which consists of individual frames sharing the same identity in a single video, allows for the validation of HTER and AUC performance metrics.

Referring to Table 5, we demonstrate enhanced performance over existing state-of-the-art models on this new dataset. SiW-Mv2's sequential frame composition within single videos enables a comprehensive assessment of not just HTER and AUC but also Variance and Bias. Models trained using the OCIM protocol and tested on the public SiW-Mv2 dataset report superior performance and generalization, surpassing that of current state-of-the-art models, thus advancing the field of FAS technology.

Method (%)	OCI→M		OMI→C		OCM→I		ICM→O		Average	
	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)	HTER (↓)	AUC (↑)
ViT-B/16 [7]	10.24	96.10	13.94	93.91	21.00	78.32	17.99	89.38	15.79	89.67
ViT-B/16 (w/Sampling)	7.38	97.60	18.0	90.20	11.6	95.39	23.03	84.91	15.00	91.52
ViT-B/16 (w/Monte Carlo Dropout)	7.14	98.19	14.66	93.53	12.5	93.88	25.46	82.47	14.19	92.76
CLIP-V [22]	7.3	96.71	2.6	98.95	5.0	99.06	4.44	98.83	4.84	98.13
CLIP-V (w/Sampling)	5.4	98.60	1.88	99.20	3.5	99.63	4.04	98.74	3.9	98.79
FLIP-MCL (ICCV' 23) [26]	4.95	98.11	0.54	99.98	4.25	99.07	2.31	99.63	3.01	99.19
ECLIPS	1.43	99.87	0.78	99.18	1.60	99.16	3.21	99.35	1.76	99.39

Table 6. Ablation studies regarding model ensembling. Compared to the previous method applying a ViT-B/16 backbone, our proposed ECLIPS achieved significant quantitative improvement. Furthermore, ECLIPS achieved superior performance compared to CLIP CLIP-based method although we only apply Visual Encoder of the CLIP, different from the other methods.

5.4. Abulation Study

As shown in Table 6, we verify the ensemble effect by utilizing only the Visual Encoder from CLIP models, confirming its efficacy in Vision Transformer (ViT)implementations as well. The CLIP model, known for its superior feature representation capabilities, forms a part of our ensemble backbone. Through rigorous evaluation on both a new benchmark dataset and established datasets, our findings reveal that this ensemble strategy, when combined with the advanced feature representation of the CLIP backbone, markedly surpasses existing state-of-the-art methods. Our model not only excels in traditional performance metrics but also shines in our newly introduced generalization metrics, offering a more comprehensive and dependable assessment of performance across different testing conditions.

5.5. Temporal robustness and Accuracy

Figure 3 illustrates the exceptional capabilities of the ECLIPS architecture in tackling overfitting, enhancing predictive performance, and demonstrating remarkable robustness against temporal and noise variations within FAS tasks. Notably, in the scatter plot comparing ensemble methods evaluated on OCIM→S, the ECLIPS model is positioned in the bottom-left quadrant. This positioning not only highlights its unparalleled accuracy but also its superior performance in terms of variance, showcasing its outstanding generalization capabilities and robustness in both accuracy and variance within the FAS domain. However, the results depicted in the figure also caution that a favorable bias does not necessarily correlate with good accuracy. Therefore, it's imperative to consider temporal robustness metrics alongside accuracy for a comprehensive evaluation.

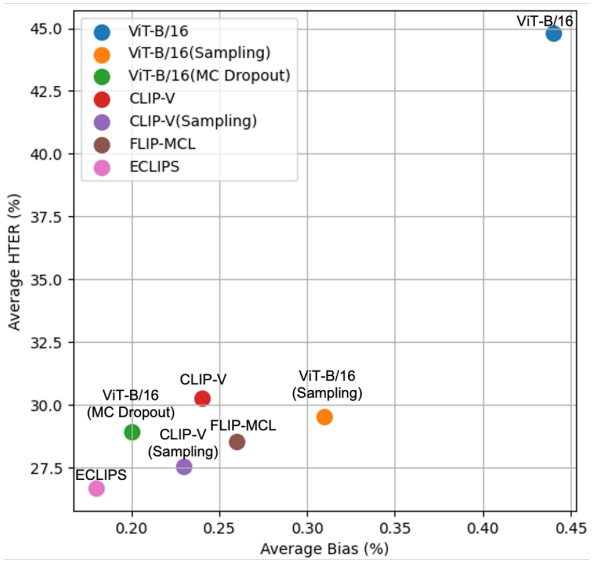

Figure 3. The scatter plot of representative FAS models from Bias to HTER. Notably, the ECLIPS model is highlighted in the bottom-left, indicating its superior accuracy and generalization ability in FAS.

6. Conclusions

In this paper, we conducted a thorough analysis of the FAS model'scapabilities in detecting presentation attacks, taking into account dataset characteristics, evaluation metrics, and design principles. Our investigation revealed significant limitations within the existing evaluation framework, especially regarding the assessment of generalization and model scalability. To verify and address these challenges, we utilized a wide range of FAS evaluation datasets reflective of real-world scenarios and introduced a robustness-aware evaluation metric designed to more accurately gauge the FAS model's generalization capabilities. Based on our analyses, we revisited and refined the ensemble strategy for smaller-sized FAS models, achieving state-of-the-art performance through extensive experimentation. Future research could explore the feasibility of analyzing spoofing scenes across diverse scenarios to enhance datasets in ways that bolster the effectiveness of the proposed ensemble method.

References

[1] Z. Boulkenafet, J. Komulainen, Lei. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. 2017. 2, 4, 6

[2] Baoliang Chen, Wenhan Yang, and Shiqi Wang. Generalized face antispoofing by learning to fuse features from high- and low-frequency domains. *IEEE MultiMedia*, 28(1):56–64, 2021. 1

[3] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1132–1139, 2021. 6

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. 2012. 2, 4, 6

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4, 7, 8

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. 2, 5

[9] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2

[10] Xiao Guo, Yaojie Liu, Anil Jain, and Xiaoming Liu. Multi-domain learning for updating face anti-spoofing models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 230–249. Springer, 2022. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 4, 7

[12] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *European Conference on Computer Vision*, pages 37–54. Springer, 2022. 1, 2, 6

[13] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 1, 2, 6

[14] Jie Jiang and Yunlian Sun. Depth-based ensemble learning network for face anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2954–2958. IEEE, 2022. 2, 3

[15] Haoliang Li, Shiqi Wang, Peisong He, and Anderson Rocha. Face anti-spoofing with deep neural network distillation. *IEEE Journal of Selected Topics in Signal Processing*, 14 (5):933–946, 2020. 1, 2

[16] Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, and Chu-Song Chen. Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6098–6107, 2023. 1, 2, 6

[17] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021. 6

[18] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021. 6

[19] Shice Liu, Shitao Lu, Hongyi Xu, Jing Yang, Shouhong Ding, and Lizhuang Ma. Feature generation and hypothesis verification for reliable face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1782–1791, 2022. 6

[20] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 2

[21] Usman Muhammad, Md Ziaul Hoque, Mourad Oussalah, and Jorma Laaksonen. Deep ensemble learning with frame skipping for face anti-spoofing. *arXiv preprint arXiv:2307.02858*, 2023. 2, 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 7, 8

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[24] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 1, 2, 6

[25] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11974–11981, 2020. 1, 2, 6

[26] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19685–19696, 2023. 1, 2, 3, 6, 7, 8

[27] 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. 1, 2, 4, 6, 7

[28] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 4, 7

[29] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20281–20290, 2022. 2, 4, 6

[30] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6678–6687, 2020. 6

[31] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2746–2754, 2021. 6

[32] Zhuo Wang, Qiangchang Wang, Weihong Deng, and Guodong Guo. Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):439–450, 2022. 1, 2

[33] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022. 2, 6

[34] Zezheng Wang, Zitong Yu, Xun Wang, Yunxiao Qin, Jiahong Li, Chenxu Zhao, Xin Liu, and Zhen Lei. Consistency regularization for deep face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 1, 2

[35] Di Wen, Anil. K Jain, and Hu Han. Face Spoof Detection with Image Distortion Analysis. *IEEE Trans. Information Forensic and Security*, 2015 (To Appear). 2, 4, 6

[36] Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You, and Yuan Zhu. Pipenet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 644–645, 2020. 2

[37] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 557–575. Springer, 2020. 1

[38] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 6

[39] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex Kot. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *arXiv preprint arXiv:2302.05744*, 2023. 1

[40] Haixiao Yue, Keyao Wang, Guosheng Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Cyclically disentangled feature translation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3358–3366, 2023. 1, 2

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 5

[42] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 641–657. Springer, 2020. 1

[43] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[44] Yuanhan Zhang, Zhenfei Yin, Jing Shao, Ziwei Liu, Shuo Yang, Yuanjun Xiong, Wei Xia, Yan Xu, Man Luo, Jian Liu, Jianshu Li, Zhijun Chen, Mingyu Guo, Hui Li, Junfu Liu, Pengfei Gao, Tianqi Hong, Hao Han, Shijie Liu, Xinhua Chen, Di Qiu, Cheng Zhen, Dashuang Liang, Yufeng Jin, and Zhanlong Hao. Celeba-spoof challenge 2020 on face anti-spoofing: Methods and results, 2021. 6

[45] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 2, 3, 4, 6

[46] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *European Conference on Computer Vision*, pages 335–356. Springer, 2022. 1, 2, 6