# Joint Physical-Digital Facial Attack Detection Via Simulating Spoofing Clues

Xianhua He[1], Dashuang Liang[1], Song Yang[1], Zhanlong Hao[1]
Hui Ma[2], Binjie Mao[1], Xi Li[1], Yao Wang[1], Pengfei Yan[1], Ajian Liu[3*]
[1]Vision AI Department, Meituan; [2]M.U.S.T, Macau; [3]MAIS, CASIA, China
[1]hexianhua@meituan.com, [3]ajian.liu@ia.ac.cn

## Abstract

*Face recognition systems are frequently subjected to a variety of physical and digital attacks of different types. Previous methods have achieved satisfactory performance in scenarios that address physical attacks and digital attacks, respectively. However, few methods are considered to integrate a model that simultaneously addresses both physical and digital attacks, implying the necessity to develop and maintain multiple models. To jointly detect physical and digital attacks within a single model, we propose an innovative approach that can adapt to any network architecture. Our approach mainly contains two types of data augmentation, which we call Simulated Physical Spoofing Clues augmentation (SPSC) and Simulated Digital Spoofing Clues augmentation (SDSC). SPSC and SDSC augment live samples into simulated attack samples by simulating spoofing clues of physical and digital attacks, respectively, which significantly improve the capability of the model to detect "unseen" attack types. Extensive experiments show that SPSC and SDSC can achieve state-of-the-art generalization in Protocols 2.1 and 2.2 of the UniAttackData dataset, respectively. Our method won first place in "Unified Physical-Digital Face Attack Detection" of the 5th Face Anti-spoofing Challenge@CVPR2024. Our final submission obtains 3.75% APCER, 0.93% BPCER, and 2.34% ACER, respectively. Our code is available at https://github.com/Xianhua-He/cvpr2024-face-anti-spoofing-challenge.*

## 1. Introduction

Facial recognition systems are widely used in our daily lives, but they are suffering from an increasing number of various types of attacks. The types of facial attacks can be mainly divided into physical attacks and digital attacks. The widely used types of high-frequency physical attacks are: Print attacks [18, 49, 50], Replay attacks [19, 20], and 3D mask attacks [4, 13, 22, 24]. Meanwhile, we can briefly summarize digital attacks as digital forgery attacks [1, 10, 11, 35, 41, 42] and adversarial attacks [3, 32, 34, 43, 44, 53]. When addressing physical attacks, central difference convolutions [46] and other methods [6, 16, 23, 27, 40, 45] can learn low-level and high-level discriminative visual features and achieve advanced performance in classifying live and physical attacks. When detecting digital attacks, deepfake methods [8, 15, 37, 52] can distinguish live and digital forgery attacks by dividing the characteristic subspace of different forgery attack types, and achieve advanced generalization in different digital forgery datasets.

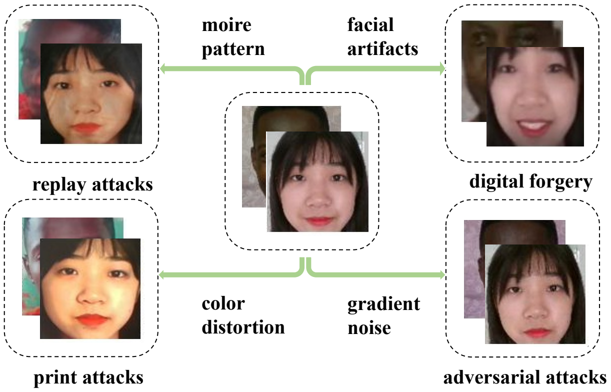

Figure 1. **Spoofing clues for four attack types.** For replay attacks, print attacks, digital forgeries, and adversarial attacks, the spoofing clues distinct from live samples are identified as moire patterns, color distortion, facial artifacts, and gradient noise, respectively.

However, few studies detect physical attacks and digital attacks simultaneously. Previous methods separately train models that detect physical and digital attack types, which means that multiple different models need to be deployed to comprehensively judge the final results. This requires more computing resources, and it is more difficult to develop and maintain multiple models. To address this issue, we discuss based on the UniAttackData dataset [5] that jointly detects physical attacks and digital attacks. The UniAttackData dataset contains 1800 subjects from 3 different races, covering two types of physical attacks, six types of digital

[*]Corresponding author.

forgery, and six types of adversarial attacks. The same face ID covers all attack types. The UniAttackData dataset defines two protocols. Protocol 1 is designed to evaluate the ability of the model to jointly detect physical attacks and digital attacks. However, there are huge intra-class differences between physical attacks and digital attacks, which brings monumental challenges to the design of the algorithm. Protocol 2 is employed to evaluate the model's ability to detect "unseen" attack types. The test set for Protocol 2.1 exclusively comprises physical attacks that were not present in the training and development sets. Similarly, the test set for Protocol 2.2 is strictly composed of digital attacks absent from the training and development phases.

We discuss the issue of jointly detecting physical and digital attacks and derive some of the following insights. The spoofing clues can improve the model's ability to distinguish between live and various types of attacks. As shown in Figure 1, the spoofing clues typically manifested in physical attacks, such as print and replay attacks, are color distortions and moire patterns. In terms of digital forgery attacks, including face-swapping or face generation, the prevalent spoofing clues are facial artifacts or distortions. Additionally, adversarial attacks frequently introduce spoofing clues by adding specified gradient noise to the original image.

Motivated by the discussions above, we introduce SPSC and SDSC to improve the model's ability to detect physical attacks and digital attacks, respectively. Specifically, in Protocol 2.1, we use SPSC to simulate color distortion and moire patterns and augment the live samples into physical attack samples for training. In Protocol 2.2, we use SDSC to simulate the artifacts caused by face swapping and augment the live samples into digital attack samples for training. SPSC and SDSC improve the ability of the models in protocols 2.1 and 2.2 to detect "unseen" attack types, respectively. Meanwhile, using SPSC and SDSC in Protocol 1 also improves the generalization performance of the model.

In summary, the main contributions of this paper are summarized as follows:

- We propose Simulated Physical Spoofing Clues augmentation (SPSC) to simulate the spoofing clues of physical attacks and address the issue of cross-attack type detection from digital to physical attacks in protocol 2.1.
- We introduce Simulated Digital Spoofing Clues augmentation (SDSC) to simulate the spoofing clues of digital attacks and overcome the barriers to how physical attacks generalize to digital attacks in protocol 2.2.

## 2. Related work

### 2.1. Physical Attack Detection.

With the development of deep learning, convolutional neural networks (CNN) have gradually become the mainstream method for solving the task of FAS. Liu et al. [29] utilize physical-based depth information as a supervisory signal instead of binary classification loss. Yu et al. [46] propose Central Difference Convolution (CDC), which is able to capture intrinsic detailed patterns via aggregating both intensity and gradient information. Feng et al. [6] propose a residual-learning framework to learn the discriminative live-spoof differences, which are defined as the spoof cues. PIFAS [23] decomposes faces into appearance information and pose codes to capture liveness and liveness-unrelated features, respectively. AA-FAS [25] regards FAS as a unified framework with the attack and defense systems, which employs adversarial training to optimize the defense system against unseen attacks. Although these algorithms have achieved astonishing results in intra-datasets experiments, their performance deteriorates severely when faced with unknown domains. To solve these limits, Domain Generalization (DG) based methods [2, 28, 36, 38] can conquer this by taking advantage of multiple source domains without seeing any target data. Jia et al. [14] propose an end-to-end single-side domain generalization framework (SSDG) to improve the generalization ability of face anti-spoofing. Sun at al. [39] encourages domain separability while aligning the live-to-spoof transition (i.e., the trajectory from live to spoof) to be the same for all domains. Huang et al. [12] introduce the ensemble adapters module and feature-wise transformation layers in the ViT to adapt to different domains for robust performance with a few samples. IADG [51] framework aligns features on the instance level, reducing sensitivity to instance-specific styles. MDIL [40] consists of an adaptive domain-specific experts (ADE) framework based on the vision transformer, and an asymmetric classifier is designed to keep the output distribution of different classifiers consistent. CFPL-FAS [27] makes use of large-scale VLMs like CLIP and leverages the textual feature to dynamically adjust the classifier's weights for exploring generalizable visual features.

Multi-modal FAS has gained significant attention due to the increasing sophistication of high-quality attacks. However, these multi-modal fusion-based algorithms require the testing phase to provide the same modal types as the training phase, severely limiting their deployment scenarios. Flexible modality-based methods [17, 21, 26, 47, 48] aim to improve the performance of any single modality by leveraging available multi-modal data.

### 2.2. Digital Attack Detection.

Some work in [33] formulate deep fake detection as a hybrid combination of supervised and reinforcement learning (RL). [33] chooses the top-k augmentations for each test sample by an RL agent in an image-specific manner and the classification scores, obtained using CNN, of all the augmentations of each test image are averaged together for final real or fake classification. Guide-Space [8] is a controllable

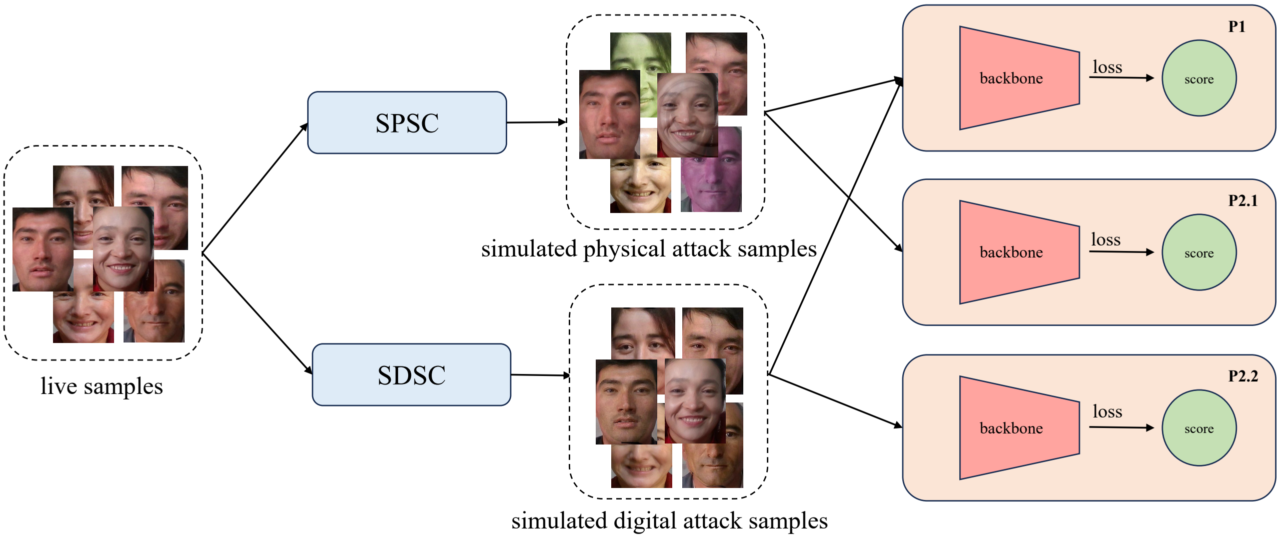

Figure 2. **The overview pipeline of our method.** We propose Simulated Physical Spoofing Clues augmentation (SPSC), which augments live samples into simulated physical attack samples for training within protocols 1 and 2.1. Concurrently, we present Simulated Digital Spoofing Clues augmentation (SDSC), converting live samples into simulated digital attack samples, tailored for training under protocols 1 and 2.2.

guide-space method to enhance the discrimination of different forgery domains. The well-designed guide space can simultaneously achieve both the proper separation of forgery domains and the large distance between real-forgery domains in an explicit and controllable manner. Self-Blending [37] presents novel synthetic training data to detect deepfakes. Self-Blending is generated by blending pseudo source and target images from single pristine images, reproducing common forgery artifacts (e.g., blending boundaries and statistical inconsistencies between source and target images).

# 3. Methodology

In this section, we introduce an overview of our method in Section 3.1. Subsequently, we elaborate on Simulated Physical Spoofing Clues augmentation and Simulated Digital Spoofing Clues augmentation in Section 3.2 and Section 3.3, respectively.

## 3.1. Overview

As depicted in Figure 2, our method principally introduces two targeted data augmentation strategies, designated as Simulated Physical Spoofing Clues augmentation (SPSC) and Simulated Digital Spoofing Clues augmentation (SDSC). We augment live samples into simulated attack instances via SPSC or SDSC, subsequently extracting features through a neural network and computing the Cross-Entropy loss to refine classification networks. The augmentations of SPSC and SDSC can be seamlessly incorporated into additional frameworks. The inference phase is consistent with established baseline methods, allowing straightforward score determination. Under Protocol 1, both SPSC and SDSC data augmentations are employed. For protocol 2.1, the model needs to be generalized to detect “unseen” physical attacks, and only SPSC is used to enhance the model’s detection performance against physical attacks. For protocol 2.2, the objective is for the model to generalize to “unseen” digital attack modalities; hence, only SDSC data augmentation is implemented to improve the model’s digital attack detection capabilities.

## 3.2. Simulated Physical Spoofing Clues

Our comprehensive analysis of physical attack characteristics has led to the development of the Simulated Physical Spoofing Clues augmentation (SPSC), which integrates both ColorJitter and moire pattern augmentation. As demonstrated in Figure 3, we simulate print attacks on live samples through a spectrum of ColorJitter adjustments, and similarly, we emulate replay attacks by applying varying degrees of moire pattern augmentation. This approach allows us to simulate the distinct visual artifacts characteristic of each attack type, thereby enriching the robustness of spoofing detection models. **ColorJitter:** Print presentation attacks frequently leave behind spoofing cues manifesting as color distortions. To emulate these distortions, ColorJitter serves as an effective tool for creating artificial spoofing cues. By applying ColorJitter, a live sample is converted into a simulated attack sample, effectively capturing the essence of a print attack scenario. For this purpose, we calibrate the ColorJitter set-

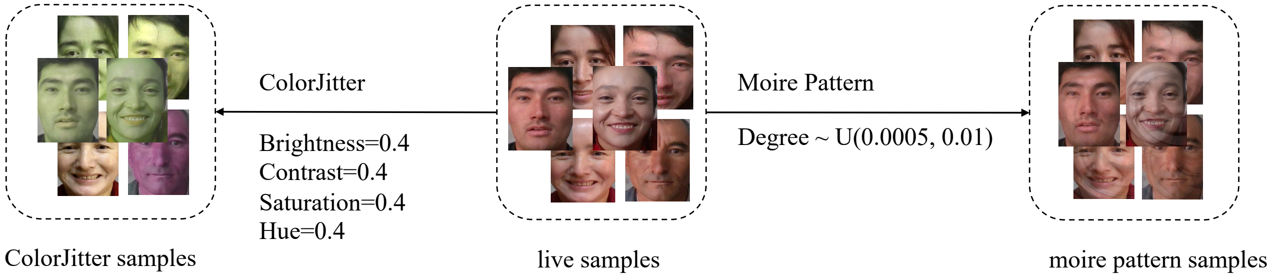

Figure 3. Live samples simulate print attacks through varying degrees of ColorJitter and replay attacks through varying degrees of moire pattern augmentation.

tings for brightness, contrast, saturation, and hue uniformly to a factor of 0.4. **Moire Pattern augmentation:** Spoofing cues in replay presentation attacks are often characterized by the presence of moire patterns. To address this issue, we design a moire pattern generation algorithm that effectively simulates these distinctive patterns. The specifics of this algorithm are presented in the form of pseudo-code, as detailed in Algorithm 1. This algorithmic approach enables us to create nuanced moire effects that closely mirror those found in actual replay attacks. First, you need to obtain the height and width of the original image, calculate the center point of the image, and randomly generate the moire intensity degree, which follows a uniform distribution from 0.0005 to 0.01. Then, create grid coordinates and generate grid coordinates $(X, Y)$ for the $(x, y)$ coordinates of each pixel. Afterward, calculate the offset and polar coordinate parameters, and calculate the $(X, Y)$ offset $(X_{offset}, Y_{offset})$ of each point relative to the center point. Subsequently, calculate the angle $\theta$ and radius $\rho$ in polar coordinates and calculate new $(X, Y)$ coordinates $(X_{new}, Y_{new})$ based on angle and radius adjustments. Finally, limit the coordinate range and map pixels and limit new coordinates $(X_{new}, Y_{new})$ to the image range. Combining the source image with the mapped pixel values generates an image with a moire pattern effect.

**Algorithm 1** Add Moire Pattern to Image

```
Require: $src$: source image
$height, width \leftarrow$ dimensions of $src$
$center \leftarrow (height/2, width/2)$
$degree \leftarrow$ random value between 0.0005 and 0.01
$x \leftarrow$ array from 0 to $width - 1$
$y \leftarrow$ array from 0 to $height - 1$
$X, Y \leftarrow$ meshgrid of $x$ and $y$
$offset_X \leftarrow X - center[0]$
$offset_Y \leftarrow Y - center[1]$
$\theta \leftarrow \text{arctan2}(offset_Y, offset_X)$
$\rho \leftarrow \sqrt{offset_X^2 + offset_Y^2}$
$new_X \leftarrow center[0] + \rho \cdot \cos(\theta + degree \cdot \rho)$
$new_Y \leftarrow center[1] + \rho \cdot \sin(\theta + degree \cdot \rho)$
$new_X \leftarrow \text{clip}(new_X, 0, width - 1)$
$new_Y \leftarrow \text{clip}(new_Y, 0, height - 1)$
$dst \leftarrow 0.8 \cdot src + 0.2 \cdot src[new_Y, new_X]$
return $dst$ as unsigned 8-bit integer
```

### 3.3. Simulated Digital Spoofing Clues

Digital forgery attacks often leverage face swapping or face generation algorithms, which typically induce distortions or artifacts within the facial region. Inspired by Self-Blending [37], we propose the Simulated Digital Spoofing Clues augmentation (SDSC). This method is designed to simulate the distortions and artifacts characteristic of digital forgery facial images. As shown in Figure 4, the process of SDSC is divided into three steps. **1)** Obtain pseudo source image and target image for blending. Copy the original Image $I$ to get $I_1$ and $I_2$. $I_1$ is used as a pseudo source image and is augmented by color transformation (e.g., Hue, Brightness, and Downscale) to obtain $O_1$. $I_2$ is used as a target image and is augmented by spatial transformation (e.g., Resize, translate) to obtain $O_2$. The boundary and landmark between $O_1$ and $O_2$ are misaligned. **2)** The original image $I$ is segmented through the face parsing algorithm to obtain the face mask. Subsequently, the face mask undergoes affine transformation through spatial transformation, and the final mask is obtained through the augmentation of mask deformation (e.g., elastic, blur). **3)** $O_1$, $O_2$ and final mask are blended according to formula 1 and the forgotten image is output.

$$O_{forgery} = O_1 \odot mask + O_2 \odot (1 - mask) \tag{1}$$

## 4. Experiments

### 4.1. Experimental Settings

**UniAttackData Datase.** As shown in Figure 5, the UniAttackData dataset [5] expands on the CASIA-SURF [20] dataset, featuring 1800 subjects of three races: African, East Asian, and Central Asian. It includes two kinds of physical attack methods (Print and Replay), along with six types of

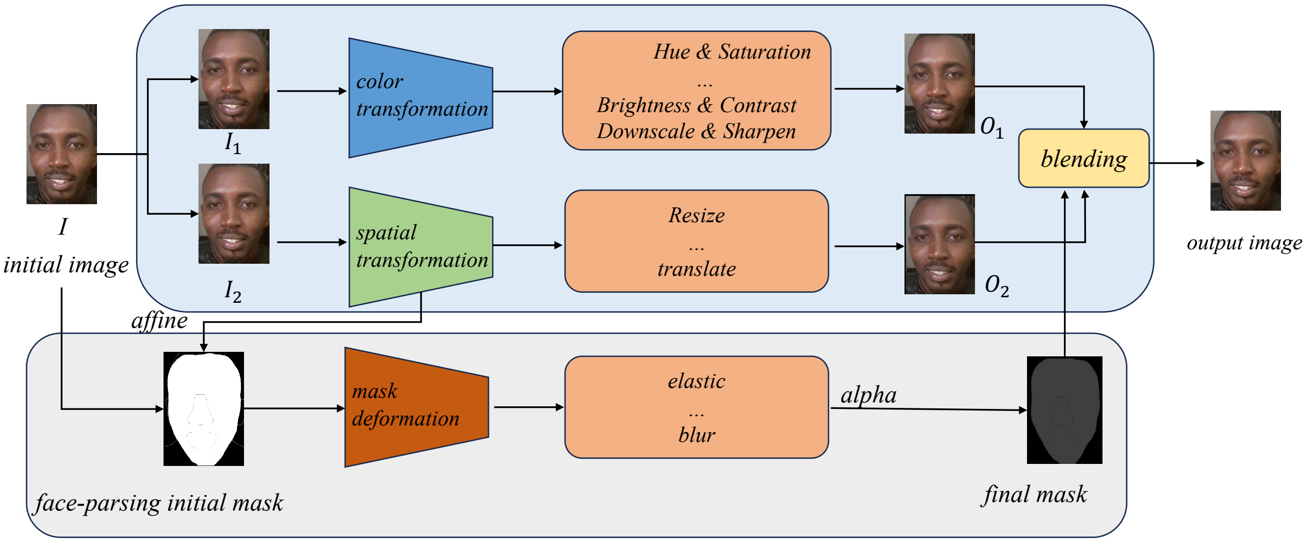

Figure 4. A live sample is transformed into a digital forgery attack sample by Simulated Digital Spoofing Clues augmentation.

digital forgery attacks and six types of adversarial attacks. UniAttackData defines two protocols to make sure that they can effectively test the system's ability to detect different types of attacks jointly. Protocol 1 is designed to test the system's ability to detect a unified attack type that encompasses both physical and digital attacks. Huge intra-class distance and diverse attacks bring more challenges to algorithm design. Protocol 2 evaluates the model's detection capabilities for "unseen" attack types, testing algorithmic adaptability across the diverse and unpredictable spectrum of physical and digital attacks. The test set of protocol 2.1 comprises exclusively novel physical attacks, while the test set of protocol 2.2 contains solely unseen digital attacks, both distinct from the training and development datasets.

**Evaluation Metrics.** The evaluation protocol used to evaluate performance follows established standards within the field of Face Anti-Spoofing (FAS). Specifically, we utilize the widely accepted metric comprising the Attack Presentation Classification Error Rate (APCER), the Bona Fide Presentation Classification Error Rate (BPCER), and the Average Classification Error Rate (ACER). They can be formulated as:

$$\begin{aligned} APCER &= \frac{FP}{FP+TN}, \\ BPCER &= \frac{FN}{FN+TP}, \\ ACER &= \frac{APCER+BPCER}{2}, \end{aligned} \tag{2}$$

where $FP$, $FN$, $TN$, and $TP$ denote the counts of false positive, false negative, true negative, and actual positive instances, respectively. ACER is used to determine the final ranking in the 5th Face Anti-spoofing Challenge@CVPR2024.

**Data Preprocess.** **Face detection:** For images larger than 700 pixels in width and height, we apply face detection and expand the bounding box by 20 pixels for face cropping. Images without detected faces undergo a center crop to yield a 500×500 input for the model. **Obtain the face mask:** Face parsing is performed on live samples to generate face masks for SDSC.

**Architecture Details.** Our method can be easily transferred to any backbone network. Since the amount of dataset from the training sample is relatively small, we chose Resnet50 [9] as the backbone of the classification network. We believe that for this task, the final results obtained by our method will not be significantly different in different backbone networks.

**Training Details.** We distinguish 3 protocols and train 3 models, respectively. In the training stage, We utilize AdamW optimizer [31] to train our model with a learning rate of $1e^{-3}$, and the weight decay is $5e^{-4}$. The cosine learning rate schedule is employed to adjust the learning rate. We resize the image to 224×224 as the input image for model training. The data augmentation of RandomResizedCrop and HorizontalFlip is used during training. We use Cross Entropy loss as the loss function and set different weights to balance the live and attack sample loss. In protocol 1, we used the complete train and dev sets. In protocols 2.1 and 2.2, we use

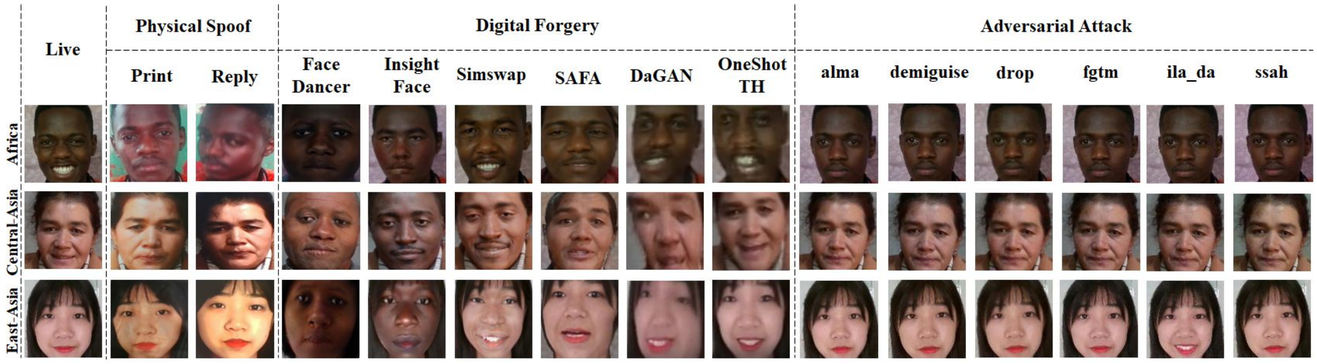

Figure 5. **The overview of UniAttackData Dataset [5].** The same face ID forges physical attack videos through two types of physical attacks (print and replay). For every live video, forge digital attacks through six digital editing algorithms and six adversarial algorithms. The attack type of each sample is indicated at the top of the graph.

| Team | AUC ↑ | APCER ↓ | BPCER ↓ | ACER ↓ |
|---|---|---|---|---|
| VAI-Face | 87.75 | 34.00 | **0.25** | 17.13 |
| BSP-Idiap | 96.33 | 23.06 | 9.36 | 16.22 |
| duileduile | 98.68 | 5.50 | 5.51 | 5.51 |
| SeaRecluse | 96.56 | 6.47 | 0.39 | 3.43 |
| Ours | **99.69** | **3.75** | 0.92 | **2.33** |

Table 1. Comparing results on the test set of the UniAttackData Dataset [5]. Our method achieves the highest performance on the AUC(%), APCER(%) and ACER (%) metrics.

| Protocol | Methods | APCER ↓ | BPCER ↓ | ACER ↓ |
|---|---|---|---|---|
| P1 | w/o SPSC&SDSC | 0.17 | 0.28 | 0.23 |
| | w/ SPSC&SDSC | 0.31 | 0.09 | 0.20 |
| P2.1 | w/o SPSC&SDSC | 76.03 | 0.06 | 38.05 |
| | w/ SPSC | 2.55 | 0.09 | 1.32 |
| P2.2 | w/o SPSC&SDSC | 88.59 | 0.11 | 44.35 |
| | w/ SDSC | 1.73 | 1.58 | 1.65 |
| All | w/o SPSC&SDSC | 54.93 | 0.15 | 27.54 |
| | w/ SPSC&SDSC | 1.53 | 0.69 | 1.06 |

Table 2. Compare the results of baseline(w/o SPSC&SDSC) and ours method(w/ SPSC&SDSC) on all protocols

the whole train set and live samples from the development set. We train the model on a single A100(80G) GPU for 200 epochs, with the batch size set to 512. Each protocol requires only one hour to train and one minute to test.

## 4.2. Comparison with SOTA Methods

We compare the performance of our proposed method with state-of-the-art (SOTA) methods (teams) on the UniAttackData dataset [5]. Table 1 summarizes the results of the comparison of four metrics: AUC, APCER, BPCER, and ACER. In the SOTA method, VAI-Face achieves the lowest BPCER with a value of 0.25%. However, our proposed method achieves the highest performance on the AUC, APCER and ACER metrics, with values of 99.69%, 3.75%, and 2.33%, respectively. In addition, our method achieves the lowest APCER with a value of 3.75%, which is significantly lower than the other methods. The experimental results validate the efficacy of our method, with SPSC and SDSC demonstrating impressive performance in detecting "unseen" attack types under Protocols 2.1 and 2.2.

## 4.3. Protocol Result

We define the baseline method without SPSC and SDSC. Then, we employ our method and retrain three models under the corresponding protocol to compare with the baseline model. The results, as shown in Table 2, indicate a marginal improvement of our method over the baseline. For protocol 1, our method has a slight improvement compared to the baseline. For Protocol 2.1, our method achieved a substantial improvement compared to the baseline, reducing the ACER from 38.05% to 1.32%. Similarly, in Protocol 2.2, our approach significantly outperformed the baseline, decreasing the ACER from 44.35% to 1.65%. By averaging the results across the three protocols, our proposed method reduced the ACER from 27.54% to 1.06%, a notable decline of 26.48%. This highlights the effectiveness of our approach. Finally, the results of our replication were slightly better than those we submitted to the competition organizers, further underscoring the robustness of our method.

## 4.4. Ablation Study

**Simulated Physical Spoofing Clues** SPSC contains ColorJitter and moire pattern augmentations, targeting print and replay attack spoofing cues, respectively. Using the complete training set and live development samples, SPSC significantly outperforms the baseline. As shown in Table 3, ColorJitter and moire pattern augmentations individually lower the baseline ACER of 38.05% to 3.62% and 6.18%, respectively. The combined augmentation of SPSC further reduces ACER to 1.32% in protocol 2.1, a substantial improvement of 36.73% over the baseline.

| Backbone | Moire | Color | APCER ↓ | BPCER ↓ | ACER ↓ |
|---|---|---|---|---|---|
| Resnet50 | ✗ | ✗ | 76.03 | 0.06 | 38.05 |
| Resnet50 | ✓ | ✗ | 11.07 | 1.28 | 6.18 |
| Resnet50 | ✗ | ✓ | 4.74 | 2.51 | 3.62 |
| Resnet50 | ✓ | ✓ | 2.55 | 0.09 | **1.32** |

Table 3. Ablation studies. Protocol 2.1 comparisons reveal that the combined moire pattern augmentation and ColorJitter(Color), termed SPSC, excel with the lowest ACER of 1.32%.

| Backbone | SDSC | Noise | APCER ↓ | BPCER ↓ | ACER ↓ |
|---|---|---|---|---|---|
| Resnet50 | ✗ | ✗ | 88.59 | 0.11 | 44.35 |
| Resnet50 | ✓ | ✗ | 1.73 | 1.58 | **1.65** |
| Resnet50 | ✗ | ✓ | 71.60 | 1.93 | 36.77 |
| Resnet50 | ✓ | ✓ | 44.0 | 1.08 | 22.57 |

Table 4. In Protocol 2.2, ablation studies contrast the effects of SDSC and GaussNoise(Noise). Employing only SDSC achieved an optimal ACER of 1.65%. In contrast, GaussNoise, designed to simulate adversarial attack noise, proved ineffective, resulting in an ACER of 36.77%.

**Simulated Digital Spoofing Clues** SDSC is developed to simulate spoofing cues from digital attacks. As shown in Table 4, using the complete training set and live samples from the development set and employing Resnet50 [9] as the backbone, SDSC significantly lowers the baseline ACER from 44.35% to 1.65%, a reduction of 42.7%, underscoring the effectiveness of SDSC. Conversely, GaussNoise, intended to mimic adversarial attack cues, was less effective, resulting in an ACER of 36.77%. The combined use of SDSC and GaussNoise yields an ACER of 22.57%.

### 4.5. Comparisons of Different Backbones

As shown in Table 5, we show the performance of our method with five different backbones: Resnet18, Resnet34, Resnet50 [9], Swin-Tiny and Swin-Base [30]. Due to the smaller size of the training set in the UniAttackData Dataset [5], our method exhibited some fluctuations in results across different backbones. However, these fluctuations remained within a reasonable range. We posit that our approach functions akin to a plug-in, capable of being easily integrated with any backbone architecture. This adaptability suggests that our method is not tightly coupled with the network structure. Our belief is that the proposed method is agnostic to the network structure, which is a significant advantage for its application in various scenarios. Ultimately, our method achieved the best ACER result of 1.06% on the Resnet50 backbone, demonstrating its effectiveness and the potential for broader applicability.

| Backbone | AUC ↑ | APCER ↓ | BPCER ↓ | ACER ↓ |
|---|---|---|---|---|
| Resnet18 [9] | 99.70 | 3.19 | 1.41 | 2.30 |
| Resnet34 [9] | 99.83 | 2.33 | 1.03 | 1.68 |
| Resnet50 [9] | 99.94 | 1.53 | 0.69 | 1.06 |
| Swin-Tiny [30] | 99.88 | 2.10 | 0.96 | 1.53 |
| Swin-Base [30] | 99.79 | 3.15 | 1.27 | 2.21 |

Table 5. Comparison of our method with different backbones.

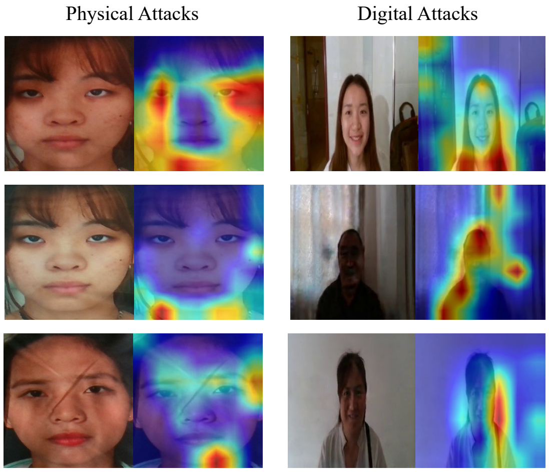

Figure 6. Visualization of attention maps for physical attacks and digital attacks.

### 4.6. Visualizations

We employ GradCam [7] to interpret and locate attack-specific features within the challenge. Figure 6 illustrates that in Protocol 2.1, the focus of the model is mainly on areas with color distortion or moire patterns, indicative of physical attacks. In Protocol 2.2, attention shifts to facial edge artifacts or distortion regions associated with digital attacks. These patterns confirm that our model has effectively learned to identify spoofing cues for both physical and digital attacks, achieving superior performance.

## 5. Conclusion

In this paper, our method introduces two novel data augmentations: Simulated Physical Spoofing Clues augmentation (SPSC) and Simulated Digital Spoofing Clues augmentation (SDSC). Extensive experimentation demonstrates their substantial enhancement of the model's detection and generalization capabilities for "unseen" attacks. Finally, our method won first place in "Unified Physical-Digital Face Attack Detection" of the 5th Face Anti-spoofing Challenge@CVPR2024.

## References

[1] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 1

[2] 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 *AAAI*, 2021. 2

[3] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021. 1

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

[5] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, et al. Unified physical-digital face attack detection. *arXiv preprint arXiv:2401.17699*, 2024. 1, 4, 6, 7

[6] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing. *arXiv preprint arXiv:2005.03922*, 2020. 1, 2

[7] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 7

[8] Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20818–20827, 2023. 1, 2

[9] 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. 5, 7

[10] Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409, 2020. 1

[11] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 1

[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. *arXiv preprint arXiv:2203.12175*, 2022. 2

[13] Shan Jia, Guodong Guo, and Zhengquan Xu. A survey on 3d mask presentation attack detection and countermeasures. *Pattern recognition*, 98:107032, 2020. 1

[14] 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. 2

[15] Minha Kim, Shahroz Tariq, and Simon S Woo. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1001–1012, 2021. 1

[16] Yunsheng Li, Lu Yuan, Yinpeng Chen, Pei Wang, and Nuno Vasconcelos. Dynamic transfer for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10998–11007, 2021. 1

[17] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186, 2022. 2

[18] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[19] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021. 1

[20] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 1, 4

[21] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021. 2

[22] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 814–823, 2021. 1

[23] Ajian Liu, Jun Wan, Ning Jiang, Hongbin Wang, and Yanyan Liang. Disentangling facial pose and appearance information for face anti-spoofing. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4537–4543. IEEE, 2022. 1, 2

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

[25] Ajian Liu, Zichang Tan, Yanyan Liang, and Jun Wan. Attack-agnostic deep face anti-spoofing. In *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6335–6344, 2023. 2

[26] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[27] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[28] 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. 2

[29] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[32] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15315–15324, 2022. 1

[33] Aakash Varma Nadimpalli and Ajita Rattani. On improving cross-dataset generalization of deepfake detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 91–99, 2022. 2

[34] Jérôme Rony, Eric Granger, Marco Pedersoli, and Ismail Ben Ayed. Augmented lagrangian adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7738–7747, 2021. 1

[35] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023. 1

[36] 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. 2

[37] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 3, 4

[38] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *ICCV*, 2023. 2

[39] 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. 2

[40] Keyao Wang, Guosheng Zhang, Haixiao Yue, Ajian Liu, Gang Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Multi-domain incremental learning for face presentation attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5499–5507, 2024. 1, 2

[41] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *2021 International Conference on 3D Vision (3DV)*, pages 679–688. IEEE, 2021. 1

[42] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[43] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *arXiv preprint arXiv:2107.01396*, 2021. 1

[44] Chiu Wai Yan, Tsz-Him Cheung, and Dit-Yan Yeung. Ila-da: Improving transferability of intermediate level attack with data augmentation. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[45] Shijie Yu, Feng Zhu, Dapeng Chen, Rui Zhao, Haobin Chen, Shixiang Tang, Jinguo Zhu, and Yu Qiao. Multiple domain experts collaborative learning: Multi-source domain generalization for person re-identification. *arXiv preprint arXiv:2105.12355*, 2021. 1

[46] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5295–5305, 2020. 1, 2

[47] Zitong Yu, Rizhao Cai, Yawen Cui, Ajian Liu, and Changsheng Chen. Visual prompt flexible-modal face anti-spoofing, 2023. 2

[48] Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin H. M. Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark, 2023. 2

[49] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. 1

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

[51] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware

domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023. 2

[52] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, pages 391–407. Springer, 2022. 1

[53] Junhua Zou, Yexin Duan, Boyu Li, Wu Zhang, Yu Pan, and Zhisong Pan. Making adversarial examples more transferable and indistinguishable. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3662–3670, 2022. 1