Unified Face Attack Detection with Micro Disturbance and a Two-Stage Training Strategy

Jiaruo Yu, Dagong Lu[*], Xingyue Shi[*], Chenfan Qu, Fengjun Guo[†]
IntSig Information Co. Ltd
Shanghai, China
{jiaruo_yu, dagong_lu, xingyue_shi, chenfan_qu, fengjun_guo}@intsig.net

Abstract

Face recognition systems are widely used in real-world scenarios but are susceptible to physical and digital attacks. Effective methods for unified detection of both physical face attacks and digital face attacks are essential to ensure the reliability of face recognition systems. However, how to obtain a unified face attack detection model that has adequate ability of fine-grained perception and cross-domain generalization ability remains an open challenge. To address this issue, we first propose a two-stage training strategy, which utilizes unlabeled face images with masked image modeling and unleashes the potential of vision transformers. Furthermore, we propose a novel method termed as Micro Disturbance, which successfully enriches the representation distribution of forged faces and increases the diversity of the training data, thereby addressing the issue of cross-domain generalization. Attribute to the effectiveness of our proposed methods, our model finally wins the third place in the 5th Face Anti-Spoofing Challenge@CVPR2024, with an impressive ACER score of 5.511.

1. Introduction

Face recognition systems have been widely applied in various scenarios such as face unlocking and face payment [52]. However, the rapid development of face manipulation and face generation methods makes it easier to fool face recognition systems, which poses serious risks to the security of social media [48]. It's crucial to develop effective methods for face image forensic.

Existing face manipulation methods can be divided into two categories, physical presentation and digital editing [7]. The former is mostly achieved using physical media, such as photo printing and video replaying [20]. The latter is mostly achieved by editing digital images with deep neural networks [43]. The board diversity of face manipulation techniques makes it challenging to verify the authenticity of face images, as shown in Figure 1.

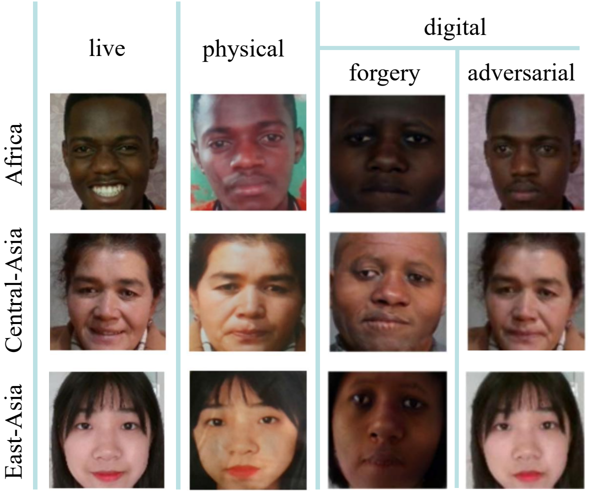

Figure 1. The samples from UniAttackData [13].

Recently, numerous methods have been proposed for face forgery detection, such as handcrafted features [3, 8] and deep forensic models [51]. Various types of external information are utilized for better performance, such as depth maps [39] and biological features [14]. Despite significant progress has been made, there are still some problems that have not been solved by previous work. Firstly, the challenging nature of face forgery detection. The difference between forged faces and real faces can be very subtle, posing significant demands on the fine-grained perceptual abilities of models [15]. Second, previous methods mostly handle the physical presentation attack and the digital editing attack separately [53], which requires large computational resources and considerably increases the inference time [13]. Third, it is almost impossible to include all types of the real-world face forgeries in the training data [41]. However, existing face forensic models trained on specific datasets gen-

[*] Equal contribution.
[†] Corresponding author.

erally struggle to perform well on unseen attack types [35]. The above problems bottleneck the real-world application of face recognition systems and cause risks for social information security.

To address the issue of fine-grained perception and unified face attack detection, we propose to harness the power of vision transformers. However, the vision transformers tend to overfit the limited training data, making it difficult to directly utilize them for face attack detection. To alleviate the data hunger of vision transformers (ViTs), we propose a two-stage training strategy. In the proposed strategy, we first pre-train the ViT model with both the real faces and fake faces in a masked image modeling manner: randomly mask 75% of the face images and train the model to predict the masked regions. Afterwards, we replace the original decoder with a single linear classifier and fine-tune the model on the target data. Despite its simplicity, this can effectively retain the knowledge learned in the pre-training stage and avoid catastrophic forgetting as much as possible.

In order to simultaneously address the issue of cross-domain generalization and fine-grained perception, we proposed a novel method, termed as Micro Disturbance. In the method, we slightly perturb both of the real faces and the forged faces with the random combination of four approaches (color jitter, screen simulation, JPEG compression and Gaussian blur), and relabel all the processed faces as forged faces despite their original labels. The processed faces can have different distributions with both of the authentic faces and the existing fake faces. Therefore, with the proposed Micro Disturbance, we significantly enrich the representation distribution of the forged faces and increase the diversity of the training data, thereby addressing the issue of cross-domain generalization. In addition, since the proposed Micro Disturbance only generates only subtle anomalies, the models' ability for fine-grained perception can also be considerably improved. As a result, models trained with the proposed Micro Disturbance achieve a significant 10 points performance gain in the 5th Face Anti-Spoofing Challenge@CVPR2024.

We conduct experiments on UniAttackData [13], the official dataset of the 5th Face Anti-Spoofing Challenge@CVPR2024. The UniAttackData contains both of the physical and digital attacks, and its digital attacks involve digital adversarial and digital forgery attacks, as shown in Figure 1. Therefore, it provides the chance to design unified attack detection frameworks. We win the third place in the 5th Face Anti-Spoofing Challenge@CVPR2024. This demonstrates the effectiveness of our methods in real-world face attack detection.

In summary, our contribution is three-fold:

- We propose a two-stage learning strategy that significantly enhances the forensic model's generalization ability on detect unseen attack types.
- We propose Micro Disturbance, a novel method to improve data diversity and address data scarcity. This method effectively improves the model's ability to generalize across unseen scenarios.
- Our method wins the third place in the 5th Face Anti-Spoofing Challenge@CVPR2024. Extensive experiments verify the effectiveness of our methods.

2. Related Work

2.1. Face anti-spoofing

Face anti-spoofing aims to detect the physical presentation attacks, which impersonates a specific live subject to achieve false acceptance [10]. There are various approaches to achieve physical presentation attack, such as photo attack, video replay attack, and 3D mask attack [2, 52], which makes the face anti-spoofing task challenging. Early works attempted to achieve face anti-spoofing with handcrafted features, such as LBP [8], HoG [50], SIFT [36] and SURF [3]. More recently, deep neural networks have accelerated the development of face anti-spoofing methods with their strong feature extraction ability [39, 51]. Deep learning based face anti-spoofing methods have demonstrated significant improvements over the conventional methods [30]. Yang et al. [51] utilized deep neural network to perform binary classification on live vs spoofed samples. Shao et al. [39] proposed to improve face anti-spoofing with depth information. Liu et al. [30] proposed a framework to train an attack system and a defense system simultaneously with adversarial training. Liu et al. [32] introduced prompt learning for generalizable face anti-spoofing. Wang et al. [46] proposed multi-domain incremental learning and Liu et al. [27] utilized adversarial cross-modality translation for better performance. Recently, many face attack detection challenges have been held, successfully promoting the development of the field of face attack detection. CASIA-SURF [54] paid attention to multi-modal face anti-spoofing [24, 55]. CeFA [26] improved cross-ethnicity face anti-spoofing [25]. HiFiMask [29] facilitated 3d high-fidelity mask face presentation attack detection [28] and SuHiFiMask [12] improved surveillance face presentation attack detection [11, 45]. Despite the progress has been made, effectively detecting spoofed faces of unseen types of attack in an unified manner remains an open challenge.

2.2. Face forgery detection

Except for physical presentation attack, digital forgery attack also causes serious risks for the security of face recognition systems [43]. In contrast to physical presentation attack that deceives with physical mediums, digital forgery attack performs digital editing at the pixel level. There are also various approaches for digital forgery attack, such as deep face synthesis [18], identity swap [42], and attribute

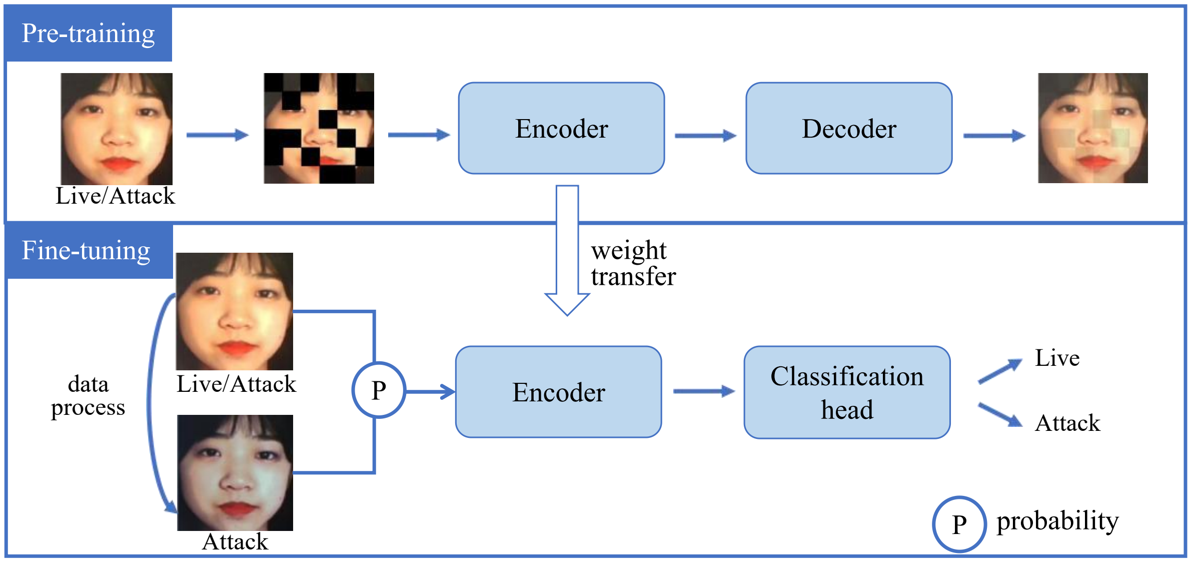

Figure 2. The pipeline of the two-training strategy. In the pre-training stage, we randomly mask the input image to predict raw RGB pixels for extracting discriminative representations. In the fine-tuning stage, we fine-tune the model using target data.

manipulation [6]. Consequently, face forgery detection becomes another challenge that increasingly attracts attention in the field of information security. Earlier methods for face forgery detection focused on capturing spatial artifacts of manipulated images, and trained a binary classifier straightforwardly [1, 4, 17, 47]. Subsequent work has used specific types of artifact clues, including frequency, blending artifacts, resolution difference [5, 22, 37]. It is an intuitive and common idea to synthesize and augment data in order to improve the diversity of datasets and thus enhance generalization. In face forgery detection, methods are proposed to synthesize forgery images by blending two images, providing data for the model to learn to detect the blending boundary artifacts [21, 40, 56].

2.3. Unified face attack detection

In the majority of existing works, the detection of physical presentation attack and digital forgery attack are regarded as two individual computer vision tasks, face anti-spoofing and face forgery detection. However, due to the diversity of real-world attacks, treating the two tasks separately requires large computational resources and significantly increases the inference time, making it a suboptimal choice [13]. Recently, researchers proposed to detect both physical presentation attacks and face forgery attacks in a single unified model. UniFAD [9] aimed to detect 25 coherent attack types from three categories, adversarial attack, digital attack, and physical attack. It employed a multi-task learning framework together with k-means clustering to learn joint representations for coherent attacks. Yu et al. [53] established the first joint benchmark for face spoofing and forgery detection using both visual appearance and physiological rPPG cues. Fang at al. [13] collected UniAttackData dataset used in the 5th Face Anti-Spoofing Challenge@CVPR2024, and further proposed a UniAttackDetection based on Vision Language Models (VLMs).

3. Methods

In this section, we introduce our methods for unified face attack detection, which consist of a two-stage training strategy and the Micro Disturbance. The Micro Disturbance is utilized in the last stage of our training strategy.

3.1. Two-Stage Training Strategy

Recently, Vision Transformers (ViTs) have demonstrated strong capabilities for fine-grained perception and long-range modeling, achieving outstanding performance in many computer vision tasks. However, ViTs suffer from data hunger and tend to overfit small training data. To address this problem, we introduce a two-stage training strategy, where the ViT model is first pre-trained on the face images using masked image modeling , and then fine-tuned on the target data, as shown in Figure 2.

Backbone. Transformer-based models have been widely adopted due to their promising performance in computer vision tasks [19]. Transformer is known for utilizing attention to model long-range dependencies in the data. We choose to use the Swin Transformer [33] as our backbone model due to its high performance and computational efficiency. It limits self-attention computation to non-overlapping local sliding windows and has hierarchical architecture that provides flexibility. The image is evenly partitioned with non-

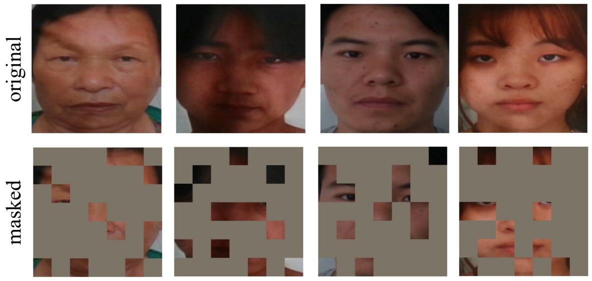

Figure 3. Masked image modeling in the pre-training stage. The first row are original images. The second row are masked images.

overlapping sliding windows. Shifted window-based self-attention computes self-attention within these windows, avoiding the need for global attention computation required by the standard Transformer architecture. This approach also allows for cross-window connections for knowledge interaction. The hierarchical architecture generates feature maps in a hierarchical structure that can detect artifact cues at different scales, which can be useful for spotting face attacks. In addition, the self-attention in Swin Transformer can help the model to pay more attention to regions with the most forgery cues, thereby achieving robust face attack detection.

Pre-training Stage. To avoid over-fitting, we first pre-train the Swin-Transformer on the face images from UniAttackData. Generally, the self-supervised pre-training techniques can be categorized into two types: contrastive-based (e.g. MoCo [16]) and reconstruction-based (e.g. SimMIM [49]). The contrastive-based methods require the construction of positive and negative pairs [16]. For face images, the negative pairs should be faces of different people to avoid confusion. However, the UniAttackData does not provide the necessary person-ID information. Therefore, we can only construct positive and negative pairs with random person-IDs. Furthermore, since the images in UniAttackData are extracted from videos, many samples have the same person-ID. As a result, many of the constructed negative pairs have the same ID, which is not semantically meaningful and confuses the model, leading to unsatisfactory performance. Therefore, we adopt reconstruction-based pre-training method Masked Image Modeling (MIM) and utilize the SimMIM framework for implementation.

Each input image is cut into 28 × 28 patches of 8 × 8 patch size. We then randomly mask 75% of the patches, as shown in Figure 3. The patches are then fed into the Swin Transformer to predict the raw pixel value, where the encoder is the Swin Transformer (Large) and the decoder is a single linear layer.

An l_1-loss is applied on the masked pixel prediction:

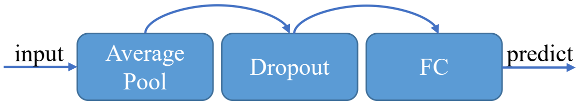

Figure 4. The decoder in the fine-tuning stage.

$$L_{pre} = \frac{1}{\Omega(x_M)} ||y_M - x_M||_1 \tag{1}$$

where $x_M, y_M \in \mathbb{R}^{3\times H\times W\times 1}$ are the input RGB pixel values and the predicted pixel values, respectively. H and W are the height and width of the image. $\Omega(\cdot)$ is the number of pixels.

Fine-tuning Stage. In the fine-tuning stage, we initialize the Swin-Transformer with the pre-trained weights, and replace the decoder with a single linear binary classifier, as shown in Figure 4. We fine-tune the model using the target data in UniAttackData and the data generated by the Micro Disturbance, which is proposed in Section 3.2. The cross-entropy loss is used to train the binary classifier:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_i log(p_i) \tag{2}$$

where labels are converted to one-hot vectors and $C = 2$ stands for live face and fake face, respectively.

The key idea of this method is to alleviate the data hunger for vision transformers by using MIM pre-training on face images, thereby unleashing their potential. This is different from previous works [23, 31] that utilize elaborated model designs.

3.2. Micro Disturbance

Motivation. To improve the model's ability for fine-grained perception and cross-domain generalization, we proposed Micro Disturbance, a novel method to expand the diversity of the existing small training data. The key idea behind the Micro Disturbance is to slightly disturb both of the live samples and attack samples, and regard all of the outputs as attack samples. The outputs of the Micro Disturbance effectively alleviate the scarcity of diverse, high-quality training data for face attack detection. After training with Micro Disturbance, the model can better identify the subtle anomaly and learn a more robust representation of live samples, thereby achieving better generalization to unseen attacks.

Implementation. As shown in Figure 5, to implement the Micro Disturbance, we process both of the live faces and fake faces in the UniAttackData with the random combination of four types of image processing techniques:

- Color Jitter. We randomly change the brightness, contrast, and saturation of the input sample with probability of 0.3 for each operation.
- Sreen Simulation. We perform random moire pattern insertion on the input sample, this well imitates the screen display spoofing. In addition, we perform random gamma correction on the input sample, which can simulate the possible color deviation on the screen. The probability for each method is 0.5.
- Image Compression. We perform random JPEG compression on the input image to achieve the quality degradation and to generate anomaly clues. The probability of JPEG compression is 0.5.
- Image Blur. We apply random Gaussian Blur to the input image, which can also generate anomaly features. The probability of the Gaussian blur is set to 0.5.

The above image processing methods are performed online, using at least one of the methods, color jitter. After processing, all output images are labeled as attack samples.

Advantages. Despite its simplicity, the clear advantages and interpretability of the proposed Micro Disturbance are manifold, as detailed below:

1. The Micro Disturbance facilitates the model's ability for fine-grained perception, which is essential for face attack detection. The operations of the proposed Micro Disturbance only slightly change the appearance of the input face. When a live face is processed with the Micro Disturbance, the output is labeled as a fake face. The model must improve its ability to detect the subtle anomaly during training. Consequently, the trained model demonstrates a strong ability for fine-grained perception and thus achieves better performance for face forensic.
2. The Micro Disturbance increases the diversity of the fake faces, thereby significantly alleviating the scarcity of high quality training data. The faces processed by the Micro Disturbance have different distributions of both authentic and existing fake faces. Therefore, the outputs can cover more types of artifacts, effectively improving the model's ability to generalize with their diversity. The model's cross-domain generalization ability can also be improved with the processed samples as these samples have diverse features and thus can prevent the model from over-fitting specific patterns in a single domain.
3. The Micro Disturbance compresses the feature distribution of live samples, so it can help the model generalise across unseen forgeries. By applying the Micro Disturbance, the features of live samples tend to be more compact and robust. The feature of the unseen type of face attack will have a more discriminative distribution than that of the live samples, so the model is better able to identify the unseen face attack.

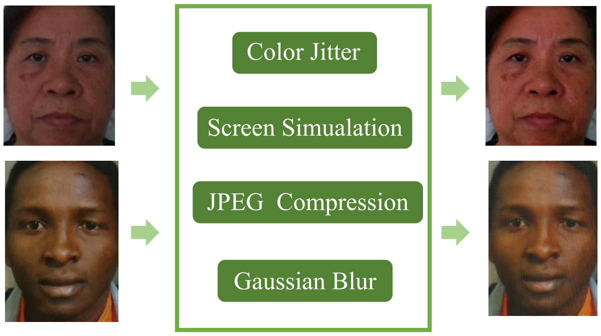

Figure 5. The pipeline of the proposed Micro Disturbance. The random combination of the four image processing techniques is applied to a random face, and the processed face is labeled as fake.

4. Experiments

4.1. Implementation Details

Our methods are implemented with the PyTorch framework. We use two NVIDIA-A100 GPUs with 80GB memory to train our model. The backbone network is Swin-Large, with the window size set to 12 [33].

In the pre-training stage, the batch size is set to 80 on each gpu. The initial learning rate is 3.125e-05, and the optimizer adopted is AdamW [34] with the weight decay set to 5e-02. The model is pre-trained for a total of 800 epochs on the UniAttackData. All images are resized to 224×224 before being fed to the model.

In the fine-tuning stage, the batch size is set to 64 per device, 4 devices in total. We keep using AdamW optimizer in this stage, and adjust the initial learning rate to 1e-04 and the weight decay parameter to 1e-02. The input image size remains 224×224.

4.2. Datasets and Evaluation Protocols

The UniAttackData dataset is from the 5th Face Anti-Spoofing Challenge@CVPR2024. In total, there are 35,906 images from 14 types of attacks, 2 types of physical attacks and 12 types of digital attacks. This enables the dataset to serve as a benchmark for unified face attack detection. The physical attacks include the most common spoofing types, printing attacks and replay attacks. The 12 digital attacks include 6 digital adversarial attacks and 6 digital forgery attacks. More details on the implementation of the attacks can be found at [13].

In the 5th Face Anti-Spoofing Challenge@CVPR2024, there are three different protocols, as shown in Table 1. In Protocol 1, models are tested on seen forgery: both of the training and testing sets include all types of face attacks. In contrast to protocol 1, in protocol 2.1 and protocol 2.2 the models are tested on unseen forgery. For example, the attack in the training set of Protocol 2.1 is digital attack, while

Pro	Set	Type				Total
		#Live	#Phys	#Adv	#Forg	
P1	Train	3000	1800	1800	1800	8400
	Val	1500	900	1800	1800	6000
	Test	4500	2700	7106	7200	21506
P2.1	Train	3000	0	9000	9000	21000
	Val	1500	0	1706	1800	5006
	Test	4500	5400	0	0	9900
P2.2	Train	3000	2700	0	0	5700
	Val	1500	2700	0	0	4200
	Test	4500	0	10706	10800	26006

Table 1. Statistics of UniAttackData [13]. Amount of train/val/test images of different types under three different protocols: P1, P2.1, and P2.2. Digital attacks are divided into finer classes digital adversarial attacks and digital forgery attacks.

Team	APCER ↓	BPCER ↓	ACER ↓	Rank
MTFace	0.9259	3.7533	2.3396	1
SeaRecluse	0.3999	6.4737	3.4369	2
Ours	5.5185	5.5037	5.5111	3
BSP-Idiap	9.3629	23.0698	16.2263	4
VAI-Face	0.2593	34.0055	17.1324	5

Table 2. Performance in the Unified Physical-Digital Face Attack Detection @CVPR2024. The mark ↓ indicates the lower the better.

in the testing set it is physical attack, similar to Protocol 2.2. As a result, Protocol 2.1 and Protocol 2.2 can evaluate the model's ability to generalize across the unseen domain.

4.3. Evaluation Metrics

We adopt three metrics to evaluate performance, the Attack Presentation Classification Error Rate (APCER), the Bona Fide Presentation Classification Error Rate (BPCER) and the Average Classification Error Rate (ACER). The definitions of APCER and BPCER are formulated as:

$$APCER = FN/(TP + FN) \tag{3}$$

$$BPCER = FP/(FP + TN) \tag{4}$$

where TP, TN, FP, and FN stand for the number of true positive, true negative, false positive, and false negative samples respectively. The ACER is defined as the average of APCER and BPCER:

$$ACER = (APCER + BPCER)/2\,. \tag{5}$$

No.	TS	MD	APCER ↓	BPCER ↓	ACER ↓
1	-	-	12.5851	13.3733	12.9792
2	√	-	0.84444	29.9455	15.3950
3	-	√	46.3481	26.6967	36.5224
4	√	√	5.5185	5.5037	5.5111

Table 3. Ablation Study on UniAttackData. 'TS' and 'MP' denote our two-stage training strategy and our proposed Micro Disturbance, respectively. The mark ↓ indicates the lower the better.

4.4. Comparison Study

We compare the performance of our methods with other teams on the testing set of UniAttackData, the comparison study is shown in Table 2. In this table , our model achieves APCER, BPCER and ACER by 5.5185, 5.5037 and 5.5111, respectively. We win the third place in the 5th Face Anti-Spoofing Challenge@CVPR2024. In Table 2, the ACER of our method is significantly better by more than 10 points than the fourth team, which confirms the effectiveness of our methods.

4.5. Ablation Study

To verify the effectiveness of the proposed methods, we conduct ablation study on the UniAttackData dataset, the results are shown in Table 3. In this table, the model with neither the proposed two-stage training strategy nor the proposed Micro Disturbance serves as the baseline (No. 1). Evidently, the model equipped with both of the proposed methods (No. 4) achieves a notable **7 points** lower ACER than the baseline (No. 1). Moreover, the proposed Micro Disturbance only works when the model is trained with our two-stage training strategy (No.3 worse than No.2), demonstrating the importance of the proposed strategy. To further find out how the proposed Micro Disturbance works, we visualize the features of the live, fake faces and the samples generated by our Micro Disturbance respectively using T-SNE [44]. As shown in Figure 6, for both protocols, **the representation distribution of the generated samples (pink) is close to both the real physical attacks (green) and the digital attacks (orange), and has more diversity.** This means that the generated samples obtained from our Micro Disturbance can well simulate various types of real-world attacks, thereby significantly alleviating the scarcity of diverse, high-quality data. It is attribute to our proposed simple-yet-effective methods that our model can achieve much lower error rate than the baseline.

4.6. Visualization

Visualization with Grad-CAM. We use Grad-CAM [38], a gradient-based visual explanation algorithm to analyse the attention maps of our model, the results are shown in

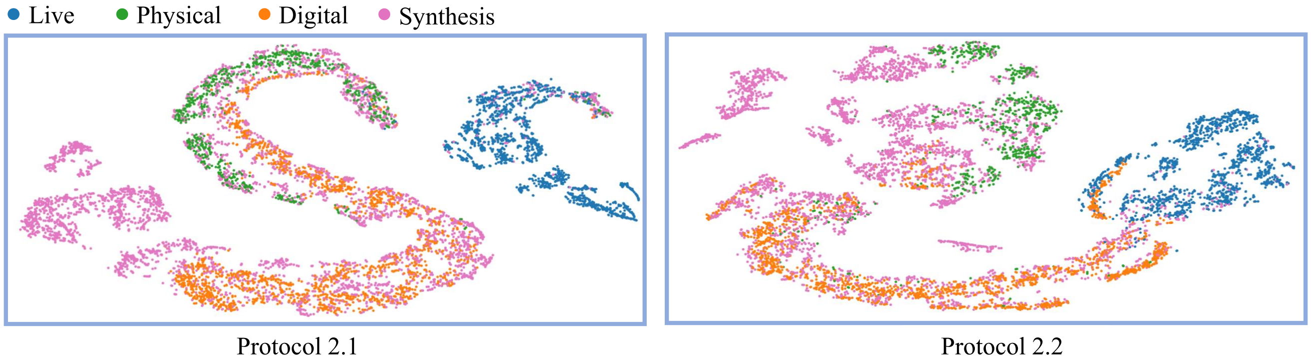

Figure 6. Visualization of feature maps obtained by the trained model under Protocol 2.1 and Protocol 2.2. The samples are come from training set of Protocol 1.

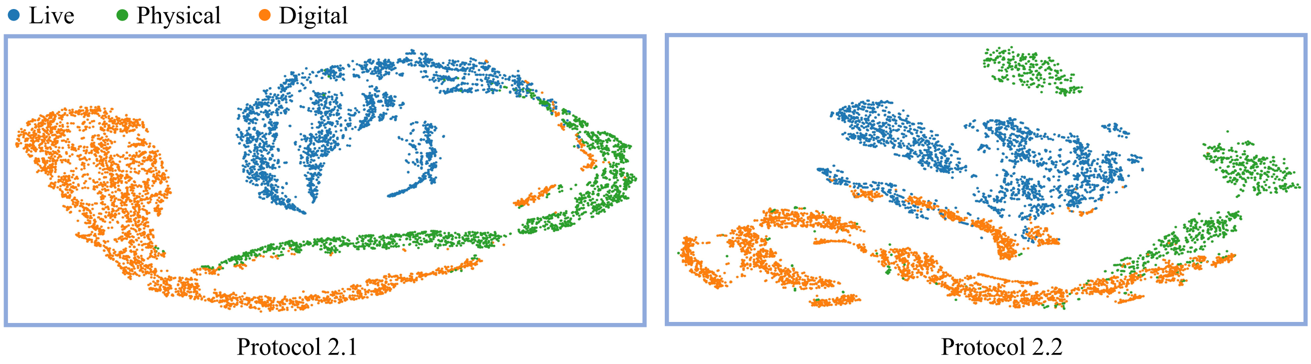

Figure 7. Visualization of feature maps obtained by the trained model under Protocol 2.1 and Protocol 2.2. The samples are come from training set of Protocol 1.

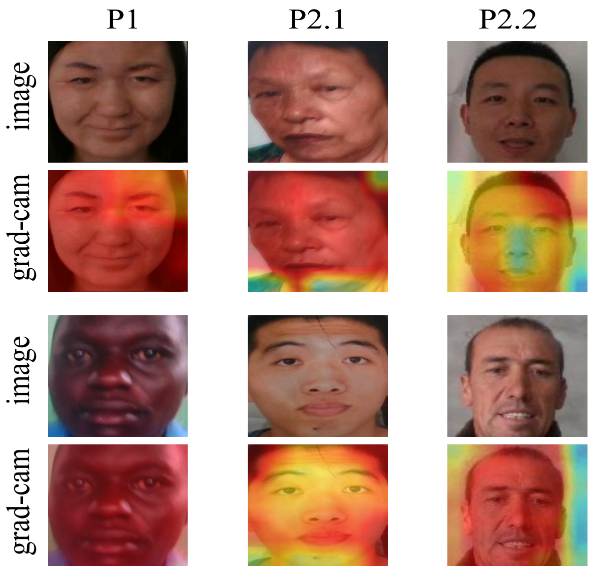

Figure 8. The visualization of the results under different protocols.

Figure 8. Evidently, even in the challenging Protocol 2.1 and Protocol 2.2 where the model is tested on unseen attack types, our model can pay attention to the regions that have the most forgery clues. This indicates that our model is robust to unseen attack types and can identify the subtle anomaly for accurate face attack detection.

Visualization with T-SNE. In order to further demonstrate the effectiveness of our model, we display the feature maps of different types of samples with the T-SNE [44] algorithm in Figure 7. In Protocol 2.1, the model is trained without seeing any physical samples. Similarly, the model trained in Protocol 2.2 has not been exposed to digital samples. Despite the absence of certain attack types, it can be observed from the visualization in Figure 7 that the models are able to clearly distinguish between most of the live samples and attack samples, further demonstrating the effectiveness of our proposed methods.

5. Conclusion

Unified face attack detection is crucial for guaranteeing the reliability of face recognition systems and ensuring social media security. In this paper, we analyse the existing problems for unified face attack detection and propose two methods to improve the performance for face forensic models. To address the problem of fine-grained perception and data scarcity, we propose to harness the power of vision transformers that trained with a two-stage strategy. The two-stage training strategy effectively alleviates the data hunger of the vision transformers. Furthermore, to simultaneously address the issues of cross-domain generalization and fine-grained perception, we propose Micro Disturbance, which compresses the representation distribution of live faces and increases the diversity of the training data. Attribute to our simple-yet-effective methods, our model achieves 7 points lower ACER than the baseline method, and we win the third place in the 5th Face Anti-Spoofing Challenge@CVPR2024. We believe that our methods can shed light on the community and promote the real-world application of unified face attack detection.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE international workshop on information forensics and security*, pages 1–7, 2018. 3

[2] Josef Bigun, Hartwig Fronthaler, and Klaus Kollreider. Assuring liveness in biometric identity authentication by real-time face tracking. In *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*, pages 104–111, 2004. 2

[3] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2016. 1, 2

[4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *IEEE European Conference on Computer Vision*, pages 103–120, 2020. 3

[5] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022. 3

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 3

[7] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[8] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In *ACCV 2012 International Workshops*, pages 121–132, 2013. 1, 2

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

[10] Nesli Erdogmus and Sebastien Marcel. Spoofing face recognition with 3d masks. *IEEE Transactions on Information Forensics and Security*, 9(7):1084–1097, 2014. 2

[11] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, and Zhen Lei. Surveillance face presentation attack detection challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6360–6370, 2023. 2

[12] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z. Li, and Zhen Lei. Surveillance face anti-spoofing, 2023. 2

[13] 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, 2, 3, 5, 6

[14] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE transactions on information forensics and security*, 15:42–55, 2019. 1

[15] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 735–743, 2022. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 4

[17] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *IEEE International Symposium on Computer, Consumer and Control*, pages 388–391, 2018. 3

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 3

[20] Sandeep Kumar, Sukhwinder Singh, and Jagdish Kumar. A comparative study on face spoofing attacks. In *International Conference on Computing, Communication and Automation*, pages 1104–1108, 2017. 1

[21] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 3

[22] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 3

[23] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *International Joint Conference on Artificial Intelligence*, pages 1180–1186, 2022. 4

[24] 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 0–10, 2019. 2

[25] 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. 2

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

[27] 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

[28] 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. 2

[29] 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. 2

[30] Ajian Liu, Zichang Tan, Yanyan Liang, and Jun Wan. Attack-agnostic deep face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 6335–6344, 2023. 2

[31] 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. 4

[32] 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. 2

[33] 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 *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3, 5

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

[35] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 2

[36] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016. 2

[37] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103, 2020. 3

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[39] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 1, 2

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

[41] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2638–2646, 2021. 1

[42] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 2

[43] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1, 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 6, 7

[45] Dong Wang, Jia Guo, Qiqi Shao, Haochi He, Zhian Chen, Chuanbao Xiao, Ajian Liu, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, Jun Wan, and Jiankang Deng. Wild face anti-spoofing challenge 2023: Benchmark and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 6380–6391, 2023. 2

[46] 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. 2

[47] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are sur-

prisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 3

[48] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019. 1

[49] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 4

[50] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics*, pages 1–6, 2013. 2

[51] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 1, 2

[52] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5609–5631, 2022. 1, 2

[53] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *IEEE Transactions on Dependable and Secure Computing*, 2024. 1, 3

[54] 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. 2

[55] 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. 2

[56] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021. 3