DiffFAS: Face Anti-Spoofing via Generative Diffusion Models

Xinxu Ge[1], Xin Liu[2*], Zitong Yu[3*], Jingang Shi[4], Chun Qi[4], Jie Li[4], and Heikki Kälviäinen[2]

[1] Tianjin University, China
[2] Lappeenranta-Lahti University of Technology LUT, Finland
[3] Great Bay University, China
[4] Xi'an Jiaotong University, China
xinxu_ge928@tju.edu.cn, linuxsino@gmail.com, yuzitong@gbu.edu.cn, {jingang, qichun, jielixjtu}@xjtu.edu.cn, heikki.kalviainen@lut.fi

Abstract. Face anti-spoofing (FAS) plays a vital role in preventing face recognition (FR) systems from presentation attacks. Nowadays, FAS systems face the challenge of domain shift, impacting the generalization performance of existing FAS methods. In this paper, we rethink about the inherence of domain shift and deconstruct it into two factors: image style and image quality. Quality influences the purity of the presentation of spoof information, while style affects the manner in which spoof information is presented. Based on our analysis, we propose DiffFAS framework, which quantifies quality as prior information input into the network to counter image quality shift, and performs diffusion-based high-fidelity cross-domain and cross-attack types generation to counter image style shift. DiffFAS transforms easily collectible live faces into high-fidelity attack faces with precise labels while maintaining consistency between live and spoof face identities, which can also alleviate the scarcity of labeled data with novel type attacks faced by nowadays FAS system. We demonstrate the effectiveness of our framework on challenging cross-domain and cross-attack FAS datasets, achieving the state-of-the-art performance. Available at https://github.com/murphytju/DiffFAS.

1 Introduction

Face recognition [16] has been developed remarkably and been widely applied in many systems. However, it is vulnerable to physical presentation attacks, such as print attacks [2, 45], replay attacks [3, 39], 3D-mask attacks [7, 29], and novel types of presentation attacks are continually emerging. Therefore, face anti-spoofing [1, 13, 32, 34, 38] plays an increasingly important role in face recognition systems.

With the emergence of academic datasets [2, 3, 7, 29, 39, 45] over the past decade, deep learning-based FAS [1] has achieved significant performance gains. In order to enhance generalization capabilities, researchers have delved into

* Corresponding authors.

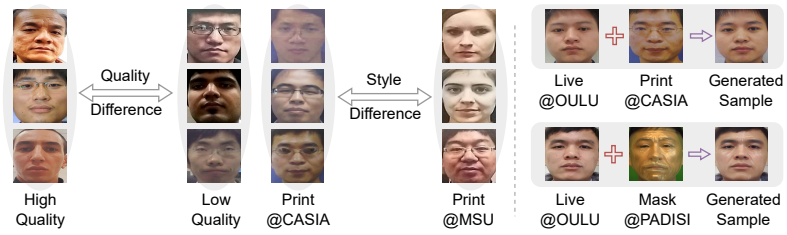


Fig. 1: DiffFAS identifies quality and style as separable elements (see the left part) in images and enables cross-domain, cross-attack generation (see the right part) to counteract style discrepancies due to domain shifts.

multi-source domain training approaches [13, 30, 32, 34, 38]. However, the utilization of multiple source domains gives rise to domain shifts, manifesting as *image style* and *image quality* challenges. As elucidated in [46], image quality is quantified by factors such as image degradation, blur, and low resolution, while image style is characterized by local image statistics [38], reflecting the color and texture differences between images of the same or different types of attacks, as shown in Fig. 1. Both quality and style introduce erroneous prior information to the network, thereby posing a formidable obstacle in learning precise spoof cues.

Existing domain generalization methods address domain shifts by incorporating human prior information, such as domain-specific details [13, 30] and ID information [34], which should be considered disturbances and need to be disregarded. However, recognizing the imprecision of these human priors, we propose employing *image quality* as a quantifiable metric [27] to directly against domain shift. For the unquantifiable *image style* factors, we advocate for the application of generative methods [10] as a viable solution.

Limited works focus on generative perspective to reduce domain shift impact. The implicit augmentation method [38] performs well on cross-domain print and replay attack. However, as a specific designed module, it is challenging to be intergrating into other FAS method. Additionally, existing FAS systems are susceptible to novel attack methods [42], but labeled data with novel type of attacks is extremely scarce in current academic datasets [7]. Therefore, explicit generation should be explored.

Existing generative methods mainly base on GAN [8] and VAE [19]. GAN-based methods, such as STDN [25], excel in transferring fine-grained attack textures, but do not address the domain shift between print and replay attacks, rendering them unsuitable for these types of attack's domain generalization. On the other hand, VAE-based methods [40] are limited by their generative capacities, often leading to local texture distortions which restrict their applicability in RGB-based methods. These approaches tend to overfit to limited data. For example, when fed a face image, the current model generates spoof data that closely resembles the original spoof face in the dataset. This generation method results in the erosion of unique identity features and provides only limited assistance for FAS tasks. Diffusion models [10] have demonstrated exceptional performance in

the realm of high-quality image generation. However, the denoising process, as discussed in [31], tends to suppress high-frequency information, encompassing vital texture cues crucial for FAS tasks. Additionally, existing diffusion models are prone to overfitting with limited data.

In this paper, we address domain shift from separate perspective of image style and image quality. For FAS tasks, live faces can be easily collected, but acquiring spoof faces is time-consuming and labor-intensive [42]. Thus, we perform live-spoof transformation to generate large number of spoof samples from live faces, as shown in Fig. 1. To better leverage diffusion models, we define the training objective as the reconstruction process of spoof images, which allows the network directly learning spoof texture to avoid texture degradation while mitigating the impact of overfitting. Additionally, we decouple spoof texture by condition guidance and design Spoofing Style Fusion Module (STFM) to suppress the extraction of identity information from the conditional branch. Finally, to preserve the identity information of live (face) images, we utilize image editing algorithms [33] to obtain spoof versions of live images. For image quality, we integrate quality quantification scores into loss function, introducing quality prior information during the training of classification model. This guides classification model to deeply mine the spoof cues presented in high-quality images, and directly mitigating the impact of image quality shift. We summarize our contributions as follows:

- We represent FAS domain shifts as quantifiable and unquantifiable components, namely image quality shift and image style shift, which provide a new perspective for multi-source domain face anti-spoofing.
- We redesign the training and inference processes to tackle texture degradation and overfitting in diffusion models, introducing STFM for decoupled texture guidance. Our DiffFAS framework delivers high-fidelity, identity-consistent spoof generation, mitigating labeled data scarcity. To our best knowledge, this is the first work to introduce diffusion models for FAS.
- We first introduce quality prior information for DG-FAS, and integrate quantifiable quality scores with the loss function to counter cross-domain image quality shifts.
- We achieve state-of-the-art performance in the cross-domain and cross-attack FAS protocol. High-fidelity generative sample improve upon the best baseline 4.4% average ACER in the setting of WMCA unseen protocol, demonstrating the effectiveness of the generative method.

2 Related Work

Domain Generalized Face Anti-Spoofing: Recently, deep learning-based methods [15,18,24,41] have demonstrated remarkable performance gains but encountered challenges in generalizing to unknown domains, posing hurdles for industrial deployment. Addressing this issue, multi-source domain FAS approaches

[13, 30, 32, 34, 38] have garnered increasing attention. SSDG [13] enhanced network performance by providing attack type information and constraining the network to learn more compact clusters for live samples. PatchNet [34], building on SSDG's priors, introduced more precise category labels and employs patch-wise learning to obfuscate identity information. SSAN [38] addressed domain shifts' image style factor by employing cross-domain implicit style shuffling. SA-FAS [36] advocated for aligning the live-spoof classification hyperplanes across different domains. Different from the aforementioned methods, we approach domain shifts by explicitly considering quality and style, enhancing the model's generalization performance. This is achieved by incorporating diverse human prior inputs specifically developed to counter quality and style shifts.

Generative Method for Face Anti-Spoofing: In recent years, there are a few Generative FAS methods [25, 36, 38, 40]. Wang *et al.* [36] found that simple data augmentation techniques like color transformations are beneficial for training anti-spoofing models. Liu *et al.* [25] trained an anti spoofing model by disentangling spoof trace from spoof samples for discrimination and simultaneously adding spoof traces on live samples to enhance classification model's training. Wu *et al.* [40] utilized VAE to learn a joint distribution of identity and spoofing pattern in latent space, thereby generating paired live and spoof images with new identity from random noise. Ho *et al.* [11] proposed classifier-free guidence, combining condition guidance with the training objective of diffusion model [10], pioneering the condition diffusion model and leading to diverse generation with multi-modal guidance [44]. However, diffusion models face challenges such as texture degradation and overfitting. To leverage the excellent generative capability of diffusion models, we redesign the generative process of diffusion models to produce identity-consistent spoof faces across multiple attack types and styles. This approach effectively addresses domain shift style factors and lessens the need for new attack type labeled data in FAS tasks.

3 Method

The proposed method, including the DiffFAS generative framework and the Sample-Level Relative Quality loss, which will be introduced next. DiffFAS generative framework consists of the following three critical properties: 1) *Precise spoofing style label*; 2) *Texture preservation*; 3) *Identity-consistency*; Labeled data helps to counter the limitations inherent in academic datasets, with texture being a crucial factor in distinguishing spoof faces. Maintaining identity consistency enables the clear observation of the advantages of authentic spoof generation, without the introduction of new identities. Then, the Relative Quality loss for classification is proposed. This strategy maximizes the extraction of discriminative cues from high-quality (HQ) images in FAS datasets, effectively preventing overfitting to low-quality (LQ) images.

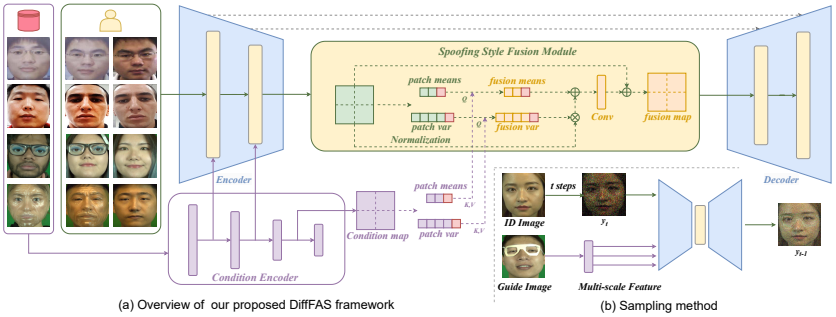


Fig. 2: The proposed DiffFAS Generative Framework is a UNet-based network composed of a Spoofing style encoder and a noise prediction module, training with Spoofing Style Pool (the first column) and Live-Spoof Pair Images (the second column). The spoofing style encoder extract texture of the random selected spoof image, and get multi-scale features from different encoder layers. We design an asymmetric Spoofing Style Fusion Module (STFM) to reduce the introduction of identity information of the conditional branch, and achieve information aggregation with the backbone through cross-attention. This allows the network to fully capture the spoof texture and achieve high-fidelity spoof synthesis. During the Inference stage, we employ image editing techniques for the sampling process, enhancing our control over the inference stage, and further improving the consistency between the generative ID and the original ID.

3.1 Preliminary

Diffusion Model. Diffusion models [10, 28] are generative models that are trained to predict an image from random noise through a gradual denoising process, which consists of the forward process and reverse process. The forward process progressively adds noise to the original image as

$$q(y_t|y_{t-1}) = \mathcal{N}(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t I), \tag{1}$$

where $t \backsim [1, T]$ and $\beta_1, \beta_2, ..., \beta_T$ is a fixed variance schedule with $\beta_t \in (0, 1)$, until the original image y_0 is degraded into Gaussian noise y_t. The true posterior $p(y_{t-1}|y_t)$ can be approximated by training a U-Net model ϵ_θ to predict the initial noise ϵ with the MSE loss as

$$\mathcal{L}_{mse} = \mathbb{E}_{t\backsim[1,T], y_0\backsim q(y_0), \epsilon}\|\epsilon - \epsilon_\theta(y_t, t)\|^2. \tag{2}$$

Then in the reverse process, we can sample from a Gaussian noise as

$$p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \epsilon_\theta(y_t, t)). \tag{3}$$

Generation Objective. We define spoofing style as various novel attack types and distinct styles of the same attack type. Consider multiple FAS datasets as $\{D_1, D_2, ...D_N\}$, with a live face $X_l^{id_1}$ from D_i and a spoof face $Y_s^{id_2}$ from D_j. Here, X and Y signify different spoofing styles. Our objective is to generate

$Y_s^{id_1}$, transforming $X_l^{id_1}$ into $Y_s^{id_1}$ while adopting the spoofing style of $Y_s^{id_2}$ as follows:

$$Y_s^{id_1} = \epsilon_\theta(Y_s^{id_2}, X_l^{id_1}). \tag{4}$$

However, due to the scarcity of cross-domain data pairs is the training process, we simplify this formula to

$$X_s^{id_1} = \epsilon_\theta(X_s^{id_2}, X_l^{id_1}), \tag{5}$$

where the training exclusively involves same-domain image pairs $(X_l^{id_1}, X_s^{id_1})$. The training and inference process is illustrated in Fig. 2.

3.2 DiffFAS Framework

Overall description. To combat texture degradation and excessive smoothing, DiffFAS merges the training objective with the generative goal through spoof image reconstruction. We use live (face) images from paired sets, matched with randomly selected spoof images of the same style, to direct the reconstruction process of spoof images. This process is formulated as follows:

$$X_t = \sqrt{\bar{\alpha}_t} X_s^{id_1} + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{6}$$

$$p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \epsilon_\theta((X_t, X_l^{id_1}), t, X_s^{id_2})), \tag{7}$$

where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. To maintain identity consistency, the live image $X_l^{id_1}$ is concatenated with the noisy version of spoof image $X_s^{id_1}$ and input into U-Net. During training progress, this operation aligns the reconstructed spoof image with the corresponding live image.

Spoofing Style Fusion Module. To effectively assimilate substantial spoof information from the randomly chosen guide image $X_s^{id_2}$, we utilize a pre-trained ResNet [9] denoted as ϕ. During the encoder training, precise spoofing style labels are provided, and the network ϕ is trained until convergence with cross-entropy loss. As the conditional control branch, feature maps are extracted from the outputs of layers to serve as conditions as

$$Cond_1, Cond_2, ..., Cond_n = \phi(X_s^{id_2}). \tag{8}$$

Employing feature maps directly as conditions in our model could influence the identity characteristics of the generated image during cross-domain or cross-attack inference scenarios. To effectively extract pertinent spoofing style features from the condition feature maps and concurrently minimizing identity distortion, we design Spoofing Style Fusion Module (STFM). Inspired by style transfer methodologies [12, 14] and face synthesis works [17], the means and variance of an image are employed to represent its style. Since we need to handle the fine-grained spoofing style carried by the spoof images, we partition the feature map into small patches, denoted as $\{P_1, P_2, ...P_N\}$. For each patch, we compute the mean and variance independently, thereby obtaining the representation

that is rich in localized information. These computed means and variances are subsequently flattened into the arrays.

In our framework, we propose to deploy asymmetric patch sizes in the means and variance branches. This approach is informed by the observation that, pooling operations, characterized by patch-wise means, tend to preserve identity information. In contrast, the n-neighborhood variance is primarily indicative of texture-related details. Consequently, opting for a larger patch size in the means branch becomes advantageous, as it effectively minimizes the influence of identity-related information. This strategic choice of patch size differentiation is pivotal in ensuring that the focus remains predominantly on the textural aspects, thereby enhancing the fidelity of spoofing style representation while concurrently attenuating identity interference. Then the global means and variance of the feature map are concatenated to incorporate comprehensive global style information into our model. The condition map $\{Cond_1, Cond_2, ..., Cond_n\}$ is processed into an array of means and variances, represented as $\{\mu_{cond}, \sigma_{cond}\}$. Parallel to this, analogous operations are executed on the feature maps derived from the U-Net backbone, generating a corresponding set of means and variances $\{\mu_x, \sigma_x\}$. The final stage involves performing cross-attention on these two distinct sets of sequences as follows:

$$\mu_f = attn(Q(\mu_x), KV(\mu_{cond})), \tag{9}$$

$$\sigma_f = attn(Q(\sigma_x), KV(\sigma_{cond})). \tag{10}$$

Once the fused feature means and variances are obtained, we remove the last element global information, resize, and upsample them to match the feature map's size with bilinear interpolation. Subsequently, we combine the processed means and variance matrix, yielding the fused spoofing style feature map as

$$X_f = X + conv(norm(X) * \sigma_f + \mu_f). \tag{11}$$

Then, we apply STFM to all U-Net blocks with dimensions less than or equal to 32×32.

3.3 Inference

Disentangled Guidance. As the model learns the mapping between image pairs in the dataset, it inherently acquires the capability to perform spoof generation. However, a critical challenge arises in controlling the intensity of the spoof strength. Excessively high spoof strength can inadvertently impact the identity information, leading to deviations from the original ID. To address this issue and enable better control over the spoofing style, we integrate classifier-free guidance [11] into our framework. Specifically, spoofing style condition will be setting to zero with a certain probability. During the sampling phase, the denoising outputs for both conditional and unconditional scenarios are computed

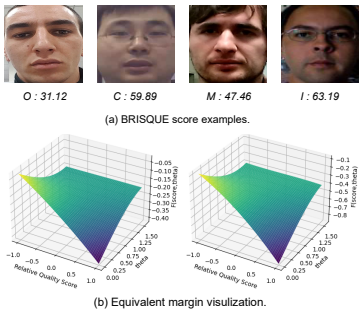


Fig. 3: (a) Examples for different domain's live sample BRISQUE score, and lower score means higher quality. (b) Visualization of the equivalent additive margin function $F(b_{rq}, \theta) = \cos(\theta + \psi) - \omega - \cos(\theta)$, with different scale coefficient 0.2, 0.4. As training progresses, the equivalent additive margin for HQ samples is larger, while LQ images are conversely.

independently at each timestep. By subtracting these outputs, we yield the specific influence of the guidance. Consequently, the sampling process is redefined as follows:

$$\begin{aligned} p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \epsilon_\theta((X_t, X_l^{id_1}), t)) + \\ \gamma * (\mathcal{N}(X_{t-1}; \epsilon_\theta((X_t, X_l^{id_1}), t, X_s^{id_2})) - \\ \mathcal{N}(X_{t-1}; \epsilon_\theta((X_t, X_l^{id_1}), t))). \end{aligned} \tag{12}$$

The influence of the spoofing style guidance on the inference stage can be controlled by adjusting the parameter γ.

Sampling. To enhance the model's ability to retain identity information and alleviate the effects of potentially problematic samples in datasets—such as image pairs with mismatched IDs—we draw inspiration from Unitune [33], namely, instead of directly sampling noise, our framework employs an image-editing sampling approach. Specifically, we introduce t steps of random noise to the original image, following Eq. 6, and then perform conditional sampling from the noisy live image, following Eq. 12.

3.4 Sample-Level Relative Quality Loss

As a primary manifestation of domain shifts, we propose that furnishing the network with quality prior information can significantly improve its performance. Instead of simple binary classification, we classify data based on specific spoofing styles. Recognizing the high similarity within each spoofing style class, we employ margin softmax loss to learn a compact clusters, similar to techniques found in cosface [35] or arcface [5]. These methods introduce an angular margin ψ or additive margin ω to the classifier, modifying the original softmax formula as:

$$\cos(\theta) \dashrightarrow g(\theta) = \cos(\theta + \psi) - \omega. \tag{13}$$

Inspired by Adaface [16], we realize that network attention towards certain images can be controlled through additive and angular margin. This allows us to integrate inherent image attributes with the loss, such as the image quality score BRISQUE [27], as illustrated in Fig. 3 (a). Notably, this score is inversely related to image quality. In FAS tasks, LQ images often contain spoof cues

Table 1: Comparisons with SoTA methods: Evaluated on four popular benchmark datasets with leave one out protocol: CASIA (**C**), Idiap Replay (**I**), MSU-MFSD (**M**), and Oulu-NPU (**O**). ↑ indicates larger values are better, and ↓ indicates smaller values are better.

Method	OCI→M		OMI→C		OCM→I		ICM→O		Average	
	HTER ↓	AUC ↑	HTER ↓	AUC ↑	HTER ↓	AUC ↑	HTER ↓	AUC ↑	HTER ↓	AUC ↑
MADDG [30]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02	23.09	84.36
SSDG-M [13]	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83	20.79	88.09
NAS-FAS [43]	19.53	88.63	16.54	90.18	14.51	93.84	13.80	93.43	16.09	91.52
DRDG [23]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75	15.67	92.04
SSAN-M [38]	10.42	94.76	16.47	90.81	14.00	94.58	19.51	88.17	15.10	92.08
SSDG-R [13]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.29	95.31
SSAN-R [38]	6.67	**98.75**	10.00	96.67	8.88	96.79	13.72	93.63	9.82	96.45
PatchNet [34]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.91	95.95
SA-FAS [32]	5.95	96.55	8.78	95.37	6.58	97.54	**10.00**	**96.23**	7.83	96.42
DiffFAS-R (ours)	**5.90**	98.10	**7.32**	**97.40**	**5.66**	**98.69**	12.38	94.33	**7.82**	**97.13**
DiffFAS-V (ours)	**2.86**	98.41	10.11	96.32	6.36	97.89	**8.11**	**97.27**	**6.86**	**97.47**

Table 2: Results of seen and unseen protocols on WMCA dataset. The values ACER(%) reported on testing sets are obtained with thresholds computed for BPCER=1% on development sets. The best results are bolded.

Method	Seen	Unseen							Average
		Glasses	Rigid Mask	Fake Head	Flexible Mask	Paper Mask	Print	Replay	
CCL [22]	27.14	35.13	15.10	21.82	**7.18**	18.91	20.53	11.79	18.64
DiVT-M [20]	6.89	34.52	4.86	10.01	22.56	6.07	2.65	23.11	14.83
EPCR [37]	–	**16.00**	3.40	0.70	49.7	**0.20**	**0.10**	**3.70**	10.50
MCDeepPixBiS [6]	5.68	24.72	3.29	0.78	24.55	2.14	1.19	20.89	11.08
DiffFAS-V	**4.25**	19.62	**2.17**	**0.00**	13.72	0.43	0.87	9.97	**6.68**

associated with local blurriness, which are not ideal for intense mining using margin softmax loss. Therefore, our goal is to focus more on mining from HQ images, while avoiding over-extraction from LQ images. First, we calculate the absolute quality scores $b_{aq} = \{b_1, b_2, ..., b_N\}$ for a batch, along with the average quality score and variance b_{mean}, b_{var} and the number of samples below or above the average, denoted as n_1, n_2. Next, we convert the distribution of these quality scores into a standard normal distribution:

$$b_{norm} = \frac{b_{mean} - b_{aq}}{(b_{std} + eps)}. \tag{14}$$

Note that normalized score is directly proportional to image quality. We then scale it according to the 3σ rule to ensure that a significant number of samples fall within a certain interval, and then limit the distribution to the range $[-n_1/N, n_2/N]$. This range approaches [-0.5,0.5] as the sample capacity tends to infinity, but extremely small scores within a batch can cause the sample mean is lower than the population mean and lead to $n_1 < n_2$, while a larger n_2/N can increase the margin applied to extreme samples, allowing adaptive handling of such extreme cases. The relative quality scores are calculated as

$$b_{rq} = torch.clip(\frac{b_{norm}}{3}, \frac{-n_1}{N}, \frac{n_2}{N}). \tag{15}$$

This ensures that the means of each batch are centered at zero on the relative quality axis, and each sample's margin is adaptively adjusted based on the

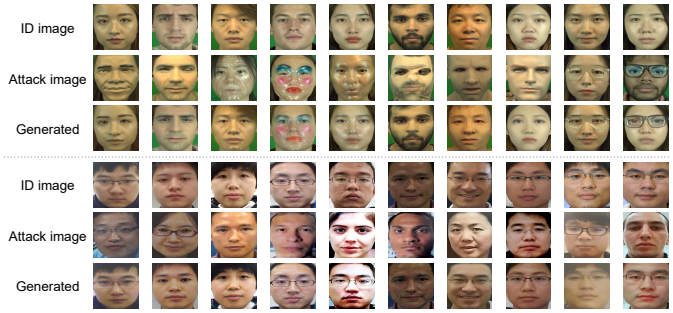


Fig. 4: Generative samples for cross-attack on PADISI (up) and cross-domain on OCIM (down). The first row is the ID image, and the second row is the randomly selected guide image.

distance between b_{mean} and their scores. Next, we use the relative quality score to calculate the margin. The loss formula is computed as:

$$\psi = -s * b_{rq}, \omega = s * (1 + b_{rq}), \tag{16}$$

$$g(\theta) = \cos(\theta - s * b_{rq}) - s * (1 + b_{rq}), \tag{17}$$

$$L_{RQ} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{m*g(\theta)}}{e^{m*g(\theta)} + \sum_{j=1, j\neq y_i}^{c} e^{\cos(\theta_j)}}, \tag{18}$$

where m is a scale coefficient. The analysis of the RQ loss is shown in Fig. 3 (b), our objective is to mine information from HQ samples. Consequently, as the training advances, the corresponding additive margin for HQ images increases, while for LQ images, it decreases. Taking inspiration from SSDG [13], where clusters formed by live and spoof classes exhibit distinct compactness, we introduce different adjustment hyper-parameters denoted as s. This ensures that the cluster formed by live images becomes more compact.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate our framework on six benchmarks, including the widely used domain generalization protocol OULU-NPU (O) [2], CASIA-MFSD (C) [45], Idiap Replay Attack (I) [3], MSU-MFSD (M) [39], containing various styles of print and replay attack; WMCA [7] and PADISI [29], which contains multiple types of novel attacks, such as flexible masks or glasses.

Metrics. The model's performance is evaluated using the following standard metrics: Half Total Error Rate (HTER), Area Under Curve (AUC), and Average Classification Error Rate (ACER).

4.2 Implementation

Generative Experiment. The goal of OCIM is cross-domain generation, specifically involving the generating of spoof samples with the ID (image) from one domain and the spoofing style from another domain. This aims to assess the generative capability of the model. In the case of WMCA and PADISI, numerous live faces lack corresponding spoof faces with specific attack types. Consequently, our objective is to facilitate cross-attack type generation, ensuring the acquisition of all types of attacks for each live face.

Classification Experiment. This experiment is to evaluates the impact of proposed methods on classification networks. Here, the input images are cropped with RetinaFace [4] and resized to 256×256. Based on our experience, MobileViT [26] is used as the backbone. Moreover, for the OCIM experiment, we conduct additional trials using ResNet-18 [9] to ensure a fair comparison with previous methods. For experiments with less training data, such as WMCA series and ICM→O, we adopt Adam optimizer with learning rate 1e-4, while others with SGD optimizer with learning rate 2e-3. In cross domains or attacks experiments, we only utilize data from visible domains or attacks for generation, ensuring that data leakage does not occur.

Table 3: Comparison on limited source domains.

Method	MI→C		MI→O	
	HTER ↓	AUC ↑	HTER ↓	AUC ↑
SSDG-R [13]	19.86	86.46	27.92	78.72
SSAN-R [38]	25.56	83.89	24.44	82.56
DiVT-M [20]	20.11	86.71	23.61	85.73
DiffFAS-V	**15.06**	**92.83**	**16.19**	**92.62**

Table 4: Comparison on WMCA, PADISI datasets.

Method	WMCA→PADISI		PADISI→WMCA	
	HTER ↓	AUC ↑	HTER ↓	AUC ↑
MCDeepPixBiS [6]	12.83	95.43	17.18	90.96
EPCR [37]	10.06	97.30	16.26	92.49
DiffFAS-V w/o generation	8.67	97.06	15.15	92.56
DiffFAS-V w/o loss	8.13	97.59	14.96	93.19
DiffFAS-V	**7.09**	**97.92**	**12.67**	**94.04**

Table 5: Comparisons with previous loss function.

Method	OCI→M	
	HTER ↓	AUC ↑
Resnet18	7.71	98.05
Resnet18 + Focal loss [21]	11.48	94.28
Resnet18 + Asymmetric softmax loss [34]	6.81	97.64
Resnet18 + Single-side triplet loss [13]	6.57	97.90
Resnet18 + RQ loss	5.90	98.10

Table 6: Ablation of RQ loss live/spoof scaling factors.

Scaling parameter	OCI→M	
	HTER ↓	AUC ↑
0.4/0.1	6.57	98.31
0.4/0.2	3.10	99.63
0.4/0.3	4.23	99.09
0.4/0.4	4.86	98.51

4.3 Results of Generative Experiment

The generative samples of cross-attack and cross-domain settings are presented in Fig. 4. For PADISI's various types of mask attacks, our method can learn specific texture features on masks with high homogeneity in the training set, and replace skin texture with attack texture while maintaining ID consistency. As such, it can enrich the diversity of masks in the dataset. For PADISI face-edit attacks, our method can perform partial replacement of the image edited part, without modifying the sampling algorithm. Note that in PADISI, our model demonstrates robust texture learning capabilities despite the constraint of having a limited number of image pairs (20-30) for training. In the case of OCIM's

domain shift, we achieve cross-domain style transfer by leveraging universal texture features, thereby underscoring the versatility and generality inherent in our texture-based approach.

4.4 Results of Classification Experiment

Cross-Domain Performance. Table 1 summarizes our OCIM cross-domain performance compared with an extensive collection of recent studies. DiffFAS achieves state-of-the-art performance on all protocols. Because of the scarcity of HQ images in ICM→O and considering that ResNet has a comparatively weaker information extraction capability compared to MobileViT, DiffFAS-R shows diminished effectiveness when compared to other protocols. The results of cross-domain testings on WMCA and PADISI are presented in Table 4, which contains several types of novel attacks. HTER% values reports on testing sets are obtained with the fixed threshold of 0.5 on testing sets. Our method shows greater benefits under testing protocols with limited data.
Cross-attack Performance. The results of cross-attack experiments on WMCA datasets are presented in Table 2, comprising the grandtest protocol and seven unseen tests for seven attack types. Leveraging our remarkable mask generation capability, DiffFAS attains a state-of-the-art performance in both seen and average unseen attack types.
Limited Source Domains Performance. Furthermore, we evaluate the performance of our method in cases of limited availability of source domains. Specifically, M and I are selected as the source domains for training while the remaining two are used as the target domains for testing, respectively. As shown in Table 3, our method achieves the lowest HTER and the highest AUC despite limited source data, which demonstrates the modeling efficiency and generalization capability of our model.

Table 7: Ablation of Framework Components.

Method	MI→C		MI→O	
	HTER ↓	AUC ↑	HTER ↓	AUC ↑
MobileVit	24.57	83.70	24.16	85.37
DiffFAS-V w/o loss	16.91	91.82	18.02	89.98
DiffFAS-V w/o generation	15.50	92.53	16.99	91.19
DiffFAS-V	15.06	92.83	16.19	92.62

Table 8: Comparisons with previous generative methods.

Method	OCI→M	
	HTER ↓	AUC ↑
Resnet18	9.67	95.96
Resnet18 + DSDG [23]	8.57	96.55
Resnet18 + DiffFAS	7.71	98.05

5 Ablation Study

5.1 Generation Method Ablation

Ablation of STFM. We establish a basic style transfer module, employing means and variance substitution of the backbone and condition as the baseline, comparing with baseline incorporating patch-wise mean and variance, STFM without patch, and proposed STFM. Generative samples are presented in Fig. 5.

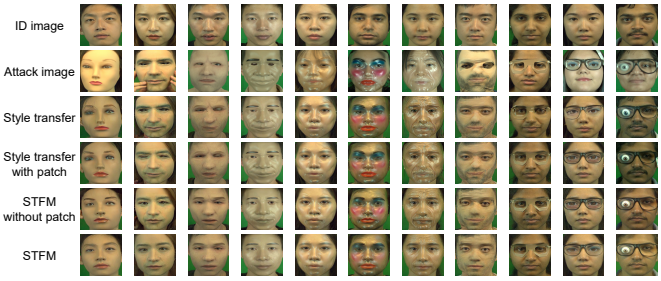


Fig. 5: Visualization with various baselines. It is evident that DiffFAS achieves the balance between identity and spoof texture.



Fig. 6: Ablation of the impact of time steps: with $\gamma = 2.0$, $t = 100$, 300, 500, 700, 900 from left to right.



Fig. 7: Ablation of γ impact: with $t = 100$, $\gamma = 1.2$, 1.5, 2.0, 2.5, 3.0 from the left to the right.

It is evident that, even with the editing algorithm in the baseline method, preserving the original identity information poses a challenge, resulting in distortion for specific attack types. The inclusion of patch information significantly improves the mitigation of distortion. Furthermore, it is observed that employing a larger mean patch aids in minimizing the impact on identity information while introducing attack texture to live faces.

Ablation of Inference Hyper-parameter. The strength of the spoof feature can be adjusted by manipulating the forward process timesteps t and CFG parameters γ. However, adding excessively strong spoof traces may potentially compromise the identity information. Ablation experiments on these two parameters are illustrated in Fig. 6 and 7. It is evident that as t increases or γ increases, the spoof feature become more significant. This capability allows us to create samples with varying levels of spoof strength, thereby enhancing the robustness of the classification network.

Comparison with existing works in preventing overfitting. In this section, we conduct comparative experiment to observe DiffFAS's capability against overfitting, which is a common challenge encountered by previous generative methods. Earlier methods suffer from overfitting due to data scarcity, leading to generative spoof images bearing a strong resemblance to the samples in the original dataset. The reduction in generative diversity diminishes the utility of synthetic datasets for classification tasks, thereby undermining the intended objective of generation. To validate our method's capability against overfitting, we conduct comparative experiments with DSDG [23]. We convert the live samples generated by DSDG into spoof samples with the same attack type, enabling us to observe the generative differences between above two methods, and results are show in Fig. 8. It is apparent that DSDG faces challenges in learning the joint distribution between live and spoof samples, which can be attributed to the extremely small scale of data in PADISI.

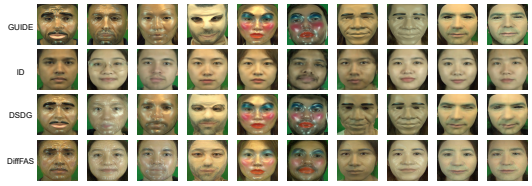


Fig. 8: Comparison with DSDG on PADISI: The ID image is sourced from DSDG's generated live samples to ensure a fair competition.

5.2 Classification Method Ablation

Effectiveness of Framework Components. The Table 4 and 7 presents both generative samples and Relative Quality loss gain for classification experiment. In different scenarios, the two major factors of domain shift: quality and style, play distinct yet complementary roles. Existing experiments confirm that these components effectively counteract the dual characteristics of domain shift, successfully mitigating quality and style shifts, thereby enhancing generalization performance. Regarding the RQ loss, we utilize different scaling factors for live and spoof samples, and the ablation study on these factors is provided in Table 6.

Table 9: Comparisons with previous generative methods.

Method	OCI→M HTER ↓	OCI→M AUC ↑
Resnet18	9.67	95.96
Resnet18 + DSDG [23]	8.57	96.55
Resnet18 + DiffFAS	7.71	98.05

Comparison with Previous Works. To facilitate a fair comparison of the impact of our method with previous generative approaches on cross-domain FAS tasks, we establish a baseline without generative data. This baseline involves binary classification with CrossEntropy loss. Subsequently, we introduce DSDG and DiffFAS data into the experiment, and the results are presented in Table 9. Additionally, to validate the effectiveness of the proposed loss, we replace various previously proposed SOTA losses with DiffFAS-R, including Focal loss [21], Asymmetric softmax loss from PatchNet [34], and the Single-side triplet loss from SSDG [13], with CrossEntropy loss as the baseline. The results are shown in Table 5. Notably, with our generative samples, SSDG and PatchNet demonstrate improved performance, highlighting the effectiveness of our generative method.

6 Conclusion

This paper proposes a novel approach to address domain shifts by separating them into measurable quality and style components. We introduce the Relative Quality loss that integrates quality scores into the training loss function, adding a quality prior to the classification model. Further, we present DiffFAS, a versatile generative model for high-fidelity cross-domain and cross-attack generation, addressing the lack of labeled data for novel attack types. Our model currently faces a constraint in open-world face reconstruction, requiring corresponding ID in the training set to generate its spoof version. Enhancing this capability is a key focus for future research. Investigating methods to broaden the model's capacity for realistic spoof generation, independent of specific training pairs, holds promise for further development.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62171309 and 62306061, Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-02).

References

1. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based cnns. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 319–328. IEEE (2017)
2. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). pp. 612–618. IEEE (2017)
3. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). pp. 1–7. IEEE (2012)
4. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
6. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2019)
7. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE transactions on information forensics and security **15**, 42–55 (2019)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
13. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8484–8493 (2020)

14. Jing, Y., Liu, X., Ding, Y., Wang, X., Ding, E., Song, M., Wen, S.: Dynamic instance normalization for arbitrary style transfer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 4369–4376 (2020)
15. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: Proceedings of the European conference on computer vision (ECCV). pp. 290–306 (2018)
16. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18750–18759 (2022)
17. Kim, M., Liu, F., Jain, A., Liu, X.: Dcface: Synthetic face generation with dual condition diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12715–12725 (2023)
18. Kim, T., Kim, Y., Kim, I., Kim, D.: Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Liao, C.H., Chen, W.C., Liu, H.T., Yeh, Y.R., Hu, M.C., Chen, C.S.: Domain invariant vision transformer learning for face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6098–6107 (2023)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
22. Liu, A., Zhao, C., Yu, Z., Wan, J., Su, A., Liu, X., Tan, Z., Escalera, S., Xing, J., Liang, Y., 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)
23. Liu, S., Zhang, K.Y., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Xie, Y., Ma, L.: Dual reweighting domain generalization for face presentation attack detection. arXiv preprint arXiv:2106.16128 (2021)
24. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4680–4689 (2019)
25. Liu, Y., Stehouwer, J., Liu, X.: On disentangling spoof trace for generic face anti-spoofing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 406–422. Springer (2020)
26. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
27. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
28. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
29. Rostami, M., Spinoulas, L., Hussein, M., Mathai, J., Abd-Almageed, W.: Detection and continual learning of novel face presentation attacks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14851–14860 (2021)
30. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 10023–10031 (2019)
31. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)
32. Sun, Y., Liu, Y., Liu, X., Li, Y., Chu, W.S.: Rethinking domain generalization for face anti-spoofing: Separability and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24563–24574 (2023)
33. Valevski, D., Kalman, M., Matias, Y., Leviathan, Y.: Unitune: Text-driven image editing by fine tuning an image generation model on a single image. arXiv preprint arXiv:2210.09477 (2022)
34. Wang, C.Y., Lu, Y.D., Yang, S.T., Lai, S.H.: 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. pp. 20281–20290 (2022)
35. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
36. Wang, W., Liu, P., Zheng, H., Ying, R., Wen, F.: Domain generalization for face anti-spoofing via negative data augmentation. IEEE Transactions on Information Forensics and Security (2023)
37. Wang, Z., Yu, Z., Wang, X., Qin, Y., Li, J., Zhao, C., Liu, X., Lei, Z.: Consistency regularization for deep face anti-spoofing. IEEE Transactions on Information Forensics and Security **18**, 1127–1140 (2023)
38. Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4123–4133 (2022)
39. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
40. Wu, H., Zeng, D., Hu, Y., Shi, H., Mei, T.: Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4626–4638 (2021)
41. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
42. Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5609–5631 (2022)
43. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: 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)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
45. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB). pp. 26–31. IEEE (2012)
46. Zou, Z., Wang, Z., Zhang, B., Xu, Y., Liu, Y., Wu, L., Guo, Z., He, Z.: Adversarial domain generalization for surveillance face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6351–6359 (2023)