Adversarial Domain Generalization for Surveillance Face Anti-Spoofing

Yongluo Liu[1,2]*, Yaowen Xu[2]*, Zhaofan Zou[2], Zhuming Wang[1,2], Bowen Zhang[1,2],
Lifang Wu[1]†, Zhizhi Guo[2]†, Zhixiang He[2]
[1]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[2]China Telecom Corporation Ltd. Data&AI Technology Company, Beijing, China
{liuyongluo, jczz1030, zhangbowen}@emails.bjut.edu.cn, lfwu@bjut.edu.cn
{xuyw1, zouzhf41, guozz2, hezx3}@chinatelecom.cn

Abstract

In traditional scenes (short-distance applications), the current Face Anti-Spoofing (FAS) methods have achieved satisfactory performance. However, in surveillance scenes (long-distance applications), those methods cannot be generalized well due to the deviation in image quality. Some methods attempt to recover lost details from low-quality images through image reconstruction, but unknown image degradation results in suboptimal performance. In this paper, we regard image quality degradation as a domain generalization problem. Specifically, we propose an end-to-end Adversarial Domain Generalization Network (ADGN) to improve the generalization of FAS. We first divide the accessible training data into multiple sub-source domains based on image quality scores. Then, a feature extractor and a domain discriminator are trained to make the extracted features from different sub-source domains undistinguishable (i.e., quality-invariant features), thus forming an adversarial learning procedure. At the same time, we have introduced the transfer learning strategy to address the problem of insufficient training data. Our method won second place in "Track Surveillance Face Anti-spoofing" of the 4th Face Anti-spoofing Challenge@CVPR2023. Our final submission obtains 9.21% APCER, 1.90% BPCER, and 5.56% ACER, respectively.

1. Introduction

Face Anti-Spoofing (FAS) technology aims to prevent face recognition systems from being vulnerable to presentation attacks, such as print attack, video attack, and 3D mask attack [16, 24, 27–29]. Due to the importance of FAS for security, both academia and industry have conducted a large amount of research and made much progress [20, 21, 23, 25, 39, 40, 47]. Compared with FAS in traditional scenes (*e.g.*, phone unlocking, face payment, and access authentication) [26, 29, 46, 48], FAS in long-distance scenes (*i.e.*, surveillance) such as station squares, parks, and self-service supermarkets are equally important [2, 5, 9]. Although recent FAS methods in traditional scenarios have achieved satisfactory performance [23, 49, 50], FAS in surveillance scenes has not yet been fully explored.

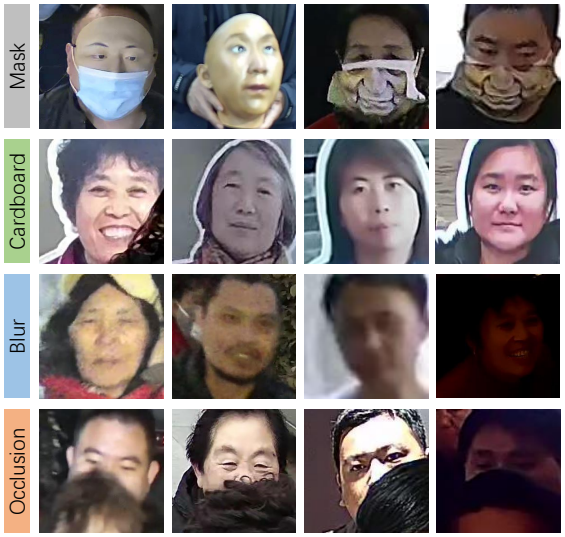


Figure 1. An overview of some characteristics of SuHiFiMask. From top to bottom: mask attacks, cardboard attacks, blur, and occlusion.

One major constraint on the performance of surveillance FAS is image quality. Under surveillance scenes, the resolution of faces is small and contains noise from motion blur, occlusion, and other bad factors, resulting in previous methods can not effectively generalize to faces with varying quality. To address the issue caused by image quality, methods [5, 9] attempt to recover high-resolution faces from low-

*These authors contributed equally to this work. Yongluo Liu, Zhuming Wang, and Bowen Zhang are interns with China Telecom Corporation Ltd. Data&AI Technology Company while doing this work.
†Corresponding authors.

resolution faces to extract informative spoofing cues. Despite some progress, unknown image degradation processes make it difficult to restore bona fide facial details. In contrast, Aravena *et al.* [2] advise discarding some low-quality samples for improving the performance. However, directly ignoring low-quality faces can not effectively address the challenge of FAS in surveillance scenes. Therefore, how to reduce the impact of image quality to further improve the generalization of FAS in surveillance scenes is still a challenging problem that remains unsolved.

To address the above problem, the widely concerned Domain Generalization (DG) methods provide us with some inspiration [19, 32, 38, 42]. Some DG-based FAS methods [17, 35, 44] minimize the difference among source domains by domain-adversarial learning for extracting domain-invariant representations. This inspires us to use multiple source domains with different qualities to learn quality-invariant features by simulating the above process. Based on the idea of domain generalization, we have to focus on the number of available surveillance FAS datasets. Recently, a large-scale surveillance FAS dataset, SuHiFiMask, is established, which includes 101 participants of different ages, 232 masks, and 200 2D attacks [9]. Some typical examples are shown in Fig. 1. The most challenging protocol 3 (image quality degradation) is utilized for the 4th Face Anti-spoofing Challenge@CVPR2023. Because we only have access to the current dataset, limited training data restrict the probability of training deep architectures from scratch. We train our model by Transfer Learning (TL) strategy [12, 33] from a pre-trained network to deal with the limited data problem.

Motivated by the discussions above, we introduce DG and TL technologies into FAS tasks in surveillance scenes simultaneously, and propose an Adversarial Domain Generalization Network (ADGN) to address the problems caused by image quality. Specifically, we first divide the training data based on quality scores to generate several sub-source domains with significant quality differences. Then, a pre-trained feature extractor is trained to compete with a domain discriminator to make the features of faces from different domains undistinguishable. Finally, we deployed a simple classifier to predict whether a face is a bona fide or malicious attack. Our contributions include:

- We treat quality degradation in FAS under surveillance scenes as a domain generalization problem, and the proposed ADGN can extract the quality-invariant features through domain-adversarial learning effectively.
- Extensive experiments are conducted on protocol 3 of SuHiFiMask to demonstrate the effectiveness of the proposed method. In addition, our method won second place in "Track Surveillance Face Anti-spoofing" of the 4th Face Anti-spoofing Challenge@CVPR2023.

2. Related Work

In this section, we first introduce some recent progress in surveillance FAS, and then demonstrate recent works on DG-based and TL-based FAS.

2.1. FAS in Surveillance Scenes

To extend FAS from traditional scenes to surveillance scenes, Chen *et al.* [5] propose a cross-device domain FAS dataset called GREAT-FASD-S. The dataset is collected by two multi-modal cameras and processed into low-quality images. Simultaneously, they propose a depth-wise separable attention module with SE-block [14] to select the most informative channels and recover the image by the nearest neighbor algorithm. Aravena *et al.* [2] reveal the impact of image quality on FAS and advise discarding some low-quality samples to improve overall performance. Recently, Fang *et al.* [9] release the first dataset collected based on real surveillance scenes called SuHiFiMask. Besides, they propose an IQV module to recover image information by combining a super-resolution network [7]. Specifically, they down-sample high-resolution images into low-resolution images as the training data to train the super-resolution network. Although these methods have made some progress, it is difficult to recover bona fide details from complex image quality degradation in surveillance scenes, and discarding low-quality samples cannot fundamentally address the challenges in surveillance scenes.

2.2. DG-Based FAS Methods

Domain generalization aims to achieve Out-Of-Distribution (OOD) generalization by using only source data for model learning [51]. In recent years, DG-based FAS methods have attracted widespread attention [4, 6, 17, 35, 41, 44]. DG-based FAS methods mainly include three categories: 1) domain-adversarial based, 2) domain-disentanglement based, and 3) domain-augmentation based. The first category of methods [17, 35, 44] usually combines Gradient Reversal Layer (GRL) [11] for domain-adversarial learning to extract domain invariant features by minimizing the difference among source domains. Methods [41, 45] of the second category disentangle domain information from feature representations through specific networks. The last category of methods [15, 18] expands the sample space by generating mixed or pseudo domains, thereby learning more general feature representations. The DG-based FAS method can effectively reduce the performance degradation caused by domain shift, which also inspires us to extend the domain generalization to surveillance scenes.

2.3. TL-Based FAS Methods

Transfer learning can effectively improve the performance of models in target domains by transferring knowl-

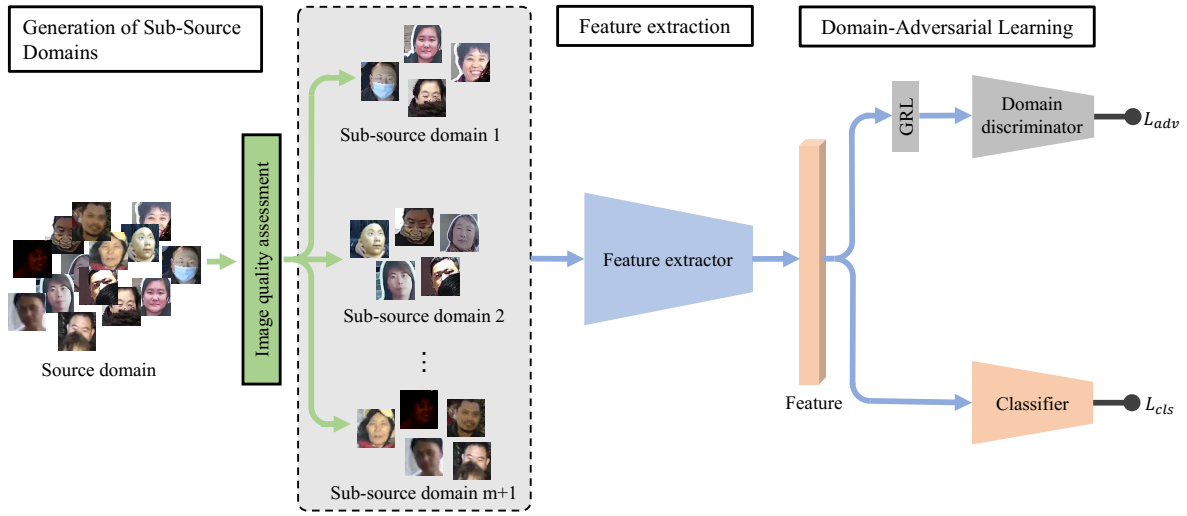


Figure 2. The overall network of ADGN. We first divide the accessible data into $m+1$ sub-source domains based on image quality. Then, a feature extractor deployed by ViT is used to extract the features of the input face. Subsequently, the domain discriminator distinguishes which sub-source domain the extracted features come from, while the feature extractor seeks to enable the extracted features to spoof the domain discriminator and form adversarial learning. Finally, a classifier is used to predict the category of input face.

edge contained in different but related source domains [52]. George *et al.* [12] first introduce the large-scale pre-trained Vision Transformer (ViT) models into FAS tasks, effectively improving the generalization of the model through fine-tuning. Liu *et al.* [22] present a noval framework, namely Modality-Agnostic Vision Transformer (MA-ViT), which effectively improves the performance of arbitrary modal attacks with the assistance of multi-modal data. Quan *et al.* [34] present a semi-supervised learning-based framework, progressively adopting the unlabeled data with reliable pseudo labels during training to enrich the variety of training data. Methods [1, 31] reduce the limitations of FAS datasets by transfer learning strategy in traditional scenes.

In order to overcome the challenges in surveillance scenes, we extend the transfer learning strategies to learn generalized quality-invariant feature representations.

3. Methodology

In this section, we first introduce the overall network of the proposed ADGN and the generation of multiple sub-source domains in Sec. 3.1 and Sec. 3.2, respectively. Then we demonstrate the adversarial domain generalization procedure in Sec. 3.3. At last the supervision signals and loss functions are presented in Sec. 3.4.

3.1. Overall Network

As shown in Fig. 2, the proposed ADGN mainly includes a feature extractor, a classifier, and a domain discriminator. For the face image $\mathbf{x}$ from an arbitrary domain, we first use the feature extractor deployed by the ViT encoder [8] to extract its feature $\mathbf{z}_{cls}$. Then, the domain discriminator is used to distinguish which domain $\mathbf{z}_{cls}$ comes from. On the other hand, we optimize the feature extractor to enforce $\mathbf{z}_{cls}$ that cannot be distinguished by the domain discriminator, so that the feature extractor and the domain discriminator form an adversarial learning process. Finally, we deploy a simple classifier to predict whether $\mathbf{x}$ is a real face. Note that in order to implement end-to-end training, we insert a GRL [11] after the feature extractor.

3.2. Generation of Sub-Source Domains

To simulate the process of domain generalization, we first subdivide the training data into multiple subsets based on quality scores, with each subset representing a sub-source domain. Specifically, we denote the accessible training data as $\mathbb{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, and the quality score $\mathbf{q}_i$ for each sample $\mathbf{x}_i$ is recorded as $\mathbb{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\}$, n is the number of samples. Then, we select a series of thresholds in $\mathbb{Q}$ and record them as $\{T_1, T_2, ..., T_m\}$. After that, the source domain $\mathbb{D}$ can be divided into $m+1$ sub-source domains based on the quality threshold, represented as follows:

$$\begin{aligned} \mathbb{D}_1 &= \{\mathbf{x}_i \in \mathbb{D} \mid 0 \leq \mathbf{q}_i < T_1\}, \\ \mathbb{D}_2 &= \{\mathbf{x}_i \in \mathbb{D} \mid T_1 \leq \mathbf{q}_i < T_2\}, \\ &..., \\ \mathbb{D}_{m+1} &= \{\mathbf{x}_i \in \mathbb{D} \mid T_m \leq \mathbf{q}_i \leq \max(\mathbb{Q})\}, \end{aligned} \tag{1}$$

where $\mathbb{D}_j$ represents the ith sub-source domain, $j \in [1, m+1]$. $\max(\mathbb{Q})$ is the highest quality score in $\mathbb{Q}$.

Through the operation described in Eq. (1), we obtain $m+1$ sub-source domains $\{\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_{m+1}\}$. For each sub-source domain $\mathbb{D}_j$, we assign a sub-source domain label $\mathbf{s}_j$ to it, and we denote all sub-source domain labels as $\mathbb{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{m+1}\}$.

3.3. Domain-Adversarial Learning

Transformers for FAS. We utilize pre-trained ViT [8] encoders to deploy feature extractors. This has two benefits: 1) transfer learning from a pre-trained model is an effective strategy to deal with limited training data; and 2) the diverse knowledge learned by a pre-trained model helps address the domain shift [12].

Specifically, for the image $\mathbf{x}$ from a sub-source domain $\mathbb{D}_j^{H\times W\times C}$, we divide $\mathbf{x}$ into $N\times N$ non-overlapped patches, and embed those patches into 1D embeddings $\mathbf{z}_i \in \mathbb{R}^{1\times C}$ after linear projections, $1 \leq i \leq N^2$. And we integrate all $\mathbf{z}_i$ as $\mathbf{z}_p \in \mathbb{R}^{N^2\times C}$, $\mathbf{z}_p = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{N^2}]$. After that, we randomly initialize classification embedding $\mathbf{z}_{cls} \in \mathbb{R}^{1\times C}$ and position embedding $\mathbf{z}_{pos} \in \mathbb{R}^{1\times C}$. Then, following [8, 43], the two embeddings are bundled with $\mathbf{z}_p$ as learnable embeddings to capture spoofing information and retain positional information, and denoted as $\mathbf{z} \in \mathbb{R}^{(N^2+1)\times C}$. $\mathbf{z}$ can be expressed as follows:

$$\mathbf{z} = \text{concat}(\mathbf{z}_{cls}, \mathbf{z}_p) + \mathbf{z}_{pos}. \tag{2}$$

The relations between the local patch and classification embeddings are captured progressively along with the cascaded transformer encoder layers. For example, the calculation process of adjacent layers can be expressed as follows:

$$\begin{aligned} \mathbf{h}^l &= \text{MSA}(\text{LN}(\mathbf{z}^l)) + \mathbf{z}^l, \\ \mathbf{z}^{l+1} &= \text{MLP}(\text{LN}(\mathbf{h}^l)) + \mathbf{h}^l, \end{aligned} \tag{3}$$

where MSA, MLP, and LN are the multi-head self-attention, multi-layer perceptron, and layer normalization in transformer encoder layers, respectively. $\mathbf{z}^l$ is the input of the l^{th} transformer encoder layer, $\mathbf{h}^l$ is the temporal variable. $\mathbf{z}^{l+1}$ denotes the output of the l^{th} encoder layer.

Finally, the output classification embedding $\mathbf{z}_{cls}$ of the last encoder layer is used for predicting.

Learn Quality-Invariant Features. For image $\mathbf{x} \in \mathbb{D}^{H\times W\times C}$, we obtain its feature representation $\mathbf{z}_{cls}$ by the feature extractor. Then, the domain discriminator is implemented based on the extracted feature $\mathbf{z}_{cls}$ to determine which sub-source domain the input features stem from. On the contrary, the feature extractor is trained to spoof the domain discriminator so that the sub-source domain labels cannot be recognized. Therefore, our method can learn a generalized feature representation by adversarial learning procedure between the feature extractor and the domain discriminator.

During the adversarial learning procedure, the feature extractor is optimized by maximizing the loss of the domain discriminator while the domain discriminator is optimized with the opposite objective. Since we divide the training data into $m+1$ sub-source domains and assign domain labels, we utilize the Cross-Entropy (CE) loss to optimize the proposed network under adversarial learning:

$$\begin{aligned} \min_{\mathcal{D}} \max_{\mathcal{G}} \mathcal{L}_{adv}(\mathcal{G}, \mathcal{D}) = \\ - \mathbb{E}_{x,s\sim D,S} \sum_{j=1}^{m+1} \mathbb{1}_{[j=\mathbf{s}]} \log \mathcal{D}(\mathcal{G}(x)), \end{aligned} \tag{4}$$

where $\mathcal{G}$ and $\mathcal{D}$ are feature extractor and domain discriminator, respectively. $\mathbb{S}$ represents the set of sub-source domain labels, $\mathbb{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{m+1}\}$.

Following [17], a GRL [11] is deployed to optimize the feature extractor and the domain discriminator simultaneously.

3.4. Loss Function

Since all the sub-source domain data contain class labels, a classifier is implemented after the feature extractor, as illustrated in Fig. 2. The feature extractor is optimized by the Binary Cross Entropy (BCE) loss, denoted as $\mathcal{L}_{cls}$.

$$\mathcal{L}_{cls} = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \log \hat{\mathbf{y}}_i + (1-\mathbf{y}_i) \log(1-\hat{\mathbf{y}}_i), \tag{5}$$

where n is the number of samples in source domain $\mathbb{D}^{H\times W\times C}$. $\mathbf{y}_i$ is the label of sample $\mathbf{x}_i$. $\hat{\mathbf{y}}_i$ is the predicted probability output by the classifier.

Integrating the $\mathcal{L}_{cls}$ and the $\mathcal{L}_{adv}$ mentioned above together, the objective of the proposed method is:

$$\mathcal{L}_{all} = \mathcal{L}_{cls} + \mathcal{L}_{adv}, \tag{6}$$

By optimizing the above overall loss, we train all the components in an end-to-end manner.

4. Experiment

In this section, extensive experiments are performed to demonstrate the effectiveness of our method. In the following, we sequentially describe the experimental settings, comparison with state-of-the-art methods, ablation study, comparison of different backbones, and feature visualizations.

4.1. Experimental Settings

SuHiFiMask dataset. SuHiFiMask [9] is collected from real surveillance scenes. This dataset includes 101 participants of different ages, 232 high-fidelity masks, 200 2D attacks, and 2 adversarial attacks. Besides, it covers 40 real surveillance scenes and different weather. To fully evaluate the performance in surveillance scenes, SuHiFiMask defines three protocols: protocol 1 is used to evaluate the comprehensive performance of the algorithm being migrated

to surveillance scenes; protocol 2 is divide into four sub-protocols by using the "leave-one-type-out testing" to evaluate the generalization of the algorithm for the "unseen" 3D facial mask type; and protocol 3 evaluates the robustness of the algorithm to image quality degradation. In this paper, our experiments are all conducted on the most challenging protocol 3 (quality degradation), which has been utilized for the 4th Face Anti-spoofing Challenge@CVPR2023.

Evaluation Metrics. The evaluation metrics are consistent with FAS tasks in traditional scenes, *i.e.*, Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) [10]. They can be formulated as:

$$\begin{aligned} \text{APCER} &= \frac{\text{FP}}{\text{TN}+\text{FP}}, \\ \text{BPCER} &= \frac{\text{FN}}{\text{FN}+\text{TP}}, \\ \text{ACER} &= \frac{\text{APCER}+\text{BPCER}}{2}, \end{aligned} \tag{7}$$

where $\text{FP}, \text{FN}, \text{TN}$ and FP are the false positive, false negative, true negative, and true positive sample numbers, respectively. ACER is used to determine the final ranking in 4th Face Anti-spoofing Challenge@CVPR2023.

Architecture Details. The proposed ADGN mainly contains a feature extractor, a classifier, and a domain discriminator. The feature extractor is deployed by the pre-trained ViT model (ViT-L/16) [8]. Specifically, ViT-L/16 has 24 transformer encoder layers and the patch size is 16×16, the patch embedding dim is 1024, and the head number of MSA is 16. The classifier consists of a simple Fully Connected layer (FC: 1024, 1). The domain discriminator is sequentially composed of FC (1024, 512), ReLU activation layer, Dropout (0.5), and FC (512, 2).

Training Settings. Our proposed method is implemented with Pytorch. In the training stage, models are trained with SGD optimizer with a momentum of 0.9 and the initial learning rate (lr) is 0.01. The cosine learning rate schedule (CosineLR) [30] is employed to adjust lr. We use warm-up for the first 100 epochs and the minimum lr is the 1% of initial lr. After the first 100 epochs, we fine-tune our model for the later 30 epochs. In the fine-tuning phase, we use different data augmentation methods to expand training data. The data augmentation methods [3] include CoarseDropout, Random Flip, RandomRotate, RandomCrop, MotionBlur, GaussianBlur, Sharpen, Downsampling, RandomlyErase, RandomFog, Posterize, etc. The augmented data are gradually added to the training set to avoid the conflicts caused by excessive augmentation methods one-time. Besides, the gradual augmentation strategy verifies the effectiveness of each augmented method. We train our model with batch size 80 on four A100 GPUs, and the maximum epoch is set to 130.

Method	APCER (%)	BPCER (%)	ACER (%)
ResNet18 [13]	21.04	13.64	17.64
ViT [8]	19.61	13.95	16.78
CDCN [48]	28.70	25.89	27.30
CQIL [9]	19.14	12.87	15.98
Ours	**9.21**	**1.90**	**5.56**

Table 1. The testing results on protocol 3 of SuHiFiMask dataset. Our method is significantly superior to the current best method CQIL [9] on all metrics.

Data Preparation. As described in Sec. 3.2, we divide accessible data into multiple sub-source domains. Please note that before dividing the accessible data, we first randomly select 10000 images from the accessible data as the verification set. Then, we select a quality threshold value, T=0.4, that is, samples with a quality score less than 0.4 are placed in sub-source domain 1, and samples with a quality score greater than 0.4 are placed in sub-source domain 2. During the training process of our ADGN, we resize samples to $224 \times 224 \times 3$, and perform some data enhancement operations, such as random erasing of regions, affine transformation, changing brightness, etc.

4.2. Comparison with State-of-the-art Methods

In this subsection, we conduct all experiments on protocol 3 of SuHiFiMask and present our final performance in "Track Surveillance Face Anti-spoofing" of the 4th Face Anti-spoofing Challenge@CVPR2023. Tab. 1 reports the testing results of the following methods: ResNet18 [13], ViT [8], CDCN [48], CQIL [9] and the proposed ADGN. As shown in Tab. 1, ResNet18, ViT, and CDCN achieve good performance in traditional scenes, but their performance in surveillance scenes is unsatisfactory. Those results indicate that quality degradation in SuHiFiMask poses a challenge to existing FAS methods. CQIL [9] improves the performance (ACER: 15.98%) in surveillance scenes by extracting quality independent features, but it still needs to be further improved. Our method significantly improves various metrics (ΔAPCER: 9.93%, ΔBPCER: 10.97%, ΔACER: 10.42%) compared to CQIL [9], which fully demonstrates the generalization of our method under different quality conditions. Overall, Table 1 fully demonstrates the effectiveness of our method.

4.3. Ablation Study

In this subsection, ablation studies are conducted to demonstrate the importance of each component and the choice of loss functions. To draw general conclusions, we implement a series of testing under different settings. and all ablation studies are conducted on protocol 3 of the SuHiFiMask dataset. We provide detailed performance under

Model	ViT-L/16	$\mathcal{L}_{cls}$/CE	$\mathcal{L}_{cls}$/BCE	$\mathcal{L}_{adv}$/CE	$\mathcal{L}_{adv}$/BCE	APCER (%)	BPCER (%)	ACER (%)
A	✓	✓				9.23	9.25	9.24
B	✓	✓		✓		10.57	2.27	6.42
C	✓		✓		✓	10.42	7.40	8.91
Ours	✓		✓	✓		**9.21**	**1.90**	**5.56**

Table 2. Ablation studies. The feature extractors of all models in the ablation study are deployed with ViT-L/16. We employ BCE loss for classifiers and CE loss for domain discriminator to achieve optimal performance.

metrics APCER (%), BPCER (%), and ACER (%).

***w/* Domain-Adversarial Learning**. As described in Sec. 3.3, we utilize the Domain-Adversarial Learning (DAL) procedure to extract quality-invariant features. To prove that the DAL procedure is effective, we conduct relevant ablation studies as shown in Tab. 2 row1 (model-A) and row2 (model-B). Specifically, compared to model-A, model-B (*w/* DAL) yields BPCER and ACER gains of 6.98%, and 2.82%, respectively. These results indicate that using a domain-adversarial learning procedure can effectively extract quality-invariant features, thereby improving the overall performance of our method in surveillance scenes.

***w/* BCE**. We explore the performance of using different loss functions for the classifier and the domain discriminator, *i.e.*, BCE, and CE loss functions. Tab. 2 reports detailed results. Compare model-B (*w/* CE, row2) and model-ours (w/ BCE, row4), our model improves by 1.36%, 0.37%, and 0.86% on APCER, BPCER, and ACER, respectively. These performance improvements indicate that BCE loss is more suitable for FAS tasks in surveillance scenes.

Similar to the classifier, we analyze the performance of the domain discriminator under different loss functions. Comparing row3 (model-C) and row4 (model-ours *w/* CE) in Tab. 2, we found that when CE loss is used for domain discriminator, the performance achieves the best. Specifically, we observe some significant improvements in performance (BPCER: 7.40%→1.90%, ACER: 8.91%→5.56%) when using the CE loss function for domain discriminator. BCE loss can be used only because we divide the accessible data into two subsets. CE loss can be applied to all data partitioning situations (*i.e.*, the number of subsets is greater than 2) because the domain discriminator is a multi-class classifier essentially. Comprehensively, we determine the optimal combination of loss functions based on the above ablation studies of loss functions.

Impact of *m*. In this paper, we first utilize SER-FIQ [36] to obtain the quality scores of the samples, then we divide the samples into $m+1$ sub-source domains to form a multi-domain adversarial procedure. In Sec. 4.2, we report the performance at $m = 1$. To further explore the effectiveness of extracting quality-invariant features from domain-adversarial learning, we report detailed results here for $m = 2$ and 3, the quality thresholds are $\{0.5, 0.8\}$ and $\{0.5, 0.7, 0.9\}$, respectively. As shown in Tab. 3, the overall performance of our method gradually improves as the number of sub-source domains increases. Comprehensively, these results demonstrate the effectiveness of our method in extracting quality-independent features.

m	APCER (%)	BPCER (%)	ACER (%)
1	9.21	1.90	5.56
2	8.76	1.85	5.36
3	8.55	1.81	5.18

Table 3. The impact of m on our method. As the number of sub-source domains increases, the performance of our method gradually improves.

Backbone	APCER (%)	BPCER (%)	ACER (%)
ViT-S/16 [8]	12.40	2.27	7.34
ViT-B/16 [8]	10.52	2.24	6.38
ViT-L/16 [8]	**9.21**	**1.90**	**5.56**

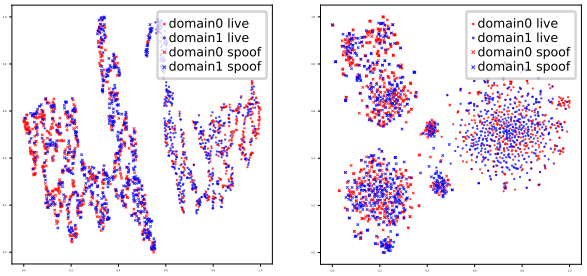
Table 4. Comparison of our method with different backbones. As the backbone scale increases, the performance of our method gradually improves.

4.4. Comparisons of Different Backbones

Here, we show the performance of the feature extractor of our method with three different backbones: ViT-S/16, ViT-B/16, and ViT-L/16 [8]. Tab. 4 reports the detailed results. Our method still achieves good performance when using ViT-S/16 as the feature extractor, which fully demonstrates the effectiveness of our proposed method (ACER: 7.34%). As the scale of feature extractors increases, the performance of our method gradually improves. As shown in Tab. 4, our method uses ViT-B/16 and ViT-L/16 as feature extractors, the ACER is 6.38% and 5.56%, respectively. To fully explore the optimal performance of the proposed method in surveillance scenes, we ultimately chose ViT-L/16 to deploy our feature extractor.

4.5. Visualizations of the Proposed Method

To analyze the feature space learned by our ADGN method, we visualize the distribution of different features using t-SNE [37], as shown in Fig. 3. For the features *wø* ADL in Fig. 3a, it can be observed that their distribution is more compact and mixed, though they may belong to multiple databases or various liveness attributions. It can be observed that the feature distribution of samples in sub-source domain 1 and sub-source domain 2 is greatly affected by image quality, and there is a significant deviation in the feature distribution of the two sub-source domains. On the contrary, Fig. 3b clearly shows that the feature distribution of the two sub-source domains is inseparable. These phenomena demonstrate that our proposed ADGN can effectively extract quality-invariant features. Comprehensively, the extensive comparing experiments, ablation studies, and feature visualization results all fully illustrate our method can effectively deal with the adverse effects of image quality degradation in surveillance scenes.



(a) The feature distribution of our method *w/o* ADL. (b) The feature distribution of our method *w/* ADL.

Figure 3. The t-SNE [37] visualizations of the extracted features.

5. Conclusion

To explore and promote the development of FAS in surveillance scenes, this paper proposes ADGN combining domain generalization and transfer learning strategies. Our ADGN effectively solves the adverse impact of image quality on FAS tasks by extracting quality-invariant features through the domain adversarial learning. In addition, transfer learning strategy effectively address the problem of limited training data. Extensive experiments show that our ADGN is effective and achieves excellent performance on the challenging protocol 3 of SuHiFiMask dataset.

Acknowledgments

This work was funded by the National Natural Science Foundation of China under Grant No.62236010, 61976010, 62106011 and 62176011, Postdoctoral Science Foundation of China under Grant No.2022M720318, and Postdoctoral Research Foundation of Beijing under Grant No.2022-zz-077.

References

[1] Faseela Abdullakutty, Eyad Elyan, Pamela Johnston, and Adamu Ali-Gombe. Deep transfer learning on the aggregated dataset for face presentation attack detection. *Cognitive computation*, 14(6):2223–2233, 2022. 3

[2] Carlos Aravena, Diego Pasmino, Juan E Tapia, and Christoph Busch. Impact of face image quality estimation on presentation attack detection. *arXiv preprint arXiv:2209.15489*, 2022. 1, 2

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 5

[4] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C Kot. Learning meta pattern for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1201–1213, 2022. 2

[5] Xudong Chen, Shugong Xu, Qiaobin Ji, and Shan Cao. A dataset and benchmark towards multi-modal face anti-spoofing under surveillance scenarios. *IEEE Access*, 9:28140–28155, 2021. 1, 2

[6] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1132–1139, 2021. 2

[7] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 605–621. Springer, 2019. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4, 5, 6

[9] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *arXiv preprint arXiv:2301.00975*, 2023. 1, 2, 4, 5

[10] International Organization for Standardization. Iso/iec jtc 1/sc 37 biometrics: Information technology biometric presentation attack detection part 1: Framework, 2016. 5

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 3, 4

[12] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2, 3, 4

[13] 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

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[15] Hanye Huang, Youjun Xiang, Guodong Yang, Lingling Lv, Xianfeng Li, Zichun Weng, and Yuli Fu. Generalized face anti-spoofing via cross-adversarial disentanglement with mixing augmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2939–2943. IEEE, 2022. 2

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

[17] 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, 4

[18] Young Eun Kim and Seong-Whan Lee. Domain generalization with pseudo-domain label for face anti-spoofing. In *Pattern Recognition: 6th Asian Conference, ACPR 2021, Jeju Island, South Korea, November 9–12, 2021, Revised Selected Papers, Part I*, pages 431–442. Springer, 2022. 2

[19] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 2

[20] Xuan Li, Jun Wan, Yi Jin, Ajian Liu, Guodong Guo, and Stan Z Li. 3dpc-net: 3d point cloud network for face anti-spoofing. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2020. 1

[21] 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

[22] 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. 3

[23] 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

[24] 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. 1

[25] 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

[26] 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

[27] 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*, pages 814–823, 2021. 1

[28] 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

[29] Yaojie Liu, Jourabloo Amin, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018. 1

[30] I Loshchilov and F Hutter. Stochastic gradient descent with warm restarts. In *Proceedings of the 5th Int. Conf. Learning Representations*, pages 1–16. 5

[31] Oeslle Lucena, Amadeu Junior, Vitor Moia, Roberto Souza, Eduardo Valle, and Roberto Lotufo. Transfer learning using convolutional neural networks for face anti-spoofing. In *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14*, pages 27–34. Springer, 2017. 3

[32] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 2

[33] Ruijie Quan, Yu Wu, Xin Yu, and Yi Yang. Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955, 2021. 2

[34] Ruijie Quan, Yu Wu, Xin Yu, and Yi Yang. Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955, 2021. 3

[35] 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

[36] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020. 6

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[38] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227, 2021. 2

[39] Jun Wan, Sergio Escalera, Hugo Jair Escalante, Guodong Guo, and Stan Z Li. Special issue on face presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):282–284, 2021. 1

[40] Jun Wan, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Multi-modal face presentation attack detection. *Synthesis Lectures on Computer Vision*, 9(1):1–88, 2020. 1

[41] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020. 2

[42] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021. 2

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

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

[45] Wenjun Yan, Ying Zeng, and Haifeng Hu. Domain adversarial disentanglement network with cross-domain synthesis for generalized face anti-spoofing. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7033–7046, 2022. 2

[46] 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*, 2022. 1

[47] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 1

[48] 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, 5

[49] 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

[50] 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

[51] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[52] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 3