Towards Unsupervised Domain Generalization for Face Anti-Spoofing

Yuchen Liu[1†], Yabo Chen[2†], Mengran Gou[3], Chun-Ting Huang[3], Yaoming Wang[1],
Wenrui Dai[2*], and Hongkai Xiong[1]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University, China
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{liuyuchen6666, chenyabo, wang_yaoming, daiwenrui, xionghongkai}@sjtu.edu.cn
[3]Qualcomm AI Research {mgou, chunting}@qti.qualcomm.com

Abstract

Generalizable face anti-spoofing (FAS) based on domain generalization (DG) has gained growing attention due to its robustness in real-world applications. However, these DG methods rely heavily on labeled source data, which are usually costly and hard to access. Comparably, unlabeled face data are far more accessible in various scenarios. In this paper, we propose the first Unsupervised Domain Generalization framework for Face Anti-Spoofing, namely UDG-FAS, which could exploit large amounts of easily accessible unlabeled data to learn generalizable features for enhancing the low-data regime of FAS. Yet without supervision signals, learning intrinsic live/spoof features from complicated facial information is challenging, which is even tougher in cross-domain scenarios due to domain shift. Existing unsupervised learning methods tend to learn identity-biased and domain-biased features as shortcuts, and fail to specify spoof cues. To this end, we propose a novel Split-Rotation-Merge module to build identity-agnostic local representations for mining intrinsic spoof cues and search the nearest neighbors in the same domain as positives for mitigating the identity bias. Moreover, we propose to search cross-domain neighbors with domain-specific normalization and merged local features to learn a domain-invariant feature space. To our best knowledge, this is the first attempt to learn generalized FAS features in a fully unsupervised way. Extensive experiments show that UDG-FAS significantly outperforms state-of-the-art methods on six diverse practical protocols.

1. Introduction

Face recognition (FR) systems [13, 39] have been widely deployed in real-world applications for person authentication, such as access control and electronic payments. However, FR systems are vulnerable to presentation attacks, e.g., print attack, video replay and 3D mask. To address this issue, face anti-spoofing (FAS) methods have been proposed, including hand-crafted descriptors [12, 22, 33] and deep learning based methods [3, 25, 28, 49]. Despite promising results in intra-dataset scenarios, these methods are dramatically degraded in cross-dataset tests due to the domain gap across datasets. To facilitate generalization on unseen target domains, domain generalization methods [9, 26, 37, 43, 55] have been introduced in FAS to alleviate domain shifts.

Despite the improved generalizability, existing DG methods rely heavily on supervised training using labeled source data. However, labeled data are laborious and costly to obtain, leading to the notorious problem of limited data in FAS. On the contrary, large amounts of unlabeled face data can be easily collected in various environments (e.g., from

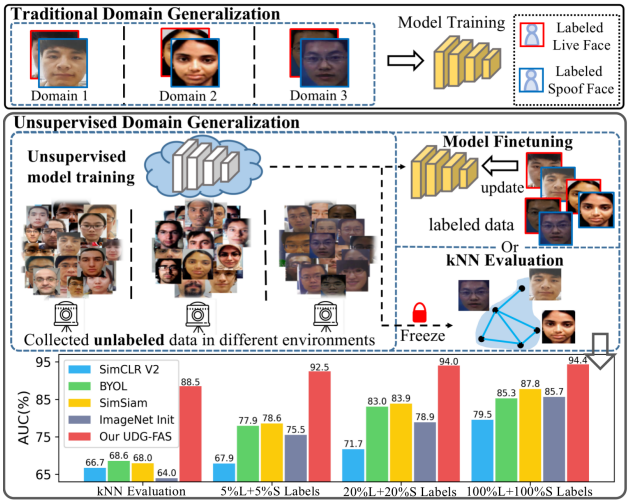


Figure 1. Different from traditional DG methods that rely on labeled data for supervised training, our *unsupervised domain generalization* can learn live/spoof features with more accessible unlabeled data in various environments. Then, we can directly deploy the unsupervised model via kNN or further finetune it with few labeled data. Our method significantly improves performance with various few labeled live (L) and spoof (S) data for I&C&M to O.

*Corresponding author: Wenrui Dai. †Equal contribution. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

(a) different classes (b) different domains (c) different identities

Figure 2. T-SNE visualization of unsupervised features learned by SimSiam [8] on Replay, CASIA, OULU. Different colors for (a) different classes(live/spoof), (b) domains and (c) identities, respectively. We randomly select 4 identities in OULU for visualization.

access control systems). Thus, we study a more practical *unsupervised domain generalization (UDG)* [52] problem that aims to exploit more accessible unlabeled data to learn discriminative representations that generalize well across domains, thus reducing the reliance on labeled data and improving the low data regime of FAS. Fig. 1 illustrates a practical application of *UDG* for FAS. A model is first pretrained on (large-scale) unlabeled data collected from various domains, and then released for deployment. In the deployment phase, we can directly employ the frozen unsupervised model via kNN or further finetune it with few labeled data. Fig. 1 shows that our method achieves superior performance, especially for more challenging real-world applications with no labeled data or few labeled data.

Besides, the prepositive unsupervised learning can be a way of pretraining. Existing DG methods neglect model initialization. It is a common practice to pretrain on ImageNet due to limited training data. However, ImageNet pretraining is not reasonable in FAS, since facial images differ significantly from natural images in the sense of data distribution, e.g., texture and context. While unsupervised pretraining on heterogeneous unlabeled FAS data from various domains is a reasonable initialization for DG. But directly applying off-of-shelf unsupervised learning methods may not achieve desirable performance on the FAS task, as shown in Fig. 1.

Recent advances in unsupervised learning focus on contrastive learning, which enforces invariances to various augmentations. To avoid trivial solutions, contrastive methods, MoCo [18] and SimCLR [6], introduce negative samples for pushing away. However, they are inherently not applicable for FAS. Due to the limited number of classes (only two of live/spoof), there are many false negative samples from the same class, which leads to a lot of noise and impairs the training process. BYOL [15] and SimSiam [8] employ the asymmetric network to eliminate the need for negative samples and succeed on natural images. However, they fail in FAS data, since facial representations contain additional complicated irrelevant factors, i.e., identity-related features. This problem exacerbates in cross-domain scenarios, where the model may learn domain-related features as shortcuts. As depicted in Fig. 2, existing methods learn identity-biased and domain-biased features rather than inherent live/spoof features. Without the actual FAS labels, it is hardly possible to regulate the model to learn live/spoof-related features under the inherent disturbances of identity and domain bias.

To solve these issues, we propose the first Unsupervised Domain Generalization framework for FAS, namely UDG-FAS. To alleviate the identity bias, we propose a novel Split-Rotation-Merge module to generate identity-agnostic local representations for mining intrinsic local spoof pattern. An input image is first split into patches and then randomly rotated. Subsequently, we merge several local embeddings encoded by different facial patches by averaging to mitigate identity-biased information while retaining the task-related one, since different facial regions usually share the same spoof cues. Besides, we propose to search the nearest neighbors in the same domain as positive samples for contrastive learning. By pulling close the images of various persons potentially from the same class, we can further mitigate identity bias across identities and learn intrinsic spoof features.

Regarding domain-related bias, we propose to search the cross-domain nearest neighbors as positive samples. By enforcing similarity on cross domain samples that may belong to the same class, we can learn domain-invariant features that generalize well on the target domain. Due to the distribution shifts, directly searching NN across domains may lead to many false matches. Thus, we normalize the features of each domain to the same reference Gaussian distribution and combine the merged local features for a more accurate search. The contributions of this paper are summarized as:

- We propose the first unsupervised domain generalization framework for face anti-spoofing, which could use more accessible unlabeled data to learn generalizable features for improving the low-data status of the FAS community.
- We design a novel Split-Rotation-Merge module to generate identity-agnostic local representations for mining intrinsic spoof features, and propose to search the nearest neighbors in the same domain as positive samples for contrastive learning to mitigate identity-related bias.
- We propose to search the cross-domain neighbors as positive samples to learn a generalized domain-invariant feature space. Domain-specific normalization with merged local features are leveraged to find more accurate neighbors for boosting performance.

To our best knowledge, we are the first attempt to mitigate identity and domain bias, and learn generalized task-aware features in a fully unsupervised manner for FAS. We build six diverse UDG FAS benchmarks for evaluation. Extensive experiments show our method achieves state-of-the-art performance on various challenging cross-domain intra-type and cross-domain cross-type protocols.

2. Related Work

Face Anti-Spoofing. Existing FAS methods focus on supervised learning, which assumes an adequate amount of

labeled data for training. Traditional FAS methods extract the frame-level features using handcrafted descriptors such as LBP [12], HOG [22] and SIFT [33]. Deep learning methods [46, 47, 50] boost the discrimination ability by employing CNNs to extract features. Auxiliary pixel-wise supervision, e.g., depth maps [27], reflection maps [48] and binary masks [28] are utilized to further explore intrinsic features.

To generalize well on unseen scenarios, domain adaptation (DA) and domain generation (DG) methods have been developed. [21] proposes a single-side adversarial learning way. [43] proposes to operate on content and style features separately. Besides, meta-learning [9, 34, 37] are proposed to regular the optimization process. Despite promising results, existing DG methods are restricted to costly labeled source data, hindering practical applications. Recently, several works [23, 30] explore unsupervised learning in FAS. However, they are based on pretext tasks, which suffer limited performance and cannot alleviate the practical domain shift. Besides, unsupervised DG benchmarks in various scenarios (e.g., cross attack types) have not been built yet.

Unsupervised Learning. Recent progresses focus on contrastive learning, which learns by enforcing similarity over augmentations while avoiding model collapse. Model collapse can be avoided by introducing negative samples for noise-contrastive [5, 18]. Shortly after that, BYOL [15] and SimSiam [8] employ an asymmetric network and eliminate the need of negatives. Besides, several methods [11, 14, 40] propose to enforce similarity among local representations for dense self-supervised learning. However, these methods rely on the i.i.d assumption, which however is not satisfied in UDG FAS, since there are identity-related and domain-related factors as biases. It is non-trivial to avoid shortcuts caused by these irrelevant factors without actual FAS labels.

Unsupervised Domain Generalization. Recently, Zhang et al. [52] present unsupervised domain generalization (UDG) on image classification and propose to select negative samples based on domain similarity. However, due to the limited number of classes, negative samples introduce a lot of noise to FAS. DN^{2}A [29] proposes a new connectivity metric to analyze the inherent problem of UDG and introduces nearest neighbors into learning a generalized feature space. BrAD [16] proposes to intentionally generate edge-like images as positive samples for learning shape-aware features, which fails to learn FAS-related low-level texture features. Thus, a UDG framework designed specifically for FAS is urged for promising results.

3. Methodology

3.1. Revisiting Vanilla Contrastive Learning

Recently, SimSiam [8] employs the asymmetric network and eliminates the need for negative samples. Inspired by this, we adopt a Siamese-like architecture with cosine similarity loss for pulling positive samples together. However, directly applying this architecture fails in UDG for FAS.

Proposition 1. *Representation Z learned by minimizing the vanilla cosine similarity loss maximizes the mutual information $I(Z; X^+)$, where X^+ is augmented positive sample.*

Proof. Please refer to the supplementary material. □

Augmented samples X^+ contain much identity/domain-related information, leading to learning biased features Z as shortcuts by maximizing $I(Z; X^+)$. To verify this, we train SimSiam on FAS datasets. Fig. 2 shows SimSiam fails to learn live/spoof-related features but learns domain-biased and identity-biased features. Specifically, samples from different domains are clustered and separable, while samples from different classes are indistinguishable. Further considering a single domain, samples of different identities are well separated. For UDG in FAS, it is challenging to learn live/spoof-related features under the disturbance of biases.

3.2. Identity-Agnostic Local Representations

Existing contrastive learning methods maximize similarity between global [6, 8, 18] or local representations [20, 40, 41], which however are identity-biased due to the contained facial structural information. Besides, spoofing cues are usually from fine-grained local information. Thus, we propose a novel Split-Rotation-Merge (SRM) strategy to generate identity-agnostic local representations. Specifically, given a cropped face x from the raw capture, two augmented views are $x_1 = t_1(x)$ and $x_2 = t_2(x)$, where $t_1, t_2 \sim \mathcal{T}$ and $\mathcal{T}$ is the sequence of non-distorted augmentation operations.

Split-Rotation. Given augmented view x_1, we first split it into a $m \times m$ grid of patches $\{x_1^p \mid p \in \{1, \dots, m^2\}\}$ as shown in Fig. 3, where p denotes the index of split patches. Then, we use random rotation $\mathcal{R}$ to augment the patches as $x_1^p = r(x_1^p)$ $(r \sim \mathcal{R})$ with the rotation invariance, which can partly destroy identity-related but not live/spoof-related information. After that, the split m^2 local patches are fed into the encoder f separately to obtain the encoded local embeddings as $\{e_1^p \mid p \in \{1, \dots, m^2\}\}$, where $e_1^p = f(x_1^p)$.

Merge. Directly using local embeddings of each patch still suffers from identity-related bias. Considering for most presentation attack types (e.g., print photo, video replay and 3D mask), each local patch contains similar spoof-related discriminative information. While identity-related information differs a lot in each patch, e.g., the patch covered by eyes is quite different from that covered by mouths. Thus, we merge multiple local embeddings e_1^p to form the merged embeddings v_1 for mitigating identity-biased information. Specifically, we select a subset $\mathbf{s}$ of n indices from the patch index set $\mathbf{p} = \{1, \dots, m^2\}$, and collect the corresponding embeddings as $\mathbf{e}_1^{\mathbf{s}} = \{e_1^p \mid p \in \mathbf{s}\}$. Then, the merged embedding v_1 is generated by averaging as $v_1 = \frac{1}{n} \sum_{p \in \mathbf{s}} e_1^p$. Taking all possible n-combinations leads to the merged embedding set $\mathbf{v}_1 = \{v_1^i \mid i \in \{1, \dots, C_{m^2}^n\}\}$, where $C_m^n =$

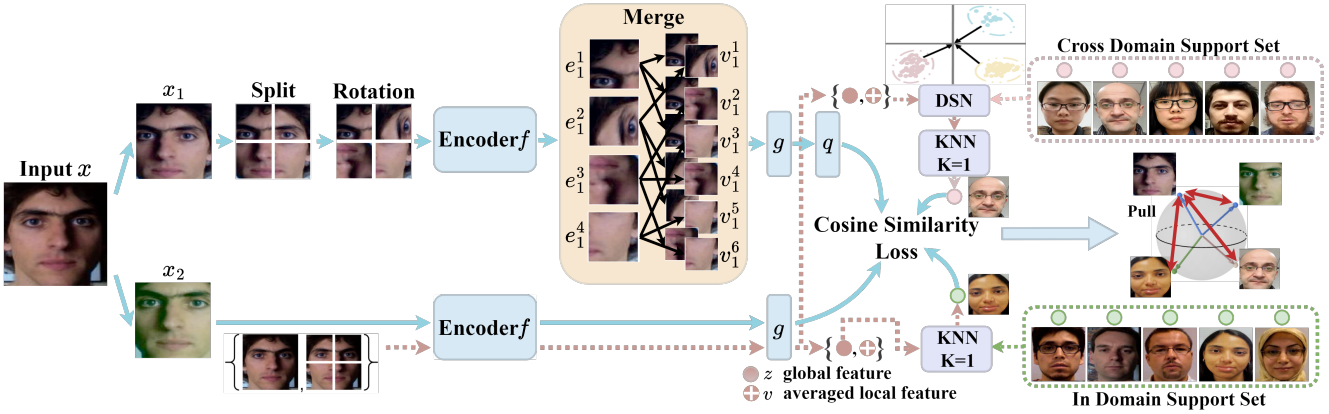


Figure 3. Overall architecture of UDG-FAS. The augmented view x_1 is split into patches and randomly rotated. Each patch is fed into the encoder f separately to obtain e_1^i, and then summed by averaging to generate merged identity-agnostic local features v_1^i. Merged features are fed into projector g, predictor q for contrastive learning. Another view x_2 is input to the encoder f and projector g. And we search the nearest neighbors in in-domain support set as positive samples for contrastive learning to mitigate identity bias. Besides, nearest neighbors in cross-domain support set are employed as positives to alleviate domain bias. To deal with distribution gap, domain-specific normalization with merged local feature is used for more robust search. Note that x_1, x_2 will be swapped and forwarded again for symmetry training.

$\frac{m!}{n!(m-n)!}$. In this way, we generate multiple merged embeddings as identity-agnostic local representations, which mitigate identity bias while retaining spoof-related information.

Similarity Loss. Then, the merged embedding $\{v_1^i\}$ is input to a projector and a predictor to generate the vector $\{p_1^i\}$. Similar to SimSiam [8], the other input view x_2 is fed into the encoder and projector to generate the global vector z_2. By alternatively feeding x_2 and x_1, we can obtain $\{p_2^i\}$ and z_1. Then, they are used to compute the similarity loss

$$\mathcal{L}_{\text{SRM}}^i = \frac{1}{2}\mathcal{D}(p_1^i, \texttt{stopgrad}(z_2)) + \frac{1}{2}\mathcal{D}(p_2^i, \texttt{stopgrad}(z_1)) \quad (1)$$

where $\mathcal{D}(p_1, z_2) = -\frac{p_1}{||p_1||_2} \cdot \frac{z_2}{||z_2||_2}$. Since there are $C_{m^2}^n$ merged embeddings, we average similarity losses from $C_{m^2}^n$ positive pairs between merged vectors p^i and the global embedding z as the final loss $\mathcal{L}_{\text{SRM}} = \sum_{i=1}^{C_{m^2}^n} \mathcal{L}_{\text{SRM}}^i / C_{m^2}^n$.

In-domain Nearest Neighbors. Whilst applying our Split-Rotation-Merge module can suppress identity-biased information within one identity, it remains unresolved how to pull close the images of different persons, which are supposed to act as positives (i.e. the same live/spoof label). To further mitigate identity-related features across identities, we resort to a simple yet effective strategy that searches the nearest neighbors (NN) in the embedding space of the same domain as positive samples for contrastive learning. Specifically, for a cropped face x and its embedding z, we have a *in-domain support set* of embeddings from the same domain $Q_z^{in} = \{z_1^{q_{in}}, ..., z_k^{q_{in}}, ..., z_{|Q_z^{in}|}^{q_{in}}\} \backslash \{z\}$, where $d = d_k^{q_{in}}$. Besides, we have the augmented view x_1 and corresponding support set $Q_{z_1}^{in}$. Then, we search z's NN in $Q_{z_1}^{in}$ as

$$id_{nn}^{q_{in}} = \underset{k \in \{1,...,|Q_z^{in}|\}}{\arg\min} \|z - z_k^{q_{in}}\|_2, z_{nn}^{q_{in}} = Q_{z_1}^{in}[id_{nn}^{q_{in}}] \quad (2)$$

Note that we do not use augmentations in x, making z, Q_z^{in} less noisy to find pure NN. Let $N(z, Q)$ denote NN of z in Q, we have in-domain NN (IDNN) as $N(z, Q_{z_1}^{in}) = z_{nn}^{q_{in}}$, which is employed as positive samples for computing loss

$$\mathcal{L}_{\text{IDNN}}^i = \frac{1}{2}\mathcal{D}(p_1^i, \texttt{stopgrad}(N(z, Q_{z_2}^{in}))) + \frac{1}{2}\mathcal{D}(p_2^i, \texttt{stopgrad}(N(z, Q_{z_1}^{in}))) \quad (3)$$

In this way, we enforce similarity over different identity samples potentially belonging to the same class (live/spoof), which can learn an identity-irrelevant representation space.

Proposition 2. *Representation Z learned by minimizing Eq. (1) and (3) minimizes the mutual information $I(Z; B)$, where B is the variable indicating the identity.*

Proof. Please refer to the supplementary material. □

Besides, using in-domain NN as positives can help to overcome intra-domain variations, e.g., material and camera quality, and better learn the intra-class compact features.

3.3. Domain-Agnostic Positive Samples

The success of existing contrastive learning methods relies on the i.i.d. assumption, which however is not satisfied in UDG due to the distribution shift across domains [52]. To avoid learning domain-biased features as shortcuts, we propose to search the nearest neighbors (NN) across domains as positive samples. However, due to huge distribution shifts, directly searching cross-domain NN may lead to many false matches, i.e., the query and its NN have different labels, which introduces noise and compromises the final result.

Considering visual domain is closely related to image style information [56], which is reflected in the feature statistics, we collect the domain-specific mean and deviation as (μ_d, σ_d^2). Then, we normalize features of each domain to the reference Gaussian distribution with zero mean and unit variance as $\hat{z} = (z - \mu_d)/\sqrt{\sigma_d^2 + \epsilon}$, where z is the

encoded features from domain d, and ϵ is a small constant to avoid numerical instability. By mapping features of various domains to the same distribution, we can search more accurate cross-domain NN. In addition to using the global feature z, we also leverage the proposed merged local features v for a more robust search, which can mitigate identity bias, e.g., ethnic and gender. With total $C_{m^2}^{n}$ merged local features, we obtain an averaged local feature as $v = \sum_i v^i$, which is then normalized to the Gaussian distribution as $\hat{v}$.

Specifically, for a given sample x_j, its normalized global feature $\hat{z}_j$ and local feature $\hat{v}_j$, we have the corresponding *cross-domain support set* of embeddings from different domains as $Q_{\hat{z}}^{cr} = \{\hat{z}_1^{q_{cr}}, ..., \hat{z}_k^{q_{cr}}, ..., \hat{z}_{|Q_{\hat{z}}^{cr}|}^{q_{cr}}\}$ and $Q_{\hat{v}}^{cr} = \{\hat{v}_1^{q_{cr}}, ..., \hat{v}_k^{q_{cr}}, ..., \hat{v}_{|Q_{\hat{v}}^{cr}|}^{q_{cr}}\}$, where $d_j \neq d_k^{q_{cr}}$. Euclidean distances between the query and support set based on the global and local feature are computed as $dist_{\hat{z}}$ and $dist_{\hat{v}}$, which are then combined as the final distance $dist = dist_{\hat{z}} + dist_{\hat{v}}$. Finally, we sort the distance matrix and obtain the index of NN as $id_{nn}^{q_{cr}}$. Besides, we have augmented view x_1 with corresponding support set $Q_{z_1}^{cr}$, and have NN as $z_{nn}^{q_{cr}} = Q_{z_1}^{cr}[id_{nn}^{q_{cr}}]$. With cross-domain NN (CDNN) as positives $N([z, v], Q_{z_1}^{cr}) = z_{nn}^{q_{cr}}$, we compute the loss as:

$$\mathcal{L}_{\text{CDNN}}^{i} = \frac{1}{2}\mathcal{D}(p_1^i, \texttt{stopgrad}(N([z, v], Q_{z_2}^{cr}))) + \frac{1}{2}\mathcal{D}(p_2^i, \texttt{stopgrad}(N([z, v], Q_{z_1}^{cr}))) \quad (4)$$

Proposition 3. *Representation Z learned by minimizing Eq. (4) minimizes the mutual information $I(Z; D)$, where D is the variable indicating the domain.*

Proof. Please refer to the supplementary material. □

In summary, with our proposed identity-agnostic local representations, in-domain and cross-domain nearest neighbors as positive samples, we have the total loss $\mathcal{L}$ as

$$\mathcal{L} = \mathcal{L}_{\text{SRM}} + \lambda_1 \cdot \mathcal{L}_{\text{IDNN}} + \lambda_2 \cdot \mathcal{L}_{\text{CDNN}} \quad (5)$$

At the start of training, the searched neighbors are unreliable. As the training proceeds, the neighbors are more and more reliable. Thus, λ_1 and λ_2 are set as time-dependent, e.g., $\lambda_1(t)$=0 in first T_1 epochs and $\lambda_1(t)$=1 when $t_1 > T_1$.

4. Experiments

4.1. Experimental Settings

Datasets. We experiment on: Idiap Replay-Attack [10] (denoted as I), OULU-NPU [2] (denoted as O), CASIA-MFSD [54] (denoted as C), MSU-MFSD [45] (denoted as M), CelebA-Spoof [53] (denoted as CA), 3DMAD [31] (denoted as D), HKBU-MARs [24] (denoted as H)[1]. Following [21, 36], Half Total Error Rate (HTER) and Area Under the Curve (AUC) are used as the evaluation metrics.

[1]Datasets were solely downloaded and evaluated by Shanghai Jiao Tong University researchers.

Unsupervised Domain Generalization FAS Protocols. We describe our proposed UDG FAS protocols as follows:

UDG-Protocol-1: We unsupervisedly pretrain the model using unlabeled data on three domains of I, O, C and M, and then finetune with labeled data, the proportion of which is 5%, 10%, 20%, 50% and 100%. Note that we split the data by identity ID. Finally, the model is evaluated on the remaining unseen target domain. In this protocol, there is almost no shortage of domain information compared to the standard DG protocol, but the amount of labeled data is relatively small. Besides, we also evaluate with full live data and few labeled spoof data of 5%, 10%, 20% and 50%.

UDG-Protocol-2: The model is pretrained using unlabeled data from three domains of I, O, C and M. Without finetuning, we perform kNN on the model to more directly evaluate the unsupervised pre-trained features on target domain. This protocol evaluates the performance under more challenging scenarios without any labeled data for training.

UDG-Protocol-3: In addition to small datasets (I, O, C, M), we include the current largest CelebA-Spoof (CA) as an additional unlabeled source dataset for pretraining. To save computational overhead, we randomly sample a subset of 100k/200k images. Besides, we extract the real faces of CA as additional source data, which are all web-crawled. After pretraining, we finetune the model with full labeled data of small datasets. This protocol evaluates the effectiveness of our method for using large-scale web-crawled face data.

UDG-Protocol-4: Two datasets among I, O, C and M are set as one group, i.e., [O, M] and [C, I] are set as two groups. The model is pretrained on one group using unlabeled data, finetuned using the labeled data, and then tested on the other unseen group. This protocol evaluates the efficiency and generalizability of models with limited source domains.

UDG-Protocol-5: In this UDG-based attack type generalization protocol, following the 'leave one attack type out' data usage in [1], we pretrain on two domains of I, C and M with partial attack type data using unlabeled data, finetune with the labeled data, and then test on the unseen domain with unseen attack types. This protocol measures the generalization of both unseen domain and 2D attack types.

UDG-Protocol-6: We evaluate the generalization on unseen 3D mask attack in this UDG-based protocol. In specific, we pretrain the model using unlabeled data on O, C, I and M, finetune using the labeled data, and then test on 3D mask dataset D and H. In addition, the model is also pretrained on O, C, M and tested on the large-scale CA dataset, which contains an unseen 3D mask attack types.

Implementation Details. For unsupervised training, we adopt ResNet-18 as the backbone. Following [8], we use a projector with three MLP layers and a predictor with two MLP layers. We adopt the SDG optimizer with lr=0.03 and a cosine decay schedule for 100 epochs of training. For our SRM module, we set m=2 and n=2. The hyperparameter

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
Random Init [19]	12.62	92.15	35.33	68.25	25.64	77.09	32.20	73.07
ImageNet Init [19]	11.43	93.99	16.44	91.25	23.57	77.25	22.31	85.65
Moco V2 [7]	12.86	93.63	17.89	88.41	16.50	87.80	27.48	79.07
SimCLR V2 [6]	12.86	93.08	17.33	90.89	15.71	87.07	26.67	79.55
BYOL [15]	14.76	86.29	22.67	84.74	14.28	90.81	22.48	85.26
SimSiam [8]	11.43	93.83	16.78	89.69	14.28	92.69	19.30	88.67
UDG-FAS (Ours)	**7.14**	**97.31**	**11.44**	**95.59**	**6.28**	**98.61**	**12.18**	**94.36**
RFM [37]	17.30	90.48	13.89	93.98	20.27	88.16	16.45	91.16
D^2AM [9]	15.43	91.22	12.70	95.66	20.98	85.58	15.27	90.87
SSDG-R [21]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
SSAN [44]	6.67	**98.75**	10.00	96.67	8.88	96.79	13.72	93.63
PatchNet [38]	7.10	98.46	11.33	94.58	13.4	95.67	11.82	95.07
UDG-FAS+SSDG (Ours)	**5.95**	98.47	**9.82**	**96.76**	**5.86**	**98.62**	**10.97**	**95.36**

Table 1. Results on *UDG-Protocol-1* with full labeled data for finetuning. The methods in the top half are firstly unsupervised pretrained and then finetuned with a baseline ResNet-18 model, while those in the lower part are DG methods with ImageNet pretraining.

Methods	Label Fraction 50% Live + 50% Spoof								Label Fraction 20% Live + 20% Spoof							
	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
ImageNet Init [19]	18.57	87.58	25.89	82.90	27.86	71.33	23.59	83.55	22.86	85.93	27.44	81.27	29.14	73.36	27.76	78.90
MoCo V2 [7]	14.52	92.41	19.33	86.75	21.43	79.37	29.32	77.52	21.43	88.77	25.33	82.43	24.43	81.24	33.73	70.22
SimCLR V2 [6]	14.05	93.18	18.67	85.87	20.71	84.28	28.89	76.98	20.24	89.78	24.67	82.73	23.64	78.94	32.62	71.71
BYOL [15]	15.95	88.97	23.33	84.42	17.86	84.84	21.79	86.45	17.38	85.80	23.33	83.92	21.50	85.08	25.83	83.05
SimSiam [8]	13.10	94.37	18.00	90.68	17.14	91.92	20.07	87.70	15.71	91.13	19.89	88.79	21.43	80.62	24.58	83.86
UDG-FAS (Ours)	**10.00**	**96.27**	**13.33**	**93.42**	**9.93**	**96.19**	**12.27**	**94.74**	**11.43**	**95.04**	**13.88**	**93.31**	**12.64**	**96.08**	**12.83**	**94.03**

Methods	Label Fraction 10% Live + 10% Spoof								Label Fraction 5% Live + 5% Spoof							
	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
ImageNet Init [19]	22.86	83.09	32.00	71.11	27.89	80.03	30.66	75.03	22.86	81.23	29.89	79.89	34.29	68.64	30.85	75.51
MoCo V2 [7]	21.43	88.05	28.56	80.38	26.21	77.22	36.79	67.26	21.43	87.47	30.11	76.72	26.43	77.98	37.22	67.85
SimCLR V2 [6]	20.71	87.82	27.22	80.21	25.14	81.19	35.99	68.46	20.23	87.10	27.89	80.17	25.28	83.21	35.83	67.88
BYOL [15]	20.00	84.69	26.67	80.24	20.71	81.59	27.93	80.89	21.43	83.93	32.67	74.14	20.71	82.85	30.54	77.93
SimSiam [8]	17.14	92.33	20.78	89.12	20.71	78.18	28.47	78.56	19.76	89.16	27.33	80.63	22.86	80.39	29.17	78.59
UDG-FAS (Ours)	**11.67**	**94.80**	**13.88**	**93.48**	**13.57**	**94.99**	**15.29**	**92.13**	**12.86**	**93.32**	**18.67**	**89.83**	**15.64**	**90.67**	**15.14**	**92.49**

Table 2. Results on *UDG-Protocol-1* with partial labeled data ranging from 5% to 50%. We split the training set by the identity ID.

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
MocoV2	32.86	69.12	45.33	58.67	36.86	65.10	40.96	62.18
SimCLRV2	31.19	82.04	46.67	57.62	37.85	66.64	37.55	66.74
BYOL	31.19	74.41	33.98	67.34	38.43	65.45	34.52	68.59
SimSiam	23.33	79.75	38.67	73.68	36.50	65.22	37.18	67.95
Ours	**19.76**	**84.34**	**23.78**	**86.69**	**12.50**	**94.75**	**17.98**	**88.52**

Table 3. Results on *UDG-Protocol-2* with kNN (k=10) evaluation.

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
SimSiam [8]	11.43	93.83	16.78	89.69	14.28	92.69	19.30	88.67
+100,000	10.95	94.78	15.33	90.91	13.57	93.12	17.89	89.19
+200,000	10.56	95.07	14.56	92.28	12.93	94.35	16.70	90.47
+Web-crawled	11.19	94.46	15.89	90.16	13.36	93.72	18.24	89.05
UDG-FAS	7.14	97.31	11.44	95.59	6.28	98.61	12.18	94.36
+100,000	6.32	97.45	9.33	96.58	5.17	98.84	10.27	96.15
+200,000	5.71	98.31	7.69	97.92	4.48	99.03	9.06	96.51
+Web-crawled	6.58	97.08	9.82	96.04	4.75	98.91	10.81	95.38

Table 4. *UDG-Protocol-3* with CA as an additional source domain.

is set as T_1=30 and T_2=60. For finetuning, we initialize a ResNet-18 encoder with unsupervised trained weight, and randomly initialize a linear classifier. The model is trained by the SGD optimizer with lr=0.001.

Methods	**O&M to C&I**		**C&I to O&M**	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)
SSDG-R [21]	20.92	88.07	22.57	85.61
DF-DM [23]	29.61	73.78	32.94	73.58
DF-DM† [23]	<u>18.96</u>	<u>89.48</u>	<u>18.60</u>	**89.76**
ImageNet Init [19]	25.65	79.14	28.14	79.05
Moco V2 [7]	22.42	83.10	33.17	72.05
SimCLR V2 [6]	23.12	80.89	32.30	72.67
BYOL [15]	22.03	88.02	27.21	80.69
SimSiam [8]	22.42	88.36	24.30	82.80
UDG-FAS (Ours)	**14.54**	**93.81**	**18.13**	<u>88.81</u>

Table 5. Results on *UDG-Protocol-4* with limited source domains. † indicates use ImageNet initialization for unsupervised training.

4.2. Experimental Results

UDG-Protocol-1. Table 1 shows our method greatly improves performance on unseen target domain, i.e., outperforms ImageNet pretraining by 9.43% average AUC gain, indicating a better initialization for FAS models. Besides, compared to recent unsupervised methods, UDG-FAS improves performance a lot by mitigating the identity and domain bias, e.g., 6.19% HTER lower than SimSiam on average. Surprisingly, without any changes on learning objec-

Methods	CASIA-MFSD			Replay-Attack			MSU			Overall
	Video	Cut photo	Warped	Video	Digital Photo	Printed	Printed	HR Video	Mobile Video	
SVM1+IMQ [1]	88.41	75.14	75.23	88.21	71.20	56.41	56.62	71.12	49.75	70.23±12.69
CDCN [50]	72.20	79.31	84.22	**97.73**	94.89	96.70	74.25	98.88	**100.00**	87.69±10.56
CDCN++ [50]	73.12	76.64	78.36	96.66	92.92	97.67	74.25	98.13	**100.00**	87.53±10.90
SSAN [44]	73.20	75.27	82.69	97.48	89.26	96.04	79.69	**99.75**	98.75	88.01±9.93
TTN-S [42]	**90.26**	79.60	**95.17**	68.81	93.82	95.88	**88.87**	95.19	99.82	89.71±9.17
ImageNet Init [19]	73.07	71.89	72.17	88.52	77.68	81.92	67.51	98.94	98.61	81.14±11.08
SimSiam [8]	79.44	74.25	75.61	93.81	82.57	97.09	75.20	99.51	99.69	86.35±10.38
UDG-FAS (Ours)	88.73	**82.80**	84.35	96.65	**96.20**	**99.26**	84.37	99.71	99.84	**92.43±6.86**

Table 6. Results on *UDG-Protocol-5* for cross domain cross 2D attack type experiments. The bottom half are pretraining methods.

Methods	O&C&I&M to D		O&C&I&M to H	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)
RFMetaFAS [37]	5.88	98.35	41.67	81.64
SSDG-R [21]	6.77	98.42	32.50	73.68
SSAN [44]	0.74	99.74	**26.98**	80.27
ImageNet init [19]	8.24	97.52	35.52	69.57
Moco V2 [7]	9.41	96.38	36.61	70.98
SimCLR V2 [6]	8.24	97.05	35.71	66.90
BYOL [15]	7.06	98.11	32.14	73.68
SimSiam [8]	6.76	98.49	32.94	70.83
UDG-FAS (Ours)	**0.29**	**99.95**	27.36	**80.31**

Table 7. Results for *UDG-Protocol-6* on unseen 3D mask attack.

Methods	M&C&O to CA	
	HTER(%)	AUC(%)
Saha *et al.* [35]	27.1	79.2
Panwar *et al.* [32]	26.1	80.0
SSDG-R [21]	25.05	82.11
CIFAS [26]	24.6	83.2
Moco V2 [7]	28.71	78.56
SimCLR V2 [6]	27.89	79.34
BYOL [15]	28.07	78.67
SimSiam [8]	26.16	81.52
UDG-FAS (Ours)	**21.35**	**87.26**

Table 8. Results for *UDG-Protocol-6* on unseen 3D attack of CA.

tives and network architectures, a baseline model, ResNet-18, can directly outperform most DG methods and achieve comparable performance to SOTA methods by using UDG-FAS as initialization. Combined with SSDG-R, UDG-FAS achieves a new SOTA DG performance, improving SSDG-R from 11.29% HTER to 8.15%. Table 2 shows that our method consistently outperforms other counterparts for all fractions of labeled data. With only 5% faces (i.e., two identities for each domain), UDG-FAS achieves 15.14% HTER for I&C&M to O, which is better than SSDG using all labeled data, showing the effectiveness of reducing label cost.

UDG-Protocol-2. As shown in Table 3, UDG-FAS outperforms other counterparts by a large margin for kNN evaluation. Compared with SimSiam, we achieve 16.63% AUC gain on average, showing the effectiveness of our unsupervised learned features for FAS. Besides, without any labeled data for training, our UDG-FAS even outperforms SimSiam finetuned with all data by 2.06% AUC for O&C&M→I.

UDG-Protocol-3. Table 4 shows using 100k unlabeled CA data improves performance with 1.49% average HTER drop. Besides, UDG-FAS consistently benefits from the increased amount of data to 200k. Meanwhile, including web-crawled real faces also improves performance with 1.27% HTER drop. Compared with SimSiam, UDG-FAS benefits more from increased data with larger accuracy gain. The results exhibit the power of UDG-FAS to use large-scale web-crawled face data for enhancing the pre-trained features.

UDG-Protocol-4. Table 5 shows that with limited source domains, our unsupervised pretraining outperforms SSDG-R by 5.41% average HTER reduction, exhibiting the data efficiency and generalizability of UDG-FAS. Besides, compared with SimSiam, UDG-FAS improves the performance by 5.73% AUC gain. Moreover, UDG-FAS outperforms DF-DM by 4.42% HTER reduction for O&M to C&I, showing the effectiveness of mitigating identity and domain bias.

UDG-Protocol-5. Under cross domain cross attack test, Table 6 shows that UDG-FAS outperforms ImageNet Init by 11.29% AUC gain. Compared with SimSiam, UDG-FAS achieves 6.08% AUC gain. Moreover, UDG-FAS even outperforms SOTA DG method, i.e., 2.72% higher AUC than TTN-S. Though unsupervised training without unseen attack types, UDG-FAS forces the model to learn an identity-irrelevant and domain-irrelevant representation space, facilitating generalization under domain and attack type shifts.

UDG-Protocol-6. As shown in Table 7, UDG-FAS outperforms ImageNet Init by 8.06% HTER reduction for unseen 3D mask attack. Compared with SimSiam, we achieve 6.03% HTER reduction on average. Moreover, UDG-FAS even outperforms SOTA DG methods, e.g., 0.25% higher AUC than SSAN and 4.08% higher AUC than SSDG-R. Table 8 shows that UDG-FAS outperforms SOTA DG methods for large-scale CA, e.g., 4.06% AUC gain compared to CIFAS, showing the validity of unsupervised pretraining.

4.3. Ablation Study

Ablation study is conducted on *UDG-Protocol-1* with full labeled data for finetuning to evaluate each component.

Effectiveness of Each Component. To verify the validity of mitigating identity-related bias, we experiment w/o Split-Rotation-Merge and in-domain NN, respectively. Table 9 shows that the performance is degraded, demonstrating the effectiveness of our SRM module and in-domain NN

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
Ours w/o Split-Rotation-Merge	10.24	96.74	14.89	93.06	11.43	94.65	17.93	89.67
Ours w/o in-domain NN	9.76	97.02	13.44	93.84	8.71	97.28	14.31	92.22
Ours w/o cross-domain NN	10.00	96.35	15.33	93.29	10.71	95.73	16.08	91.31
UDG-FAS (Ours)	**7.14**	**97.31**	**11.44**	**95.59**	**6.28**	**98.61**	**12.18**	**94.36**

Table 9. Evaluations of different components of the proposed method on four cross-dataset testing protocols.

	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
$k=1$	**7.14**	**97.31**	11.44	**95.59**	**6.28**	**98.61**	**12.18**	**94.36**
$k=2$	**7.14**	96.45	**10.77**	94.98	7.07	98.36	12.91	93.92
$k=4$	7.38	96.04	12.00	93.74	7.85	97.51	14.24	92.87

Table 10. Ablation on k in our in-domain and cross-domain NN.

Methods	50%labels	20%labels	10%labels	5%labels	kNN
Ours w/o NNs +SSDG	17.10	19.91	23.78	24.93	31.15
Our UDG-FAS	12.27	12.83	15.29	15.14	17.98

Table 11. HTER on I&C&M to O without in/cross-domain NNs.

for alleviating identity bias. Besides, to prove the importance of mitigating domain-related information, we experiment w/o cross-domain NN. The results in Table 9 indicate that learning a domain-irrelevant feature space is beneficial to improve the generalizability for cross-domain FAS tasks. Moreover, our unsupervised model can be deployed with no labels or finetuned with few labels, in which practical cases finetuning with regularization is inapplicable or inferior due to the lack of sufficient labels. As labels become fewer, Table 11 shows UDG-FAS obtains higher performance gain over Ours w/o NNs finetuning with SSDG.

Effectiveness of Split-Rotation-Merge Module. We dive into SRM module to inspect the influence of each part. As shown in Table 12, the performance degrades if any part of the module is removed due to the less suppressed identity bias. We also compare with patch shuffle (PS) augmentation [51], which is a way to mitigate identity-related information at the input level. UDG-FAS outperforms PS by 3.49% HTER reduction, showing that SRM module is more effective in mitigating identity bias. Besides, we examine the choice of the number of split and merged patches (i.e., m and n), where the split number controls the patch size. Fig. 4 shows that when $m = 2$, selecting $n = 2$ patches for merging is best, since there is information gap with half of patches, which filters the identity-biased information. While small patch size (m=3) may fragment spoofing cues and degrade the performance.

Effectiveness of Cross-domain Nearest Neighbor. Table 13 shows that ours w/o DSN and w/o local degrade the performance, showing the effectiveness of domain-specific normalization (DSN) and combining local distances for the more accurate cross domain search. Ours w GT denotes using ground-truth labels to construct cross-domain samples as positives, which is the upper bound performance.

Effects of k in Nearest Neighbors. In experiments, we

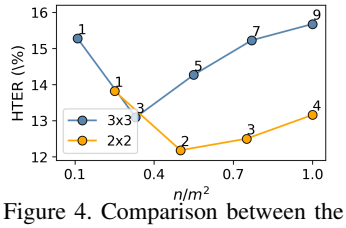


Figure 4. Comparison between the number of split samples and merged samples on I&C&M to O.

Methods	HTER(%)
Ours w/o split	16.25
Ours w/o rotation	14.09
Ours w/o merge	13.82
Ours w PS	15.67
Ours	**12.18**

Table 12. Ablation study on our Split-Rotation-Merge on I&C&M to O.

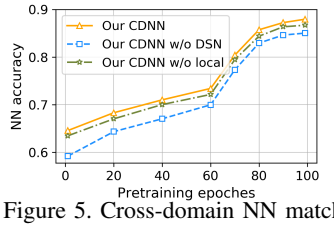
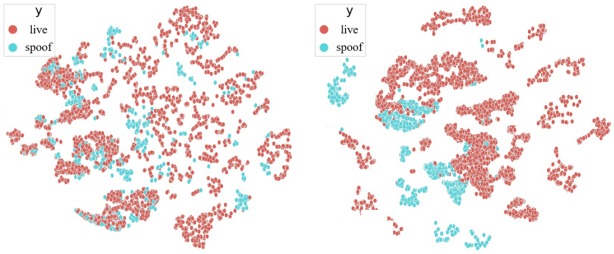


Figure 5. Cross-domain NN match accuracy on O&C&M to I.

Methods	HTER(%)
Ours w/o DSN	13.94
Ours w/o local	12.97
Ours w GT	10.83
Ours	**12.18**

Table 13. Ablation study on our cross-domain NN search strategy on I&C&M to O.



(a) Vanilla SimSiam (b) Our UDG-FAS

Figure 6. T-sne visualization of unsupervised features learned by SimSiam and our approach. Samples of each category (live/spoof) tend to be grouped together in our learned feature space (though not perfect as our model is unsupervised trained without labels).

select the top-1 ranked neighbor as the positive samples. Table 10 shows that, UDG-FAS is somewhat robust to changing the value of k, but increasing beyond $k = 1$ results in slight degradation due to the brought noise.

4.4. Visualization and Analysis

Visualization of Feature Space. Fig. 6 shows SimSiam fails to learn live/spoof-related features and samples of different classes are closely entangled. By contrast, samples of each class are separable in our learned feature space.

Class Activation Map (CAM). Fig. 7 shows UDG-FAS focuses on the facial region for live samples and attaches importance to photo cut position and holding hand for predicting spoof samples. While SimSiam focuses on the land-

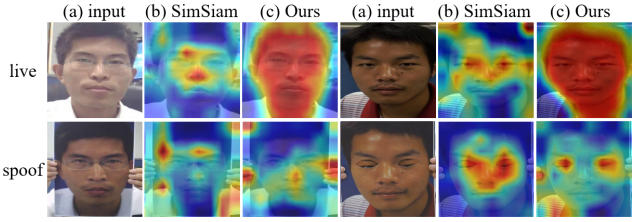


Figure 7. Grad-CAM visualizations for O&M&I to C. (a) Input images. Visualizations for (b) SimSiam and (c) our UDG-FAS.

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
DINO	10.24	95.36	15.33	92.14	12.64	93.46	17.22	90.64
MAE	9.05	96.21	13.67	92.48	11.50	94.27	15.28	91.35
Ours	**7.14**	**97.31**	**11.44**	**95.59**	**6.28**	**98.61**	**12.18**	**94.36**
DINO+SSDG	9.76	96.38	14.00	92.94	11.36	94.27	16.39	90.95
MAE+SSDG	8.33	96.63	12.56	93.69	10.64	95.18	14.16	92.03
Ours+SSDG	**5.95**	**98.47**	**9.82**	**96.76**	**5.86**	**98.62**	**10.97**	**95.36**

Table 14. Comparison of transformer-based SSL methods.

mark areas in faces that contain identity-biased features.

Nearest Neighbor Match Accuracy. Fig. 5 shows the accuracy of searched cross-domain NN for three strategies (ours, ours w/o DSN and w/o local). Ours w/o DSN and w/o local have the lower NN search accuracy, demonstrating the effectiveness of our method for searching accurate NN.

Searched Nearest Neighbor. Fig. 8 shows the nearest neighbors retrieved with our unsupervised features. Cross-domain NNs searched by UDG-FAS are from the same live/spoof class, where the spoof ones are even from the same attack type, e.g., video or print attacks. Besides, the searched in-domain NNs are also accurate and have different personal traits (identity/gender), demonstrating that the learned feature contains less identity-biased information.

Comparison with Transformer-based SSL Methods. We compare our method (based on ResNet-18) with SOTA transformer-based methods DINO [4] and MAE [17] (using ViT-Small as backbone). Table 14 shows our method is superior with fewer parameters and FLOPs, and achieves more significant gains when combined with SSDG. This means using SSL methods without taking the properties of FAS tends to learn identity/domain-biased features and degrades the performance.

Identity Retrieval Performance. To further evaluate whether our method can effectively mitigate identity-biased information in an unsupervised fashion, we use the unsupervised pre-trained network to extract the facial features for identity retrieval. Table 15 shows that the identity retrieval performance degrades with our unsupervised training, in line with our objective of removing identity-related information to avoid shortcuts in FAS.

5. Conclusion

In this paper, we propose the first unsupervised domain generalization framework for face anti-spoofing, which can exploit large amounts of more accessible unlabeled data to learn generalizable features for enhancing the low-data status of FAS. Regarding the inherent identity and domain biases, we propose a novel SRM module to explore identity-agnostic local representations. Besides, in-domain nearest neighbors are employed as positives to further mitigate identity bias. Moreover, cross-domain nearest neighbors are searched with the domain-specific normalization to learn domain-invariant features. Extensive experiments validate the effectiveness of our method statistically and visually. Even with 5% labeled data, UDG-FAS can still achieve promising results without much performance degradation.

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	P@1↓	P@5↓	P@1↓	P@5↓	P@1↓	P@5↓	P@1↓	P@5↓
MocoV2	33.39	15.78	34.41	16.05	36.61	18.42	29.19	13.38
SimCLRV2	34.56	16.07	37.28	16.73	36.89	18.49	30.13	13.41
SimSiam	28.04	15.03	29.57	15.36	33.34	18.09	26.63	13.27
Ours	**17.82**	**9.96**	**19.69**	**10.29**	**21.11**	**11.48**	**10.37**	**4.70**

Table 15. Identity retrieval performance of different unsupervised learning methods.

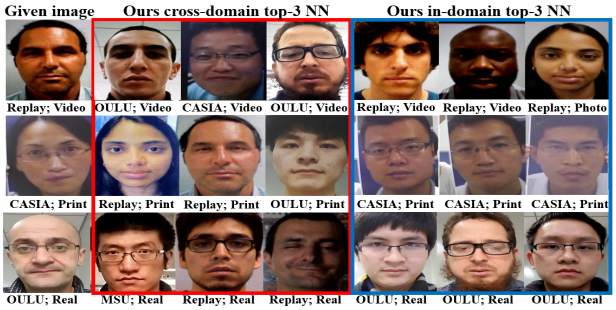


Figure 8. The searched cross-domain and in-domain NN by UDG-FAS, where spoof ones are from the same fine grained attack types.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62250055, Grant 61932022, Grant 62120106007, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 20511100100.

References

[1] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE access*, 5:13868–13882, 2017. 5, 7

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618. IEEE, 2017. 5

[3] Rizhao Cai, Yawen Cui, Zhi Li, Zitong Yu, Haoliang Li, Yongjian Hu, and Alex Kot. Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In

Internetional Conference on Computer Vision (ICCV), 2023. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640. IEEE, 2021. 9

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2, 3, 6, 7

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6, 7

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 3, 4, 5, 6, 7

[9] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *AAAI*, pages 1132–1139, 2021. 1, 3, 6

[10] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7, 2012. 5

[11] Yuanzheng Ci, Chen Lin, Lei Bai, and Wanli Ouyang. Fastmoco: Boost momentum-based contrastive learning with combinatorial patches. *arXiv preprint arXiv:2207.08220*, 2022. 3

[12] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. LBP-TOP based countermeasure against face spoofing attacks. In *ACCV*, pages 121–132. Springer, 2012. 1, 3

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[14] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022. 3

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3, 6, 7

[16] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, et al. Unsupervised domain generalization by learning a bridge across domains. *arXiv preprint arXiv:2112.02300*, 2021. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 9

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[19] 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. 6, 7

[20] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Learning where to learn in cross-view self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14451–14460, 2022. 3

[21] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8484–8493. IEEE, 2020. 3, 5, 6, 7

[22] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *BTAS*. IEEE, 2013. 1, 3

[23] Haozhe Liu, Zhe Kong, Raghavendra Ramachandra, Feng Liu, Linlin Shen, and Christoph Busch. Taming self-supervised learning for presentation attack detection: In-image de-folding and out-of-image de-mixing. *arXiv preprint arXiv:2109.04100*, 2021. 3, 6

[24] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016. 5

[25] Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In *ECCV*, 2022. 1

[26] Yuchen Liu, Yabo Chen, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Causal intervention for generalizable face anti-spoofing. In *ICME*, 2022. 1, 7

[27] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398. IEEE, 2018. 3

[28] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, pages 4680–4689. IEEE, 2019. 1, 3

[29] Yuchen Liu, Yaoming Wang, Yabo Chen, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3510–3519, 2023. 3

[30] Usman Muhammad, Zitong Yu, and Jukka Komulainen. Self-supervised 2d face presentation attack detection via temporal sequence sampling. *Pattern Recognition Letters*, 156:15–22, 2022. 3

[31] Erdogmus Nesli and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS'13)*, pages 1–8, 2013. 5

[32] Ankush Panwar, Pratyush Singh, Suman Saha, Danda Pani Paudel, and Luc Van Gool. Unsupervised compound domain adaptation for face anti-spoofing. In *FG*. IEEE, 2021. 7

[33] Keyurkumar Patel, Hu Han, and Anil K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016. 1, 3

[34] Yunxiao Qin, Chenxu Zhao, Xiangyu Zhu, Zezheng Wang, Zitong Yu, Tianyu Fu, Feng Zhou, Jingping Shi, and Zhen Lei. Learning meta model for zero-and few-shot face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11916–11923, 2020. 3

[35] Suman Saha, Wenhao Xu, Menelaos Kanakis, Stamatios Georgoulis, Yuhua Chen, Danda Pani Paudel, and Luc Van Gool. Domain agnostic feature learning for image and video based face anti-spoofing. In *CVPR Workshops*, pages 802–803. IEEE, 2020. 7

[36] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031. IEEE, 2019. 5

[37] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981. AAAI Press, 2020. 1, 3, 6, 7

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

[39] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1

[40] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[41] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16590–16599, 2022. 3

[42] Zhuo Wang, Qiangchang Wang, Weihong Deng, and Guodong Guo. Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1254–1269, 2022. 7

[43] 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 (CVPR)*, pages 4123–4133, June 2022. 1, 3

[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. 6, 7

[45] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 5

[46] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *ACPR*, pages 141–145. IEEE, 2015. 3

[47] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 3

[48] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *ECCV*, pages 557–575. Springer, 2020. 3

[49] 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, 2021. 1

[50] 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 *CVPR*, pages 5295–5305. IEEE, 2020. 3, 7

[51] K.-Y. Zhang, Taiping Yao, Jian Zhang, Shice Liu, Bangjie Yin, Shouhong Ding, and Jilin Li. Structure destruction and content combination for face anti-spoofing. In *IJCB*. IEEE, 2021. 8

[52] Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Domain-irrelevant representation learning for unsupervised domain generalization. In *CVPR*, 2022. 2, 3, 4

[53] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. CelebA-Spoof: Large-scale face anti-spoofing dataset with rich annotations. In *ECCV*, pages 70–85. Springer, 2020. 5

[54] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31. IEEE, 2012. 5

[55] Guanghao Zheng, Yuchen Liu, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Learning causal representations for generalizable face anti spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[56] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 4