Generalized Face Anti-Spoofing via Finer Domain Partition and Disentangling Liveness-Irrelevant Factors

Jingyi Yang[1], Zitong Yu[2], Xiuming Ni[3], Jia He[3] and Hui Li[1,*]

[1]Dept. EEIS, University of Science and Technology of China
The CAS Key Laboratory of Wireless-Optical Communications
[2]School of Computing and Information Technology, Great Bay University
[3]Anhui Tsinglink Information Technology Co.,Ltd.

Abstract. Face anti-spoofing techniques based on domain generalization have recently been studied widely. Adversarial learning and meta-learning techniques have been adopted to learn domain-invariant representations. However, prior approaches often consider the dataset gap as the primary factor behind domain shifts. This perspective is not fine-grained enough to reflect the intrinsic gap among the data accurately. In our work, we redefine domains based on identities rather than datasets, aiming to disentangle liveness and identity attributes. We emphasize ignoring the adverse effect of identity shift, focusing on learning identity-invariant liveness representations through orthogonalizing liveness and identity features. To cope with style shifts, we propose Style Cross module to expand the stylistic diversity and Channel-wise Style Attention module to weaken the sensitivity to style shifts, aiming to learn robust liveness representations. Furthermore, acknowledging the asymmetry between live and spoof samples, we introduce a novel contrastive loss, Asymmetric Augmented Instance Contrast. Extensive experiments on four public datasets demonstrate that our method achieves state-of-the-art performance under cross-dataset and limited source dataset scenarios. Additionally, our method has good scalability when expanding diversity of identities. Code is available at https://github.com/yjyddq/DLIF.

1 Introduction

Face recognition (FR) technology has been widely used in various applications such as access control systems and mobile payments. Unfortunately, FR systems are vulnerable to various presentation attacks, including printing attacks, video replays, and 3D masks among others. To counter such risks, researchers have proposed face anti-spoofing (FAS) techniques, which have become a popular research topic in recent years. A series of face anti-spoofing methods have been developed, including those based on hand-crafted features [22, 7, 21, 3], as well as deeply-learned features [8, 36, 15], both of which have shown promising results in intra-dataset scenarios. However, they often exhibit poor generalization ability to unknown domains, largely due to assumptions about the stationary settings of liveness-irrelevant factors such as lighting, resolution of capture devices. They struggle to overcome dataset bias (domain shifts) resulting from real-world scenarios. Consequently, a significant chal-
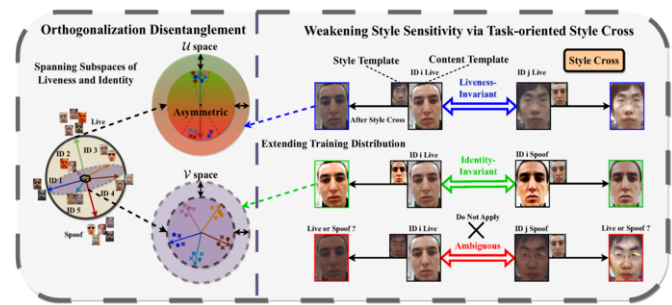
Figure 1. **(Left)** Orthogonalization of liveness and identity attributes. The earth's axis represents the subspace $\mathcal{U}$ associated with the liveness component, where the green and red arrows indicate "live" and "spoof". The equatorial plane represents the subspace $\mathcal{V}$ belonging to the identity component, and colored arrows represent different identities. **(Right)** In $\mathcal{U}$ space, the liveness of the content template and the style template should be consistent. While in $\mathcal{V}$ space, the identity invariance is guaranteed.

lenge for FAS systems is the inability to effectively transfer the anti-spoofing models, learned within one or several domains, to an unseen domain.

To overcome the domain shifts, recent studies have leveraged domain generalization (DG) techniques to enhance the generalization ability of FAS system to unknown domains. The majority of DG-based FAS (DG-FAS) approaches utilize adversarial learning [14, 24, 12] or meta-learning [25, 16] to learn domain-invariant representations. However, these approaches consider the dataset gap as the intrinsic divergence among the data and employ dataset partition as domain. The domain labels they used are coarse, and cannot comprehensively reflect the latent correlation from data. Because, even within the same dataset, inconsistencies, such as identities, illuminations, and resolutions may exist. Though D^{2}AM [5] attempts to assign pseudo-domain labels through clustering, it roughly aggregates the source data into a few clusters which does not solve the problem in essence. Moreover, another considerable drawback is that when the number of source datasets is limited, it is not conducive to learning domain-invariant representations. Under the limited source dataset scenarios, the performance of dataset partition is significantly inferior. In extreme situations, such methods are powerless when only one source dataset is available.

In this paper, we refine the factors that cause the domain shift into identity, style, and unseen spoof patterns, rather than vague

dataset gaps. Additionally, we adopt a more finer domain partition according to identities instead of datasets. Our framework, named *Disentangling Liveness-irrelevant Factors* (**DLIF**). Concretely, we employ two networks to extract liveness and identity features separately. These features are then treated as dissimilar and expressed as orthogonal from a subspace perspective, instead utilizing generative adversarial network and pixel reconstruction approaches like [30, 40], which exhibit heavy computational overheads. To enhance the efficiency and scalability of our framework, we propose two plug-and-play modules: Style Cross (SC) and Channel-wise Style Attention (CWSA). Specifically, SC is a feature-level style augmentation technique, and we explore in detail the effectiveness of executing it at different levels of the network. Meanwhile, in order to prevent label ambiguity caused by uncontrolled random SC in specific tasks, we implement Liveness-invariant Style Cross for FAS network and Identity-invariant Style Cross for FR network. CWSA, designed to adaptively generate style-insensitive features based on channel styles, is introduced specifically for FAS to further mitigate the impact of style shifts. Furthermore, we propose a Asymmetric Augmented Instance Contrastive loss which consider the asymmetry of live and spoof samples to learn robust liveness representation distribution. Simultaneously, our method exhibits excellent scalability. Building on the success of face recognition, we can leverage the knowledge gained from well-trained FR networks to provide an auxiliary supervision for FAS networks. Our main contributions are four-fold:

- We propose a novel perspective involving finer domain partition corresponding to identity. Here, liveness features and identity features are considered orthogonal and disentangled through orthogonality, as illustrated in Figure 1 (Left).
- We propose a plug-and-play Style Cross module for style augmentation, as shown in Figure 1 (Right), along with a Channel-wise Style Attention module to learn style-insensitive features.
- We propose an Asymmetric Augmented Instance Contrast loss that asymmetrically treats live (homogeneity-aware) and spoof (heterogeneity-aware) instances which substantially improves the generalization capability.
- Our framework is compatible with most existing face recognition models, enabling us to leverage well-trained face recognition models for disentangling rather than training from scratch. The scalability study demonstrates that after utilizing the well-trained FR models, there is a significant improvement in performance. Furthermore, if we increase the identity diversity of the training data, the performance will be further improved.

2 Related Work

Face Anti-spoofing Methods. Early handcrafted methods, such as SIFT [21], LBP [7], and HOG [35] were utilized to address the FAS problem. Subsequently, with the emergence of CNN-based deep networks, binary classification-based approaches gain popularity [36, 15], or leveraging auxiliary supervised signals that contain rich anti-spoofing information, such as pseudo-depth maps, reflection maps, and rPPG signals [2, 18]. Furthermore, novel custom operators [39, 37] were introduced, suggesting their effectiveness for the FAS task. Recent efforts have delved into domain adaptation (DA) [31, 32], and domain generalization (DG) [12, 16, 26] to achieve generalization on unseen domains.

Generalizing Domain-specific Styles. Previous work [27] indicates that implementing feature-level style crossing can expand the

training distribution and enhance the generalization ability against to domain shift. Inspired by AdaIN [11], SSAN [33] first proposed the shuffle style assembly (SSA) which randomly swap and mix source styles among source contents for the FAS problem. Additionally, Zhou et al. [44] declare that the feature covariance stores domain-specific features, and instance whitening is effective in removing such domain-specific styles in image translation. They propose asymmetric instance adaptive whitening to align instances at a finer granularity. Those suggest that it is worthwhile to consider feature-level style augmentation and style sensitivity weakening methods that are more applicable to the FAS problem.

Disentangling Liveness-Irrelevant Representation. Disentangled representation learning (DRL) focuses on extracting features that can effectively capture correlations between different datasets, to solve the problem that features related to different tasks can be easily coupled with each other when there is no clear guidance. [42, 40] divide the representation of an image into content and liveness parts to solve FAS problems. Wang et al. [30] explicitly disentangles identity with liveness. Liu et al. [19] disentangles a spoof face into a live counterpart and spoof trace and aims to explicitly extract the spoof traces from faces. These methods entail pixel-level image reconstruction and adversarial training of GANs, which involves significant training cost and difficulty. Instead, we consider the liveness and identity to be dissimilar, a relationship that can be expressed as orthogonal in terms of cosine similarity. It is simpler and more efficient to execute.

3 Proposed Method

Our architecture is illustrated in Figure 2. We employ encoder U for the FAS task and encoder V for the FR task, respectively. Both encoders employ ResNet18 as their backbone and are equipped with task-oriented style cross modules at different levels. First, the disentanglement of liveness and identity is introduced in Sec 3.2. Next, we present the part on weakening style sensitivity in Sec 3.3. Finally, the Asymmetry Augment Instance Contrast is detailed in Sec 3.4.

3.1 Problem Definition and Notations

We have source (training) and target (testing) samples, denoted as $S = \{(x_s^i, y_{fas}^i, y_{id}^i)\}_{n_s, n_{id}}^i$, $T = \{(x_t^i, y_{fas}^i)\}_{n_t}^i$, where $x_s^i, x_t^i \in \mathbb{R}^{H \times W \times 3}$. y_{fas}^i is a one-hot label corresponding to liveness. y_{id}^i is a one-hot label corresponding to identity. n_{id} denotes the number of source subject IDs, and n_s, n_t denotes the number of source, target samples. The goal of encoder U_θ and classifier C_θ is to categorize between live and spoof. The goal of encoder V_θ and discriminator D_θ is to assist U_θ and C_θ for disentangling. For simplicity, in cases where there is no ambiguity, we generally omit the subscript θ of U_θ, V_θ, C_θ, and D_θ.

3.2 Orthogonalization Identity Disentanglement

In our scheme of learning task-relevant representations, U generates the liveness feature $f_u = U(x) \in \mathbb{R}^N$ in space $\mathcal{U}$ for live/spoof classification, while V generates the identity feature $f_v = V(x) \in \mathbb{R}^N$ in space $\mathcal{V}$. Here N represents the dimension of features. If we don't impose any constraints, space $\mathcal{U}$ and space $\mathcal{V}$ may be intertwined, and the f_u and f_v are not orthogonal. However, In this work, we assume that the liveness is irrelevant to the identity, they are treated as dissimilar and expressed as orthogonal through cosine similarity. Thus that we aim to orthogonalize all f_u and f_v by minimizing the square
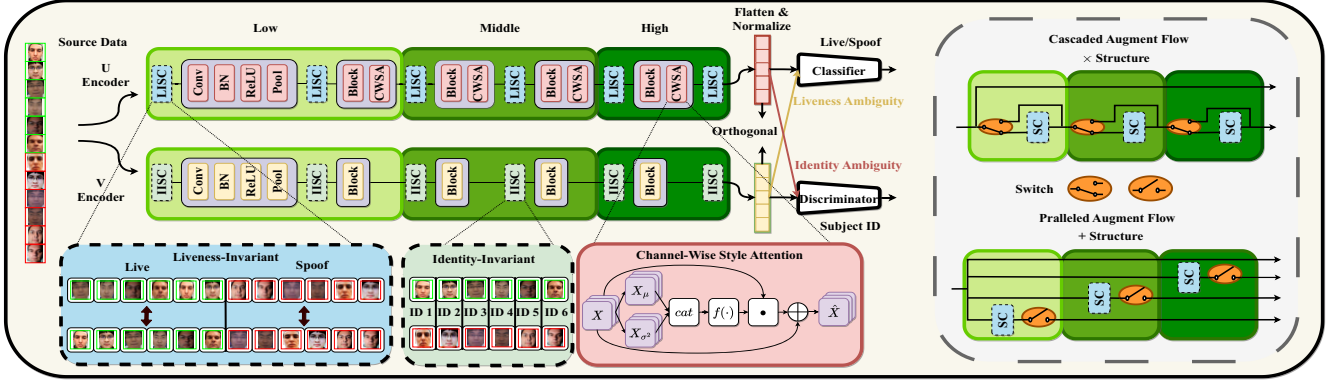
Figure 2. **(Left)** The architecture mainly consists of two encoders: encoder U and V. U extracts the liveness feature, and V extracts the identity feature. The SC implements two types of mode in U and V which are liveness-invariant and identity-invariant, the dashed line indicates detachable, and we use colors from light to dark to represent the low, middle, and high levels of the encoder. The CWSA is utilized to weaken the sensitivity of the model for style variation. In addition, **(Right)** shows the style augmented flow of ($\times$) and ($+$) structures.

of cosine similarity between them. To construct $\mathcal{U} \perp \mathcal{V}$, a similarity matrix M_{uv} for each batch is constructed as:

$$M_{uv} = F_u F_v^T, \tag{1}$$

Here, $F_u = \mathrm{Norm}(f_u^1, f_u^2, \cdots, f_u^B)^T \in \mathbb{R}^{B \times N}$ and $F_v = \mathrm{Norm}(f_v^1, f_v^2, \cdots, f_v^B)^T \in \mathbb{R}^{B \times N}$ where B denotes the batch size, Norm denotes the normalization. The shape of M_{uv} is $B \times B$ and $M_{uv}(i, j)$ represents the cosine similarity between f_u^i and f_v^j. Accordingly, the orthogonal loss in each batch is defined as follows:

$$L_{ortho} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|M_{uv}(i, j)\|_2^2, \tag{2}$$

Moreover, it is crucial for classifier C to exhibit liveness ambiguity (li-amb) towards f_v, while discriminator D should demonstrate identity ambiguity (id-amb) towards f_u. Consequently, we introduce two losses, L_{liamb} and L_{idamb}, to account for these characteristics:

$$L_{idamb} = \frac{1}{B} \sum_{i}^{B} \left\| D(\mathrm{Norm}(U(x_s^i))) - \frac{1}{n_{id}} \right\|_2^2, \tag{3}$$

$$L_{liamb} = \frac{1}{B} \sum_{i}^{B} \left\| C(\mathrm{Norm}(V(x_s^i))) - \frac{1}{2} \right\|_2^2, \tag{4}$$

The above two loss functions ensure that the normal vector of hyperplane used for determining liveness property in space $\mathcal{U}$ is orthogonal to space $\mathcal{V}$. Likewise, the normal vectors of multiple classification boundaries in space $\mathcal{V}$, are orthogonal to space $\mathcal{U}$, as depicted in Figure 1. For some succinct and intuitive theoretical interpretations, please refer to the *Supplementary Materials* [1].

3.3 Weakening Style Sensitivity

Aiming to expand the diversity of style, we propose the task-oriented Style Cross (SC). In the context of the FAS task, the application of SC between live and spoof samples may introduce uncertainty regarding liveness. To address this concern, we limit the implementation of SC between samples with identical liveness labels, termed as Liveness-invariant Style Cross (LISC) in Eqn 5, 6. Similarly, for FR task, to alleviate potential ambiguities in identity, we adopt Identity-invariant Style Cross (IISC) between samples that share the same identity, in

Eqn 7. Moreover, unlike shuffle style assembly [33] which involves parameter layers, we chose a lightweight design, directly exchanging styles channel by channel between samples. Style Cross is enabled during training and disabled during testing.

$$\mathrm{LISC}(f_{u(l)}^a, f_{u(l)}^b, \mathrm{Livness=Live}) = \sigma_{u(l)}^b \frac{f_{u(l)}^a - \mu_{u(l)}^a}{\sigma_{u(l)}^a} + \mu_{u(l)}^b \tag{5}$$

$$\mathrm{LISC}(f_{u(s)}^a, f_{u(s)}^b, \mathrm{Liveness=Spoof}) = \sigma_{u(s)}^b \frac{f_{u(s)}^a - \mu_{u(s)}^a}{\sigma_{u(s)}^a} + \mu_{u(s)}^b \tag{6}$$

$$\mathrm{IISC}(f_{v(i)}^a, f_{v(i)}^b, \mathrm{ID}=i) = \sigma_{v(i)}^b \frac{f_{v(i)}^a - \mu_{v(i)}^a}{\sigma_{v(i)}^a} + \mu_{v(i)}^b \tag{7}$$

where $f_{u(l)}^a$ and $f_{u(l)}^b$ represent two different live samples, however, $f_{u(s)}^a$ and $f_{u(s)}^b$ represent two different spoof samples. $f_{u(i)}^a$ and $f_{v(i)}^b$ represent two samples with the same identity. μ and σ represent channel-wise mean and standard deviation respectively.

Additionally, we conduct an extensive investigation to determine the optimal level of models for performing SC, as well as the most effective augmentation flow. The right of Figure 2 illustrates the various augment flows that we classify them into: Low (L), Middle (M), High (H), L $\times$ M, L $\times$ H, M $\times$ H, L $\times$ M $\times$ H, L $+$ M, L $+$ H, M $+$ H, L $+$ M $+$ H. Here the symbol $\times$ denotes a flow that undergoes multiple levels SC (Cascaded), while the symbol $+$ represents the augmented features obtained from different levels considered as individual output flow (Pralleled). Our definitions of Low, Middle, and High levels can be found on the left of Figure 2. For a discussion on this part of the motivation, please refer to the *Supplementary Materials* [1].

In order to further reduce the sensitivity to style shift, we introduce the Channel-Wise Style Attention, which is a SE-like [10] module. First, calculating the mean μ and variance σ^2 of each channel within the feature map. These channel-wise styles are aggregated through a nonlinear operation, and finally, the feature response of each channel is obtained through a sigmoid activation:

$$a = \mathrm{Sigmoid}(W_2(\mathrm{ReLU}(W_1(\mathrm{cat}(\mu_X, \sigma_X^2))))), \tag{8}$$

$$\hat{X} = a \cdot X + X, \tag{9}$$

Adaptive scaling those channels with domain-specific style variation, thereby maximizing the extraction of domain-independent valid
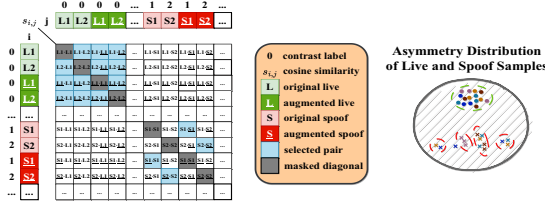
Figure 3. AAIC results in a compact cluster of live samples, scatter pattern of spoofs.

information and mitigating the adverse impact of style shift.

3.4 Asymmetric Augment Instance Contrast

When considering the FAS problem from the perspective of structural materials, it is observed that all live samples exhibit highly similar surface materials and surface reflectance properties. In contrast, spoofs exhibit greater diversity. Homogenize all spoof samples (Binary Contrast) or treating spoofs with dataset-gap-aware (Asymmetric Triplet) is obviously not optimal, but if each type of attack is labeled to perform a fine attack discrimination task, that will increase labor costs and is not beneficial for the generalization to unseen attacks. To address this issue, we aim to exclusively maximize the similarity between the original version spoofs and its Style Cross version spoofs within a batch, due to their liveness representations should be more consistent compared to other samples in the same batch (instance-pair-aware). Concurrently, we bring all original and augmented live instances closer together. This dual focus not only accommodates the asymmetry but also mitigates the model's sensitivity to the stylistic variations, as shown in Figure 3. We term our contrast strategy as asymmetric augmented instances contrast (AAIC). The AAIC is defined in Equation 10:

$$L_{aaic} = -\frac{1}{B}\sum_{i}^{B}\sum_{\substack{i \neq j \\ y_c^i = y_c^j}}^{B} \frac{\exp(s_{i,j}/\tau)}{\sum_{i \neq k}^{B}\exp(s_{i,k}/\tau)} \quad (10)$$

where τ is a hyper parameter, y_c represents the contrast label. Elements on the diagonal are masked. For the FR task, we assign the same contrast labels to both the original and augmented samples from identical identities.

3.5 Overall Objective Function

For the FAS task, we use the asymmetric am-softmax loss (denoted as aams) in [29] as follows:

$$L_{cls} = \frac{1}{B}\sum_{i}^{B} L_{aams}(C(\text{Norm}(f_u^i)), y_{fas}^i), \quad (11)$$

Then the total loss function of the FAS task is:

$$L_{FAS} = L_{cls} + \lambda_{aaicU}L_{aaicU} + \lambda_{idamb}L_{idamb} + \lambda_{orthoU}L_{orthoU}, \quad (12)$$

For the FR task, we use the cross-entropy loss:

$$L_{id} = \frac{1}{B}\sum_{i}^{B} \text{CrossEntropy}(D(\text{Norm}(f_v^i)), y_{id}^i), \quad (13)$$

the total loss function of the FR task is:

$$L_{FR} = L_{id} + \lambda_{aaicV}L_{aaicV} + \lambda_{liamb}L_{liamb} + \lambda_{orthoV}L_{orthoV}, \quad (14)$$

The optimization process is presented in Algorithm 1.

Algorithm 1 The optimization strategy of **DLIF** network

Input: Source Data $S = \{x^j, y_{fas}^j, y_{id}^j\}_{n_s, n_{id}}^j$, Target Data $T = \{x^j, y_{fas}^j\}_{n_t}^j$, $U_\theta(\cdot), C_\theta(\cdot), V_\theta(\cdot), D_\theta(\cdot)$
Output: $U_{\theta*}(\cdot), C_{\theta*}(\cdot), V_{\theta*}, D_{\theta*}(\cdot)$
1: **while** not end of iteration **do**
2:   Sampling a mini-batch B samples with M identities: $X_s = \{x^i, y_{fas}^i, y_{id}^i\}_{B,M}^i$
3:   For FAS task:$f_u^i = U_\theta(x^i), \hat{y}_{fas}^i = C_\theta(\text{Norm}(f_u^i)),$ $f_{uAug}^i = U_\theta(LISC(x^i)), \hat{y}_{liamb}^i = D_\theta(\text{Norm}(f_u^i))$
4:   For FR task:$f_v^i = V_\theta(x^i), \hat{y}_{id}^i = D_\theta(\text{Norm}(f_v^i)), f_{vAug}^i = V_\theta(IISC(x^i)), \hat{y}_{idamb}^i = C_\theta(\text{Norm}(f_v^i))$
5:   Compute $L_{cls}, L_{orthoU}, L_{aaicU}, L_{idamb}$, Update θ of U and C with Loss (12)
6:   Compute $L_{id}, L_{orthoV}, L_{aaicV}, L_{liamb}$, Update θ of V and D with Loss (14)
7:   Evaluate $U_\theta(\cdot), C_\theta(\cdot)$ on $T = \{x^j, y_{fas}^j\}_{n_t}^j$
8:   **if** performance better **then**
9:     Update $\theta*$ of $U_{\theta*}(\cdot), C_{\theta*}(\cdot), V_{\theta*}(\cdot), D_{\theta*}(\cdot)$ with θ
10:   **end if**
11: **end while**
Return $U_{\theta*}(\cdot), C_{\theta*}(\cdot), V_{\theta*}(\cdot), D_{\theta*}(\cdot)$

---

4 Experiment

4.1 Datasets and Protocol

We use four public datasets: Oulu-NPU (denoted as O) [4], CASIA-FASD (denoted as C) [43], Idiap Replay-Attack(denoted as I) [6], MSU-MFSD (denoted as M) [34], and follow the cross-datasets protocol same as previous DG-based methods [12, 5, 33, 29, 44, 26, 20] to evaluate the effectiveness of our method. The evaluation metrics are Half Total Error Rate (HTER) and Area Under the Curve (AUC).

4.2 Implementation Details

We employ ResNet-18 [9] as our backbone. Utilizing MTCNN [41] for face detection, followed by cropping and resizing the facial area to a size of 256×256. We apply random resized cropping and rotation for data augmentation. In training, each batch involves the random selection of four different IDs from each dataset and randomly choosing four live faces and four spoof faces from each ID. We use Adam optimizer [13], and the initial learning rate is set to 0.0005, which is decayed by 2 after every 50 epochs, the total training epoch is 200. We set the weight-decay as 5e-4, τ=0.07, and set λ_{aaicU}=λ_{idamb}=λ_{orthoU}=λ_{aaicV}=λ_{liamb}=λ_{orthoV}=1, U, C, V, D share the same hyper-parameter, optimizer, learning rate decay scheduler. Our method is implemented under the Pytorch framework.

4.3 Comparison with state-of-art methods

In accordance with the commonly applied protocols used in DG-FAS methods, we perform the Leave-One-Out (LOO) and limited source domains evaluation protocols.

4.3.1 Leave-One-Out (LOO).

In the LOO setting, we employ O, C, M, and I datasets, randomly selecting three of them as source domains for training, while the remaining one is held as the unseen target domain for testing. As shown in Table 1, our method demonstrates superior performance compared to most of the other methods that use datasets as domain concepts.

Table 1. Comparison to state-of-the-art FAS methods under the LOO setting. The bold numbers indicate the SOTA, while the underline indicates close to SOTA, second only to SOTA. HTER ↓ indicates smaller values are better, and AUC ↑ indicates larger values are better.

Method	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 ↑ (%)
MMD-AAE [14]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [24]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
NAS-FAS [38]	19.53	88.63	16.54	90.18	14.51	93.84	13.80	93.43
RFM [25]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SSDG [12]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
D^{2}AM [5]	12.70	95.66	20.98	85.58	15.43	91.22	15.27	90.87
DRDG [17]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75
ANRL [16]	10.83	96.75	17.83	89.26	16.03	91.04	15.67	91.90
SSAN [33]	6.67	**98.75**	10.00	96.67	8.88	96.79	13.72	93.63
PatchNet [29]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07
IADG [44]	5.41	98.19	8.70	96.44	10.62	94.50	**8.86**	**97.14**
SA-FAS [26]	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23
UDG-FAS [20]	5.95	98.47	9.82	96.76	<u>5.86</u>	**98.62**	10.97	95.36
DLIF (Ours)	**3.75**	98.33	**6.67**	**97.27**	5.82	<u>98.13</u>	<u>8.89</u>	<u>96.36</u>

Table 2. Comparison under limited source domains setting

Method	M&I to C		M&I to O	
	HTER ↓ (%)	AUC ↑ (%)	HTER ↓ (%)	AUC ↑ (%)
MADDG [24]	41.02	64.33	39.35	65.10
SSDG-M [12]	31.89	71.29	36.01	66.88
D^{2}AM [5]	32.65	72.04	27.70	75.36
DRDG [17]	31.28	71.50	33.35	69.14
ANRL [16]	31.06	72.12	30.73	74.10
SSAN-M [33]	30.00	76.20	29.44	76.62
IADG [44]	24.07	85.13	18.47	**90.49**
DLIF (Ours)	**11.11**	**92.77**	**16.97**	89.64

Table 3. Ablation of each component on O&C&I to M

Baseline	V	CWSA	SC (M+H)	HTER ↓ (%)	AUC ↑ (%)
✓	-	-	-	12.08	94.21
✓	-	-	✓(LISC)	9.17	97.31
✓	-	✓	-	10.00	96.14
✓	✓	-	-	7.91	95.79
✓	✓	-	✓(LI-IISC)	8.66	96.92
✓	✓	✓	-	7.92	97.08
✓	-	✓	✓(LISC)	6.70	97.83
✓	✓	✓	✓(LI-IISC)	**3.75**	**98.33**

These results demonstrate the generalization ability of our method since we employ identity partition, which is more fine-grained and shrinks the scope of the domain. It is conducive to learning domain-invariant representations, as using the dataset as the domain partition will bring more significant intra-domain variations. Compared to SSAN, we propose a task-oriented style cross that ensures the rationality of style augmentation, and also attempt various style augmentation flows. In addition, our proposed AAIC loss is more effective than previous binary contrastive loss and triplet loss.

4.3.2 Limited source domains.

As shown in Table 2, we evaluate our method under the limited source domains. Following prior research, we select M and I as source domains, while C and O, are respectively utilized as the unseen target domain. Our method achieves a significant improvement compared to previous state-of-the-art methods. It is proved that eliminating the effect of identity shift and employing a finer domain partition is extremely effective for the unseen target domain in the case of limited source data and source identities.

4.4 Ablation and Discussion

In this subsection, we conduct ablation studies on individual contribution of components. Additionally, we compare various contrast losses and style augmentation strategies. All ablation and comparison studies are performed under the O&C&I to M setting.

4.4.1 Contribution of each component.

Table 3 shows the ablation studies. For the baseline configuration, we employ ResNet-18 as the backbone. Only aam-softmax loss and binary contrast loss are utilized, without any additional components.

Given that there are three components in total, a total of seven combinations are formed. We conduct experiments on all combinations, and M+H flow in SC is consistently used. In the case of introducing SC, the contrast loss for the FAS task is replaced with AAIC. Furthermore, the introduction of V involves employing the orthogonal loss for both U and V. The results show that incorporating each component individually leads to performance improvements compared to the baseline. Moreover, when including two components simultaneously, further enhancements are observed. Notably, our model achieves state-of-the-art when all three components are integrated concurrently, thus confirming the effectiveness of each component.

4.4.2 Comparisons of different contrast strategies.

The performance of various contrast losses is presented on the upper side of Table 6. In comparison to Binary and Triplet, AAIC achieves superior generalization ability, because the Binary disregards the inherent differences in density and diversity between live and spoof samples. Although Triplet notices this distinction which is better than Binary, the domain-gap-aware is not fine-grained enough. In contrast, AAIC focuses on asymmetry, ie., considering that live instances exhibit homogeneity while spoof instances display heterogeneity, besides, adopting instance-pair-aware is more refined than domain-gap-aware for spoofs.

4.4.3 Comparisons of different style augmentation strategies.

Table 7 shows the impact of implementing SC at various levels (L, M, H) and employing different augmentation flows ($\times$, $+$). Several key findings can be drawn from the results: 1) Solely Performing SC at a low level could lead to a slight performance decrease. 2) Irrespective of whether it is $\times$ or $+$, the multi-level SC is more effective

Table 4. Scalability 1) : Effectiveness of well-trained FR networks. * means that the encoder V with pre-trained weights. **Bold** represents the best result. HTER ↓ indicates smaller values are better, and AUC ↑ indicates larger values are better. Red font represents ascent and Blue font indicates descent.

Backbone	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 ↑ (%)
ResNet18	3.75	98.33	6.67	97.27	5.82	98.13	8.89	96.36
FaceNet	5.47	97.22	8.70	95.14	10.17	94.51	9.49	94.86
FaceNet*	3.13 (-2.34)	98.45 (+1.23)	8.88	96.40 (+1.26)	5.52 (-4.65)	97.50 (2.99)	9.03 (-0.46)	96.17 (+1.31)
CosFace	4.91	98.16	6.66	96.58	9.64	95.55	9.27	96.60
CosFace*	4.30 (-0.61)	98.68 (+0.52)	5.26 (-1.4)	97.04 (+0.46)	7.43 (-2.21)	95.39	8.64 (-0.63)	97.11 (+0.51)

Table 5. Scalability 2) : Effectiveness of identity diversity. Gray fill indicates the baseline. HTER ↓ indicates smaller values are better, and AUC ↑ indicates larger values are better. Red font represents ascent and Blue font indicates descent.

Backbone	O&C&I to M		CS&C&I to M		O&CS&I to M		O&C&CS to M	
	HTER ↓ (%)	AUC ↑ (%)	HTER ↓ (%)	AUC ↑ (%)	HTER ↓ (%)	AUC ↑ (%)	HTER ↓ (%)	AUC ↑ (%)
DLIF (FaceNet*)	3.13	98.45	2.42 (-0.71)	99.44 (+0.99)	1.79 (-1.34)	99.67 (+1.22)	1.84 (-1.29)	99.69 (+1.24)
DLIF (CosFace*)	4.30	98.68	1.92 (-2.38)	99.68 (+1.00)	2.03 (-2.27)	99.52 (+0.84)	1.64 (-2.66)	99.81 (+1.13)

Table 6. Comparison of contrast and augment on O&C&I to M

Method	Option	HTER ↓ (%)	AUC ↑ (%)
Contrast Strategy	Binary	9.70	97.31
	Triplet	5.42	96.85
	Ours(AAIC)	**3.75**	**98.33**
Augment Strategy	SSA	10.00	95.54
	Only LISC	7.41	95.60
	LISC & IISC	**3.75**	**98.33**

Figure 4. Feature distribution of different contrast strategies via t-SNE visualization.

than the single-level, and the combination M+H yields the best results. Additionally, the lower side of Table 6 reveals that LISC outperforms SSA, indicating that LISC is more suitable for the FAS task than SSA. Moreover, implementing the IISC in V is also indispensable. The unbalanced expansion of the $\mathcal{U}$ space alone may cause the $\mathcal{V}$ space to be squeezed, which may lead to the leakage of certain identity information.

4.4.4 Scalability Study.

Benefiting from the success of face recognition, numerous well-trained open-source FR models are available, then V no longer needs to be trained from scratch. In this regard, we delve into how our model can benefit from a well-trained FR network, given that we requires identity features in essence, not identity labels. We outline the scalability as follows: **Scalability 1):** Leveraging well-trained FR networks to provide disentanglement guidance for FAS networks. We replace the ResNet18 with FaceNet and CosFace, respectively. In Table 4, we employ them as the backbone of encoders U and V. The results demonstrates that if well-trained weights are loaded, most protocols have improved under the LOO setting. This suggests that our framework is compatible with most FR models, highlighting the advantages of utilizing well-trained FR models to extract identity representations. In addition, concerning training efficiency, utilizing ResNet18 as the backbone for two encoders, requiring only 2-3 hours to complete 200 epochs on two 1080Ti GPUs. Furthermore, freezing a well-trained encoder V during training can significantly improve training efficiency. **Scalability 2):** Increasing the diversity of identities in the source domain will significantly improve the performance. We conduct experiments to demonstrate that increasing the identity diversity in training phase will improve the generalization capability. Given the scarcity of identities in OCIM, where O has 40 IDs, C has 50 IDs, I has 35 IDs, and M has 55 IDs. However, CelebA Spoof (denoted as CS) is an open-source FAS dataset with identity diversity, consisting of 27260 IDs. The IDs in CS surpass the total number of OCIM identities by more than 150 times. The specific experimental
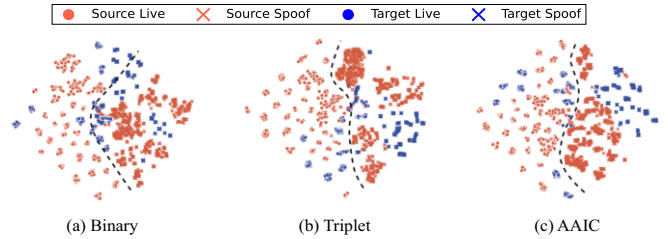
setup is as follows: we sequentially replace the three source datasets in the O&C&I to M protocol with CS, designating M as the target domain dataset (excluding M because it contains the most IDs in OCIM). The results presented in Table 5 highlight that increasing identity diversity significantly improves model performance.

5 Visualization and Analysis

### 5.1 Feature distributions of different contrast losses.

The distribution of features optimized by different contrast loss is visualized by t-SNE [28], as shown in Figure 4. In (a), the source spoofs are compact, while the target spoofs appear relatively scattered. Regarding target samples, the distance between live and spoof samples is closer. In (b), source spoofs cluster into three cliques, but the target spoofs do not belong to any of them, and the distance between the live and spoof samples of the target is not far enough. For (c), the spoof distribution is loose, and the distance between the live and spoof samples of the target is farther than the (a) and (b). The reason is that AAIC is equivalent to adopting a more refined clustering for spoofs, where each spoof instance pair has a unique contrast label in a batch, resulting in a more dispersed distribution of spoofs, while the instance-pair-aware can also weaken the style sensitivity.

5.2 Orthogonality derives from orthogonal constrain.

To prove that introducing V, orthogonal, ambiguous loss can help U learn identity-invariant representations, we use U_o, C_o, V_o which are optimized simultaneously with orthogonal and ambiguous loss, another U_{no}, C_{no} having the same setting as U_o, C_o except for the orthogonal and ambiguous constraint. As shown in Figure 6 (a), we randomly select 12 IDs from source data and compute the absolute value of cosine similarity of f_{vo} and f_{uo} or f_{uno}. The points of each column represent the similarity between the f_{vo} of the corresponding abscissa's ID and the f_{uo} (pink points) or f_{uno} (purple points) of all IDs. Due to orthogonal loss and FR task-relevant

Table 7. Ablation of different SC flows on O&C&I to M. The bold indicates the best and the underlined means the worst

Level	L	M	H	L × M	L × H	M × H	L × M × H	L + M	L + H	M + H	L + M + H
HTER ↓ (%)	12.92	10.00	7.92	5.00	6.80	**5.00**	7.08	10.21	7.50	**3.75**	7.50
AUC ↑ (%)	93.97	94.03	96.13	97.92	97.82	**98.00**	98.19	97.02	97.51	**98.33**	97.12
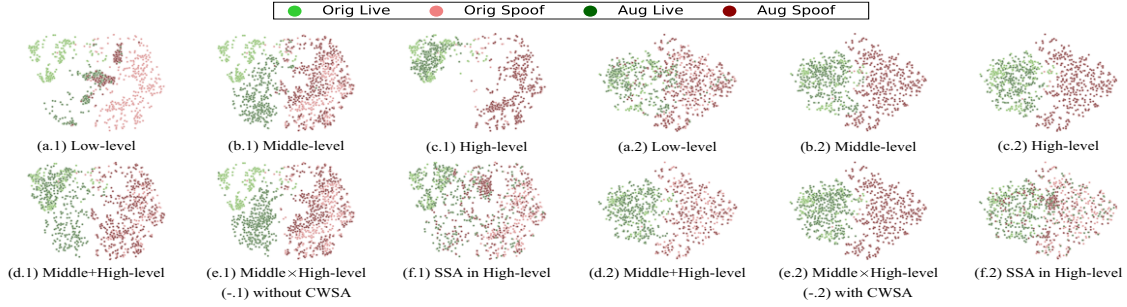
Figure 5. (a.-) (b.-), (c.-), (d.-), (e.-), and (f.-) correspond to the feature distribution of L, M, H, M+H, M ×H, SSA these six style augmentation methods respectively. The (-.1) and (-.2) indicate whether U is equipped with the CWSA.
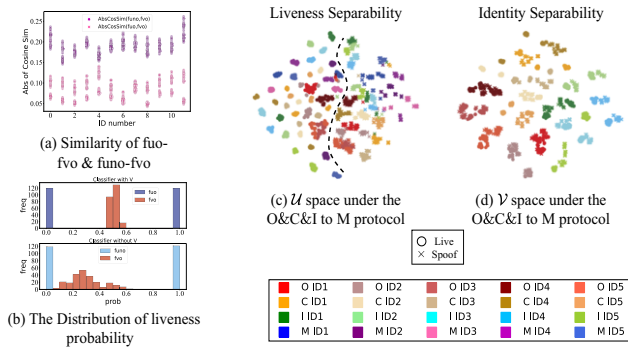
Figure 6. The orthogonality of space $\mathcal{U}$ & $\mathcal{V}$.

loss, V_o can get identity attributes and filter out liveness attributes as much as possible, thus the pink points are smaller around 0.075, but purple points are larger around 0.2. The upper subgraph of Figure 6 (b) shows the $C_o(\text{Norm}(f_{vo}))$ (liveness probability) concentrated around 0.5, and $C_o(\text{Norm}(f_{uo}))$ locates in 0 or 1, however, the lower subgraph of Figure 6 (b) shows the $C_{no}(\text{Norm}(f_{vo}))$ is more scattered, and $C_{no}(\text{Norm}(f_{uno}))$ locates in 0 or 1. This phenomenon proves that the classifier without introducing orthogonal and ambiguous loss does not exhibit liveness ambiguity to f_{vo}. As shown in Figure 6 (c), (d), colorful points represent different identities from various datasets, we can observe that $\mathcal{U}$ space has liveness separability and $\mathcal{V}$ space has identity separability when introducing the orthogonality and ambiguity.

5.3 The effectiveness of different SC flows and CWSA.

We visualize the all augmented features obtained from various style augmentation flows alongside the original features in the same coordinate system by t-SNE. As shown in Figure 5 (-.1), where are not equipped with CWSA, the following three results can be observed: 1) multiple levels style cross yields a more style diverse feature, as in (d.1) (e.1) versus (a.1) (b.1) (c.1). 2) In (a.1) we find that the augmented features approach or even cross the classification hyperplane, such that L-level SC degrades the performance. Thus for + and × Aug Flow, augmentation without L is preferable indicating that style cross is recommended at middle and high levels. 3) As shown in (f.1),
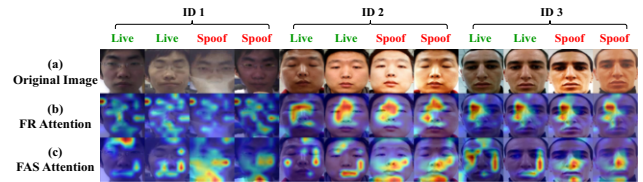
Figure 7. Grad-CAM visualizations of activation areas.

we can observe the liveness variation, which can explain why LISC is more suitable for the FAS task than SSA. Moreover, in Figure 5 (-.2), after equipping with CWSA, similar feature distributions were ultimately obtained for different style augmentation flows. This indicates that CWSA has the robustness to diverse style shifts.

5.4 Attention visualization of GradCAM.

We use Grad-CAM [23] to visualize the activation map of the last layer of U and V. In Figure 7, for faces with the same identity, the regions of **FR** attention are similar, while the difference in **FAS** attention regions is reflected in liveness, and it can be observed that **FAS** attention exhibits overlaps in certain extent for live samples and highly overlapping for spoofs with similar attack patterns.

6 Conclusion

In this work, we propose a novel perspective to learn identity-invariant spoof representations, by simultaneously training two networks for the FAS task and the FR task and constraining them through orthogonal loss to disentangle the liveness and identity. We also utilize the task-oriented style augmentation, as well as the CWSA to weaken the style sensitivity and design the AAIC. Through extensive experiments, we demonstrate the effectiveness of our method which achieves SOTA performance on prevalent benchmarks. Especially in the case of limited source data, the advantage is obvious. Furthermore, our method has strong scalability.

Acknowledgments

References

[1] Supplementary material of generalized face anti-spoofing via finer domain partition and disentangling liveness-irrelevant factors. https://github.com/yjyddq/Supp-of-ECAI-2024.

[2] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017.

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

[4] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017.

[5] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. 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.

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[7] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Lbp-top based countermeasure against face spoofing attacks. In *Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I 11*, pages 121–132. Springer, 2013.

[8] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451–460, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[12] Y. Jia, J. Zhang, S. Shan, and X. 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.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[15] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[16] S. Liu, K.-Y. Zhang, T. Yao, M. Bi, S. Ding, J. Li, F. Huang, and L. Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021.

[17] S. Liu, K.-Y. Zhang, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, Y. Xie, and L. Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021.

[18] Y. Liu, A. Jourabloo, and X. 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.

[19] Y. Liu, J. Stehouwer, and X. Liu. On disentangling spoof trace for generic face anti-spoofing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 406–422. Springer, 2020.

[20] Y. Liu, Y. Chen, M. Gou, C.-T. Huang, Y. Wang, W. Dai, and H. Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20654–20664, 2023.

[21] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016.

[22] B. Peixoto, C. Michelassi, and A. Rocha. Face liveness detection under bad illumination conditions. In *2011 18th IEEE International Conference on Image Processing*, pages 3557–3560. IEEE, 2011.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[24] R. Shao, X. Lan, J. Li, and P. 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.

[25] R. Shao, X. Lan, and P. C. Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11974–11981, 2020.

[26] Y. Sun, Y. Liu, X. Liu, Y. Li, and W.-S. Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24563–24574, 2023.

[27] Z. Tang, Y. Gao, Y. Zhu, Z. Zhang, M. Li, and D. N. Metaxas. Crossnorm and selfnorm for generalization under distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 52–61, 2021.

[28] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[29] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. 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.

[30] G. Wang, H. Han, S. Shan, and X. 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.

[31] G. Wang, H. Han, S. Shan, and X. Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2020.

[32] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2746–2754, 2021.

[33] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. 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.

[34] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[35] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013.

[36] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014.

[37] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 557–575. Springer, 2020.

[38] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. 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.

[39] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. 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.

[40] H. Yue, K. Wang, G. Zhang, H. Feng, J. Han, E. Ding, and J. Wang. Cyclically disentangled feature translation for face anti-spoofing. *arXiv preprint arXiv:2212.03651*, 2022.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[42] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma. Face anti-spoofing via disentangled representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 641–657. Springer, 2020.

[43] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012.

[44] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, R. Yi, S. Ding, and L. Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023.