Supplementary Material
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

A. Proof of the Proposition

A.1 Proof of Proposition 1

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. The mutual information $I(Z;X^+)$ is decomposed as

$$\begin{aligned} I\left(Z;X^+\right) &= \mathbb{E}_{Z,X^+}\log\frac{p\left(Z|X^+\right)}{p(Z)} \\ &= \mathbb{E}_{X^+}\mathbb{E}_{Z|X^+}\left[\log p\left(Z|X^+\right)\right] - \mathbb{E}_{Z,X^+}[\log p(Z)] \\ &= -\mathbb{E}_{X^+}\left[H\left(Z|X^+\right)\right] + H(Z) \end{aligned} \tag{S-1}$$

The first term $\mathbb{E}_{X^+}\left[H\left(Z|X^+\right)\right]$ measures the uncertainty of $Z|X^+$, which is minimized when Z can be completely determined by X^+. The second term $H(Z)$ measures the uncertainty of Z itself and it is minimized when outcomes of Z are equally likely.

Then, we firstly show Z can be completely determined by X^+ when the cosine similarity loss achieves the minimum. Based on the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\mathbb{E}_{X,X^+}[-\frac{p}{||p||_2}\cdot\frac{z^+}{||z^+||_2} - \frac{p^+}{||p^+||_2}\cdot\frac{z}{||z||_2}] \\ &= \mathbb{E}_{X,X^+}[-\tilde{p}\cdot\tilde{z}^+ - \tilde{p}^+\cdot\tilde{z}] \\ &\geq \mathbb{E}_{X,X^+}[-||\tilde{p}||_2\cdot||\tilde{z}^+||_2 - ||\tilde{p}^+||_2\cdot||\tilde{z}||_2] = -2 \end{aligned} \tag{S-2}$$

The equality holds when $\tilde{p}=\tilde{z}^+$ and $\tilde{p}^+=\tilde{z}$ for all x, x^+ from the augmentations of the same image. For any augmentations x_1^+, x_2^+ from the same image x, we have:

$$F\left(x_1^+\right) = G\left(x\right), \quad F\left(x_2^+\right) = G\left(x\right) \tag{S-3}$$

where $F = g(f(\cdot))$ and $G = q(g(f(\cdot)))$. Note that Eq. (S-3) is equivalent to *perfect alignment* [11], which is a common assumption in analyzing the behavior of contrastive learning methods. We can find $F(x_1^+) = F(x_2^+)$ for any images x_1^+, x_2^+ from the same image x. The result can be extended to the general case: $F(X_1^+) = F(X_2^+)$ for any $(X_1^+, X) \sim P(X^+, X)$, $(X_2^+, X) \sim P(X^+, X)$ with the same image X. Thus Z can be determined by X^+ with the equation $Z = F(X^+)$, which minimizes $\mathbb{E}_{X^+}\left[H\left(Z|X^+\right)\right]$.

When $p(Z = c_y|X) = \frac{1}{|\mathcal{Y}|}$, where $\mathcal{Y}$ denotes the total number of classes ($\mathcal{Y}$=2 in our FAS case), the entropy $H(Z)$ is maximized. With the asymmetric architecture and parameter updates, we assume the collapsed solutions are avoided. By this assumption, the model learns different clusters c_y for different representations with different labels. Thus, for a class-balanced dataset, the outcomes of Z are equally likely and it maximizes the second term $H(Z)$. Totally, the learned representations by minimizing the vanilla cosine similarity loss maximizes the mutual information $I\left(Z;X^+\right)$. □

A.2 Proof of Proposition 2

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. Eq. (1) can mitigate the identity-biased information within one identity. When Eq. (1) achieves the minimum, based on the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\mathbb{E}_{X,X^+}[-\frac{p^i}{||p^i||_2}\cdot\frac{z^+}{||z^+||_2} - \frac{p^{i+}}{||p^{i+}||_2}\cdot\frac{z}{||z||_2}] \\ &= \mathbb{E}_{X,X^+}[-\tilde{p}^i\cdot\tilde{z}^+ - \tilde{p}^{i+}\cdot\tilde{z}] \\ &\geq \mathbb{E}_{X,X^+}[-||\tilde{p}^i||_2\cdot||\tilde{z}^+||_2 - ||\tilde{p}^{i+}||_2\cdot||\tilde{z}||_2] = -2 \end{aligned} \tag{S-4}$$

The equality holds when $\tilde{p}^i=\tilde{z}^+$ and $\tilde{p}^{i+}=\tilde{z}$ for all $i \in \{1,..,C_m^n\}$ and x^+ is from the augmentations of the same

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

image. Thus, we have:

$$F\left(x^{+}\right)=\tilde{p}^{1}=\tilde{p}^{2}=\cdots=\tilde{p}^{C_{m}^{n}} \tag{S-5}$$

By merging the local patch embeddings by averaging, $\tilde{p}^{i}$ contains little identity-related information while retaining the live/spoof-related information, i.e., $I(P;B)\rightarrow 0$. Since $P=Z$ in Eq. (S-5), we have $I(Z;B)\rightarrow 0$.

Eq. (3) can further migigate the identity-biasd information across different identities. Note that the positive sample is the in-domain nearest neigbhors of X in Eq. (3) as $X^{+}=N_{in}(X)$. When Eq. (3) achieves the minimum, based on the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathbb{E}_{X,X^{+}}[-\frac{p^{i}}{||p^{i}||_{2}}\cdot\frac{z^{+}}{||z^{+}||_{2}}-\frac{p^{i+}}{||p^{i+}||_{2}}\cdot\frac{z}{||z||_{2}}] \\ &=\mathbb{E}_{X,X^{+}}[-\tilde{p}^{i}\cdot\tilde{z}^{+}-\tilde{p}^{i+}\cdot\tilde{z}] \\ &\geq\mathbb{E}_{X,X^{+}}[-||\tilde{p}^{i}||_{2}\cdot||\tilde{z}^{+}||_{2}-||\tilde{p}^{i+}||_{2}\cdot||\tilde{z}||_{2}]=-2\end{aligned} \tag{S-6}$$

The equality holds when $\tilde{p}^{i}=\tilde{z}^{+}$ and $\tilde{p}^{i+}=\tilde{z}$ for all $i\in\{1,..,C_{m}^{n}\}$ and x^{+} is from the same class with different identities as x. Thus, we have:

$$F\left(x^{+}\right)=\tilde{p}^{1}=\tilde{p}^{2}=\cdots=\tilde{p}^{C_{m}^{n}} \tag{S-7}$$

For one thing, by averaging the local embeddings, $\tilde{p}^{i}$ contains little identity-related information, i.e., $I(P;B)\rightarrow 0$. Since $P=Z$ in Eq. (S-7), we have $I(Z;B)\rightarrow 0$.

For another, since x and x^{+} are from the different identities, they contain different identity-related information. If $I(Z;B)>0$, we also have $I(P;B)>0$. Due to the different identity-related information, we have $P\neq Z$, which fails to match Eq. (S-7). Thus, we have $I(Z;B)\rightarrow 0$. □

A.3 Proof of Proposition 3

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. Note that the positive sample is the cross-domain nearest neigbhors of X in Eq. (4) as $X^{+}=N_{cr}(X)$. When Eq. (4) achieves the minimum, based on the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathbb{E}_{X,X^{+}}[-\frac{p^{i}}{||p^{i}||_{2}}\cdot\frac{z^{+}}{||z^{+}||_{2}}-\frac{p^{i+}}{||p^{i+}||_{2}}\cdot\frac{z}{||z||_{2}}] \\ &=\mathbb{E}_{X,X^{+}}[-\tilde{p}^{i}\cdot\tilde{z}^{+}-\tilde{p}^{i+}\cdot\tilde{z}] \\ &\geq\mathbb{E}_{X,X^{+}}[-||\tilde{p}^{i}||_{2}\cdot||\tilde{z}^{+}||_{2}-||\tilde{p}^{i+}||_{2}\cdot||\tilde{z}||_{2}]=-2\end{aligned} \tag{S-8}$$

The equality holds when $\tilde{p}^{i}=\tilde{z}^{+}$ and $\tilde{p}^{i+}=\tilde{z}$ for all $i\in\{1,..,C_{m}^{n}\}$ and x^{+} is from the same class with different domains as x. Thus, we have:

$$F\left(x^{+}\right)=\tilde{p}^{1}=\tilde{p}^{2}=\cdots=\tilde{p}^{C_{m}^{n}} \tag{S-9}$$

Since x and x^{+} are from the different domains, they contain different domain-related information. If $I(Z;D)>0$, we also have $I(P;D)>0$. Due to the different domain-related information, we have $P\neq Z$, which fails to match Eq. (S-9). Thus, we have $I(Z;D)\rightarrow 0$. □

B. Extensive Experiments

B.1 Experiments on Few Labeled Spoof Faces

Since labeled spoof faces are more hard to access in practical applications, we conduct experiments with few labeled spoof faces. Specifically, we unsupervised pretrain on all the faces of three domains among O, C, M and I. Then, we finetune the pretrained model with full labeled live data and few labeled spoof data, i.e., ranging from 5% to 50%, which is more relevant to the practical scenarios, since labeled spoof faces are more costly to obtain. As shown in Table S-1, given 50% labeled spoof data, our UDG-FAS achieves 17.88% AUC gain compared with Random Init. For 20% labeled spoof data, we outperforms ImageNet Init by 12.30% HTER reduction. Besides, with 10% labeled spoof data, our UDG-FAS reduces HTER by 7.74% and increases AUC by 6.96% in comparison to SimSiam. Moreover, with only 5% labeled spoof data (i.e., only spoof faces of 2 subjects for each domain), our UDG-FAS achieves 13.62% HTER for I&C&M to O, which is without much performance degradation compared with using full labeled spoof data. Finally, Table S-2 shows, when combined with SSDG, our method yields improved performance by finetuning with few (10% and 50%) spoof data, despite the performance gain reducing with fewer spoof data due to imbalanced classification.

B.2 Unsupervised Training with No Spoof Faces

Considering that for real-world application scenarios in FAS, spoof faces are expensive to collect and usually scarce and unavailable. Comparably, we can collect very large amounts of live faces, which are easily obtained and cheap. Thus, we study a more challenging and practical scenarios to investigate that whether unsupervised pretraining only on live faces can help to improve the performance. Specifically, we only pretrain on all the live faces of three domains among O, C, M and I. Then, we finetune the pretrained model with full labeled live and spoof data, and test on the remaining unseen target domain.

Table S-3 shows that even with only live faces for unsupervised pretraining, our UDG-FAS outperforms ImageNet Init by a considerable margin, e.g., 7.87% AUC gain on average. Thus, even using easily and cheaply collected live faces, our UDG-FAS can provide a promising initialization for FAS models. Besides, existing contrastive learning methods cannot achieve satisfactory performance with only live faces for pretraining, e.g., SimSiam suffers 0.81% AUC degradation on average compared with ImageNet Init.

Methods	Label Fraction 100% **Live** + 50% **Spoof**								Label Fraction 100% **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
Random Init	14.05	92.47	34.78	69.27	26.43	74.05	31.04	75.18	14.29	91.56	34.89	71.29	29.86	73.31	33.61	71.07
ImageNet Init	18.33	87.34	22.56	87.15	23.64	77.14	22.62	85.02	19.76	90.74	27.33	81.86	26.43	74.25	24.86	82.65
MoCo V2 [4]	14.29	92.63	18.67	86.42	20.07	88.56	28.06	79.47	15.71	92.13	26.11	81.33	25.71	76.41	29.76	76.09
SimCLR V2 [3]	13.10	90.66	18.00	88.68	19.21	91.60	27.62	79.39	14.52	91.59	25.33	82.47	24.29	78.22	29.03	78.74
BYOL [8]	14.52	88.17	22.67	84.37	16.43	90.21	23.47	85.18	15.71	89.98	22.67	84.26	20.07	80.93	24.54	84.69
Simsiam [5]	12.86	93.60	18.56	91.85	17.14	89.98	19.84	87.76	13.10	92.78	20.78	88.03	19.14	85.46	21.04	86.07
Ours	**9.76**	**96.40**	**13.89**	**94.14**	**8.71**	**96.69**	**12.18**	**95.26**	**11.43**	**94.86**	**14.67**	**93.25**	**10.71**	**95.18**	**12.37**	**94.24**
Methods	Label Fraction 100% **Live** + 10% **Spoof**								Label Fraction 100% **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
Random Init	16.91	92.20	35.33	70.17	31.50	69.15	35.50	67.71	17.14	88.59	35.89	71.63	32.86	70.58	33.75	71.04
ImageNet Init	21.43	87.22	26.67	82.47	27.79	72.87	26.46	81.03	20.95	84.93	27.33	80.90	29.28	70.12	28.63	78.86
MoCo V2 [4]	17.38	91.82	28.56	80.25	26.43	73.52	34.55	70.43	18.57	90.06	34.33	74.09	26.43	72.07	35.61	68.82
SimCLR V2 [3]	16.91	92.09	27.89	79.85	25.71	72.73	32.59	73.08	17.38	89.45	33.89	72.12	25.64	75.07	34.29	71.09
BYOL [8]	16.91	88.62	23.33	83.57	19.28	84.27	24.08	84.71	18.81	86.85	27.44	79.12	21.50	82.49	27.92	79.64
Simsiam [5]	14.05	93.32	21.44	85.19	19.44	89.78	25.69	81.83	15.95	92.40	25.33	81.75	21.14	77.58	26.28	81.59
Ours	**10.47**	**96.08**	**14.89**	**91.96**	**12.00**	**95.48**	**12.29**	**94.45**	**10.00**	**94.79**	**16.44**	**91.48**	**15.00**	**93.49**	**13.62**	**92.77**

Table S-1: Results on *UDG-Protocol-1* with full labeled live data and partial labeled spoof data ranging from 5% to 50%. We split the training set by the subject ID.

	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(L+50%S)	9.76	96.40	13.89	94.14	8.71	96.69	12.18	95.26
+SSDG	8.57	97.05	12.56	94.57	7.85	96.93	11.53	95.71
Ours(L+10%S)	10.47	96.08	14.89	91.96	12.00	95.48	12.29	94.45
+SSDG	9.76	96.44	14.00	92.51	11.43	95.73	11.78	94.49

Table S-2: Combined with SSDG on *UDG-Protocol-1* using full labeled live (L) and few-shot labeled spoof (S) data for finetuning.

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	12.62	92.15	35.33	68.25	25.64	77.09	32.20	73.07
ImageNet Init	11.43	93.99	16.44	91.25	23.57	77.25	22.31	85.65
MocoV2	15.95	91.39	20.67	85.93	18.57	85.89	28.89	78.16
SimCLRV2	15.71	90.26	20.11	86.43	18.50	91.35	26.94	80.24
BYOL	16.91	88.76	23.78	86.68	17.79	91.69	26.11	81.63
SimSiam	15.95	90.67	23.33	84.48	17.86	89.99	26.35	79.77
Ours	8.33	96.92	12.67	94.35	5.64	98.50	17.54	89.83
Ours+SSDG	**7.14**	**97.31**	**11.33**	**94.67**	**5.43**	**98.79**	**15.96**	**91.52**

Table S-3: Results on unsupervised pretraining using only live faces.

Comparably, our UDG-FAS improves SimSiam by a large margin, e.g. 9.83% HTER reduction on average. Though unsupervised pretraining without spoof faces, our UDG-FAS forces the model to learn an identity-irrelevant and domain-irrelevant representation space, facilitating the generalization capability when finetuned with full labeled data.

B.3 Experiments with Various Domain Information

Following previous work [9, 12], we take each dataset as a domain for a fair comparison. However, dataset like CASIA uses several capture devices and very different environments to collect the data, which can be further divided into fine-grained sub-domains. In specific, we further divide CASIA into three sub-domains based on capture devices. C subdivision shows a slight gain in Table S-4. Thus, domain information only serves to separate the support set of NN, and our UDG-FAS is somewhat tolerable to the noise. Moreover, our method can work flexibly even without domain information. We use pretrained ResNet to extract features for K-means clustering to obtain κ domains, and Table S-4 shows UDG-FAS works well with various κ.

	-	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	C subdivision
UDG-FAS	12.18	13.09	12.43	12.71	11.83

Table S-4: HTER (%) on I&C&M to O for various kinds of domain separation.

	5% labels	10% labels	20% labels	50% labels
SSDG-R [9]	24.98	23.61	20.43	17.54
SSAN-R [12]	23.04	21.25	17.10	15.12
UDG-FAS	**15.14**	**15.29**	**12.83**	**12.27**

Table S-5: HTER (%) on I&C&M to O with partial labeled live and spoof data, ranging from 5% to 50%.

B.4 Evaluation with Other FAS SOTA Methods

We perform the evaluation on limited labeled data with SOTA FAS methods, i.e., SSDG [9] and SSAN [12]. Table S-5 shows that our method consistently outperforms SOTA FAS methods under various portions of labeled data ranging from 5% to 50%, especially with fewer data. For example, our UDG-FAS outperforms SSDG by 5.27%

HTER drop with 50% labels, and the gap further increases to 9.84% HTER drop with 5% labels.

B.5 Discussion on Hyperparameters Setting

The choice of the split number m controls the patch size. Fig. 5 shows m=2 outperforms m=3, since a small patch size (m=3) may fragment spoofing cues and degrade the performance. Besides, n determines the number of merged patches. Directly averaging all features of patches (n=m^2) produces only one ($C_{m^2}^{m^2}$) positive pair. While averaging n local features from subsets of m^2 patches generates $C_{m^2}^{n}$ positive pairs, which significantly provides more supervision signals. Besides, averaging all local features would weaken the power of some discriminative local features. Fig. 5 shows C_4^2 is better than C_4^4. There are many false matches at random initialization. Thus, at the start of training, we do **not** use searched neighbors as positive samples to compute the loss (Eq.(5) in the main text). After training T_1=30 and T_2=60 epochs for warm-up, we employ reliable in-domain and cross-domain NN as positives, respectively.

C. More Visualization Analysis

C.1 Searched Nearest Neighbors

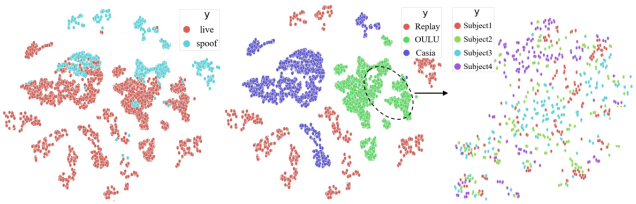
In Fig. S-2, we show the nearest neighbors searched using features encoded by our self-supervised model, which is unsupervised trained on three datasets among O, C, M and I. Cross-domain NNs picked by our method are from the same live/spoof class, where the spoof ones are even from the same fine grained attack types, e.g., video and print attacks. Besides, the searched in-domain NNs are also accurate with the same fine grained attack types, and are from quite different identities (e.g., different ethnics and genders).

C.2 Visualization of Feature Space

Fig. S-1 shows the t-sne visualization of our unsupervised learned features. As shown in Fig. S-1 (a), samples of each class (live/spoof) are separable in our learned feature space, indicating the learning of live/spoof-related features, though not perfect since our model is unsupervised trained without labels. Fig. S-1 (b) and (c) show that samples from different domains and different identities are closely entangled and inseparable in our learned feature space, indicating the learning of domain-irrelevant and identity-irrelevant representation space. Thus, our unsupervised learned features can effectively mitigate domain-biased information and identity-biased information.

D. Main Algorithms

The unsupervised training algorithm of our UDG-FAS with unlabeled data collected under various domains is elaborated in Algorithm 1. After unsupervised training, we finetune the model with few labeled live and spoof data, as described in Algorithm 2.



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

Figure S-1: T-sne visualization of unsupervised features learned by our UDG-FAS 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. Our UDG-FAS could learn live/spoof-related features to separate the feature space well. While samples from different domains and different identities are closely entangled, indicating the effectiveness of mitigating domain bias and identity bias.

Symbol	Meaning
x_i, y_i, d_i	image/category label/domain label
B, D	random variables indicating identity/domain
m, n	number of split/merged patches
e_1^p	encoded local embeddings
v_1^i	merged local embeddings
p_1^i	vectors of merged local embeddings
z	embedding for the input x
f, g, q	encoder, projector and predictor
Q_z^{in}, Q_z^{cr}	in-domain/cross domain support set for z
$N(z, Q)$	nearest neighbor of z in Q
$z_{nn}^{q_{in}}$, $z_{nn}^{q_{cr}}$	in-domain/cross domain nearest neighbor of z
$N(z, Q_z^{in})$	in-domain nearest neighbor of z
$N(z, Q_z^{cr})$	cross domain nearest neighbor of z

Table S-6: The meaning of the main symbols defined in the paper.

E. The Symbol Table

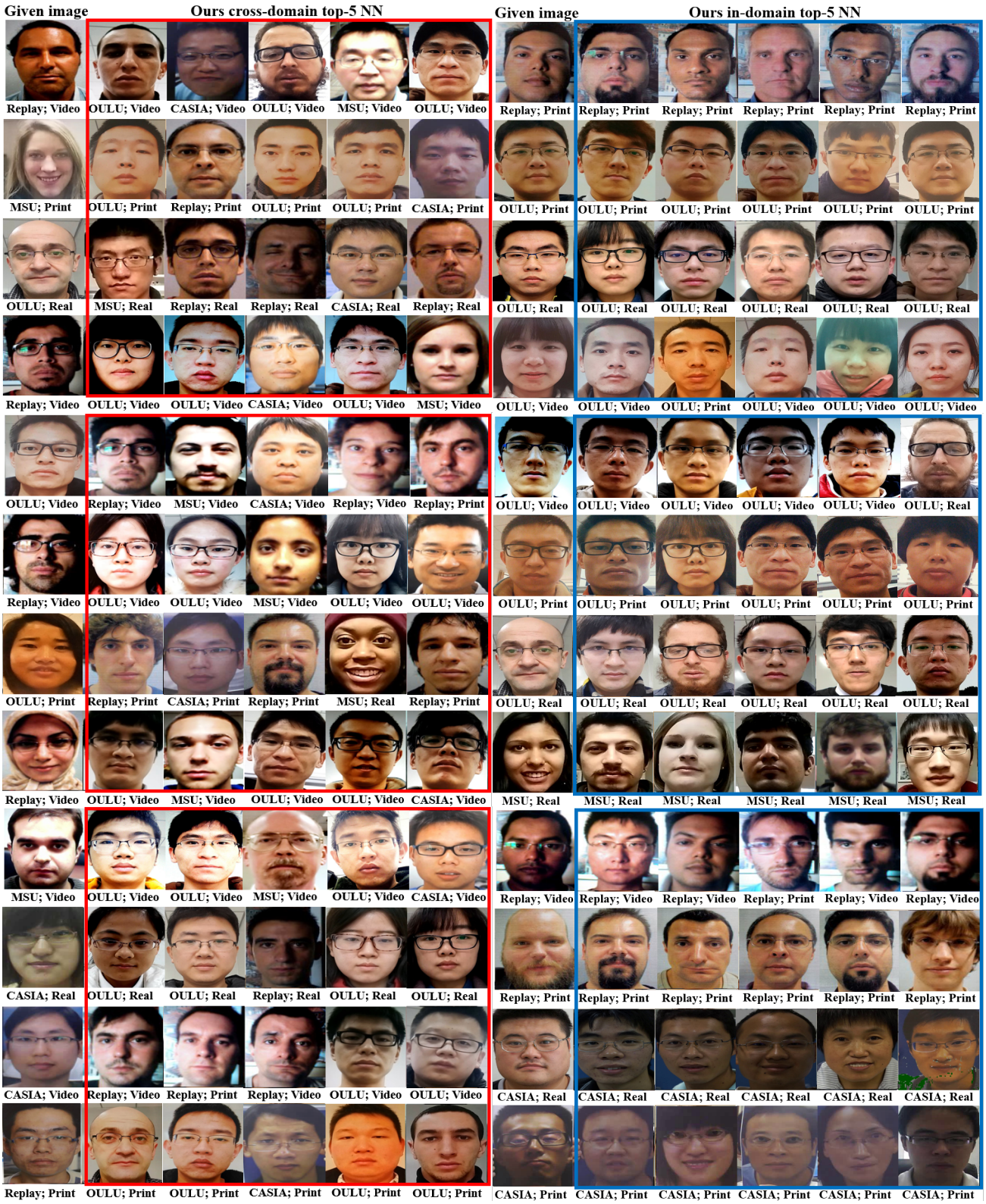
We add a symbol table as shown in Table S-6 to clarify the meaning of the math symbols used in the paper.

F. Datasets and Protocols

F.1 Datasets

Experiments are conducted on five publicly available datasets: Idiap Replay-Attack [6] (denoted as I), OULU-NPU [2] (denoted as O), CASIA-MFSD [15] (denoted as C), MSU-MFSD [13] (denoted as M) and CelebA-Spoof [14] (denoted as CA). Basic information of these datasets is summarized in Table S-7.

- **Idiap Replay-Attack** (abbr. I) captures all live and spoof faces from 50 clients under two different lighting conditions in 1,200 videos. Five attack types consist of four kinds of replayed faces and one kind of printed face.



((a)) Cross-domain Nearest Neighbors

((b)) In-domain Nearest Neighbors

Figure S-2: Cross-domain and in-domain nearest neighbors searched by our UDG-FAS method.

Datasets	Subjects	Data	Sensors	Spoof Types
Idiap Replay-Attack (I) [6]	50	1,200 videos	2	1 Print, 2 Video-replay
OULU-NPU (O) [2]	55	4,950 videos	6	2 Print, 2 Video-replay
CASIA-MFSD (C) [15]	50	600 videos	3	2 Print, 1 Video-replay
MSU-MFSD (M) [13]	35	280 videos	2	1 Print, 2 Video-replay
3DMAD (D) [7]	17	255 videos	5	1 3D Mask
HKBU-MARs (H) [10]	12	1008 videos	6	2 3D Mask
CelebA-Spoof (CA) [14]	10177	625,537 images	>10	3 Print, 3 Replay, 3 Paper Cut, 1 3D Mask

Table S-7: A summary of the FAS datasets used in our experiments.

Figure S-3: Sample frames from CASIA-MFSD [15], Idiap Replay-Attack [6], MSU-MFSD [13], and OULU-NPU [2] datasets. The figures with red border represent the real faces, while the ones with green border represent the video replay attacks. From these examples, it can be seen that large cross-dataset variations due to the differences on materials, illumination, background, resolution and so on, cause significant domain shift among these datasets.

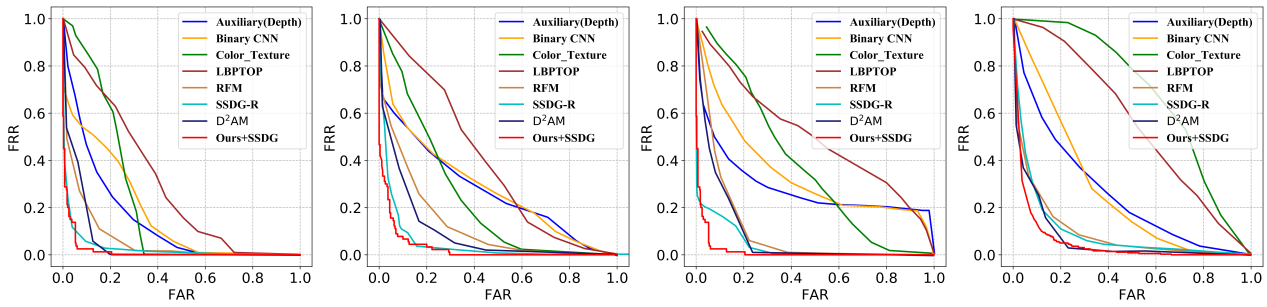
- **OULU-NPU** (abbr. O) is a high-resolution dataset with 3,960 spoof face videos and 990 live face videos, containing two kinds of printed spoof faces and two kinds of replayed spoof faces captured under six cameras and three sessions.
- **CASIA-MFSD** (abbr. C) consists of 50 subjects and each subject has 12 videos. Three attack types (printed photo attack, cut photo attack, and video attack) are used to create spoof faces, and each face image is recorded with three kinds of imaging qualities.
- **MSU-MFSD** (abbr. M) consists of totally 280 videos for 35 subjects under two different cameras. Three spoof types include two kinds of replayed faces and one kind of printed face.
- **3DMAD** (abbr. D) is collected in 3 different sessions for all subjects and for each session 5 videos of 300 frames are captured, which contain high-fidelity 3D mask attacks.
- **HKBU-MARs** (abbr. H) contains 1008 videos from 12 subjects and masks. 2 types of 3D masks are included with different illumination conditions.
- **CelebA-Spoof** (abbr. CA) is the current largest scale FAS dataset with rich and diverse annotations, which comprises 625,537 pictures of 10,177 subjects covering four spoofing types (i.e., print, paper-cut, replay, and 3D mask) captured under eight scenes.

F.2 Protocols

UDG-Protocol-1: We unsupervisedly pretrain the model using unlabeled data on three domains of I, O, C and M, and then finetune the unsupervised trained model with the labeled data, the proportion of which varies from 5% to 100%. 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 deliberately evaluate the performance of models with few labeled spoof data (from 5% to 50%), which is more relevant to the practical situation (as shown in Table S-1). The training data is split by the subject ID, and we select part of labeled data for finetuning. For example, Label fraction 5% denotes that 5% of the live and spoof data in the order of subject ID is labeled. Label fraction Spoof 5% denotes that 100% of the live data and 5% of the spoof data in the order of subject ID is labeled. To make a further analysis, we draw ROC curves of ***UDG-Protocol-1*** with full labeled live and spoof data for finetuning, as shown in Fig. S-4.

UDG-Protocol-2: The model is unsupervised pretrained using unlabeled data from three domains of I, O, C and M. Without any labeled data for supervised finetuning, we perform kNN on the model to evaluate the unsupervised pre-trained features more directly. Note that we conduct kNN (k=10) with 10% labeled data for evaluation and test. This protocol is proposed to evaluate the performance under more challenging scenarios without any labeled data for training.

UDG-Protocol-3: Besides small traditional FAS datasets (I, O, C, M), we exploit more large-scale unlabeled data for pretraining to demonstrate the advantages of our method. In



((a)) O&C&I to M ((b)) O&M&I to C ((c)) O&C&M to I ((d)) I&C&M to O

Figure S-4: ROC curves of our proposed UDG-FAS and state-of-the-art face anti-spoofing methods on four testing scenarios.

Algorithm 1 Unsupervised Domain Generalization for FAS

Input: Number of pretraining epochs N_T, encoder network f, projector g, predictor q, data augmentation $\mathcal{T}$ and loss balanced parameters λ_1 and λ_2.

Output: Encoder network f, and throw away g, q.

for *epoch* = 1 to N_T **do**
  Given a sampled minibatch x, draw two augmentation functions $t_1, t_2 \sim \mathcal{T}$, and generate two augmented views as $x_1 = t_1(x)$, $x_2 = t_2(x)$.
  Input x_1, x_2 alternatively to Split-Rotation-Merge module and obtain merged local vectors $\{p_1^i\}$,$\{p_2^i\}$.
  Input x_1, x_2 directly to the encoder f and projector g to obtain the global vectors as z_1, z_2.
  Calculate cosine similarity loss $\mathcal{L}_{\mathrm{SRM}}$ via Eq. (1).
  With torch.no grad():
    Input x to the encoder and projector to obtain z.
    Gather in-domain global vectors as Q_z^{in}.
    Gather cross-domain global vectors as Q_z^{cr}.
    Input split x to encoder and projector to build v.
    Gather cross-domain local vectors as Q_v^{cr}.
    Search in-domain nearest neighbors as $id_{nn}^{q_{in}}$.
    Normalize Q_z^{cr},Q_v^{cr} to the Gaussian distribution.
    Search cross-domain nearest neighbors as $id_{nn}^{q_{cr}}$.
  Obtain the in-domain nearest neighbors as $z_1[id_{nn}^{q_{in}}]$ and $z_2[id_{nn}^{q_{in}}]$, respectively. Obtain the cross-domain nearest neighbors as $z_1[id_{nn}^{q_{cr}}]$ and $z_2[id_{nn}^{q_{cr}}]$.
  Using in-domain nearest neighbors as positives for computing cosine similarity loss $\mathcal{L}_{\mathrm{IDNN}}$ via Eq. (3).
  Using cross-domain nearest neighbors as positives for computing the cosine similarity loss $\mathcal{L}_{\mathrm{CRNN}}$ via Eq. (4).
  Obtain the overall loss $\mathcal{L}$ via Equation (5).
  Update f, g and q via $\mathcal{L}$ by gradient descent.
end for
Return encoder network f, and throw away g, q.

Algorithm 2 Finetuning for UDG-FAS

Input: Labeled source domain dataset $\mathcal{D}_S = \{\mathbf{x}_S, \mathbf{y}_S\}$, number of training epochs N_S, the pre-trained feature encoder f and a randomly initialized one-layer linear classifier h_s.

Output: Trained model $h \circ f$.

for *epoch* = 1 to N_S **do**
  Input a sampled minibatch $\mathbf{x}_S$.
  Obtain the model output $\tilde{\mathbf{y}}_S = f_s(\mathbf{x}_S)$.
  Calculate the loss $\mathcal{L}_{ce} = \mathrm{BCE}(\tilde{\mathbf{y}}_S, \mathbf{y}_S)$.
  Update the parameters of $f_s(\cdot)$ via $\mathcal{L}_{ce}$.
end for
Return Trained model $h \circ f$.

specific, we include the current largest CelebA-Spoof (CA) dataset as an additional unlabeled source dataset with three domains of I, O, C, M for unsupervised pretraining. To save computational overhead, we randomly sample a subset of 100k/200k images. Moreover, we extract the real faces of CA as additional source data for evaluation, which are all web-crawled. After unsupervised pretraining, full labeled data of three domains are used to finetune the model for evaluation. This protocol is proposed to evaluate the effectiveness of our method for using large-scale web-crawled face data to enhance the pre-trained features and improve the low-data regime of the FAS community.

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 unsuperivsed pretrained on one group using unlabeled data, finetuned using the labeled data, and then tested on the unseen target data in the other group. This protocol evaluates the efficiency and generalizability of models with limited source domains. In other words, this protocol can validate the performance of models with limited training data.

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 live data and partial attack type data using unlabeled data, finetune with the labeled data. Subsequently, the samples with the remaining one attack type are set as the target set for test. This protocol measures the generalization of the model on both unseen domain and 2D attack types. For example, samples with 'Live' and 2 attack types ('Video' and 'Digital Photo' in Replay, 'HR Video' and 'Mobile Video' in MSU) are set as labeled training set, while samples with another one attack type ('Warped Photo') in CASIA are set as unlabeled target testing set.

UDG-Protocol-6: We evaluate the generalization on unseen 3D mask attack based on UDG in this protocol. Following the 'leave on attack type out' testing, we pretrained on unlabeled data with 2D attack types, finetune using the labeled 2D attack data, and then test on the unseen 3D mask attach types data. 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. Besides, we also pretrain on O, C and M, and test on the large-scale CA dataset, which also contain unseen 3D mask attack types.

G. Implementation Details

For unsupervised training, we adopt ResNet-18 as the backbone. Following SimSiam [5], we add a projector with three MLP layers and a predictor with two MLP layers, which are discarded after unsupervised pretraining. We adopt SDG optimizer with base $lr = 0.03$ and a cosine decay schedule for 100 epochs unsupervised pretraining. For O, C, I and M, the scale of data is too small, so we pretrained the model for 100 epochs on a subset of 50,000 images from CelebA-Spoof. For our SRM module, we set m=2 and n=2. For unsupervised pretraining, we leverage random crop, random light/contrast, random erasing and random horizontal flip as the data augmentation to augment the input image. The hyperparameter is $T_1 = 30$ and $T_2 = 60$. For finetuning, we initialize a ResNet-18 encoder with unsupervised pretrained weight, and randomly initialize a one-layer linear classifier. The model is trained by SGD optimizer with $lr = 0.001$ for 1000 iterations. Following previous works [9, 12] for a fair comparison, we select the best model based on the test set.

H. Limitations and Future Work

While our work shows promising results, there are still some limitations.

i) Considering the cross-domain NN accuracy, we manage to get 87.9% in 100 epochs. This suggests that there is still a possibility of improving performance with a better NN picking strategy, although it might be hard to design one that works in a purely unsupervised way;

ii) Domain labels are assumed to be accessible in our method to divide in-domain support set and cross-domain set for NN search. For a more practical scenario, we may obtain a mixture domain dataset, where the domain label is unknown. We can leverage more advanced unsupervised clustering methods to construct domain partition, which could be further explored in future work;

iii) Our current approach is built on convolutional neural networks for a fair comparison with existing works. A direct extension could be employing our approach on top of more powerful vision transformers.

iv) Experiments with large-scale CA data have demonstrated the potential of our method to use large-scale web-crawled face data to enhance the pre-trained features and improve the low-data regime of the FAS community. We would validate on other large-scale web-crawled face data later.

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. 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. 4, 6

[3] 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. 3

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[5] 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. 3, 8

[6] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7, 2012. 4, 6

[7] Nesli Erdogmus and Sebastien Marcel. Spoofing face recognition with 3d masks. *IEEE transactions on information forensics and security*, 9(7):1084–1097, 2014. 6

[8] 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. 3

[9] 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, 8

[10] 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. 6

[11] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on

the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1

[12] 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. 3, 8

[13] 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. 4, 6

[14] 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. 4, 6

[15] 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. 4, 6