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

\[
I(Z; X^+) = E_{Z,X^+} \log \frac{p(Z|X^+)}{p(Z)}
\]

\[
= E_{X^+} E_{Z|X^+} \left[ \log p(Z|X^+) \right] - E_{Z,X^+} \left[ \log p(Z) \right]
\]

\[
= -E_{X^+} \left[ H(Z|X^+) \right] + H(Z) \quad (S-1)
\]

The first term \( E_{X^+} \left[ H(Z|X^+) \right] \) 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

\[
E_{X,X^+} \left[ -\frac{p}{||p||_2} \cdot \frac{z^+}{||z^+||_2} - \frac{p^+}{||p^+||_2} \cdot \frac{z}{||z||_2} \right]
\]

\[
= E_{X,X^+} \left[ -\tilde{p} \cdot \tilde{z}^+ - \tilde{p}^+ \cdot \tilde{z} \right]
\]

\[
\geq E_{X,X^+} \left[ -||\tilde{p}||_2 \cdot ||\tilde{z}^+||_2 - ||\tilde{p}^+||_2 \cdot ||\tilde{z}||_2 \right] = -2 \quad (S-2)
\]

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:

\[
F(x_1^+) = G(x), \quad F(x_2^+) = G(x) \quad (S-3)
\]

**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

\[
E_{X,X^+} \left[ -\frac{p^i}{||p^i||_2} \cdot \frac{z^+}{||z^+||_2} - \frac{p^i}{||p^i||_2} \cdot \frac{z}{||z||_2} \right]
\]

\[
= E_{X,X^+} \left[ -\tilde{p}^i \cdot \tilde{z}^+ - \tilde{p}^i \cdot \tilde{z} \right]
\]

\[
\geq E_{X,X^+} \left[ -||\tilde{p}^i||_2 \cdot ||\tilde{z}^+||_2 - ||\tilde{p}^i||_2 \cdot ||\tilde{z}||_2 \right] = -2
\]

The equality holds when \( \tilde{p}^i = \tilde{z}^+ \) and \( \tilde{p}^i = \tilde{z} \) for all \( i \in \{1, \ldots, C_m\} \) and \( x^+ \) is from the augmentations of the same
image. Thus, we have:

\[ F(x^+) = \tilde{p}^1 = \tilde{p}^2 = \cdots = \tilde{p}^{C_m} \quad (S-5) \]

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

\[ E_{X,X^+}[-\frac{p_i^+}{\|p_i^+\|_2} \cdot \frac{z^+}{\|z^+\|_2} - \frac{p_i^+}{\|p_i^+\|_2} \cdot \frac{z}{\|z\|_2}] \]
\[ = E_{X,X^+}[-\frac{\tilde{p}^i}{\|\tilde{p}^i\|_2} \cdot \frac{\tilde{z}^i}{\|\tilde{z}^i\|_2} - \frac{\tilde{p}^i}{\|\tilde{p}^i\|_2} \cdot \frac{\tilde{z}}{\|\tilde{z}\|_2}] \]
\[ \geq E_{X,X^+}[-||\tilde{p}^i||_2 \cdot ||\tilde{z}^i||_2 - ||\tilde{p}^i||_2 \cdot ||\tilde{z}||_2] = -2 \]

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

\[ F(x^+) = \tilde{p}^1 = \tilde{p}^2 = \cdots = \tilde{p}^{C_m} \quad (S-7) \]

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-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 RandomInit. For 20% labeled spoof data, we outperform 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.
Table S-1: Results on UDG-Protocol-I with full labeled live data and partial labeled spoof data ranging from 5% to 50%. We split the training set by the subject ID.

<table>
<thead>
<tr>
<th>Methods</th>
<th>O&amp;C&amp;I to M</th>
<th>O&amp;M&amp;I to C</th>
<th>O&amp;C&amp;M to I</th>
<th>I&amp;C&amp;M to O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Init</td>
<td>O&amp;C&amp;I to M</td>
<td>O&amp;M&amp;I to C</td>
<td>O&amp;C&amp;M to I</td>
<td>I&amp;C&amp;M to O</td>
</tr>
<tr>
<td>HTER</td>
<td>auc</td>
<td>HTER</td>
<td>auc</td>
<td>HTER</td>
</tr>
<tr>
<td>Random Init</td>
<td>14.05</td>
<td>92.47</td>
<td>34.78</td>
<td>69.27</td>
</tr>
<tr>
<td>ImageNet Init</td>
<td>18.33</td>
<td>87.34</td>
<td>22.56</td>
<td>87.15</td>
</tr>
<tr>
<td>MoCo V2 [4]</td>
<td>14.29</td>
<td>92.63</td>
<td>18.87</td>
<td>86.42</td>
</tr>
<tr>
<td>SimCLR V2 [3]</td>
<td>13.10</td>
<td>90.66</td>
<td>18.00</td>
<td>88.68</td>
</tr>
<tr>
<td>BYOL [5]</td>
<td>14.52</td>
<td>88.17</td>
<td>22.67</td>
<td>84.37</td>
</tr>
<tr>
<td>SimSiam [5]</td>
<td>12.86</td>
<td>93.60</td>
<td>18.56</td>
<td>91.85</td>
</tr>
<tr>
<td>Ours</td>
<td>9.76</td>
<td>96.40</td>
<td>13.89</td>
<td>94.14</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>O&amp;C&amp;I to M</th>
<th>O&amp;M&amp;I to C</th>
<th>O&amp;C&amp;M to I</th>
<th>I&amp;C&amp;M to O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Init</td>
<td>O&amp;C&amp;I to M</td>
<td>O&amp;M&amp;I to C</td>
<td>O&amp;C&amp;M to I</td>
<td>I&amp;C&amp;M to O</td>
</tr>
<tr>
<td>HTER</td>
<td>auc</td>
<td>HTER</td>
<td>auc</td>
<td>HTER</td>
</tr>
<tr>
<td>Ours(L+50%S) +SSDG</td>
<td>9.76</td>
<td>96.40</td>
<td>13.89</td>
<td>94.14</td>
</tr>
<tr>
<td>Ours(L+10%S) +SSDG</td>
<td>9.76</td>
<td>96.40</td>
<td>13.89</td>
<td>94.14</td>
</tr>
</tbody>
</table>

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

Comaparably, our UDG-FAS improves SimSiam by a large margin, e.g., 9.63% 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

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

**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 ($C^2_{m^2}$) positive pair. While averaging $n$ local features from subsets of $m^2$ patches generates $C^2_n$ 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 $C^2_2$ is better than $C^2_4$. 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-relevant and identity-relevant representation space. Thus, our unsupervised learned features can effectively mitigate domain-biased information and identity-biased information.

**D. Main Algorithms**


- **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.
Figure S-2: Cross-domain and in-domain nearest neighbors searched by our UDG-FAS method.
<table>
<thead>
<tr>
<th>Datasets</th>
<th>Subjects</th>
<th>Data</th>
<th>Sensors</th>
<th>Spoof Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiap Replay-Attack (I) [6]</td>
<td>50</td>
<td>1,200</td>
<td>2</td>
<td>1 Print, 2 Video-replay</td>
</tr>
<tr>
<td>OULU-NPU (O) [2]</td>
<td>55</td>
<td>4,950</td>
<td>6</td>
<td>2 Print, 2 Video-replay</td>
</tr>
<tr>
<td>CASIA-MFSD (C) [15]</td>
<td>50</td>
<td>600</td>
<td>3</td>
<td>2 Print, 1 Video-replay</td>
</tr>
<tr>
<td>MSU-MFSD (M) [13]</td>
<td>35</td>
<td>280</td>
<td>2</td>
<td>1 Print, 2 Video-replay</td>
</tr>
<tr>
<td>3DMAD (D) [7]</td>
<td>17</td>
<td>255</td>
<td>5</td>
<td>1 3D Mask</td>
</tr>
<tr>
<td>HKBU-MARs (H) [10]</td>
<td>12</td>
<td>1008</td>
<td>6</td>
<td>2 3D Mask</td>
</tr>
<tr>
<td>CelebA-Spoof (CA) [14]</td>
<td>10177</td>
<td>625,537</td>
<td>&gt;10</td>
<td>3 Print, 3 Replay, 3 Paper Cut, 1 3D Mask</td>
</tr>
</tbody>
</table>

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 unsupervisedly pretrain 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
specific, we include the current largest CelebA-Spoof (CA) dataset as an additional unlabeled source dataset with three domains of I, O, C and 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 unsupervised pretrainned 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 attack 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 $l r = 0.03$ 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 $l r = 0.001$ 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**


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


