Supplementary Material of 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.

1 Intra-Dataset Protocol

We evaluate our model on the Oulu dataset (intra-dataset protocol). The evaluation includes four challenging protocols: 1) evaluate the model's robustness against the unseen environment, 2) unseen spoof mediums, 3) unseen capture devices, and 4) a combination of all the above. As depicted in Table 1, our method achieves state-of-the-art (SOTA) comparable performance.

Table 1. Intra-dataset protocol on Oulu

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Disentangle [8]	1.7	0.8	1.3
	SpoofTrace [2]	0.8	1.3	1.1
	BCN [5]	**0.0**	1.6	0.8
	CDCN [7]	0.4	1.7	1.0
	NAS-FAS [6]	0.4	**0.0**	0.2
	PatchNet [3]	**0.0**	**0.0**	**0.0**
	DLIF (Ours)	**0.0**	**0.0**	**0.0**
2	Disentangle [8]	1.1	3.6	2.4
	SpoofTrace [2]	2.3	1.6	1.9
	BCN [5]	2.6	**0.8**	1.7
	CDCN [7]	1.5	1.4	1.5
	NAS-FAS [6]	1.5	**0.8**	1.2
	PatchNet [3]	1.1	1.2	1.2
	DLIF (Ours)	**1.0**	1.2	**1.1**
3	Disentangle [8]	2.8±2.2	1.7±2.6	2.2±2.2
	SpoofTrace [2]	1.6±1.6	4.0±5.4	2.8±3.3
	BCN [5]	2.8±2.4	2.3±2.8	2.5±1.1
	CDCN [7]	2.4±1.3	2.2±2.0	2.3±1.4
	NAS-FAS [6]	2.1±1.3	1.4±1.1	1.18±1.26
	PatchNet [3]	1.8±1.47	0.56±1.24	2.8±2.2
	DLIF (Ours)	**1.4±1.2**	0.8±1.5	**1.1±1.3**
4	Disentangle [8]	5.4±2.9	3.3±6.0	4.4±3.0
	SpoofTrace [2]	2.3±3.6	5.2±5.4	3.8±4.2
	BCN [5]	2.9±4.0	7.5±6.9	5.2±3.7
	CDCN [7]	4.6±4.6	9.2±8.0	6.9±2.9
	NAS-FAS [6]	4.2±5.3	**1.7±2.6**	2.9±2.8
	PatchNet [3]	2.5±3.81	3.33±3.73	2.9±3.0
	DLIF (Ours)	**2.1±4.1**	2.3±4.2	**2.2±3.5**

* Corresponding Author. This work was supported by the National Science Foundation of China, under Grant No. 62171425 and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037).

2 Succinct and Intuitive Theoretical Basis

We give a succinct and intuitive theoretical basis for the effectiveness of our method. **Motivation:** Our primary motivation is to address the limitations of existing domain generalization (DG) approaches that utilize the dataset as the domain concept, which we argue is not fine-grained enough. Common inconsistent factors within the same dataset, such as identity variances, can hinder the learning of truly domain-invariant representations. In contrast, we propose a novel perspective that employs identity as the domain concept, allowing for the learning of identity-invariant representations. It is more finer compared to datasets, as typically a dataset contains many identities. Furthermore, it narrows the scope of the domain and reduces inconsistent factors within the same domain, which helps capture domain invariance. **Inspired by this motivation, we aim to disentangle the liveness and identity. To offer theoretical insight, we draw parallels to principal component analysis (PCA):** We hypothesize the principle components of faces can be factorized into identity, liveness, and the sum of other components that are difficult to quantify. It can be represented by Eqn 1, 2:

$$F = F_{id}\alpha_{id} + F_{liveness}\alpha_{liveness} + \sum F_i\alpha_i\,, \quad F_{id} \perp F_{liveness} \perp F_i \tag{1}$$

$$F = F_{id}\alpha_{id} + F_{liveness}\alpha_{liveness} + \bar{F}_{style}\,, \quad F_{id} \perp F_{liveness} \tag{2}$$

where F represents the deeply-learned representation, F_{id} is the identity component, $F_{liveness}$ is the liveness component, and α is the coefficient of principle component. Due to the orthogonality of each principle component in the PCA, we apply orthogonal constrain during the optimization of neural network to disentangle identity and liveness components. The CrossEntropy and DistanceMetric losses utilized for classification ensure the maximum separability of F_{id} and $F_{liveness}$. We denote other difficult to quantify components as $\bar{F}_{style}$ and propose that the SC and CWSA modules filter them and weaken their effects before disentanglement. The optimization goal is:

$$\begin{gathered}\text{Minimize } L = \text{CrossEntropy} + \text{DistanceMetric} \quad \text{rst. } F_{id}^T F_{liveness} = 0 \\ L = \text{CrossEntropy} + \text{DistanceMetric} + \lambda(F_{id}^T F_{liveness})^2\end{gathered} \tag{3}$$

Metric Learning in Asymmetric Augmented Instance Contrast (AAIC): For general formula of supervised contrastive learning, the

contrastive loss can evolve into Eqn 4:

$$
\begin{aligned}
L &= -\log \frac{\exp(z_a z_p/\tau)}{\exp(z_a z_p/\tau) + \exp(z_a z_n/\tau)} = \log(1 + \exp((z_a z_n - z_a z_p)/\tau)) \\
&\approx \exp((z_a z_n - z_a z_p)/\tau) \quad \text{(Taylor expansion of log)} \\
&\approx 1 + (z_a z_n - z_a z_p)/\tau \approx 1 - (\|z_a - z_n\|^2 - \|z_a - z_p\|^2)/2\tau \\
&\propto \|z_a - z_p\|^2 - \|z_a - z_n\|^2
\end{aligned} \tag{4}
$$

where a denotes anchor, p, n represent positive and negative, respectively. Assuming that $z_a \cdot z_p \gg z_a \cdot z_n$, the high-order infinitesimal terms in the Taylor expansion are ignored. The contrastive loss ultimately evolves to optimize the distance between anchor relative to the positives and negatives. In fact, the main differences in contrast losses are specific to the definitions of anchor, positive, and negative. As depicted in Section 5.1 and Figure 3, 4, whether treats positives and negatives equally (Binary) or highlights the differences between inter-domain spoofs while ignoring the differences in intra-domain spoofs (Triplet). Both of them are not very consistent with the distribution of live and spoof samples in the real-world. However our AAIC only pull close the augmented spoof samples generated by the SC module with original spoofs, which is the instance-pair-aware metric learning, resulting the distribution of spoofs more scattered. This is more in line with the sample distribution that in the real world, spoofs are more diverse than live samples. Therefore, AAIC exhibits stronger generalization ability.

3 Motivation of Feature-level Style Cross

3.1 Why limit Style Cross according to task?

In Figure 1, when we straightforwardly exchange the RGB channels' style of the original image, it can be observed that the spoof pattern of video-replay attacks can be easily transferred or replaced. A live sample appears as if a video-replay after being equipped with the RGB's style of video-replay. Moreover, when video-replay attacks are equipped with the style of live samples, they appear to be one step closer to a live face. However, printed attacks are entirely opposite. For example, the spoof traces on a printed paper with eyes gouged out will not be transferred or replaced through simple RGB style exchanges.

Building upon the observations of the aforementioned phenomena, we introduce the concept of imposing task-oriented constraints on Style Cross (SC). Specifically, in face anti-spoofing tasks, we cannot guarantee whether liveness information can be transferred when implementing SC between live and spoof samples. In essence, there arises a controversy over the definition of liveness for samples that have undergone SC in this scenario. This ambiguity is also applicable to face recognition tasks, where ensuring identity consistency is necessary. Consequently, we propose Liveness-invariant Style Cross (LISC) and Identity-invariant Style Cross (IISC) to mitigate this potential adverse effect.

3.2 Why employ various levels and flows?

In AdaIN [1], style transfer is implemented at the top of the feature encoder utilizing high-level semantic style. On the other hand, SSAN [4] accumulates multiple hierarchical style features and implements shuffle style assembly (SSA) at the top of feature extractor. The question of whether style in low-level semantic information or high-level semantic information is more crucial for FAS tasks is a topic worthy of exploration.

In our consideration, we recognize that applying SC at different levels may have different impacts on FAS systems. Therefore, We

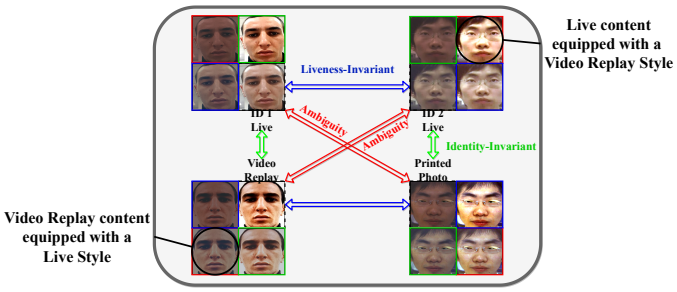

Figure 1. Low-level semantic (RGB channels) Style Cross.

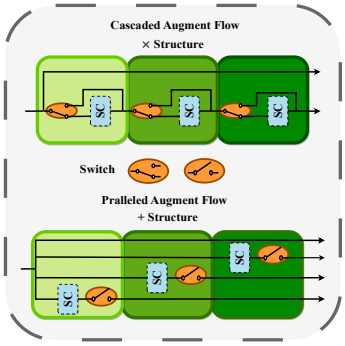

Figure 2. Structure (cascaded & paralleled) of augment flow.

can naively divide the network into three levels: Low (L), Middle (M), and High (H), corresponding to light green to dark green. In addition to these easily thought of simple forms, there are also composite forms such as cascaded ($\times$) and paralleled ($+$) structure, as illustrated in Figure 2. We can obtain various augmented flows through toggling switches, including: L, M, H, L $\times$ M, L $\times$ H, M $\times$ H, L $\times$ M $\times$ H, L $+$ M, L $+$ H, M $+$ H, L $+$ M $+$ H. The original sample is obtained from the topmost flow. After a large number of comparative experiments, we found that for FAS tasks, Style Cross at the middle and high levels has a gain in the generalization of the model. The implementation results and visualization are provided in the main text. It is important to note that our conclusions may hinge on the design of our simple Style Cross module. On one hand, our definitions of low, middle, and high levels are intuitive. On the other hand, style augmentation techniques with complex designs may not necessarily be applicable to our conclusions.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.