Style-Guided Domain Adaptation for Face Presentation Attack Detection

Young-Eun Kim¹, Woo-Jeoung Nam², Kyungseo Min³ and Seong-Whan Lee¹∗

¹Department of Artificial Intelligence, Korea University, Republic of Korea
²Department of Computer and Radio Communications Engineering, Korea University, Republic of Korea
³Department of Brain and Cognitive Engineering, Korea University, Republic of Korea

{ye.kim,nwj0612,ks.min,sw.lee}@korea.ac.kr

Abstract

Domain adaptation (DA) or domain generalization (DG) for face presentation attack detection (PAD) has attracted attention recently with its robustness against unseen attack scenarios. Existing DA/DG-based PAD methods, however, have not yet fully explored the domain-specific style information that can provide knowledge regarding attack styles (e.g., materials, background, illumination and resolution). In this paper, we introduce a novel Style-Guided Domain Adaptation (SGDA) framework for inference-time adaptive PAD. Specifically, Style-Selective Normalization (SSN) is proposed to explore the domain-specific style information within the high-order feature statistics. The proposed SSN enables the adaptation of the model to the target domain by reducing the style difference between the target and the source domains. Moreover, we carefully design Style-Aware Meta-Learning (SAML) to boost the adaptation ability, which simulates the inference-time adaptation with style selection process on virtual test domain. In contrast to previous domain adaptation approaches, our method does not require either additional auxiliary models (e.g., domain adaptors) or the unlabeled target domain during training. To verify our experiments, we utilize the public datasets: MSU-MFSD, CASIA-FASD, OULU-NPU and Idiap REPLAYATTACK. In most assessments, the result demonstrates a notable gap of performance compared to the conventional DA/DG-based PAD methods.

1 Introduction

With the increasing interests in face recognition on the daily life application (e.g., checking-in and mobile payment), detecting face presentation attacks (PA) such as printed photo, video replay, and 3D-mask has emerged as an important issue. To cope with PAs, recent research on face presentation attack detection has been actively carried out in the name of face anti-spoofing (FAS) for the security of authentication systems. Various FAS methods are mainly divided into two categories, temporal-based and texture-based methods. Temporal-based methods utilize temporal cues from an image sequence, such as optical flow [Anjos et al., 2014] and rPPG [Liu et al., 2018a]. Texture-based methods exploit texture differences between live and spoof faces, such as color texture [Boulkenafet et al., 2017a], image distortion [Wen et al., 2015] and deep feature [Yang et al., 2014]. Although existing methods show promising results in intra-dataset testing scenarios, they perform inadequately in cross-dataset testing, where the training datasets (source domains) and the test datasets (target domains) are different. The main reason of this issue is the domain shift — the changes in the data distribution between training and test datasets. To resolve the performance degradation on cross-dataset testing, recent FAS works focus on domain adaptation (DA) or domain generalization (DG) techniques and exhibits promising performance.

Despite the significant success on DA/DG-based methods, there still remain several limitations. Firstly, previous DA-based methods commonly perform training by leveraging the labeled source data and the unlabeled target data [Li et al., 2018b; Wang et al., 2019]. In real-world scenarios, how-
ever, it is often difficult to acquire target domain data during training. Although DG-based studies [Shao et al., 2019; Jia et al., 2020] try to address this issue by utilizing the multiple source domains to learn a generalized feature space without including any target data, there are still some drawbacks occurred by the lack of target data information.

To overcome these limitations, we present a novel Style-Guided Domain Adaptation (SGDA) method that leverages the domain-specific style information of both target and source domains. To this end, Style-Selective Normalization (SSN) is proposed to construct the domain-adaptive model by utilizing multiple sets of normalization parameters to learn the domain-specific style information. At inference, SSN helps the network to be adaptive to the target domain with style information so that the optimal SSN parameters are selected considering the domain shift. Besides, we propose an optimization strategy, Style-Aware Meta-Learning (SAML), which simulates the domain shift with style features. Based on the estimated style features at meta-train step, SAML optimizes the adaptive network with the style features of the virtual target domain, which leads to ensuring the performance of inference-time adaptation.

Our main contributions can be concluded as follows:

- We propose a new Style-Guided Domain Adaptation framework, which is an inference-time domain adaptation for face presentation attack detection. The proposed framework is designed based on a Style-Selective Normalization, which fully explores the domain-specific style features to select the optimal normalization path of the adaptive network.
- We carefully design Style-Aware Meta-Learning to boost the adaptation ability of the network embedded with style selective normalization by simulating the inference-time style selection process on the virtual target domain.
- Extensive experiments is performed with four public FAS datasets: MSU-MFSD, CASIA-FASD, OULU-NPU and Idiap REPLAYATTACK, demonstrating the effectiveness of our method relative to the state-of-the-art competitors.

3 Proposed Method

In this section, we introduce our Style-Guided Domain Adaptation (SGDA), which aims to adapt the model for the target data in the guidance of style features within target and source domains. A novel Style-Selective Normalization and Style-Aware Meta-Learning are addressed in Section 3.1 and 3.2, respectively.

3.1 Style-Selective Normalization

Suppose that there are $N$ source domains, denoted as $S = \{S_1, S_2, ..., S_N\}$. To separate the domain-specific information, we extensively designed Style-Selective Normalization (SSN) by constructing $N$ sets of normalization parameters reserved for each source domain. Then, SSN is formulated as:

$$SSN (x_k; \beta_k, \gamma_k) = \gamma_k \frac{x_k - \mu_k}{\sigma_k} + \beta_k \quad (k = 1, \ldots, N),$$  \hspace{1cm} (1)

where $x_k$ denotes activations for source domain $S_k$, $\mu_k$ and $\sigma_k$ are the mean and standard deviation of activations.
Expanding on the general normalization estimating the feature mean and variance during training, SSN also estimates the feature covariances, since the feature correlation between the feature mean and variance during training, SSN also estimates the feature covariance during training, SSN also estimates the feature covariance during training. The feature covariance during training, SSN also estimates the feature covariance during training. The feature covariance during training, SSN also estimates the feature covariance during training. The feature covariance during training, SSN also estimates the feature covariance during training.

Formally, the style distance is calculated using Eq. (2)-(3). The former part of Eq. (5) is sum of squared Wasserstein-2 distance of the first-order feature statistics across the channel. The latter part is squared L-2 distance of the feature covariance.

To get the optimal index \( k^* \) of SSN parameters, we take the index \( k \) which minimizes the \( D_{\text{style}} \) as follows:

\[
D_{\text{style}}(X_k, X_i) = \begin{cases} 
\frac{1}{HW} \sum_{c=1}^{C} \left( \| \mu_{t,c} - \mu_{k,c} \|^2 + \| \sigma_{t,c} - \sigma_{k,c} \|^2 \right), & k \neq i \\
0, & k = i
\end{cases}
\] (5)

where \( X_i \) is the target feature map, \( \mu_t \) and \( \sigma_t \) denote the mean and standard deviation of the target feature map, \( c \) is a channel index and \( C_t \) is the target feature covariance calculated using Eq. (2)-(3). The former part of Eq. (5) is sum of squared Wasserstein-2 distance of the first-order feature statistics across the channel. The latter part is squared L-2 distance of the feature covariance.

To get the optimal index \( k^* \) of SSN parameters, we take the index \( k \) which minimizes the \( D_{\text{style}} \) as follows:

\[
k^* = \arg \min_k D_{\text{style}}(X_k, X_i) (k = 1, \ldots, N).
\] (6)

By selecting the optimal normalization parameters at each layer in the guidance of style distance, the network can be adaptively reduce the style difference between target and source domains.

### 3.2 Style-Aware Meta-Learning

In this section, we introduce Style-Selective Meta-Learning (SAML) as shown in algorithm 1. SAML offers us the advantage of enhancing the adaptation ability of our feature extractor by simulating the inference-time adaptation during training.

The network proposed in our framework is composed of a feature extractor \( F \), a classifier \( C \) and a depth estimator \( D \) as shown in Figure 2. In order to simulate the real scenarios, \( N - 1 \) source domains are randomly chosen from \( S \) as the meta-train domain \( S_{tr} \) and remaining one domain is selected as the meta-test domain \( S_{ts} \) in every iteration.

#### Meta-Train

In the meta-train loop, we sample batch in each domain from \( S_{tr} \) as \( S_i \) for \( i = 1, \ldots, N \). The cross-entropy classification loss is conducted based on the binary class labels in \( S_i \) as follows:

\[
L_{\text{cls}}(S_i; \theta_F, \theta_C) = \sum_{(x,y) \sim S_i} 1[y = q] \log C(F_{S_i}(x)),
\] (7)

where \( \theta_F \) and \( \theta_C \) are the parameters of the feature extractor and the classifier. \( F_{S_i} \) is the feature extractor \( F \) com-
posed with the corresponding SSN branch reserved for domain $S_i$. The updated classifier $C_{i}^+$ can be calculated as

$$
\theta_{C_{i}^+} = \theta_C - \alpha \nabla_{\theta_C} \mathcal{L}_{cls}(\hat{S}_i)(\theta_F, \theta_C).
$$

Meanwhile, following [Wang et al., 2021b], we also incorporate face depth information as auxiliary supervision in the learning process as follows:

$$
\mathcal{L}_{Dep}(\hat{S}_i; \theta_F, \theta_D) = \sum_{(x, I) \sim \hat{S}_i} \|D(F_{\hat{S}_i}(x)) - I_d\|^2, \quad (8)
$$

where $\theta_D$ is the parameter of the depth estimator and $I_d$ is the pre-calculated face depth maps of input face images.

**Meta-Test**

In the meta-test loop, we evaluate the model’s performance on (virtual) testing domain $S_{ts}$. By evaluating the feature extractor adapted via style-selection, we encourage our model that is trained on every meta-train domain with its corresponding SSN branch, can simultaneously perform well on the meta-test domain. Thus, the model should have good classification performance on the target domain by minimize the cross-entropy loss defined over the updated $\theta_{C_{i}^+}$ as:

$$
\sum_{i=1}^{N-1} \mathcal{L}_{cls}(\hat{S}_i; \theta_F, \theta_{C_{i}^+}) = \sum_{i=1}^{N-1} \sum_{S} 1[y = c] \log C_{i}^+(F_{\hat{S}_i}(x))
$$

The face depth information is also incorporated:

$$
\mathcal{L}_{Dep}(\hat{S}_i; \theta_F, \theta_D) = \sum_{S} \|D(F_{\hat{S}_i}(x)) - I_d\|^2 \quad (9)
$$

**Meta-Optimization**

We jointly optimize the parameters of three modules in a meta-learning framework as follows:

$$
(\theta_C, \theta_F, \theta_D) \leftarrow (\theta_C, \theta_F, \theta_D) - \nabla_{\theta_C, \theta_F, \theta_D} \left( \mathcal{L}_{Dep}(\hat{S}_i; \theta_F, \theta_D) + \sum_{i=1}^{N-1} (\mathcal{L}_{cls}(\hat{S}_i; \theta_F, \theta_C) + \mathcal{L}_{Dep}(\hat{S}_i; \theta_F, \theta_D) + \mathcal{L}_{cls}(\hat{S}_i; \theta_F, \theta_{C_{i}^+})) \right) \quad (11)
$$

As we minimize the overall loss for the feature extractor with style selection process, this can make the model focus on better-adaptive learning direction.

Once the update procedure is completed, we keep all the parameters of the model fixed, and test the target domain by selecting the SSN layerwisely using Eq. (6) at inference.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets**

Four public face anti-spoofing datasets are utilized to evaluate the effectiveness of our method: OULU-NPU [Boulkenafet et al., 2017b] (denoted as O), CASIA-MFSD [Zhang et al., 2012] (denoted as C), Idiap Replay-Attack [Chingovska et al., 2012] (denoted as I), and MSU-MFSD [Wen et al., 2015] (denoted as M). Following the setting in a previous study [Wang et al., 2021b], a single dataset is treated as one domain in our experiment. We randomly select three datasets as the source domains for training, and the remaining one as the target domain for testing. Thus, we have four testing sets in total: O&C&I to M, O&M&I to C, O&C&M to I, and I&C&M to O.

### Implementation Details

Our method is implemented using PyTorch and trained with SGD optimizer with momentum of 0.9 and weight decay of 5e-4. The MTCNN algorithm [Zhang et al., 2016] is adopted for face detection for data pre-processing. We use the face alignment network named PRNet [Feng et al., 2018] for depth maps of real faces. The size of face image is $256 \times 256 \times 3$, where we extract only RGB channels of each face image to further reduce the network complexity. We strictly follow the popular evaluation metrics, which are Half Total Error Rate (HTER) and the Area Under Curve (AUC).

### 4.2 Discussion

**Comparison with the State-of-the-art PAD Methods**

As shown in Figure 3 and Table 1, our method outperforms most of state-of-the-art face anti-spoofing methods on four testing sets. We can observe that our method not only performs better than conventional methods but also performs better than DA/DG-based methods. Specifically, SSDG-M and MADDG utilize the domain discriminator to learn the domain-invariant features, however, it is difficult to optimize and can affect the discriminative feature space. Self-DA and RFM adopt a meta-learning strategy results in better performance than the adversarial learning-based methods. Our
method focuses on adapting the model by leveraging domain-specific style information existing within both the target and source domains, which is optimized through the proposed meta-learning strategy resulting in better performance. Furthermore, compared to Self-DA [Wang et al., 2021b] which is the most recent DA-based PAD method, SGDA gains 4.6% and 4.4% improvement on the task O&C&I to M and O&C&M to I respectively with HTER.

Comparison with the Inference-time DA Methods
As shown in Table 2, we compare our SGDA with the related inference-time domain adaptation methods. AdaBN and TENT modulate the parameters to adapt the model to target domain. 3C-GAN incorporates generator and discriminator to generate target-style data, which collaboratively enhance the adaptation ability of the prediction model. However, the results show that none of them perform adequately on the task of PAD. Self-DA framework is designed for PAD utilizing an additional adaptor network and an autoencoder. In contrast, our SGDA framework is constructed without additional network training by directly leveraging the domain-related knowledge within the feature statistics, which outperforms all the inference-time domain adaptation methods.

Analysis and Interpretation of Style Selection
To analyze the inference-time style-selection, we conducted quantitative experiments using the target dataset split by face presentation attack types as shown in Figure 4. Specifically, we experimented with each category of face attack under O&M&I to C task and produced the statistics of the selected index of SSN parameters (abbreviated as style index). Some examples are shown in the upper part of Figure 4, which are randomly selected from each category of face attacks in CASIA-FASD dataset. The bottom part shows corresponding results of the aggregated ratio of the selected style index across the whole dataset. Each color denotes a style index. As from the chart, we can observe that, 1) most of the layer shows the dominance of one specific style index and the set of most dominant style index is significantly different by the attack categories. This demonstrates that our framework can identify the different style of attack and properly adapts to the target data depending on their attack style. 2) It can be
seen that the proportion of the selected style index is evenly distributed as it goes to the deeper layer of the network. This observation shows that style information mainly exists in the early layers of the network, which justifies the adoption of our style-selective normalization on the feature extractor.

**Influences of Each Component**

As shown in Table 3, we performed the ablation study to evaluate the performance gained by each component. SGDA w/o SSN denotes our proposed method without SSN, in which the model is constructed with a single branch of normalization parameters. SGDA w/o SAML is our model without SAML, which means that the model equipped with SSN is optimized by solely passing the source domain to its assigned branch of SSN and does not simulate the style selection with virtual test-domain during training. SGDA w/o Adaptation denotes the proposed model without adaptation process at inference. In this setting, we do not select the optimal SSN parameters and pass through a single branch which gives the best performance among all sets. Specifically, the results of SGDA w/o Adaptation validate that adaptation by leveraging the style feature of the test domain is efficient in improving the performance. The comparison results verify that our method exhibits degraded performance if any component is excluded.

**Impact of Weighting Feature Covariance**

As shown in Figure 5, we evaluate the impacts of considering the feature covariance as domain-specific style information. As discussed in Section 3.1, λ controls the contribution of the feature correlation, i.e., the higher λ, the more feature correlation information reflected to style distance. As illustrated in Fig. 5(a), when λ = 0.2, it achieves the best performance at 11.29% HTER and 95.00% AUC, while the performance dropped when λ gets higher or lower. Compared with our method without feature covariance (i.e., λ = 0), our method with covariance gains 9.14% and 6.14% improvement with HTER and AUC, respectively. It is worth noting that, benefited from the feature covariance with the first-order statistics of feature distribution, SGDA could obtain remarkable performance gains on O&C&M to I set.

**5 Conclusion**

In this paper, we proposed a new Style-Guided Domain Adaptation framework that automatically adapts the model to the target data in the guide of the domain-specific style information. To achieve this, Style-Selective Normalization is designed to construct the inference-time adaptive model by en-
courage the target data to explore the optimal normalization parameters. Besides, a Style-Aware Meta-Learning facilitates a better domain adaptation by simulating the inference-time style-selection process with the virtual test domain during training. Extensive experiments and analyses on public PAD dataset have demonstrated the effectiveness of our proposed methods against the conventional DA/DG-based methods.

References

[Anjos et al., 2014] André Anjos, Murali Chakka, and Sébastien Marcel. Motion-based counter-measures to photo attacks in face recognition. IET biometrics, 3, 2014.

[Boulkenafet et al., 2017a] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofer using speeded-up robust features and fisher vector encoding. IEEE SPL, 24, 2017.

[Boulkenafet et al., 2017b] Zinelabidine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoqi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In IEEE FG, 2017.

[Chingovska et al., 2012] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofer. In BIOSIG, 2012.

[Choi et al., 2021] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban scene segmentation via instance selective whitening. In CVPR, 2021.

[Feng et al., 2018] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In ECCV, 2018.

[Jia et al., 2020] Yunpei Jia, Jie Zhang, Shiguang Shang, and Xinlin Chen. Single-Side Domain Generalization for FaceAnti-Spoofing. In CVPR, 2020.

[Li et al., 2017a] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In IJCAI, 2017.

[Li et al., 2017b] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In NIPS, 2017.

[Li et al., 2018a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In AAAI, 2018.

[Li et al., 2018b] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C. Kot. Unsupervised domain adaptation for face anti-spoofing. IEEE TIFS, 13, 2018.

[Li et al., 2018c] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In CVPR, 2018.

[Li et al., 2018d] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. Pattern Recognition, 80, 2018.

[Li et al., 2020] Rui Li, Qianfen Jiao, Wenming Cao, Huai-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In CVPR, 2020.

[Liu et al., 2018a] Si-Qi Liu, Xiangyuan Lan, and Pong C. Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In ECCV, 2018.

[Liu et al., 2018b] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In CVPR, 2018.

[Matsuura and Harada, 2020] Toshihiko Matsuura and Takeyuki Harada. Domain generalization using a mixture of multiple latent domains. In AAAI, 2020.

[Määttä et al., 2011] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In IJCB, 2011.

[Park et al., 2020] KwanYong Park, Sanghyun Woo, Inkyu Shin, and In So Kweon. Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. In NIPS, 2020.

[Patel et al., 2015] Keyurkumar Patel, Hu Han, Anil. K. Jain, and Greg Ott. Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In ICB, 2015.

[Pereira et al., 2014] Tiago Pereira, Jukka Komulainen, André Anjos, José De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. EURASIP Journal on Image and Video Processing, 2014, 2014.

[Shao et al., 2019] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In CVPR, 2019.

[Shao et al., 2020] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized fine-grained meta face anti-spoofing. In AAAI, 2020.

[Wang et al., 2019] Guoqing Wang, Hu Han, Shiguang Shang, and Xinlin Chen. Improving cross-database face presentation attack detection via adversarial domain adaptation. In ICB, 2019.

[Wang et al., 2021a] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In ICLR, 2021.

[Wang et al., 2021b] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In AAAI, 2021.

[Wen et al., 2015] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. IEEE TIFS, 10, 2015.

[Yang et al., 2014] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing. CoRR, abs/1408.5601, 2014.
[Zhang et al., 2012] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In ICB, 2012.

[Zhang et al., 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE SPL, 23, 2016.