Dual-Cross Central Difference Network for Face Anti-Spoofing

Zitong Yu\textsuperscript{1}, Yunxiao Qin\textsuperscript{2}, Hengshuang Zhao\textsuperscript{3}, Xiaobai Li\textsuperscript{1} and Guoying Zhao\textsuperscript{1*}

\textsuperscript{1}CMVS, University of Oulu \textsuperscript{2}Northwestern Polytechnical University \textsuperscript{3}University of Oxford

\{zitong.yu, xiaobai.li, guoying.zhao\}@oulu.fi, qyxqyx@mail.nwpu.edu.cn, hengshuang.zhao@eng.ox.ac.uk

\section*{Abstract}
Face anti-spoofing (FAS) plays a vital role in securing face recognition systems. Recently, central difference convolution (CDC) \cite{Yu et al., 2020d} has shown its excellent representation capacity for the FAS task via leveraging local gradient features. However, aggregating central difference clues from all neighbors/directions simultaneously makes the CDC redundant and sub-optimized in the training phase. In this paper, we propose two Cross Central Difference Convolutions (C-CDC), which exploit the difference of the center and surround sparse local features from the horizontal/vertical and diagonal directions, respectively. It is interesting to find that, with only five ninth parameters and less computational cost, C-CDC even outperforms the full directional CDC. Based on these two decoupled C-CDC, a powerful Dual-Cross Central Difference Network (DC-CDN) is established with Cross Feature Interaction Modules (CFIM) for mutual relation mining and local detailed representation enhancement. Furthermore, a novel Patch Exchange (PE) augmentation strategy for FAS is proposed via simply exchanging the face patches as well as their dense labels from random samples. Thus, the augmented samples contain richer live/spoof patterns and diverse domain distributions, which benefits the intrinsic and robust feature learning. Comprehensive experiments are performed on four benchmark datasets with three testing protocols to demonstrate our state-of-the-art performance.

\section{Introduction}
Face recognition technology has widely used in many interactive intelligent systems due to their convenience and remarkable accuracy. However, face recognition systems are still vulnerable to presentation attacks (PAs) ranging from print, replay and 3D-mask attacks. Therefore, both the academia and industry have recognized the critical role of face anti-spoofing (FAS) for securing the face recognition system.

In the past decade, both traditional \cite{de Freitas Pereira et al., 2012; Komulainen et al., 2013; Patel et al., 2016} and deep learning-based \cite{Yu et al., 2020d; Yu et al., 2020a; Liu et al., 2018; Jourabloo et al., 2018; Yang et al., 2019; Yu et al., 2020b} methods have shown effectiveness for presentation attack detection (PAD). On one hand, some classical local descriptors (e.g., local binary pattern (LBP) \cite{Boulkenafet et al., 2015} and histogram of gradient (HOG) \cite{Komulainen et al., 2013}) are robust for describing the detailed invariant information (e.g., color texture, moiré pattern and noise artifacts) from spoofing faces. However, the shallow and coarse feature extraction procedure limits the discriminative capacity of these local descriptors.

On the other hand, convolutional neural networks (CNNs) focus on representing deeper semantic features to distinguish the bonafide and PAs, which are weak in capturing fine-grained intrinsic patterns (e.g., lattice artifacts) between the live and spoof faces, and easily influenced under diverse scenarios. Although central difference convolution (CDC) \cite{Yu et al., 2020d} introduced the central-oriented local gradient features to enhance models’ generalization and discrimination capacity, it still suffers from two disadvantages. In CDC, central gradients from all neighbors are calculated, which is 1) inefficient in both inference and back-propagation stages; and 2) redundant and sub-optimized due to the discrepancy among diverse gradient directions. Thus, to study the impacts and relations among central gradients is not a trivial work.

One key challenge in the FAS task is how to learn representation with limited data as existing FAS datasets (e.g., OULU-NPU \cite{Boulkenafet et al., 2017} and SiW-M \cite{Liu et al., 2019}) do not have large amount of training data due to the high collection cost for both spoofing generation and video recording. Although generic augmentation manners (e.g., horizontal flip, color jitter and Cutout \cite{DeVries and Taylor, 2017}) are able to expand the scale and diversity of live/spoof samples, it still contributes not much performance improvement. Thus, it is worth rethinking the augmentation for FAS and design task-dedicated augmentation paradigm.

Motivated by the discussions above, we propose a novel convolution operator family called Cross Central Difference
Convolution (C-CDC), which decouples the central gradient features into cross-directional combination (horizontal/vertical or diagonal) thus more efficient and concentrated for message aggregation. Furthermore, in order to mimic more general attacks (e.g., partial print and mask attacks) and learn to distinguish spoofing in both global and patch level, Patch Exchange (PE) augmentation is proposed for mixed sample as well as corresponding dense label generation. To sum up, our contributions include:

- We design a sparse convolution family called Cross Central Difference Convolution (C-CDC), which decouples the vanilla CDC into two cross (i.e., horizontal/vertical and diagonal) directions, respectively. Compared with CDC, our proposed C-CDC could achieve better performance for FAS with only five ninth parameters and less computational cost.
- We propose a Dual-Cross Central Difference Network (DC-CDN), consisting of two-stream backbones with horizontal/vertical and diagonal C-CDC, respectively. Moreover, we also introduce Cross Feature Interaction Modules (CFIM) between two streams of DC-CDN for mutual neighbor relation mining and local detailed representation enhancement.
- We propose the first FAS-dedicated data augmentation method, Patch Exchanges (PE), to synthesize mixed samples with diverse attacks and domains, which is able to plug and play in not only DC-CDN but also existing FAS methods for performance improvement.
- Our proposed method achieves state-of-the-art performance on four benchmark datasets with intra-dataset, cross-dataset, and cross-type testing protocols.

2 Related Work

Face Anti-Spoofing. Traditional face anti-spoofing methods usually extract handcrafted features from the facial images to capture the spoofing patterns. Some classical local descriptors such as LBP [Boulkenafet et al., 2015] and HOG [Komulainen et al., 2013] are utilized for handcrafted features. More recently, a few deep learning based methods are proposed for face anti-spoofing. On the one hand, FAS can be naturally treated as a binary classification task, between live and spoof faces, dense pixel-wise supervisions [Yu et al., 2020d; Yu et al., 2020c; Yu et al., 2020b] is proposed for invariant and detailed features extraction, which is suitable for the FAS task. In this paper, we devote to improving the vanilla CDC in terms of both performance and efficiency.

3 Methodology

In this section we first briefly review the CDC [Yu et al., 2020d], and then introduce the novel Cross Central Difference Convolution (C-CDC) family in Sec. 3.2. Based on the C-CDC operators, we propose Dual-Cross Central Difference Networks in Sec. 3.3. Finally we present the novel data augmentation strategy in Sec. 3.4.

3.1 Preliminary

As the basic operator in deep networks, the vanilla 2D convolution consists of two main steps: 1) sampling local neighbor region \( R \) over the input feature map \( x \); and then 2) aggregating the sampled values via learnable weights \( w \). As a result, the output feature map \( y \) can be formulated as

\[
y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n),
\]

where \( p_0 \) denotes the current location on both input and output feature maps while \( p_n \) enumerates the locations in \( R \). For instance, local receptive field region for convolution operator with \( 3 \times 3 \) kernel and dilation 1 is \( R = \{(-1, -1), (-1, 0), \ldots, (0, 1), (1, 1)\} \).

Different from the vanilla convolution, the CDC introduces central gradient features to enhance the representation and generalization capacity, which can be formulated as

\[
y(p_0) = \sum_{p_n \in R} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)).
\]

As both intensity-level semantic information and gradient-level detailed clues are crucial for robust FAS, the generalized
Aggregation more concentrated and intrinsic features for FAS. Difference Convolution (C-CDC) family, intending to learn optimized. We assume that exploiting sparse center-oriented difference region.

\[ \theta \in [0, 1] \] trade-offs the contribution between intensity-level and gradient-level information. The higher value of \( \theta \) means the more importance of central gradient features. Please note that \( w(p_n) \) is shared between vanilla convolution and CDC, thus no extra parameters are added. Henceforth, the generalized CDC will be referred as CDC directly for clear statement.

### 3.2 Cross Central Difference Convolution

As can be seen from Eq. (3) that CDC aggregates both the vanilla and central gradient features from entire local neighbor region \( R \), which might be redundant and hard to be optimized. We assume that exploiting sparse center-oriented difference features would alleviate local competition and overfitting issues. Therefore, we propose the sparse Cross Central Difference Convolution (C-CDC) family, intending to learn more concentrated and intrinsic features for FAS.

\[
y(p_0) = \theta \cdot \sum_{p_n \in R} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)) + (1 - \theta) \cdot \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n),
\]

\[
= \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) + \theta \cdot \sum_{p_n \in R} w(p_n),
\]

where hyperparameter \( \theta \in [0, 1] \) trade-offs the contribution between intensity-level and gradient-level information. The higher value of \( \theta \) means the more importance of central gradient features. Please note that \( w(p_n) \) is shared between vanilla convolution and CDC, thus no extra parameters are added. Henceforth, the generalized CDC will be referred as CDC directly for clear statement.

| Output Feature Map | Conv | Horizontal \& Vertical Cross Central Difference | Sampling | Aggregation |
|-------------------|------|----------------------------------------|----------|-------------|
| Input Feature Map | Expand | | | |

\[ y(p_0) = \sum_{p_n \in S} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)) \] (4)

To be specific, we decouple \( R \) into two cross neighbor regions, including 1) horizontal \& vertical (HV) cross neighbor regions \( S_{HV} = \{-1, 0, 1\} \times \{-1, 0, 1\} \); and 2) diagonal (DG) cross neighbor regions \( S_{DG} = \{-1, -1, 1, 1\} \times \{-1, 1, 1, 1\} \). In this way, horizontal \& vertical CDC and diagonal CDC can be represented when \( S = S_{HV} \) and \( S = S_{DG} \), respectively. Figure 5 illustrates the workflow of the C-CDC(HV) and C-CDC(DG). Similarly, the generalized C-CDC can be easily formulated when replacing \( R \) with \( S \) in Eq. (3). We will use this generalized C-CDC henceforth.

In terms of designs of the sparse local region \( S \), there are also some other solutions with different neighbor locations or fewer neighbors. The reason that we consider the cross (i.e., HV and DG) fashions for \( S \) derives from their symmetry, which is beneficial for model convergence and robust feature representation. The studies about the sparsity and neighbor locations are shown in Section 4.3 and Appendix A.

### 3.3 Dual-Cross Central Difference Network

Pseudo depth-based supervision takes advantage of the discrimination between live and spoof faces based on 3D shape, which is able to provide pixel-wise detailed clues to enforce FAS model to capture intrinsic features. Following the similar depth-supervised backbone as “DepthNet” [Liu et al., 2018] and “CDCN” [Yu et al., 2020d], we replace all the \( 3 \times 3 \) convolution operators with our proposed C-CDC to form the Cross Central Difference Network (C-CDN). Given a single RGB face image with size \( 3 \times 256 \times 256 \), multi-level (low-level, mid-level, and high-level) fused features are extracted for predicting the grayscale facial depth with size \( 32 \times 32 \). The details of C-CDN are shown in Table 1. It can be seen that

| Output | DepthNet | CDCN | C-CDN (Ours) |
|--------|----------|------|--------------|
| 256 \times 256 | \( 3 \times 3 \) conv, 64 | \( 3 \times 3 \) CDC, 64 | \( 3 \times 3 \) C-CDC, 64 |
| 128 \times 128 (Low) | \( 3 \times 3 \) conv, 128 | \( 3 \times 3 \) CDC, 128 | \( 3 \times 3 \) C-CDC, 128 |
| 64 \times 64 (Mid) | \( 3 \times 3 \) conv, 128 | \( 3 \times 3 \) CDC, 128 | \( 3 \times 3 \) C-CDC, 128 |
| 32 \times 32 (High) | \( 3 \times 3 \) conv, 128 | \( 3 \times 3 \) CDC, 128 | \( 3 \times 3 \) C-CDC, 128 |
| # params | \( 2.25 \times 10^6 \) | \( 2.25 \times 10^6 \) | \( 1.25 \times 10^6 \) |

Table 1: Architectures of DepthNet, CDCN, and the proposed C-CDN. Inside the brackets are the filter sizes and feature dimensions. 'conv' suggests the vanilla convolution. All convolutional layers are with stride=1 and are followed by a BN-ReLU layer while pooling layers are with stride=2.

Compared with CDC operating on \( R \), C-CDC prefers to sample a sparser local region \( S \), which can be formulated as

\[
y(p_0) = \sum_{p_n \in S} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)).
\]

The studies about the sparsity and neighbor locations are shown in Section 4.3 and Appendix A.
with the similar architecture (e.g., network depth and width), C-CDN has only five ninth parameters and two hundredth computational cost compared with DepthNet and CDCN due to the sparse local sampling mechanism in C-CDC. We use $\theta = 0.8$ as the default setting, and the corresponding study about $\theta$ will be discussed in Section 4.3.

Although the C-CDC decouples and learns the local gradient features with particular views (HV and DG), it still suffers from information loss compared with CDC operating on the full local neighbors. In order to fully exploit the local features and interact between HV and DG views, a Dual-Cross Central Difference Network (DC-CDN)\footnote{https://github.com/ZitongYu/CDCN} is proposed. As shown in Figure 2, two independent (unshared) networks respectively assembled with C-CDC(HV) and C-CDC(DG) are used. Then the extracted dual-stream features from different views are fused for final depth prediction. In this way, the full neighbor aggregation step can be disentangled into two sub-steps: 1) sparse neighbor aggregation for individual stream; and 2) dual-stream fusion.

**Cross Feature Interaction Module.** With only simple late fusion, the performance improvement might be limited due to the lack of message passing from the preceding (i.e., low-level, mid-level, and high-level) stages. In order to effectively mine the relations across dual streams and enhance local detailed representation capacity, we propose the Cross Feature Interaction Module (CFIM) to fuse dual-stream multi-level features adaptively. To be specific, given the two-stream features $F_{HV}$ and $F_{DG}$, the CFIM enhanced features $F_{HV}'$ and $F_{DG}'$ can be formulated as

\[
F_{HV}' = \varphi(\alpha) \cdot F_{HV} + (1 - \varphi(\alpha)) \cdot F_{DG},
F_{DG}' = \varphi(\beta) \cdot F_{DG} + (1 - \varphi(\beta)) \cdot F_{HV},
\]

where $\varphi()$ denotes the Sigmoid function for $[0, 1]$ range mapping. $\alpha$ and $\beta$ are the attention weights for $F_{HV}$ and $F_{DG}$, respectively. In our default setting, both $\alpha$ and $\beta$ are initialized to 0 and learnable, which could be adaptively adjusted during the training iterations. We also investigate the manually fixed $\alpha$ and $\beta$ in Appendix B. As illustrated in Fig. 2, here we respectively plug three CFIMs (with learnable $\alpha_{low}$, $\alpha_{mid}$, $\alpha_{high}$, $\beta_{low}$, $\beta_{mid}$, $\beta_{high}$) in the output features from low-mid-high levels before multi-level concatenation.

### 3.4 Patch Exchange Augmentation

Due to the high collection cost for spoof attacks, there are limited data scale and diversity in public FAS datasets. In this paper, we also propose a FAS-dedicated data augmentation method, named Patch Exchanges (PE), to synthesize mixed samples with diverse attacks and domains. There are three advantages for PE augmentation: 1) face patches from different domains (e.g., recorded scenario, sensor, and subject) are introduced for enriching data distribution; 2) random live and PA patches are exchanged to mimic arbitrary partial attacks; and 3) the exchanged patches with corresponding dense labels enforce the model to learn more detailed and intrinsic features for spoofing detection. The complete algorithm of PE is summarized in Algorithm 1. The two hyperparameters $\gamma$ and $\rho$ control the augmentation ratio and intensity, respectively. As a tradeoff, we use empirical settings $\gamma = 0.5$ and $\rho = 2$ for experiments.

Note that as the face images $I$ for PE are coarsely aligned, the exchanged patches would have similar semantic content (e.g., cheek, nose and mouth) but with diverse live/spoof

---

**Algorithm 1** Patch Exchange Augmentation

**Input:** Face images $I$ with batchsize $N$, pseudo depth map labels $D_i$, augmented ratio $\gamma \in [0, 1]$, step number $\rho$

1: for each $I_i$ and $D_i$, $i = 1, \ldots, [\gamma \cdot N]$ do
2:     for each step $\rho$ do
3:         Randomly select a patch region $P$ within $I_i$
4:         Randomly select a batch index $j$, $j \leq N$
5:         Exchange the image patch $I_i(P) = I_j(P)$ and label patch $D_i(P) = D_j(P)$
6:     end
7: end
8: return augmented $I$ and $D$
Table 2: The ablation study about DC-CDN, CFIM, and PE augmentation on Protocol 1 of OULU-NPU.

| Model          | CFIM         | PE Augmentation | ACER(%)     |
|----------------|--------------|-----------------|-------------|
| C-CDN(HV)      |              |                 | 0.7         |
| C-CDN(DG)      |              |                 | 0.9         |
| C-CDN (average)|              |                 | 1.0         |
| DC-CDN (concat)|              |                 | 0.8         |
| C-CDN (concat) |              |                 | 0.5         |
| C-CDN(HV)      | √            |                 | 0.6         |
| C-CDN(DG)      | √            |                 | 0.7         |
| DC-CDN (concat)| √            |                 | 0.4         |

We train models with maximum 800 epochs while lr halves in the 500th epoch. Similar to [Yu et al., 2020d], all the models are supervised by mean square error (MSE) and contrastive depth loss (CDL). In the testing stage, we calculate the mean value of the predicted depth map as the final score.

4.3 Ablation Study

In this subsection, all ablation studies are conducted on the Protocol-1 (different illumination condition and scenario between training and testing sets) of OULU-NPU dataset.

Impact of θ in C-CDN. As discussed in Section 3.2, θ controls the contribution of the gradient-based features, i.e., the higher θ, the more local detailed information included. As illustrated in Fig. 4(a), when θ > 0.3, C-CDN(HV) and C-CDN(DG) always achieve better performance than their vanilla counterpart (i.e., θ = 0), indicating the effectiveness of local gradient features for FAS task. As the best performance (ACER=0.7% and 0.9% for C-CDN(HV) and C-CDN(DG), respectively) are obtained when θ = 0.8, we use this setting for the following experiments.

C-CDN vs. CDC. Here we evaluate the impacts of the neighbor sparsity for both vanilla and gradient-based convolutions. As shown in Fig. 4(b), “VanillaConv(HV)” and “VanillaConv(DG)” performs more poorly than “VanillaConv”, which might be caused by the limited semantic representation capacity from the sparse neighbor sampling. In contrast, it is interesting to see that their central difference-based counterparts perform inversely. Compared with CDC, C-CDN(HV) and C-CDN(DG) have only five ninth parameters, and even achieve better performance (-0.3% and -0.1% for protocols with θ = 0.8, respectively). This suggests that our proposed method can effectively learn the salient features for face authentication while reducing the model size.
ACER, respectively), indicating the efficiency of cross gradient features for FAS. More discussion about the sparsity and neighbor locations will be shown in Appendix A.

Effectiveness of DC-CDN and CFIM. The upper part of Table 2 shows the ablation results of DC-CDN w/ and w/o CFIM. It can be seen from the third and fourth rows that DC-CDN w/o CFIM achieves worse when using simple final depth map averaging or late fusion via concatenation. In contrast, assembled with CFIM, DC-CDN obtains 0.2% and 0.4% ACER decrease compared with one-stream C-CDN(HV) and C-CDN(DG), respectively. It demonstrates the importance of adaptive mutual feature interaction during the training stage. More discussion about the MFIM and the learned $\alpha, \beta$ visualization are shown in Appendix B.

Effectiveness of PE augmentation. It can be seen from the lower part of Table 2 that PE augmentation improves the performance of both single-stream (C-CDN) and dual-stream (DC-CDN) models on Protocol 1 (with slight domain shifts) of OULU-NPU. It is worth noting that, benefited from the augmented data with rich and diverse domain and attack clues, DC-CDN could obtain remarkable performance gains on more challenging scenarios, including cross-type (-1.2% EER, see Table 4) and cross-dataset (respective -2.5% and -2.2% HTER, see Table 5) testings. We also evaluate the generalization of PE for other architectures in Appendix C.

4.4 Comparison with State of the Arts

Intra Testing on OULU-NPU. As shown in Table 3, our proposed DC-CDN ranks first on the first three protocols (0.4%, 1.3% and 1.9% ACER, respectively), which indicates the proposed method performs well at the generalization of the external environment, attack mediums and input camera variation. It is clear that the proposed DC-CDN consistently outperforms CDCN [Yu et al., 2020d] on all protocols with -0.6%, -0.2%, -0.4%, and -2.9%, respectively, indicating the superiority of DC-CDN. In the most challenging Protocol 4, DC-CDN also achieves comparable performance with state-of-the-art SpoofTrace [Liu et al., 2020] but more robustness with smaller ACER standard deviation across six kinds of unseen scenarios.

Cross-type Testing on SiW-M. Following the leave-one-type-out (total 13 attack types) protocol on SiW-M, we compare the proposed methods with several recent FAS methods to validate the generalization capability of unseen attacks. As shown in Table 4, our method achieves the best EER performance and can perform more robustly in several challenging attacks including ‘Silicone Mask’, ‘Transparent Mask’, ‘Mannequin’, ‘Impersonation’, ‘Funny Eye’ and ‘Paper Glasses’. Note that our DC-CDN could achieve comparable performance with the state-of-the-art method BCN [Yu et al., 2020a], which needs the supervision signals from three kinds of pixel-wise labels (i.e., binary mask, pseudo depth and reflection map). Moreover, it is reasonable to see from the last two rows of Table 4 that PE augmentation helps to improve generalization ability of unknown Partial Attacks (i.e., Funny Eye, Paper Glasses and Partial Paper) obviously.

Cross-dataset Testing. Here cross-dataset testing is conducted to further testify the generalization ability of our models under unseen scenarios. There are two cross-dataset testing protocols. One is that training on the CASIA-MFSD and testing on Replay-Attack, which is named as protocol CR; the second one is exchanging the training dataset and the testing dataset, named protocol RC. As shown in Table 5, our proposed DC-CDN achieves 6.0% HTER on protocol CR, outperforming the prior state-of-the-art by a convincing margin of 9.5%. For protocol RC, we also slightly outperform state-of-the-art frame-level methods (e.g., CDCN, BCN and Disentangled [Zhang et al., 2020]). Note that inspired by the dynamic clues usage in Auxiliary [Liu et al., 2018] and FASSGTD [Wang et al., 2020], the performance of DC-CDN might be further boosted via introducing the similar temporal dynamic features, which will be explored in our future work.

5 Conclusions

In this paper, we propose a sparse operator family called Cross Central Difference Convolution (C-CDC) for face anti-spoofing task. Moreover, we design a Dual-Cross Central Difference Network with Cross Feature Interaction Modules for dual-stream feature enhancement. Besides, a novel Patch Exchange augmentation strategy is proposed for enriching the
training data. Extensive experiments are performed to verify the effectiveness of the proposed methods. In the future, we will extend the C-CDC family to a 3D version for fine-grained video understanding tasks.

Appendix

A. Impacts of Sparsity and Neighbor Locations of C-CDC

As already shown in the Section 4.3, C-CDC(HV) and C-CDC(DG) adopt the symmetric sparse local neighbor sampling and even achieve better performance than CDC. Thus, two questions arise: 1) the sparser, the better? 2) is symmetry helpful for feature representation?

In order to tackle these questions, we compare C-CDC with three CDC with sparse/asymmetric neighbor sampling (i.e., “S1-CDC”, “S2-CDC”, “S3-CDC”). The results intra-tested on Protocol 1 of OULU-NPU are illustrated in Figure 5. In terms of the sparsity, the “S2-CDC” and “S3-CDC” respectively consider only three and two local neighbors for sampling, which is much sparser than both CDC and C-CDC. However, they suffer from sharp performance drops due to the limited local representation capacity. As for the symmetry, we compare the C-CDC with “S1-CDC” as the latter has the same sparsity but asymmetric neighbor locations. It is interesting to see that “S1-CDC” still has comparable performance (1% ACER) with CDC but slightly poorer results than the symmetric C-CDC(HV) and C-CDC(DG).

Despite with such coarse examples for comparison, it still provides evidence that the designed C-CDC family is an excellent tradeoff in terms of performance and efficiency.

B. Discussion about CFIM

As CFIM plays a vital role in mutual relation mining and dual-stream feature enhancement, it is necessary to explore more details about CFIM, e.g., 1) learned $\alpha$, $\beta$ visualization; 2) impacts of learnable/fixed setting and initialization.

As illustrated in Figure 6, with the defaulted initialization to all zeros (i.e., $\alpha = 0, \beta = 0, \varphi(\alpha), \varphi(\beta) = 0.5$), the trends of $\alpha_{low}, \alpha_{mid}, \alpha_{high}, \beta_{mid}, \beta_{high}$ are relatively steady except $\beta_{low}$. To be specific, $\alpha_{low}, \beta_{mid}, \beta_{high}$ prefer to hold the positive values while $\alpha_{mid}, \alpha_{high}, \beta_{low}$ keep the negative values in most times. The dramatic decrease of $\beta_{low}$ denotes that contributions from the low-level features of the C-CDC(DG) stream might be small in DC-CDN.

Moreover, it can be seen from Figure 7 that the initialization values influence the CFIM a lot especially with learnable setting. Small initial values (e.g., $\alpha, \beta = [0, 1]$) seem to be more suitable for subsequent dual-stream feature learning. The best performance could be achieved (0.5% ACER in Protocol 1 of OULU-NPU) when $\alpha, \beta$ are learnable and initialized as zeros.

C. Analysis of PE Augmentation

In the paper, although the PE Augmentation is proven to be helpful for DC-CDN performance improvement, it is still necessary to evaluate its generalization ability for different architectures under various scenarios. Besides DC-CDN, here we also give comparisons of two classical FAS architectures, i.e., DepthNet [Liu et al., 2018], CDCN [Yu et al., 2020d] on intra- and cross-type testings. As shown in Figure 8, PE augmentation consistently improves DepthNet and CDCN when the training data are even small-scale (Protocol 4 of OULU-NPU) or testing under unseen attack types (SiW-M). All these results demonstrate that PE augmentation is promising for the FAS task.

As partial attacks (e.g., partial print, mask and funny eye glasses) have been used for hacking the face recognition system, here we show a visualization about the predictions from DC-CDN trained w/ and w/o PE augmentation. The models are trained and tested on the sub-Protocol 4-1 of OULU-NPU. As illustrated in the first three columns of Figure 9, both the models (trained w/ or w/o PE augmentation) perform well for the normal live face and print attacks. However, we can see from the last column that w/o PE augmentation, it fails to capture the spoofing regions from the synthesized face with partial attacks. In contrast, the model trained w/ PE could clearly discover and localize them.
Figure 7: Comparison of the learnable and fixed settings of $\alpha$, $\beta$ in CFIM with different initialized values on Protocol 1 of OULU-NPU.

Figure 8: Ablation study about the PE augmentation for different architectures and testing protocols.

Figure 9: Visualization of the predicted results from DC-CDN w/ and w/o PE augmentation. When given the inputs with partial print attacks, the model trained w/ PE is able to discover the spoofing patches while that trained w/o PE fails.

References

[Boulkenafet et al., 2015] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdennour Hadid. Face anti-spoofing based on color texture analysis. In ICIP, 2015.

[Boulkenafet et al., 2017] Zinelabidine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdennour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In FGR, 2017.

[Chingovska et al., 2012] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In Biometrics Special Interest Group, 2012.

[Dai et al., 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In CVPR, 2017.

[De Freitas Pereira et al., 2012] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In ACCV, 2012.

[DeVries and Taylor, 2017] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

[George and Marcel, 2019] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In ICIE, number CONF, 2019.

[Guo et al., 2020] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In ECCV, 2020.

[Hu et al., 2019] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In CVPR, 2019.

[ISO, 2016] ISO. Iso/iec jtc 1/sc 37 biometrics: Information technology biometric presentation attack detection part 1: Framework. In https://www.iso.org/obp/ui/iso, 2016.

[Jourabloo et al., 2018] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In ECCV, 2018.

[Juefei-Xu et al., 2017] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In CVPR, 2017.

[Komulainen et al., 2013] Jukka Komulainen, Abdennour Hadid, and Matti Pietikainen. Context based face anti-spoofing. In BTAS, 2013.

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

[Liu et al., 2019] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In CVPR, pages 4680–4689, 2019.

[Liu et al., 2020] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In ECCV. Springer, 2020.
[Luan et al., 2018] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu. Gabor convolutional networks. TIP, 27(9), 2018.

[Parmar et al., 2019] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In NeurIPS, 2019.

[Patel et al., 2016] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. TIFS, 11(10), 2016.

[Qin et al., 2020] Yunxiao Qin, Chenxu Zhao, Xiangyu Zhu, Zezheng Wang, Zitong Yu, Tianyu Fu, Feng Zhou, Jingping Shi, and Zhen Lei. Learning meta model for zero-and few-shot face anti-spoofing. In AAAI, 2020.

[Wang et al., 2020] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In CVPR, 2020.

[Yang et al., 2019] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhihong Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In CVPR, 2019.

[Yu and Koltun, 2015] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[Yu et al., 2020a] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In ECCV. Springer, 2020.

[Yu et al., 2020b] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In CVPRW, 2020.

[Yu et al., 2020c] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. IEEE TPAMI, 2020.

[Yu et al., 2020d] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In CVPR, 2020.

[Yu et al., 2021] Zitong Yu, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Revisiting pixel-wise supervision for face anti-spoofing. IEEE TBIOM, 2021.

[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., 2020] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In ECCV. Springer, 2020.

[Zhao et al., 2020] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In CVPR, 2020.