Dual-Cross Central Difference Network for Face Anti-Spoofing

Zitong Yu[1] , **Yunxiao Qin**[2] , **Hengshuang Zhao**[3] , **Xiaobai Li**[1] and **Guoying Zhao**[1*]
[1]CMVS, University of Oulu
[2]Northwestern Polytechnical University
[3]University of Oxford
{zitong.yu, xiaobai.li, guoying.zhao}@oulu.fi, qyxqyx@mail.nwpu.edu.cn,
hengshuang.zhao@eng.ox.ac.uk

Abstract

Face anti-spoofing (FAS) plays a vital role in securing face recognition systems. Recently, central difference convolution (CDC) [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.

1 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 [de Freitas Pereira *et al.*, 2012; Komulainen *et al.*, 2013; Patel *et al.*, 2016] and deep learning-based [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) [Boulkenafet *et al.*, 2015] and histogram of gradient (HOG) [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) [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 [Boulkenafet *et al.*, 2017] and SiW-M [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 [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

*Corresponding author. This work was supported by the Academy of Finland for project MiGA (grant 316765), ICT 2023 project (grant 328115), Infotech Oulu, project 6+E (grant 323287) funded by Academy of Finland, and project PhInGAIN (grant 200414) funded by The Finnish Work Environmental Fund. The authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

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, thus binary cross entropy loss is used for model supervision. On the other hand, according to the physical discrepancy between live and spoof faces, dense pixel-wise supervisions [Yu *et al.*, 2021] such as pseudo depth map [Liu *et al.*, 2018; Yu *et al.*, 2020d; Wang *et al.*, 2020], reflection map [Yu *et al.*, 2020a], texture map [Zhang *et al.*, 2020] and binary map [George and Marcel, 2019] are designed for fine-grained learning. In this work, we supervise the deep networks with pseudo depth map due to its effectiveness.

Due to the high collection cost for spoof attacks, there are limited data scale and diversity in public datasets. Supervised with small-scale predefined scenarios and PAs, most existing FAS methods are easy to overfit and vulnerable to domain shift and unseen attacks. In order to detect unseen attacks successfully, deep tree network [Liu *et al.*, 2019] and adaptive inner-update meta learning [Qin *et al.*, 2020] are developed for zero-shot FAS. However, it is still urgent to provide larger-scale and richer live/spoof data for deep models training. Here we consider novel data augmentation for FAS to tackle this challenge.

Convolution Operators. The convolution operator is commonly used for local feature representation in modern deep learning framework. Recently, a few extensions to the vanilla convolution operator have been proposed. In one direction, pre-defined or learnable local relation is embedded in the convolution operator. Representative works include Local Binary Convolution [Juefei-Xu *et al.*, 2017] and Gabor Convolution [Luan *et al.*, 2018], which is proposed for local invariance preservation and enhancing the resistance to spatial changes, respectively. Besides, self-attention layer [Parmar *et al.*, 2019], self-attention block [Zhao *et al.*, 2020] and local relation layer [Hu *et al.*, 2019] are designed for mining the local relationship flexibly. Another direction is to modify the spatial scope for aggregation. Two related works are dialated convolution [Yu and Koltun, 2015] and deformable convolution [Dai *et al.*, 2017]. However, these convolution operators may not be suitable for FAS task because of the limited representation capacity for invariant fine-grained features. In contrast, CDC [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 $\mathcal{R}$ 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

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n), \tag{1}$$

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

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

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)). \tag{2}$$

As both intensity-level semantic information and gradient-level detailed clues are crucial for robust FAS, the generalized

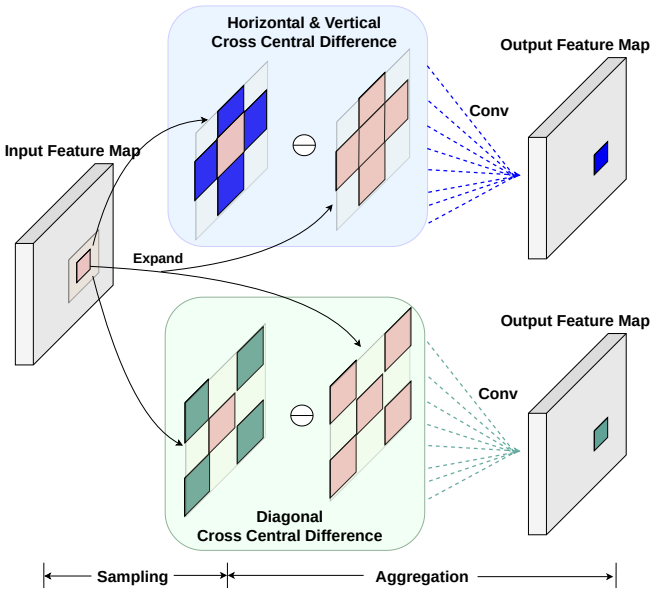

Figure 1: Cross central difference convolution. The C-CDC(HV) in the upper part calculates the central gradients from the horizontal & vertical neighbors while the C-CDC(DG) in the lower part from the diagonal neighbors.

CDC operator can be represented by combination of vanilla convoluiton and CDC

$$
\begin{aligned}
y(p_0) &= \theta \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}} \\
&\quad + (1-\theta) \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}}, \\
&= \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}} + \underbrace{\theta \cdot (-x(p_0) \cdot \sum_{p_n \in \mathcal{R}} w(p_n))}_{\text{central difference term}},
\end{aligned} \tag{3}
$$

where hyperparameter $\theta \in [0, 1]$ trade-offs the contribution between intensity-level and gradient-level information. The higher value of θ 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 $\mathcal{R}$, 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.

Output	DepthNet	CDCN	C-CDN (Ours)
256×256	3×3 conv, 64	3×3 CDC, 64	3×3 **C-CDC**, 64
128×128 (Low)	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ CDC}, 196 \\ 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \textbf{ C-CDC}, 196 \\ 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
64×64 (Mid)	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ CDC}, 196 \\ 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \textbf{ C-CDC}, 196 \\ 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
32×32 (High)	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ CDC}, 196 \\ 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \textbf{ C-CDC}, 196 \\ 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
32×32	[concat (Low, Mid, High), 384]		
32×32	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 64 \\ 3 \times 3 \text{ conv}, 1 \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ CDC}, 128 \\ 3 \times 3 \text{ CDC}, 64 \\ 3 \times 3 \text{ CDC}, 1 \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \textbf{ C-CDC}, 128 \\ 3 \times 3 \textbf{ C-CDC}, 64 \\ 3 \times 3 \textbf{ C-CDC}, 1 \end{bmatrix}$
# params	2.25×10^6	2.25×10^6	1.25×10^6
# FLOPs	4.8×10^{10}	4.8×10^{10}	1.78×10^8

Table 1: Architectures of DepthNet, CDCN, and the proposed C-CDN. Inside the brackets are the filter sizes and feature dimensionalities. '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 $\mathcal{R}$, C-CDC prefers to sample a sparser local region $\mathcal{S}$, which can be formulated as

$$
y(p_0) = \sum_{p_n \in \mathcal{S}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)). \tag{4}
$$

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

In terms of designs of the sparse local region $\mathcal{S}$, 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 $\mathcal{S}$ 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.

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 3×3 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 32×32. The details of C-CDN are shown in Table 1. It can be seen that

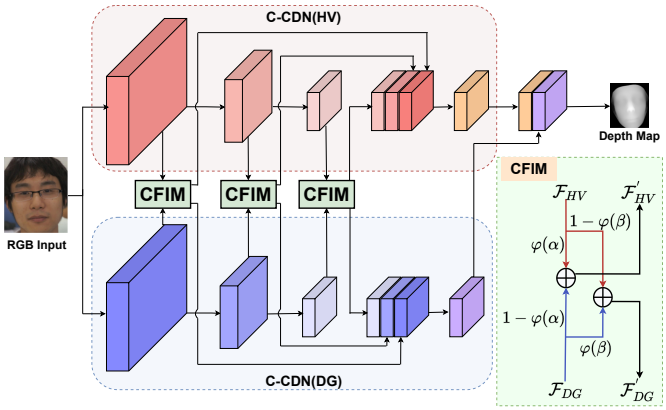

Figure 2: Dual-cross central difference network with Cross Feature Interaction Modules for low-mid-high level feature enhancement.

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 θ 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) 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 $\mathcal{F}_{HV}$ and $\mathcal{F}_{DG}$, the CFIM enhanced features $\mathcal{F}^{'}_{HV}$ and $\mathcal{F}^{'}_{DG}$ can be formulated as

$$\begin{aligned} \mathcal{F}^{'}_{HV} &= \varphi(\alpha) \cdot \mathcal{F}_{HV} + (1 - \varphi(\alpha)) \cdot \mathcal{F}_{DG}, \\ \mathcal{F}^{'}_{DG} &= \varphi(\beta) \cdot \mathcal{F}_{DG} + (1 - \varphi(\beta)) \cdot \mathcal{F}_{HV}, \end{aligned} \tag{5}$$

where $\varphi()$ denotes the Sigmoid function for [0,1] range mapping. α and β are the attention weights for $\mathcal{F}_{HV}$ and $\mathcal{F}_{DG}$, respectively. In our default setting, both α and β are initialized to 0 and learnable, which could be adaptively adjusted during the training iterations. As illustrated in Fig. 2, here we respectively plug three CFIMs (with learnable α_{low}, α_{mid}, α_{high}, β_{low}, β_{mid}, β_{high}) in the output features from low-mid-high levels before multi-level concatenation.

Algorithm 1 Patch Exchange Augmentation

Input: Face images I with batchsize N, pseudo depth map labels D, augmented ratio $\gamma \in [0, 1]$, step number ρ
1 : for each I_i and D_i, $i = 1, ..., \lfloor \gamma * N \rfloor$ **do**
2 : **for** each step ρ **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

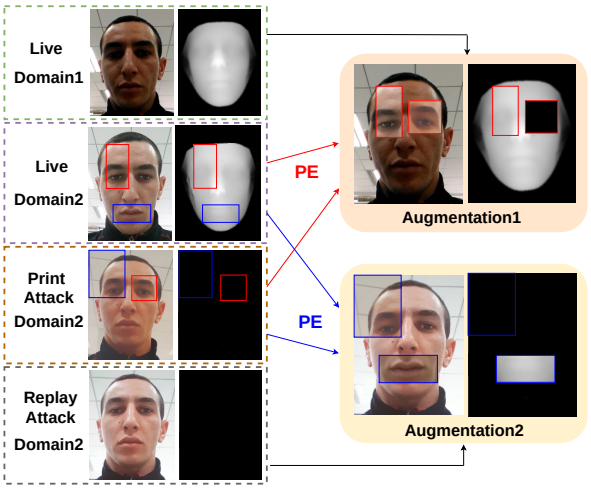

Figure 3: Patch Exchange augmentation. Two augmented samples are synthesized via exchanging the RGB patches as well as corresponding pseudo depth labels from "Live Domain2" and "Print Attack Domain2". Thus, the samples "Augmentation1" and "Augmentation2" contain diverse domains and attack types, respectively.

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. 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 clues. Thus, the augmented live/spoof faces are still realistic and even more challenging to be distinguished. Some typical samples with PE augmentation are visualized in Figure 3.

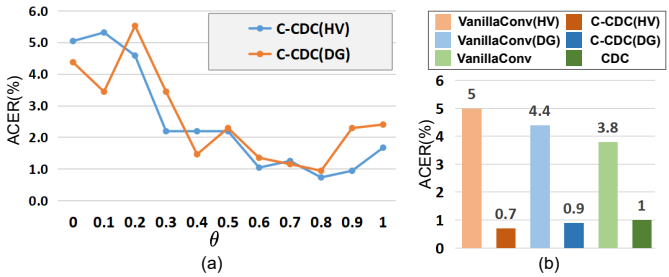

Figure 4: (a) Impact of θ in C-CDN. (b) Comparison among various convolutions. The lower ACER, the better performance.

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

Table 2: The ablation study about DC-CDN, CFIM, and PE augmentation on Protocol 1 of OULU-NPU.

4 Experiments

4.1 Datasets and Metrics

Databases. Four databases OULU-NPU [Boulkenafet *et al.*, 2017], CASIA-MFSD [Zhang *et al.*, 2012], Replay-Attack [Chingovska *et al.*, 2012] and SiW-M [Liu *et al.*, 2019] are used in our experiments. OULU-NPU is a high-resolution database, containing four protocols to evaluate the generalization (e.g., unseen illumination and attack medium) of models respectively, which is used for intra testing. CASIA-MFSD and Replay-Attack are small-scale databases with low-resolution videos, which are used for cross testing. SiW-M is designed for cross-type testing for unseen attacks as there are rich (13) attack types inside.

Performance Metrics. In OULU-NPU dataset, we follow the original protocols and metrics, i.e., Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and ACER for a fair comparison. Half Total Error Rate (HTER) is adopted in the cross testing between CASIA-MFSD and Replay-Attack. For the cross-type test on SiW-M, ACER and Equal Error Rate (EER) are employed for evaluation.

4.2 Implementation Details

Depth Generation. Dense face alignment [Guo *et al.*, 2020] is adopted for pseudo depth generation. To clearly distinguish live and spoof faces, at the training stage, we follow [Liu *et al.*, 2018] to normalize the live depth maps in a range of $[0, 1]$, while leaving spoof depth maps to all zeros.

Training and Testing Setting. Our proposed method is implemented with Pytorch. In the training stage, models are trained with batch size 8 and Adam optimizer on a single V100 GPU. Data augmentations including horizontal flip, color jitter and Cutout are used as baseline. The initial learning rate (lr) and weight decay are 1e-4 and 5e-5, respectively. 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.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	STASN [Yang *et al.*, 2019]	1.2	2.5	1.9
	Auxiliary [Liu *et al.*, 2018]	1.6	1.6	1.6
	FaceDs [Jourabloo *et al.*, 2018]	1.2	1.7	1.5
	SpoofTrace [Liu *et al.*, 2020]	0.8	1.3	1.1
	Disentangled [Zhang *et al.*, 2020]	1.7	0.8	1.3
	FAS-SGTD [Wang *et al.*, 2020]	2.0	0.0	1.0
	CDCN [Yu *et al.*, 2020d]	0.4	1.7	1.0
	BCN [Yu *et al.*, 2020a]	0.0	1.6	0.8
	DC-CDN (Ours)	0.5	0.3	**0.4**
2	FaceDs [Jourabloo *et al.*, 2018]	4.2	4.4	4.3
	Auxiliary [Liu *et al.*, 2018]	2.7	2.7	2.7
	Disentangled [Zhang *et al.*, 2020]	1.1	3.6	2.4
	STASN [Yang *et al.*, 2019]	4.2	0.3	2.2
	BCN [Yu *et al.*, 2020a]	2.6	0.8	1.7
	SpoofTrace [Liu *et al.*, 2020]	2.3	1.6	1.9
	FAS-SGTD [Wang *et al.*, 2020]	2.5	1.3	1.9
	CDCN [Yu *et al.*, 2020d]	1.5	1.4	1.5
	DC-CDN (Ours)	0.7	1.9	**1.3**
3	FaceDs [Jourabloo *et al.*, 2018]	4.0±1.8	3.8±1.2	3.6±1.6
	Auxiliary [Liu *et al.*, 2018]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [Yang *et al.*, 2019]	4.7±3.9	0.9±1.2	2.8±1.6
	SpoofTrace [Liu *et al.*, 2020]	1.6 ±1.6	4.0±5.4	2.8±3.3
	FAS-SGTD [Wang *et al.*, 2020]	3.2±2.0	2.2±1.4	2.7±0.6
	BCN [Yu *et al.*, 2020a]	2.8±2.4	2.3±2.8	2.5±1.1
	CDCN [Yu *et al.*, 2020d]	2.4±1.3	2.2±2.0	2.3±1.4
	Disentangled [Zhang *et al.*, 2020]	2.8±2.2	1.7±2.6	2.2±2.2
	DC-CDN (Ours)	2.2±2.8	1.6±2.1	**1.9±1.1**
4	Auxiliary [Liu *et al.*, 2018]	9.3±5.6	10.4±6.0	9.5±6.0
	STASN [Yang *et al.*, 2019]	6.7±10.6	8.3±8.4	7.5±4.7
	CDCN [Yu *et al.*, 2020d]	4.6±4.6	9.2±8.0	6.9±2.9
	FaceDs [Jourabloo *et al.*, 2018]	1.2±6.3	6.1±5.1	5.6±5.7
	BCN [Yu *et al.*, 2020a]	2.9±4.0	7.5±6.9	5.2±3.7
	FAS-SGTD [Wang *et al.*, 2020]	6.7±7.5	3.3±4.1	5.0±2.2
	Disentangled [Zhang *et al.*, 2020]	5.4±2.9	3.3±6.0	4.4±3.0
	SpoofTrace [Liu *et al.*, 2020]	2.3±3.6	5.2±5.4	**3.8±4.2**
	DC-CDN (Ours)	5.4±3.3	2.5±4.2	4.0±3.1

Table 3: The results of intra testing on the OULU-NPU dataset.

4.3 Ablation Study

In this subsection, all ablation studies are conducted on the Protocol-1 of OULU-NPU dataset.

Impact of θ in C-CDC. 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 $\theta \geqslant 0.3$, C-CDC(HV) and C-CDC(DG) always achieve better performance than their vanilla counterpart (i.e., $\theta = 0$), indicating the effectivenss of local gradient features for FAS task. As the best performance (ACER=0.7% and 0.9% for C-CDC(HV) and C-CDC(DG), respectively) are obtained when $\theta = 0.8$, we use this setting for the following experiments.

C-CDC 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-CDC(HV) and C-CDC(DG) have only five ninth parameters, and even achieve better performance (-0.3% and -0.1%

Method	Metrics(%)	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Average
				Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Im.	Cos.	Fun.	Glasses	Partial	
Auxiliary [Liu *et al.*, 2018]	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	23.6±18.5
	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	**9.4**	21.4	18.6	4.0	17.0±17.7
DTN [Liu *et al.*, 2019]	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	**19.3**	19.8	8.5	16.8 ±11.1
	EER	10.0	**2.1**	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	16.1± 12.2
CDCN [Yu *et al.*, 2020d]	ACER	8.7	7.7	11.1	9.1	20.7	4.5	5.9	44.2	2.0	15.1	25.4	19.6	3.3	13.6 ±11.7
	EER	8.2	7.8	8.3	**7.4**	20.5	5.9	5.0	47.8	1.6	14.0	24.5	18.3	**1.1**	13.1± 12.6
SpoofTrace [Liu *et al.*, 2020]	ACER	**7.8**	7.3	**7.1**	12.9	**13.9**	4.3	6.7	53.2	4.6	19.5	20.7	21.0	5.6	14.2 ±13.2
	EER	**7.6**	3.8	8.4	13.8	14.5	**5.3**	4.4	35.4	**0.0**	19.3	21.0	20.8	1.6	12.0± 10.0
BCN [Yu *et al.*, 2020a]	ACER	12.8	**5.7**	10.7	10.3	14.9	**1.9**	2.4	**32.3**	0.8	12.9	22.9	16.5	**1.7**	**11.2 ±9.2**
	EER	13.4	5.2	**8.3**	9.7	13.6	5.8	2.5	**33.8**	**0.0**	14.0	23.3	16.6	1.2	11.3± 9.5
DC-CDN w/o PE (Ours)	ACER	12.9	10.2	9.7	9.1	16.5	5.3	**1.6**	44.6	0.8	14.0	22.9	17.3	3.8	12.9 ±11.6
	EER	12.3	8.7	12.6	**7.4**	13.6	5.9	**0.0**	43.4	**0.0**	14.0	19.5	16.7	2.3	12.0± 11.3
DC-CDN (Ours)	ACER	12.1	9.7	14.1	**7.2**	14.8	4.5	**1.6**	40.1	**0.4**	**11.4**	20.1	**16.1**	2.9	11.9±10.3
	EER	10.3	8.7	11.1	**7.4**	**12.5**	5.9	**0.0**	39.1	**0.0**	12.0	**18.9**	**13.5**	1.2	**10.8±10.1**

Table 4: Results of the cross-type testing on the SiW-M dataset.

Method	Train	Test	Train	Test
	CASIA-MFSD	Replay-Attack	Replay-Attack	CASIA-MFSD
FaceDs [Jourabloo *et al.*, 2018]	28.5		41.1	
STASN [Yang *et al.*, 2019]	31.5		30.9	
Auxiliary [Liu *et al.*, 2018]	27.6		28.4	
Disentangled [Zhang *et al.*, 2020]	22.4		30.3	
FAS-SGTD [Wang *et al.*, 2020]	17.0		**22.8**	
BCN [Yu *et al.*, 2020a]	16.6		36.4	
CDCN [Yu *et al.*, 2020d]	15.5		32.6	
DC-CDN w/o PE (Ours)	8.5		32.3	
DC-CDN (Ours)	**6.0**		30.1	

Table 5: The results of cross-dataset testing between CASIA-MFSD and Replay-Attack. The evaluation metric is HTER(%).

ACER, respectively), indicating the efficiency of cross gradient features for FAS.

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.

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 senarios, including cross-type (-1.2% EER, see Table 4) and cross-dataset (respective -2.5% and -2.2% HTER, see Table 5) testings.

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 among 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 capacity 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 SOTA method BCN [Yu *et al.*, 2020a], which is supervised by three kinds of pixel-wise labels. 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 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 and BCN).

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.

References

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

[Boulkenafet *et al.*, 2017] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour 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 *ICB*, 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.

[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, Abdenour 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 *NeurPIS*, 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, Zhifeng 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.