Searching Central Difference Convolutional Networks for Face Anti-Spoofing

Zitong Yu[1], Chenxu Zhao[2], Zezheng Wang[3], Yunxiao Qin[4],
Zhuo Su[1], Xiaobai Li[1], Feng Zhou[3], Guoying Zhao[1*]

[1]CMVS, University of Oulu   [2]Mininglamp Academy of Sciences, Mininglamp Technology
[3]Aibee   [4]Northwestern Polytechnical University

{zitong.yu, zhuo.su, xiaobai.li, guoying.zhao}@oulu.fi,
{zhaochenxu}@mininglamp.com,{zezhengwang, fzhoug}@aibee.com, {qyxqyx}@mail.nwpu.edu.cn

Abstract

Face anti-spoofing (FAS) plays a vital role in face recognition systems. Most state-of-the-art FAS methods 1) rely on stacked convolutions and expert-designed network, which is weak in describing detailed fine-grained information and easily being ineffective when the environment varies (e.g., different illumination), and 2) prefer to use long sequence as input to extract dynamic features, making them difficult to deploy into scenarios which need quick response. Here we propose a novel frame level FAS method based on Central Difference Convolution (CDC), which is able to capture intrinsic detailed patterns via aggregating both intensity and gradient information. A network built with CDC, called the Central Difference Convolutional Network (CDCN), is able to provide more robust modeling capacity than its counterpart built with vanilla convolution. Furthermore, over a specifically designed CDC search space, Neural Architecture Search (NAS) is utilized to discover a more powerful network structure (CDCN++), which can be assembled with Multiscale Attention Fusion Module (MAFM) for further boosting performance. Comprehensive experiments are performed on six benchmark datasets to show that 1) the proposed method not only achieves superior performance on intra-dataset testing (especially 0.2% ACER in Protocol-1 of OULU-NPU dataset), 2) it also generalizes well on cross-dataset testing (particularly 6.5% HTER from CASIA-MFSD to Replay-Attack datasets). The codes are available at https://github.com/ZitongYu/CDCN.

1. Introduction

Face recognition has been widely used in many interactive artificial intelligence systems for its convenience. However, vulnerability to presentation attacks (PA) curtail its reliable deployment. Merely presenting printed images or videos to the biometric sensor could fool face recognition
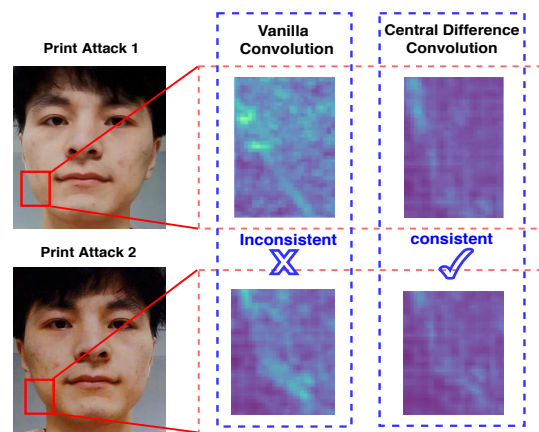
---
*denotes corresponding author

Figure 1. Feature response of vanilla convolution (VanillaConv) and central difference convolution (CDC) for spoofing faces in shifted domains (illumination & input camera). VanillaConv fails to capture the consistent spoofing pattern while CDC is able to extract the invariant detailed spoofing features, e.g., **lattice artifacts**.

systems. Typical examples of presentation attacks are print, video replay, and 3D masks. For the reliable use of face recognition systems, face anti-spoofing (FAS) methods are important to detect such presentation attacks.

In recent years, several hand-crafted features based [7, 8, 15, 29, 45, 44] and deep learning based [49, 64, 36, 26, 62, 4, 19, 20] methods have been proposed for presentation attack detection (PAD). On one hand, the classical hand-crafted descriptors (e.g., local binary pattern (LBP) [7]) leverage local relationship among the neighbours as the discriminative features, which is robust for describing the detailed invariant information (e.g., color texture, moiré pattern and noise artifacts) between the living and spoofing faces. On the other hand, due to the stacked convolution operations with nonlinear activation, the convolutional neural networks (CNN) hold strong representation abilities to distinguish the bona fide and PA. However, CNN based methods focus on the deeper semantic features, which are weak in describing detailed fine-grained information between liv-

ing and spoofing faces and easily being ineffective when the environment varies (e.g., different light illumination). **How to integrate local descriptors with convolution operation for robust feature representation is worth exploring.**

Most recent deep learning based FAS methods are usually built upon image classification task based backbones [61, 62, 20], such as VGG [54], ResNet [22] and DenseNet [23]. The networks are usually supervised by binary cross-entropy loss, which easily learns the arbitrary patterns such as screen bezel instead of the nature of spoofing patterns. In order to solve this issue, several depth supervised FAS methods [4, 36], which utilize pseudo depth map label as auxiliary supervised signal, have been developed. However, all these network architectures are carefully designed by human experts, which might not be optimal for FAS task. Hence, **to automatically discover best-suited networks for FAS task with auxiliary depth supervision should be considered.**

Most existing state-of-the-art FAS methods [36, 56, 62, 32] need multiple frames as input to extract dynamic spatio-temporal features (e.g., motion [36, 56] and rPPG [62, 32]) for PAD. However, long video sequence may not be suitable for specific deployment conditions where the decision needs to be made quickly. Hence, frame level PAD approaches are advantageous from the usability point of view despite inferior performance compared with video level methods. **To design high-performing frame level methods is crucial for real-world FAS applications.**

Motivated by the discussions above, we propose a novel convolution operator called Central Difference Convolution (CDC), which is good at describing fine-grained invariant information. As shown in Fig. 1, CDC is more likely to extract intrinsic spoofing patterns (e.g., lattice artifacts) than vanilla convolution in diverse environments. Furthermore, over a specifically designed CDC search space, Neural Architecture Search (NAS) is utilized to discover the excellent frame level networks for depth supervised face anti-spoofing task. Our contributions include:

- We design a novel convolution operator called Central Difference Convolution (CDC), which is suitable for FAS task due to its remarkable representation ability for invariant fine-grained features in diverse environments. Without introducing any extra parameters, CDC can replace the vanilla convolution and plug and play in existing neural networks to form Central Difference Convolutional Networks (CDCN) with more robust modeling capacity.

- We propose CDCN++, an extended version of CDCN, consisting of the searched backbone network and Multiscale Attention Fusion Module (MAFM) for aggregating the multi-level CDC features effectively.

- To our best knowledge, this is the first approach that searches neural architectures for FAS task. Different

from the previous classification task based NAS supervised by softmax loss, we search the well-suited frame level networks for depth supervised FAS task over a specifically designed CDC search space.

- Our proposed method achieves state-of-the-art performance on all six benchmark datasets with both intra- as well as cross-dataset testing.

2. Related Work

Face Anti-Spoofing. Traditional face anti-spoofing methods usually extract hand-crafted features from the facial images to capture the spoofing patterns. Several classical local descriptors such as LBP [7, 15], SIFT [44], SURF [9], HOG [29] and DoG [45] are utilized to extract frame level features while video level methods usually capture dynamic clues like dynamic texture [28], micro-motion [53] and eye blinking [41]. More recently, a few deep learning based methods are proposed for both frame level and video level face anti-spoofing. For frame level methods [30, 43, 20, 26], pre-trained deep CNN models are fine-tuned to extract features in a binary-classification setting. In contrast, auxiliary depth supervised FAS methods [4, 36] are introduced to learn more detailed information effectively. On the other hand, several video level CNN methods are presented to exploit the dynamic spatio-temporal [56, 62, 33] or rPPG [31, 36, 32] features for PAD. Despite achieving state-of-the-art performance, video level deep learning based methods need long sequence as input. In addition, compared with traditional descriptors, CNN overfits easily and is hard to generalize well on unseen scenes.

Convolution Operators. The convolution operator is commonly used in extracting basic visual features in deep learning framework. Recently extensions to the vanilla convolution operator have been proposed. In one direction, classical local descriptors (e.g., LBP [2] and Gabor filters [25]) are considered into convolution design. Representative works include Local Binary Convolution [27] and Gabor Convolution [38], which is proposed for saving computational cost and enhancing the resistance to the spatial changes, respectively. Another direction is to modify the spatial scope for aggregation. Two related works are dialated convolution [63] and deformable convolution [14]. However, these convolution operators may not be suitable for FAS task because of the limited representation capacity for invariant fine-grained features.

Neural Architecture Search. Our work is motivated by recent researches on NAS [11, 17, 35, 47, 68, 69, 60], while we focus on searching for a depth supervised model with high performance instead of a binary classification model for face anti-spoofing task. There are three main categories of existing NAS methods: 1) Reinforcement learning based [68, 69], 2) Evolution algorithm based [51, 52], and 3) Gradient based [35, 60, 12]. Most of NAS approaches search

networks on a small proxy task and transfer the found architecture to another large target task. For the perspective of computer vision applications, NAS has been developed for face recognition [67], action recognition [46], person ReID [50], object detection [21] and segmentation [65] tasks. To the best of our knowledge, no NAS based method has ever been proposed for face anti-spoofing task.

In order to overcome the above-mentioned drawbacks and fill in the blank, we search the frame level CNN over a specially designed search space with the new proposed convolution operator for depth-supervised FAS task.

3. Methodology

In this section, we will first introduce our Central Difference Convolution in Section 3.1, then introduce the **C**entral **D**ifference **C**onvolutional **N**etworks (CDCN) for face anti-spoofing in Section 3.2, and at last present the searched networks with attention mechanism (CDCN++) in Section 3.3.

3.1. Central Difference Convolution

In modern deep learning frameworks, the feature maps and convolution are represented in 3D shape (2D spatial domain and extra channel dimension). As the convolution operation remains the same across the channel dimension, for simplicity, in this subsection the convolutions are described in 2D while extension to 3D is straightforward.

Vanilla Convolution. As 2D spatial convolution is the basic operation in CNN for vision tasks, here we denote it as vanilla convolution and review it shortly first. There are two main steps in the 2D convolution: 1) *sampling* local receptive field region $\mathcal{R}$ over the input feature map x; 2) *aggregation* of sampled values via weighted summation. Hence, the output feature map y can be formulated as

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

where p_0 denotes 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 operation with 3×3 kernel and dilation 1 is $\mathcal{R} = \{(-1, -1), (-1, 0), \cdots, (0, 1), (1, 1)\}$.

Vanilla Convolution Meets Central Difference. Inspired by the famous local binary pattern (LBP) [7] which describes local relations in a binary central difference way, we also introduce central difference into vanilla convolution to enhance its representation and generalization capacity. Similarly, central difference convolution also consists of two steps, i.e., *sampling* and *aggregation*. The sampling step is similar to that in vanilla convolution while the aggregation step is different: as illustrated in Fig. 2, central difference convolution prefers to aggregate the center-oriented gradient of sampled values. Eq. (1) becomes
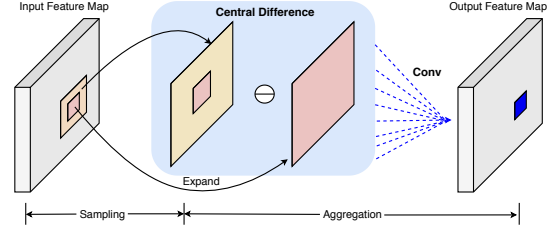

Figure 2. Central difference convolution.

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

When $p_n = (0, 0)$, the gradient value always equals to zero with respect to the central location p_0 itself.

For face anti-spoofing task, both the intensity-level semantic information and gradient-level detailed message are crucial for distinguishing the living and spoofing faces, which indicates that combining vanilla convolution with central difference convolution might be a feasible manner to provide more robust modeling capacity. Therefore we generalize central difference convolution as

$$y(p_0) = \theta \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}}$$
$$+ (1 - \theta) \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}}, \qquad (3)$$

where hyperparameter $\theta \in [0, 1]$ tradeoffs the contribution between intensity-level and gradient-level information. The higher value of θ means the more importance of central difference gradient information. We will henceforth refer to this generalized **C**entral **D**ifference **C**onvolution as **CDC**, which should be easy to identify according to its context.

Implementation for CDC. In order to efficiently implement CDC in modern deep learning framework, we decompose and merge Eq. (3) into the vanilla convolution with additional central difference term

$$y(p_0) = \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}} + \theta \cdot \underbrace{\left(-x(p_0) \cdot \sum_{p_n \in \mathcal{R}} w(p_n)\right)}_{\text{central difference term}}.$$
$$(4)$$

According to the Eq. (4), CDC can be easily implemented by a few lines of code in PyTorch[42] and TensorFlow[1]. The derivation of Eq. (4) and codes based on Pytorch are shown in ***Appendix A***.

Relation to Prior Work. Here we discuss the relations between CDC and vanilla convolution, local binary convolution[27] and gabor convolution[38], which share

similar design philosophy but with different focuses. The ablation study is in Section 4.3 to show superior performance of CDC for face anti-spoofing task.

Relation to Vanilla Convolution. CDC is more generalized. It can be seen from Eq. (3) that vanilla convolution is a special case of CDC when $\theta = 0$, i.e., aggregating local intensity information without gradient message.

Relation to Local Binary Convolution[27]. Local binary convolution (LBConv) focuses on computational reduction so its modeling capacity is limited. CDC focuses on enhancing rich detailed feature representation capacity without any additional parameters. On the other side, LBConv uses pre-defined filters to describe the local feature relation while CDC can learn these filters automatically.

Relation to Gabor Convolution[38]. Gabor convolution (GaborConv) devotes to enhancing the representation capacity of spatial transformations (i.e., orientation and scale changes) while CDC focuses more on representing fine-grained robust features in diverse environments.

3.2. CDCN

Depth-supervised face anti-spoofing methods [36, 4] take advantage of the discrimination between spoofing and living faces based on 3D shape, and provide pixel-wise detailed information for the FAS model to capture spoofing cues. Motivated by this, a similar depth-supervised network [36] called "DepthNet" is built up as baseline in this paper. In order to extract more fine-grained and robust features for estimating the facial depth map, CDC is introduced to form **C**entral **D**ifference **C**onvolutional **N**etworks (CDCN). Note that DepthNet is the special case of the proposed CDCN when $\theta = 0$ for all CDC operators.

The details of CDCN are shown in Table 1. Given a single RGB facial 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. We use $\theta = 0.7$ as the default setting and ablation study about θ will be shown in Section 4.3.

For the loss function, mean square error loss $\mathcal{L}_{MSE}$ is utilized for pixel-wise supervision. Moreover, for the sake of fine-grained supervision needs in FAS task, contrastive depth loss $\mathcal{L}_{CDL}$ [56] is considered to help the networks learn more detailed features. So the overall loss $L_{overall}$ can be formulated as $\mathcal{L}_{overall} = \mathcal{L}_{MSE} + \mathcal{L}_{CDL}$.

3.3. CDCN++

It can be seen from Table 1 that the architecture of CDCN is designed coarsely (e.g., simply repeating the same block structure for different levels), which might be suboptimized for face anti-spoofing task. Inspired by the classical visual object understanding models [40], we propose an extended version CDCN++ (see Fig. 5), which consists of a NAS based backbone and Multiscale Attention Fusion

Level	Output	DepthNet [36]	CDCN ($\theta = 0.7$)
	256×256	3×3 conv, 64	3×3 **CDC**, 64
Low	128×128	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ \textbf{CDC}}, 196 \\ 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
Mid	64×64	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ \textbf{CDC}}, 196 \\ 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
High	32×32	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 196 \\ 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ \textbf{CDC}}, 196 \\ 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ max pool} \end{bmatrix}$
	32×32	[concat (Low, Mid, High), 384]	
	32×32	$\begin{bmatrix} 3 \times 3 \text{ conv}, 128 \\ 3 \times 3 \text{ conv}, 64 \\ 3 \times 3 \text{ conv}, 1 \end{bmatrix}$	$\begin{bmatrix} 3 \times 3 \text{ \textbf{CDC}}, 128 \\ 3 \times 3 \text{ \textbf{CDC}}, 64 \\ 3 \times 3 \text{ \textbf{CDC}}, 1 \end{bmatrix}$
# params		2.25×10^6	2.25×10^6

Table 1. Architecture of DepthNet and CDCN. Inside the brackets are the filter sizes and feature dimensionalities. "conv" and "CDC" suggest vanilla and central difference convolution, respectively. All convolutional layers are with stride=1 and followed by a BN-ReLU layer while max pool layers are with stride=2.

Module (MAFM) with selective attention capacity.

Searching Backbone for FAS task. Our search algorithm is based on two gradient-based NAS methods [35, 60], and more technical details can be referred to the original papers. Here we mainly state the new contributions about searching backbone for FAS task.

As illustrated in Fig. 3(a), the goal is to search for cells in three levels (low-level, mid-level and high-level) to form a network backbone for FAS task. Inspired by the dedicated neurons for hierarchical organization in human visual system [40], we prefer to search these multi-level cells freely (i.e., cells with varied structures), which is more flexible and generalized. We name this configuration as "**Varied Cells**" and will study its impacts in Sec. 4.3 (see Tab. 2). Different from previous works [35, 60], we adopt only one output of the latest incoming cell as the input of the current cell.

As for the cell-level structure, Fig. 3(b) shows that each cell is represented as a directed acyclic graph (DAG) of N nodes $\{x\}_{i=0}^{N-1}$, where each node represents a network layer. We denote the operation space as $\mathcal{O}$, and Fig. 3(c) shows eight designed candidate operations (none, skip-connect and CDCs). Each edge (i, j) of DAG represents the information flow from node x_i to node x_j, which consists of the candidate operations weighted by the architecture parameter $\alpha^{(i,j)}$. Specially, each edge (i, j) can be formulated by a function $\tilde{o}^{(i,j)}$ where $\tilde{o}^{(i,j)}(x_i) = \sum_{o \in \mathcal{O}} \eta_o^{(i,j)} \cdot o(x_i)$. Softmax function is utilized to relax architecture parameter $\alpha^{(i,j)}$ into operation weight $o \in \mathcal{O}$, that is $\eta_o^{(i,j)} = \frac{exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} exp(\alpha_{o'}^{(i,j)})}$. The intermediate node can be denoted as $x_j = \sum_{i<j} \tilde{o}^{(i,j)}(x_i)$ and the output node x_{N-1} is represented by weighted summation of all intermediate nodes $x_{N-1} = \sum_{0<i<N-1} \beta_i(x_i)$. Here we propose a node at-
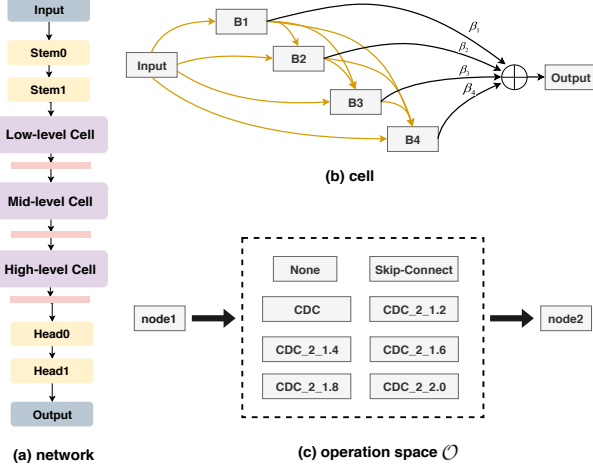
Figure 3. Architecture search space with CDC. (a) A network consists of three stacked cells with max pool layer while stem and head layers adopt CDC with 3×3 kernel and $\theta = 0.7$. (b) A cell contains 6 nodes, including an input node, four intermediate nodes B1, B2, B3, B4 and an output node. (c) The edge between two nodes (except output node) denotes a possible operation. The operation space consists of eight candidates, where **CDC_2_r** means using two stacked CDC to increase channel number with ratio r first and then decrease back to the original channel size.

tention strategy to learn the importance weights β among intermediate nodes, that is $\beta_i = \frac{exp(\beta_i')}{\sum_{0<j<N-1} exp(\beta_j')}$, where β_i is the softmax of the original learnable weight β_i' for intermediate node x_i.

In the searching stage, $\mathcal{L}_{train}$ and $\mathcal{L}_{val}$ are denoted as the training and validation loss respectively, which are all based on the depth-supervised loss $\mathcal{L}_{overall}$ described in Sec. 3.2. Network parameters w and architecture parameters α are learned with the following bi-level optimization problem:

$$\begin{aligned} \min_{\alpha} \quad & \mathcal{L}_{val}(w^*(\alpha), \alpha), \\ s.t. \quad & w^*(\alpha) = arg\min_{w} \mathcal{L}_{train}(w, \alpha) \end{aligned} \quad (5)$$

After convergence, the final discrete architecture is derived by: 1) setting $o^{(i,j)} = arg\,max_{o\in\mathcal{O},o\neq none}\, p_o^{(i,j)}$; 2) for each intermediate node, choosing one incoming edge with the largest value of $max_{o\in\mathcal{O},o\neq none}\, p_o^{(i,j)}$, and 3) for each output node, choosing the one incoming intermediate node with largest value of $max_{0<i<N-1}\,\beta_i$ (denoted as "**Node Attention**") as input. In contrast, choosing last intermediate node as output node is more straightforward. We will compare these two settings in Sec. 4.3 (see Tab. 2).

MAFM. Although simply fusing low-mid-high levels features can boost performance for the searched CDC architecture, it is still hard to find the important regions to focus, which goes against learning more discriminative features. Inspired by the selective attention in human visual system [40, 55], neurons at different levels are likely to have stim-
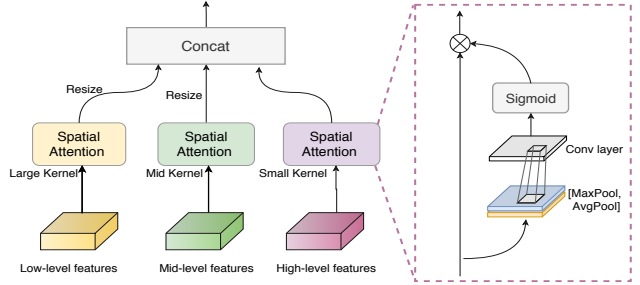

Figure 4. Multiscale Attention Fusion Module.

uli in their receptive fields with various attention. Here we propose a Multiscale Attention Fusion Module (MAFM), which is able to refine and fuse low-mid-high levels CDC features via spatial attention.

As illustrated in Fig. 4, features $\mathcal{F}$ from different levels are refined via spatial attention [58] with receptive fields related kernel size (i.e., the high/semantic level should be with small attention kernel size while low level with large attention kernel size in our case) and then concatenate together. The refined features $\mathcal{F}'$ can be formulated as

$$\mathcal{F}_i' = \mathcal{F}_i \odot (\sigma(\mathcal{C}_i([\mathcal{A}(\mathcal{F}_i), \mathcal{M}(\mathcal{F}_i)]))), i \in \{low, mid, high\}, \quad (6)$$

where $\odot$ represents Hadamard product. $\mathcal{A}$ and $\mathcal{M}$ denotes avg and max pool layer respectively. σ means the sigmoid function while $\mathcal{C}$ is the convolution layer. Vanilla convolutions with 7×7, 5×5 and 3×3 kernels are utilized for $\mathcal{C}_{low}$, $\mathcal{C}_{mid}$ and $\mathcal{C}_{high}$, respectively. CDC is not chosen here because of its limited capacity of global semantic cognition, which is vital in spatial attention. The corresponding ablation study is conducted in Sec. 4.3.

4. Experiments

4.1. Datasets and Metrics

Databases. Six databases OULU-NPU [10], SiW [36], CASIA-MFSD [66], Replay-Attack [13], MSU-MFSD [57] and SiW-M [37] are used in our experiments. OULU-NPU and SiW are high-resolution databases, containing four and three protocols to validate the generalization (e.g., unseen illumination and attack medium) of models respectively, which are utilized for intra testing. CASIA-MFSD, Replay-Attack and MSU-MFSD are databases which contain 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) attacks types inside.

Performance Metrics. In OULU-NPU and SiW 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 [24] for a fair comparison. Half Total Error Rate (HTER) is adopted in the cross testing between CASIA-MFSD and Replay-Attack. Area Under Curve (AUC) is uti-
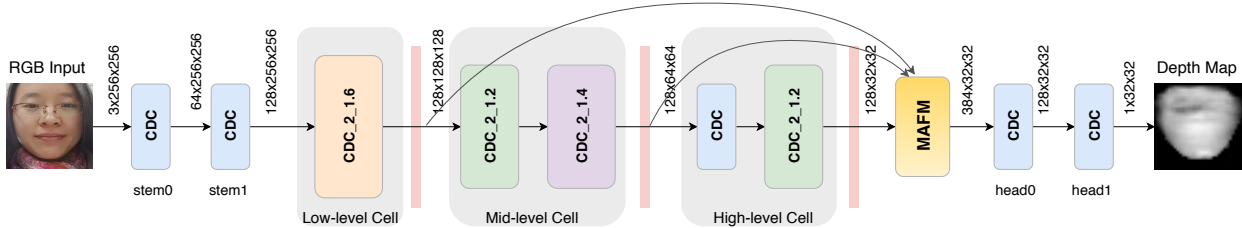
Figure 5. The architecture of CDCN++. It consists of the searched CDC backbone and MAFM. Each cell is followed by a max pool layer.

lized for intra-database cross-type test on CASIA-MFSD, Replay-Attack and MSU-MFSD. For the cross-type test on SiW-M, APCER, BPCER, ACER and Equal Error Rate (EER) are employed.

4.2. Implementation Details

Depth Generation. Dense face alignment PRNet [18] is adopted to estimate the 3D shape of the living face and generate the facial depth map with size 32×32. More details and samples can be found in [56]. To distinguish living faces from spoofing faces, at the training stage, we normalize living depth map in a range of $[0, 1]$, while setting spoofing depth map to 0, which is similar to [36].

Training and Testing Setting. Our proposed method is implemented with Pytorch. In the training stage, models are trained with Adam optimizer and the initial learning rate (lr) and weight decay (wd) are 1e-4 and 5e-5, respectively. We train models with maximum 1300 epochs while lr halves every 500 epochs. The batch size is 56 on eight 1080Ti GPUs. In the testing stage, we calculate the mean value of the predicted depth map as the final score.

Searching Setting. Similar to [60], partial channel connection and edge normalization are adopted. The initial number of channel is sequentially $\{32, 64, 128, 128, 128, 64, 1\}$ in the network (see Fig. 3(a)), which doubles after searching. Adam optimizer with lr=1e-4 and wd=5e-5 is utilized when training the model weights. The architecture parameters are trained with Adam optimizer with lr=6e-4 and wd=1e-3. We search 60 epochs on Protocol-1 of OULU-NPU with batchsize 12 while architecture parameters are not updated in the first 10 epochs. The whole process costs one day on three 1080Ti.

4.3. Ablation Study

In this subsection, all ablation studies are conducted on Protocol-1 (different illumination condition and location between train and test sets) of OULU-NPU [10] to explore the details of our proposed CDC, CDCN and CDCN++.

Impact of θ in CDCN. According to Eq. (3), θ controls the contribution of the gradient-based details, i.e., the higher θ, the more local detailed information included. As illustrated in Fig. 6(a), when $\theta \geqslant 0.3$, CDC always achieves better performance than vanilla convolution ($\theta = 0$, ACER=3.8%), indicating the central difference based
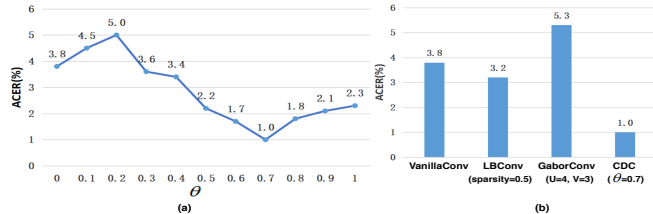

Figure 6. (a) Impact of θ in CDCN. (b) Comparison among various convolutions (only showing the hyperparameters with best performance). The lower ACER, the better performance.

Table 2. The ablation study of NAS configuration.

Model	Varied cells	Node attention	ACER(%)
NAS_Model 1			1.7
NAS_Model 2	$\checkmark$		1.5
NAS_Model 3		$\checkmark$	1.4
NAS_Model 4	$\checkmark$	$\checkmark$	**1.3**

fine-grained information is helpful for FAS task. As the best performance (ACER=1.0%) is obtained when $\theta = 0.7$, we use this setting for the following experiments. Besides keeping the constant θ for all layers, we also explore an adaptive CDC method to learn θ for every layer, which is shown in **Appendix B**.

CDC vs. Other Convolutions. As discussed in Sec. 3.1 about the relation between CDC and prior convolutions, we argue that the proposed CDC is more suitable for FAS task as the detailed spoofing artifacts in diverse environments should be represented by the gradient-based invariant features. Fig. 6(b) shows that CDC outperforms other convolutions by a large margin (more than 2% ACER). It is interesting to find that LBConv performs better than vanilla convolution, indicating that the local gradient information is important for FAS task. GaborConv performs the worst because it is designed for capturing spatial invariant features, which is not helpful in face anti-spoofing task.

Impact of NAS Configuration. Table 2 shows the ablation study about the two NAS configurations described in Sec. 3.3, i.e., varied cells and node attention. Compared to the baseline setting with the shared cells and last intermediate node as output node, both these two configurations can boost the searching performance. The reason behind is twofold: 1) with more flexible searching constraints, NAS is able to find dedicated cells for different levels, which is more similar to human visual system [40], and 2) taking the last intermediate node as output might not be optimal while choosing the most important one is more reasonable.

Table 3. The ablation study of NAS based backbone and MAFM.

Backbone	Multi-level Fusion	ACER(%)
w/o NAS	w/ multi-level concat	1.0
w/ NAS	w/ multi-level concat	0.7
w/ NAS	w/ MAFM (3x3,3x3,3x3 CDC)	1.2
w/ NAS	w/ MAFM (3x3,3x3,3x3 VanillaConv)	0.6
w/ NAS	w/ MAFM (5x5,5x5,5x5 VanillaConv)	1.1
w/ NAS	w/ MAFM (7x7,5x5,3x3 VanillaConv)	**0.2**

Table 4. The results of intra testing on four protocols of OULU-NPU. We only report the results "STASN [62]" trained without extra datasets for a fair comparison.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	GRADIANT [6]	1.3	12.5	6.9
	STASN [62]	1.2	2.5	1.9
	Auxiliary [36]	1.6	1.6	1.6
	FaceDs [26]	1.2	1.7	1.5
	FAS-TD [56]	2.5	0.0	1.3
	DeepPixBiS [20]	0.8	0.0	0.4
	CDCN (Ours)	0.4	1.7	1.0
	CDCN++ (Ours)	0.4	0.0	**0.2**
2	DeepPixBiS [20]	11.4	0.6	6.0
	FaceDs [26]	4.2	4.4	4.3
	Auxiliary [36]	2.7	2.7	2.7
	GRADIANT [6]	3.1	1.9	2.5
	STASN [62]	4.2	0.3	2.2
	FAS-TD [56]	1.7	2.0	1.9
	CDCN (Ours)	1.5	1.4	1.5
	CDCN++ (Ours)	1.8	0.8	**1.3**
3	DeepPixBiS [20]	11.7±19.6	10.6±14.1	11.1±9.4
	FAS-TD [56]	5.9±1.9	5.9±3.0	5.9±1.0
	GRADIANT [6]	2.6±3.9	5.0±5.3	3.8±2.4
	FaceDs [26]	4.0±1.8	3.8±1.2	3.6±1.6
	Auxiliary [36]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [62]	4.7±3.9	0.9±1.2	2.8±1.6
	CDCN (Ours)	2.4±1.3	2.2±2.0	2.3±1.4
	CDCN++ (Ours)	1.7±1.5	2.0±1.2	**1.8±0.7**
4	DeepPixBiS [20]	36.7±29.7	13.3±14.1	25.0±12.7
	GRADIANT [6]	5.0±4.5	15.0±7.1	10.0±5.0
	Auxiliary [36]	9.3±5.6	10.4±6.0	9.5±6.0
	FAS-TD [56]	14.2±8.7	4.2±3.8	9.2±3.4
	STASN [62]	6.7±10.6	8.3±8.4	7.5±4.7
	FaceDs [26]	1.2±6.3	6.1±5.1	5.6±5.7
	CDCN (Ours)	4.6±4.6	9.2±8.0	6.9±2.9
	CDCN++ (Ours)	4.2±3.4	5.8±4.9	**5.0±2.9**

Effectiveness of NAS Based Backbone and MAFM. The proposed CDCN++, consisting of NAS based backbone and MAFM, is shown in Fig. 5. It is obvious that cells from multiple levels are quite different and the mid-level cell has deeper (four CDC) layers. Table 3 shows the ablation studies about NAS based backbone and MAFM. It can be seen from the first two rows that NAS based backbone with direct multi-level fusion outperforms (0.3% ACER) the backbone without NAS, indicating the effectiveness of our searched architecture. Meanwhile, backbone with MAFM achieves 0.5% ACER lower than that with direct multi-level fusion, which shows the effectiveness of MAFM. We also analyse the convolution type and kernel size in MAFM and find that vanilla convolution is more suitable for capturing the semantic spatial attention. Besides, the attention kernel size should be large (7x7) and small (3x3) enough for low-level and high-level features, respectively.

4.4. Intra Testing

The intra testing is carried out on both the OULU-NPU and the SiW datasets. We strictly follow the four protocols

Table 5. The results of intra testing on three protocols of SiW [36].

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Auxiliary [36]	3.58	3.58	3.58
	STASN [62]	–	–	1.00
	FAS-TD [56]	0.96	0.50	0.73
	CDCN (Ours)	0.07	0.17	**0.12**
	CDCN++ (Ours)	0.07	0.17	**0.12**
2	Auxiliary [36]	0.57±0.69	0.57±0.69	0.57±0.69
	STASN [62]	–	–	0.28±0.05
	FAS-TD [56]	0.08±0.14	0.21±0.14	0.15±0.14
	CDCN (Ours)	0.00±0.00	0.13±0.09	0.06±0.04
	CDCN++ (Ours)	0.00±0.00	0.09±0.10	**0.04±0.05**
3	STASN [62]	–	–	12.10±1.50
	Auxiliary [36]	8.31±3.81	8.31±3.80	8.31±3.81
	FAS-TD [56]	3.10±0.81	3.09±0.81	3.10±0.81
	CDCN (Ours)	1.67±0.11	1.76±0.12	**1.71±0.11**
	CDCN++ (Ours)	1.97±0.33	1.77±0.10	1.90±0.15

on OULU-NPU and three protocols on SiW for the evaluation. All compared methods including STASN [62] are trained without extra datasets for a fair comparison.

Results on OULU-NPU. As shown in Table 4, our proposed CDCN++ ranks first on all 4 protocols (0.2%, 1.3%, 1.8% and 5.0% ACER, respectively), which indicates the proposed method performs well at the generalization of the external environment, attack mediums and input camera variation. Unlike other state-of-the-art methods (Auxiliary [36], STASN [62], GRADIANT [6] and FAS-TD [56]) extracting multi-frame dynamic features, our method needs only frame-level inputs, which is suitable for real-world deployment. It's worth noting that the NAS based backbone for CDCN++ is transferable and generalizes well on all protocols although it is searched on Protocol-1.

Results on SiW. Table 5 compares the performance of our method with three state-of-the-art methods Auxiliary [36], STASN [62] and FAS-TD [56] on SiW dataset. It can be seen from Table 5 that our method performs the best for all three protocols, revealing the excellent generalization capacity of CDC for (1) variations of face pose and expression, (2) variations of different spoof mediums, (3) cross/unknown presentation attack.

4.5. Inter Testing

To further testify the generalization ability of our model, we conduct cross-type and cross-dataset testing to verify the generalization capacity to unknown presentation attacks and unseen environment, respectively.

Cross-type Testing. Following the protocol proposed in [3], we use CASIA-MFSD [66], Replay-Attack [13] and MSU-MFSD [57] to perform intra-dataset cross-type testing between replay and print attacks. As shown in Table 6, our proposed CDC based methods achieve the best overall performance (even outperforming the zero-shot learning based method DTN [37]), indicating our consistently good generalization ability among unknown attacks. Moreover, we also conduct the cross-type testing on the latest SiW-M [37] dataset and achieve the best average ACER (12.7%) and EER (11.9%) among 13 attacks. The detailed results are shown in ***Appendix C***.

Table 6. AUC (%) of the model cross-type testing on CASIA-MFSD, Replay-Attack, and MSU-MFSD.

Method	CASIA-MFSD [66]			Replay-Attack [13]			MSU-MFSD [57]			Overall
	Video	Cut Photo	Wrapped Photo	Video	Digital Photo	Printed Photo	Printed Photo	HR Video	Mobile Video	
OC-SVM$_{RBF}$+BSIF [3]	70.74	60.73	95.90	84.03	88.14	73.66	64.81	87.44	74.69	78.68±11.74
SVM$_{RBF}$+LBP [10]	91.94	91.70	84.47	99.08	98.17	87.28	47.68	99.50	97.61	88.55±16.25
NN+LBP [59]	94.16	88.39	79.85	99.75	95.17	78.86	50.57	99.93	93.54	86.69±16.25
DTN [37]	90.0	97.3	97.5	99.9	99.9	99.6	**81.6**	99.9	97.5	95.9±6.2
CDCN (Ours)	**98.48**	**99.90**	**99.80**	**100.00**	99.43	99.92	70.82	**100.00**	99.99	96.48±9.64
CDCN++ (Ours)	98.07	**99.90**	99.60	99.98	**99.89**	**99.98**	72.29	**100.00**	99.98	**96.63±9.15**

Table 7. The results of cross-dataset testing between CASIA-MFSD and Replay-Attack. The evaluation metric is HTER(%). The **multiple-frame** based methods are shown in the upper half part while **single-frame** based methods in bottom half part.

Method	Train	Test	Train	Test
	CASIA-MFSD	Replay-Attack	Replay-Attack	CASIA-MFSD
Motion-Mag [5]	50.1			47.0
LBP-TOP [16]	49.7			60.6
STASN [62]	31.5			30.9
Auxiliary [36]	27.6			28.4
FAS-TD [56]	17.5			**24.0**
LBP [7]	47.0			39.6
Spectral cubes [48]	34.4			50.0
Color Texture [8]	30.3			37.7
FaceDs [26]	28.5			41.1
CDCN (Ours)	15.5			32.6
CDCN++ (Ours)	**6.5**			29.8

Cross-dataset Testing. In this experiment, 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 7, our proposed CDCN++ has 6.5% HTER on protocol CR, outperforming the prior state-of-the-art by a convincing margin of 11%. For protocol RC, we also outperform state-of-the-art frame-level methods (see bottom half part of Table 7). The performance might be further boosted via introducing the similar temporal dynamic features in Auxiliary [36] and FAS-TD [56].

4.6. Analysis and Visualization.

In this subsection, two perspectives are provided to demonstrate the analysis why CDC performs well.

Robustness to Domain Shift. Protocol-1 of OULU-NPU is used to verify the robustness of CDC when encountering the domain shifting, i.e., huge illumination difference between train/development and test set. Fig. 7 shows that the network using vanilla convolution has low ACER on development set (blue curve) while high ACER on test set (gray curve), which indicates vanilla convolution is easily overfitted in seen domain but generalizes poorly when illumination changes. In contrast, the model with CDC is able to achieve more consistent performance on both development (red curve) and test set (yellow curve), indicating the robustness of CDC to domain shifting.

Features Visualization. Distribution of multi-level fused features for the testing videos on Protocol-1 OULU-NPU is shown in Fig. 8 via t-SNE [39]. It is clear that the
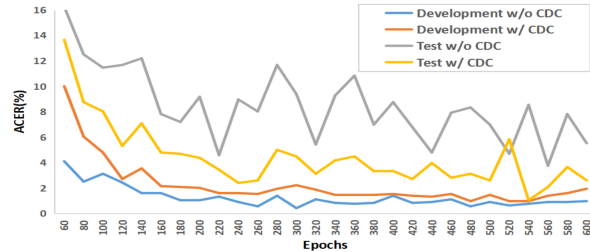

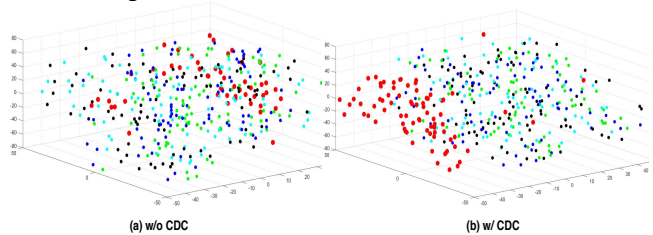Figure 7. The performance of CDCN on development and test set when training on Protocol-1 OULU-NPU.

Figure 8. 3D visualization of feature distribution. (a) Features w/o CDC. (b) Features w/ CDC. Color code used: red=live, $green$=printer1, $blue$=printer2, $cyan$=replay1, $black$=replay2.

features with CDC (Fig. 8(a)) presents more well-clustered behavior than that with vanilla convolution (Fig. 8(b)), which demonstrates the discrimination ability of CDC for distinguishing the living faces from spoofing faces. The visualization of the feature maps (w/o or w/ CDC) and attention maps of MAFM can be found in ***Appendix D***.

5. Conclusions and Future Work

In this paper, we propose a novel operator called Central Difference Convolution (CDC) for face anti-spoofing task. Based on CDC, a Central Difference Convolutional Network is designed. We also propose CDCN++, consisting of a searched CDC backbone and Multiscale Attention Fusion Module. Extensive experiments are performed to verify the effectiveness of the proposed methods. We note that the study of CDC is still at an early stage. Future directions include: 1) designing context-aware adaptive CDC for each layer/channel; 2) exploring other properties (e.g., domain generalization) and applicability on other vision tasks (e.g., image quality assessment [34] and FaceForensics).

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 3

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006. 2

[3] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, 5:13868–13882, 2017. 7, 8

[4] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328, 2017. 1, 2, 4

[5] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 105–110, 2013. 8

[6] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 688–696. IEEE, 2017. 7

[7] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *IEEE international conference on image processing (ICIP)*, pages 2636–2640, 2015. 1, 2, 3, 8

[8] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. 1, 8

[9] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2017. 2

[10] 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*, pages 612–618, 2017. 5, 6, 8

[11] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017. 2

[12] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *International Conference on Learning Representations (ICLR)*, 2019. 2

[13] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Biometrics Special Interest Group*, pages 1–7, 2012. 5, 7, 8

[14] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[15] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132, 2012. 1, 2

[16] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *2013 international conference on biometrics (ICB)*, pages 1–8, 2013. 8

[17] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1761–1770, 2019. 2

[18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2017. 6

[19] Junying Gan, Shanlu Li, Yikui Zhai, and Chengyun Liu. 3d convolutional neural network based on face anti-spoofing. In *ICMIP*, pages 1–5, 2017. 1

[20] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *International Conference on Biometrics*, number CONF, 2019. 1, 2, 7

[21] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7036–7045, 2019. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[24] international organization for standardization. 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. 5

[25] Anil K Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186, 1991. 2

[26] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face despoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018. 1, 2, 7, 8

[27] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19–28, 2017. 2, 3, 4

[28] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection using dynamic texture. In *Asian Conference on Computer Vision*, pages 146–157. Springer, 2012. 2

[29] Jukka Komulainen, Abdenour Hadid, and Matti Pietikainen. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013. 1, 2

[30] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *IPTA*, pages 1–6, 2016. 2

[31] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4244–4249. IEEE, 2016. 2

[32] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face liveness detection by rppg features and contextual patch-based cnn. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, pages 61–68. ACM, 2019. 2

[33] Chen Lin, Zhouyingcheng Liao, Peng Zhou, Jianguo Hu, and Bingbing Ni. Live face verification with multiple instantialized local homographic parameterization. In *IJCAI*, pages 814–820, 2018. 2

[34] Suiyi Ling, Patrick Le Callet, and Zitong Yu. The role of structure and textural information in image utility and quality assessment tasks. *Electronic Imaging*, 2018(14):1–13, 2018. 8

[35] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *International Conference on Learning Representations (ICLR)*, 2019. 2, 4

[36] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 1, 2, 4, 5, 6, 7, 8

[37] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 5, 7, 8

[38] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu. Gabor convolutional networks. *IEEE Transactions on Image Processing*, 27(9):4357–4366, 2018. 2, 3, 4

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[40] Thomas J Palmeri and Isabel Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5(4):291, 2004. 4, 5, 6

[41] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 2

[42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[43] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition*, pages 611–619, 2016. 2

[44] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016. 1, 2

[45] Bruno Peixoto, Carolina Michelassi, and Anderson Rocha. Face liveness detection under bad illumination conditions. In *ICIP*, pages 3557–3560. IEEE, 2011. 1, 2

[46] Wei Peng, Xiaopeng Hong, and Guoying Zhao. Video action recognition via neural architecture searching. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 11–15. IEEE, 2019. 3

[47] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *International Conference on Machine Learning (ICML)*, 2018. 2

[48] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12):4726–4740, 2015. 8

[49] 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. *AAAI*, 2020. 1

[50] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. *IEEE International Conference on Computer Vision*, 2019. 3

[51] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4780–4789, 2019. 2

[52] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2902–2911. JMLR. org, 2017. 2

[53] Talha Ahmad Siddiqui, Samarth Bharadwaj, Tejas I Dhamecha, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Face anti-spoofing with multifeature videolet aggregation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1035–1040. IEEE, 2016. 2

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[55] Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000. 5

[56] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv preprint arXiv:1811.05118*, 2018. 2, 4, 6, 7, 8

[57] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 5, 7, 8

[58] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5

[59] Fei Xiong and Wael AbdAlmageed. Unknown presentation attack detection with face rgb images. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2018. 8

[60] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient differentiable architecture search. *arXiv preprint arXiv:1907.05737*, 2019. 2, 4, 6

[61] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 2

[62] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 7, 8

[63] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[64] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. Auto-fas: Searching lightweight networks for face anti-spoofing. *ICASSP*, 2020. 1

[65] Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, and Tao Mei. Customizable architecture search for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11641–11650, 2019. 3

[66] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31, 2012. 5, 7, 8

[67] Ning Zhu and Xiaolong Bai. Neural architecture search for deep face recognition. *arXiv preprint arXiv:1904.09523*, 2019. 3

[68] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2017. 2

[69] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 2