Revisiting Pixel-Wise Supervision for Face Anti-Spoofing

Zitong Yu^{ID}, Graduate Student Member, IEEE, Xiaobai Li^{ID}, Member, IEEE, Jingang Shi^{ID},
Zhaoqiang Xia^{ID}, Member, IEEE, and Guoying Zhao^{ID}, Senior Member, IEEE

Abstract—Face anti-spoofing (FAS) plays a vital role in securing face recognition systems from the presentation attacks (PAs). As more and more realistic PAs with novel types spring up, it is necessary to develop robust algorithms for detecting unknown attacks even in unseen scenarios. However, deep models supervised by traditional binary loss (e.g., ‘0’ for bonafide vs. ‘1’ for PAs) are weak in describing intrinsic and discriminative spoofing patterns. Recently, pixel-wise supervision has been proposed for the FAS task, intending to provide more fine-grained pixel/patch-level cues. In this paper, we firstly give a comprehensive review and analysis about the existing pixel-wise supervision methods for FAS. Then we propose a novel pyramid supervision, which guides deep models to learn both local details and global semantics from multi-scale spatial context. Extensive experiments are performed on five FAS benchmark datasets to show that, without bells and whistles, the proposed pyramid supervision could not only improve the performance beyond existing pixel-wise supervision frameworks, but also enhance the model’s interpretability (i.e., locating the patch-level positions of PAs more reasonably). Furthermore, elaborate studies are conducted for exploring the efficacy of different architecture configurations with two kinds of pixel-wise supervisions (binary mask and depth map supervisions), which provides inspirable insights for future architecture/supervision design.

Index Terms—Face anti-spoofing, pixel-wise supervision, pyramid supervision.

I. INTRODUCTION

DUE TO its convenience and remarkable accuracy, face recognition technology [1] has applied in a few interactive intelligent applications such as checking-in and mobile payment. However, most existing face recognition systems are vulnerable to presentation attacks (PAs) ranging from print, replay, makeup and 3D-mask attacks [2].

Manuscript received November 24, 2020; revised February 13, 2021; accepted March 8, 2021. This work was supported in part by the Academy of Finland for project MiGA under Grant 316765; in part by ICT 2023 Project under Grant 328115; and in part by Infotech Oulu. The work of Jingang Shi was supported by the National Natural Science Foundation of China under Grant 62002283. This article was recommended for publication by Associate Editor J. Wan upon evaluation of the reviewers’ comments. (Corresponding author: Guoying Zhao.)

Zitong Yu, Xiaobai Li, and Guoying Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland (e-mail: zitong.yu@oulu.fi; xiaobai.li@oulu.fi; guoying.zhao@oulu.fi).

Jingang Shi is with the School of Software Engineering, Xi’an Jiaotong University, Xi’an 710049, China (e-mail: jingang.shi@xjtu.edu.cn).

Zhaoqiang Xia is with the School of Electronics and Information, Northwestern Polytechnical University, Xi’an 710072, China (e-mail: zxia@nwpu.edu.cn).

Therefore, both the academia and industry have paid attention on the development of face anti-spoofing (FAS) technology for securing the face recognition system.

In the past two decades, both traditional [3], [4], [5], [6], [7] and deep learning-based [8], [9], [10], [11], [12], [13], [14] methods have shown their effectiveness for presentation attack detection (PAD). Most traditional algorithms focus on human liveness cues and handcrafted features, which need rich task-aware prior knowledge. In term of the methods based on the liveness cues, eye-blinking [3], [15], [16], face and head movement [17], [18] (e.g., nodding and smiling), gaze tracking [19], [20] and remote physiological signals (e.g., rPPG [4], [10], [21], [22]) are explored as dynamic discrimination. However, these physiological liveness cues are usually captured from long-term interactive face videos and easily mimicked from the video attacks, making them less reliable and inconvenient for practical deployment. On the other hand, classical handcrafted descriptors (e.g., LBP [5], [23], SIFT [7], HOG [6] and DoG [24]) are designed for extracting effective spoofing patterns from various color spaces (RGB, HSV, and YCbCr). Although such handcrafted features could be cascaded with a trained classifier (e.g., SVM [25]) efficiently, they still suffer from limited representation capacity and are vulnerable under unseen scenarios and unknown PAs.

In contrast, a few deep learning based methods are proposed for both frame and video level face liveness detection. Most works [26], [31], [32], [33], [34], [35], [36] treat FAS as a binary classification problem (e.g., ‘0’ for live while ‘1’ for spoofing faces) thus supervised by a simple binary cross-entropy loss. However, convolutional neural networks (CNNs) with binary loss might discover arbitrary cues that are able to separate the two classes (e.g., screen bezel), but not the faithful spoofing patterns. Recently, pixel-wise supervision [9], [10], [12], [26], [29], [37] attracts more attention as it provides more fine-grained context-aware supervision signals, which is beneficial for deep models learning intrinsic spoofing cues. As can be seen from Fig. 1, pseudo depth labels [10], [12], reflection maps [9], [29], binary mask label [26], [27], [28] and 3D point cloud maps [30] are typical pixel-wise supervision signals, which describes the local live/spoofing cues in pixel/patch level. Despite with gained performance, existing pixel-wise supervisions still have the representation gaps between local context and global semantics, and are easily influenced by the local disturbance.

For the perspective of evaluation protocols, most existing FAS methods are supervised by predefined scenarios and PAs.

Fig. 1. Visualization of different kinds of pixel-wise supervision for FAS. Given an RGB live/spoof face image as input, the deep models could be supervised by binary mask label [26], [27], [28], pseudo reflection maps [9], [29], pseudo depth labels [10], [12] and 3D point cloud maps [30].

Thus, the trained models probabilistically overfit on several specific domains and attacks, which are vulnerable to domain shift and unseen attacks. As FAS is actually an open-set problem in practice, the uncertain gaps (e.g., environments and attack types) between training and testing conditions should be considered. On one hand, maximum mean discrepancy based distribution alignment [38], adversarial learning [39], fine-grained meta learning [40] and multi-domain disentangled learning [41] are utilized to learn robust features for domain generalization in FAS. On the other hand, in order to detect unseen attacks successfully, one class SVM [42], deep tree network [27] and adaptive inner-update meta learning [43] are developed. Despite enhancing the generalization capacity via learning strategies [39], [40], [41], [43], they are still hard to explicitly learn detailed intrinsic spoofing patterns with coarse supervision signals. Moreover, it is not trivial to explore and understand how the pixel-wise supervision impacts the mainstream architectures on various FAS evaluation protocols.

Motivated by the discussions above, we firstly investigate existing pixel-wise supervision methods for FAS, which aims to help FAS researchers conveniently compare and understand pixel-wise supervision based state of the arts. Furthermore, we propose a novel pyramid supervision, which is able to plug-and-play into the current pixel-wise supervision framework flexibly and provide multi-scale patch/global signals. To sum up, the main contributions of this paper are listed:

- We give a comprehensive review of the existing pixel-wise supervision methods in the FAS task.
- We propose a novel plugged-and-played pyramid supervision, which is able to provide richer multi-scale spatial context for fine-grained learning. The proposed pyramid supervision could not only improve the performance beyond the existing pixel-wise supervision framework, but also enhance the model’s interpretability.
- We conduct elaborate studies for exploring the efficacy of different architecture configurations with pixel-wise supervisions, which provides the insights for future architecture/supervision design.

In the rest of the paper, Section II provides a review of pixel-wise supervision for FAS. Section III introduces the pyramid supervision for multiple kinds of pixel-wise supervisions (e.g., binary mask maps and pseudo depth maps). Section IV provides rigorous ablation studies and evaluates

the performance of the proposed pyramid supervision on five benchmark datasets. Finally, conclusions and future works are given in Section V.

II. REVIEW OF PIXEL-WISE SUPERVISION

In this section, we will review existing FAS methods with pixel-wise supervision including depth map, binary mask and so on. The relevant summary is shown in Table I.

Supervised by Depth Map: According to the human prior knowledge of FAS, most PAs (e.g., plain printed paper and electronic screen) merely have no real facial depth information, which could be utilized as discriminative supervision signals. As a result, some recent works [8], [12], [44], [45] adopt pixel-wise pseudo depth labels (forth column in Fig. 1) to guide the deep models, enforcing them predict the real depth for live samples while zero maps for spoof ones. Atoum *et al.* [12] first leveraged pseudo depth labels to guide the multi-scale fully convolutional network (in this paper we call it as ‘DepthNet’ for simplicity). Thus, the well-trained DepthNet is able to predict holistic depth maps as decision evidence. Based on the generated pseudo depth labels, Wang *et al.* [45] designed a contrastive depth loss (CDL) for exploiting fine-grained local depth cues, which also has been applied in training the central difference convolutional networks (CDCN) [8]. Similarly, Peng *et al.* [44] fused the depth supervised stream with another color (RGB,HSV, YCbCr) stream to obtain more robust representation.

Despite with better interpretability and performance compared with traditional binary label, pixel-wise depth label still exists two issues: 1) synthesis of 3D shapes for every training sample is costly and not accurate enough; and 2) it lacks the reasonability for some PAs with real depth (e.g., 3D mask and Mannequin).

Supervised by Binary Mask: Compared with pseudo depth map, binary mask label [26], [27], [47], [48], [49], [50], [51] is easier to be generated and more generalizable to all PAs. To be specific, the binary supervision would be provided for the deep embedding features in each spatial position (second column in Fig. 1). In other words, through the binary mask label, we can find whether PAs occur in the corresponding patches, which are attack-type-agnostic and spatially interpretable. George and Marcel [26] were the first to introduce deep pixel-wise binary supervision (PixBis), which assisted to predict the intermediate confidence map for the cascaded final binary classification. Similarly, Liu *et al.* [27] constrained the leaf nodes with both binary classification and pixel-wise mask regression. Hossain *et al.* [47] proposed to add an attention module for feature refinement before calculating the deep pixel-wise binary loss. Yu *et al.* [48] searched lightweight FAS architectures with pixel-wise binary supervision. Liu *et al.* [50] utilized Early Spoof Regressor with pixel-wise binary supervision to enhance discriminativeness of the generator. Ma *et al.* [49] proposed a multi-regional CNN with the local binary classification loss to local patches. Yu *et al.* [13] utilized pixel-wise binary label to supervise the multimodal CDCN and won the first place in the ChaLearn multi-modal face anti-spoofing attack detection challenge @CVPR2020 [74].

TABLE I
SUMMARY OF THE REPRESENTATIVE FACE ANTI-SPOOFING METHODS WITH **PIXEL-WISE SUPERVISION**. ‘S/D’ IS SHORT FOR STATIC/DYNAMIC. ‘NAS’ DENOTES NEURAL SEARCHED ARCHITECTURE. NOTE THAT SOME METHODS ALSO CONSIDER CLASSIFICATION LOSS (E.G., BINARY CROSS ENTROPY LOSS, TRIPLET LOSS, AND ADVERSARIAL LOSS) WHICH ARE NOT LISTED IN THE ‘SUPERVISION’ COLUMN

Method	Year	Supervision	Backbone	Input	S/D	Description
Depth&Patch [12]	2017	Depth	PatchNet DepthNet	YCbCr HSV	S	local patch features and holistic depth maps extracted by two-stream CNNs
Auxiliary [10]	2018	Depth rPPG spectrum	DepthNet	RGB HSV	D	local temporal features learned from CNN-RNN model with pixel-wise depth and sequence-wise rPPG supervision
De-Spoof [34]	2018	Depth FourierMap	DSNet DepthNet	RGB HSV	S	subtle spoof noise estimated from CNN with proper supervisions
BASN [29]	2019	Depth Reflection	DepthNet Enrichment	RGB HSV	S	generalizable features via bipartite auxiliary supervision
Reconstruction [46]	2019	RGB Input (live) ZeroMap (spoof)	U-Net	RGB	S	multi-level semantic features from autoencoder
DTN [27]	2019	BinaryMask	Tree Network	RGB HSV	S	partition the spoof samples into semantic sub-groups in an unsupervised fashion
PixBiS [26]	2019	BinaryMask	DenseNet161	RGB	S	deep pixel-wise binary supervision without trivial depth synthesis
A-PixBiS [47]	2020	BinaryMask	DenseNet161	RGB	S	incorporate a variant of binary cross entropy that enforces a margin in angular space for attentive pixel wise supervision
Auto-FAS [48]	2020	BinaryMask	NAS	RGB	S	well-suitable lightweight networks searched for mobile-level FAS
MRCNN [49]	2020	BinaryMask	Shallow CNN	RGB	S	introduces local losses to patches, and constraints the entire face region to avoid over-emphasizing certain local areas
STDN [50]	2020	BinaryMask	U-Net PatchGAN	RGB	S	disentangled spoof trace via adversarial learning
FCN-LSA [51]	2020	BinaryMask	DepthNet	RGB	S	high frequent spoof cues from lossless size adaptation module
CDCN [8]	2020	Depth	DepthNet	RGB	S	intrinsic detailed patterns via aggregating both intensity and gradient information from stacked central difference convolutions.
CDCN++ [8]	2020	Depth	NAS	RGB	S	searched architecture with central difference convolution and multiscale attention fusion module.
NAS-FAS [37]	2020	Depth or BinaryMask	NAS	RGB	D	static-dynamic features from searched domain/type-aware architectures with central difference convolution and pooling.
MM-CDCN [13]	2020	BinaryMask	DepthNet	RGB NIR Depth	S	multimodal features from central difference convolutional networks
FAS-SGTD [45]	2020	Depth	DepthNet STPM	RGB	D	detailed discriminative dynamics cues from stacked Residual Spatial Gradient Block and Spatio-Temporal Propagation Module
TS-FEN [44]	2020	Depth	ResNet34 FCN	RGB YCbCr HSV	S	discriminative fused features from depth-stream and chroma-stream networks
SAPLC [28]	2020	TernaryMap	DepthNet	RGB HSV	S	accurate image-level decision via spatial aggregation of pixel-level local classifiers even with insufficient training samples
BCN [9]	2020	BinaryMask Depth Reflection	DepthNet	RGB	S	intrinsic material-based patterns captured via aggregating multi-level bilateral macro- and micro- information
Disentangled [52]	2020	Depth TextureMap	DepthNet	RGB	S	liveness and content features via disentangled representation learning
AENet [53]	2020	Depth Reflection	ResNet18	RGB	S	rich semantic features using Auxiliary Information, Embedding Network with multi-task learning framework
LGSC [54]	2020	ZeroMap (live)	U-Net(ResNet18)	RGB	S	discriminative live-spoof differences learned within a residual-learning framework with the perspective of anomaly detection
3DPC-Net [30]	2020	3D Point Cloud	ResNet18	RGB	S	discriminative features via fine-grained 3D Point Cloud supervision

With the help of spatially positional knowledge, binary mask label not only boosts the models’ discrimination, but also benefits neural architecture search. As the types of PAs could be defined as fine-grained multiple classes, it is also worth exploring whether extending binary maps to multi-class maps are beneficial for supervising the PAs detector.

Supervised by Others: Besides the mainstream depth map and binary mask labels, there are several informative pixel-wise supervision (e.g., pseudo reflection map [9], [29], [53], 3D point cloud map [30], and ternary map [28]). According to the discrepancy of facial material-related albedo between

the live skin and spoof mediums, Kim *et al.* [29] proposed to supervise deep models with both depth and reflection labels. Moreover, Yu *et al.* [9] trained the bilateral convolutional networks with multiple pixel-wise supervisions (binary mask, reflection map, and depth map) simultaneously. Similarly, Zhang *et al.* [53] adopted pseudo reflection and depth labels to guide the networks to learn rich semantic features. Unlike binary mask label considering all spatial positions, Sun *et al.* [28] removed the face-unrelated parts and left the entire face regions as a refined binary mask called ‘ternary map’, which eliminated the noise outside the face. In

Fig. 2. Pyramid supervision with multi-scale binary mask labels. In this framework, the deep features are firstly extracted from ResNet50, and downsampled into multi-scale representation. Unshared 1×1 convolutions cascaded with ‘Sigmoid’ function are utilized for features mapping and predicting the multi-scale binary masks. Then multi-scale binary mask labels would provide sufficient pyramid supervision signals for training. Finally, the predicted multi-scale masks are flattened and concatenated for final live/spoof classification.

order to reduce the redundancy from the dense depth map, Li *et al.* [30] used a sparse 3D point cloud map to efficiently supervise the lightweight models. In addition, encoder-decoder networks with the supervision from Fourier map [34], LBP texture map [52], zero map for live [54], and original RGB inputs [46], also showed their excellent detailed representation capacities.

Although existing pixel-wise supervisions could provide fine-grained local information, there are still two limitations: 1) feature representation gap between pixel-wise supervision (local detailed features) and binary classification (semantic features); and 2) variant face resolution as well as PAs size among training/testing samples, which needs strict and exhausted pixel-wise annotated labels for quality assurance. In this paper, we propose the concept of pyramid supervision, which could be applied to existing pixel-wise supervision to alleviate the above-mentioned drawbacks.

III. METHODOLOGY

As can be seen from Table I, the mainstream backbones for pixel-wise supervision based FAS could be divided into two categories: 1) classical classification based networks (e.g., ResNet [55] and DenseNet [56]) with binary mask supervision; and 2) multi-scale fully convolutional networks (e.g., DepthNet [10]) with pseudo depth supervision. In this section, we introduce the novel **pyramid supervision**, which is able to plug in these two categories flexibly. Here we adopt famous ResNet50 [55] and CDCN [8] as the baselines for these two categories, respectively. The ablation studies of other architectures will be conducted in Section IV-C.

A. Pyramid Binary Mask Supervision

Compared with binary scalar label, binary mask label [26], [27], [47], [48], [49], [50], [51] contains rich spatial context, which benefits the models’ cognition about the position of spoofing attacks. In order to exploit multi-scale

spatial information from binary mask label, we propose the concept of pyramid supervision, which decomposes the original pixel-wise label into multiple spatial scales for supervising multi-scale features. There are three main advantages of pyramid supervision: 1) with observation and guidance of the pyramid labels, the models are able to learn features from multiple perspectives (from local details to global semantics) simultaneously; 2) the decision relies on multi-scale predicted results instead of any specific spatial level, which is more robust due to rich context evidence; and 3) the model could predict the multi-scale binary maps on-the-fly, which enhances the interpretability and localizes the spoofing attacks in different granularities.

As shown in Fig. 2, given a detected RGB face image with size $3 \times 256 \times 256$ (Channel $\times$ Height $\times$ Width) as input X , the deep features $\mathcal{F}_8$ with size $C \times 8 \times 8$ could be extracted via forwarding the backbone networks φ (e.g., ResNet50). Then average pooling operators with different kernel size and strides are performed on $\mathcal{F}_8$ to generate multi-scale features $\mathcal{F}_4$, $\mathcal{F}_2$ and $\mathcal{F}_1$, respectively. Subsequently, respective feature-to-mask mappings (Θ_8 , Θ_4 , Θ_2 and Θ_1) with 1×1 convolution are conducted on multi-scale features ($\mathcal{F}_8$, $\mathcal{F}_4$, $\mathcal{F}_2$ and $\mathcal{F}_1$) to predict the corresponding binary masks ($\mathcal{M}_8$, $\mathcal{M}_4$, $\mathcal{M}_2$ and $\mathcal{M}_1$), which are then concatenated for final binary classification. The procedure of multi-scale binary mask prediction could be formulated as:

$$\mathcal{M}_i = \text{Sigmoid}(\Theta_i(\text{AvgPool}_i(\varphi(X)))) , i = \{8, 4, 2, 1\}, \quad (1)$$

where the setting $\text{kernel} \& \text{stride} = 1, 2, 4, 8$ is utilized for AvgPool_8 , AvgPool_4 , AvgPool_2 , AvgPool_1 , respectively. The *Sigmoid* function guarantees the predicted masks ranging from [0, 1].

Loss Function: In terms of the pixel-wise ground truth Y , we could use the already-annotated binary mask labels directly or the generated coarse binary masks. Then we decompose original mask labels into multi-scale mask labels Y_8 , Y_4 , Y_2 ,

Fig. 3. Pyramid supervision with multi-scale depth map labels. The pyramid depth loss could be calculated from the pair-wise errors between the predicted and ground truth depth maps with the same scale.

and Y_1 with the following criterion: as multi-scale labels are sub-sampled from Y , the positional values are determined by whether the corresponding local receptive fields have the spoofing attacks or not (see Fig. 2 bottom for instance). As the predicted multi-scale masks and the ground truth have the same size, the pyramid loss $\mathcal{L}_{pyramid}$ can be calculated via accumulating the binary cross-entropy loss (BCE) from each position within each scale:

$$\mathcal{L}_{pyramid} = \sum_{i=8,4,2,1} -(Y_i \log(\mathcal{M}_i) + (1 - Y_i) \log(1 - \mathcal{M}_i)). \quad (2)$$

In the training stage, the overall loss $\mathcal{L}_{overall}$ can be formulated as $\mathcal{L}_{overall} = \mathcal{L}_{pyramid} + \mathcal{L}_{binary}$, where the latter is the BCE loss for final classification. In the evaluation stage, for simplicity, only the final binary score is utilized for decision.

Relation to Spatial Pyramid Pooling [57]: Here we discuss the relation between pyramid supervision and spatial pyramid pooling (SPP) [57], which share similar design philosophy but with different focuses. On one hand, SPP concatenates the deep multi-scale features directly for classification, which has huge feature dimension thus easily overfits in the FAS task (small-scale datasets). In contrast, in the framework of pyramid supervision, the predicted multi-scale compact maps are concatenated for final decision. On the other hand, SPP only provides the scalar-level supervision signal at the tail, while the proposed method could give multi-scale pixel-wise supervision before final classification. The ablation study is in Section IV-C to show superior performance of pyramid supervision for FAS.

B. Pyramid Depth Map Supervision

Besides assembling deep models with pyramid binary mask supervision, we also show an example about applying pyramid depth map supervision in the recent CDCN. As shown in Fig. 3, CDCN extracts multi-level features from the face input with size $3 \times 256 \times 256$, and predicts the grayscale facial depth with size 32×32 . Similar to pyramid binary mask supervision, both the predicted depth map $\mathcal{D}_{32}$ and generated pseudo depth label Y^{depth} are downsampled and resized to the same scales (32×32 , 16×16 , etc.). Here we adopt pyramid supervision with two scales (i.e., 32×32 and 16×16) as

Fig. 4. Visualization of the live and spoofing faces from the OULU-NPU [58], CASIA-MFSD [59], Replay-Attack [60], and MSU-MFSD [61]. It can be seen that serious domain shifts occur among these four datasets.

default setting, as we find that more scales make no contributions for performance improvement (see ablation study in Section IV-C). In terms of loss function, pyramid depth loss $\mathcal{L}_{pyramid}^{depth}$ can be formulated as

$$\mathcal{L}_{pyramid}^{depth} = \sum_{i=32,16} \left(\mathcal{L}_{MSE}(\mathcal{D}_i, Y_i^{depth}) + \mathcal{L}_{CDL}(\mathcal{D}_i, Y_i^{depth}) \right), \quad (3)$$

where $\mathcal{D}_i$ means the predicted depth map with scale i . $\mathcal{L}_{MSE}$ and $\mathcal{L}_{CDL}$ denotes mean square error loss and contrastive depth loss [45], respectively. In the training stage, only $\mathcal{L}_{pyramid}^{depth}$ is used for supervision. In the testing stage, we calculate the mean value of the predicted depth maps from all scales as the final score.

IV. EXPERIMENTS

In this part, extensive experiments are conducted to demonstrate the effectiveness of the proposed pyramid supervision. In the following, we sequentially describe the employed datasets & metrics (Section IV-A), implementation details (Section IV-B), results (Section IV-C–IV-F) and visualization (Section IV-G).

A. Databases and Metrics

Datasets: Five datasets including OULU-NPU [58], CASIA-MFSD [59], Replay-Attack [60], MSU-MFSD [61] and SiW-M [27] are used in our experiments. OULU-NPU is a high-resolution database, containing four sub-protocols to validate the generalization (e.g., unseen illumination and attack medium) of models, which is used for intra-dataset testing. The videos in CASIA-MFSD, Replay Attack, MSU-MFSD and OULU-NPU are recorded under quite different scenarios with variant cameras and subjects, which are used for cross-dataset domain generalization protocol. As the SiW-M dataset has rich (13) attack types inside, we follow the original intra-dataset cross-type protocol on SiW-M to evaluate the generalization ability for unseen attack types. Some typical bonafide and PAs samples from these datasets are showed in Fig. 4.

Evaluation Metrics: In the OULU-NPU dataset, we follow the original sub-protocols and metrics, i.e., Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and ACER [62] for a fair comparison. Area Under Curve (AUC) and Half Total Error Rate (HTER) are utilized for the cross-dataset domain generalization protocol on CASIA-MFSD, Replay-Attack, MSU-MFSD and OULU-NPU. For the intra-dataset cross-type protocol on SiW-M, both ACER and Equal Error Rate (EER) are employed.

B. Implementation Details

Ground Truth Generation: The pseudo facial depth map label is generated by the off-the-shelf 3D face model [63]. Apart from the SIW-M (already-annotated labels), the binary mask labels are generated simply by filling each position with corresponding binary label for other four datasets. The generated binary and depth maps are then downsampled into multiple resolution (32×32 , 16×16 , 8×8 , etc.) The live depth map is normalized in a range of $[0, 1]$, while the spoof one is all 0 at the training stage, which is beneficial for learning discriminative patterns for the FAS task.

Pre-processing: For each video frame, MTCNN [64] is used for face bounding box generation. As for the data augmentation strategies, random horizontal flip, color jitter on brightness, contrast, and saturation, and Cutout [65] are utilized. Moreover, face scales are randomly expanded or shrunk with times $[0.8, 1.3]$ to enrich the face/background context.

Experimental Setting: Our proposed method is implemented with Pytorch. The ImageNet pretrained classification based models (e.g., ResNet and DenseNet) are trained on the FAS task using SGD optimizer with the initial learning rate (lr), momentum, and weight decay (wd) are $1e-3$, 0.9 , and $5e-5$, respectively. We train models with maximum 80 epochs and batchsize 32 on single Nvidia V100 GPU. As for the multi-scale FCN models (e.g., CDCN with $\theta = 0.7$ and DepthNet), Adam optimizer with initial $lr=1e-4$ and $wd=5e-5$ is used. The models are trained with batchsize 8 for maximum 800 epochs while lr halves at the 500th epoch.

C. Ablation Study

In this subsection, all ablation studies are conducted on the most challenging Protocol-4 (different illumination conditions, locations, cameras and attacks between training and testing sets) of OULU-NPU [58] to explore the properties of pixel-wise supervision as well as the details of our proposed pyramid supervision.

Impact of Pixel-Wise Supervision for Deep Models: Here we fully evaluate the performance of two models (ResNet50 and CDCN) with variant supervision signals, which are shown in Table II. It can be seen from the 1st, 5th and 8th rows that binary mask supervision sharply improves the performance of ResNet50 compared with using either binary score or depth map supervision. It is surprising that supervision with depth map label performs the worst. This might be caused by the contradiction between semantic representation learning and the

TABLE II
THE ABLATION STUDY ABOUT SCALES OF PYRAMID SUPERVISION WITH BINARY MASK AND DEPTH MAP LABELS. ‘SPP’ DENOTES SPATIAL PYRAMID POOLING [57]

Model	Pixel-wise Label	Scale	ACER(%)
ResNet50	-	-	7.1 $\pm$ 3.4
ResNet50	-	2x2, 1x1 (SPP)	6.8 $\pm$ 4.2
ResNet50	-	4x4, 2x2, 1x1 (SPP)	7.6 $\pm$ 4.1
ResNet50	-	8x8, 4x4, 2x2, 1x1 (SPP)	8.9 $\pm$ 4.5
ResNet50	depth map	8x8	9.8 $\pm$ 5.1
ResNet50	depth map	8x8, 4x4	9.2 $\pm$ 4.8
ResNet50	depth map	8x8, 4x4, 2x2	8.9 $\pm$ 4.9
ResNet50	binary mask	8x8	6.3 $\pm$ 3.7
ResNet50	binary mask	8x8, 4x4	5.4 $\pm$ 3.0
ResNet50	binary mask	4x4, 1x1	5.6 $\pm$ 2.5
ResNet50	binary mask	2x2, 1x1	5.8 $\pm$ 1.5
ResNet50	binary mask	8x8, 4x4, 2x2	4.8 $\pm$ 2.8
ResNet50	binary mask	8x8, 2x2, 1x1	4.4 $\pm$ 3.0
ResNet50	binary mask	4x4, 2x2, 1x1	5.2 $\pm$ 2.4
ResNet50	binary mask	8x8, 4x4, 2x2, 1x1	4.4$\pm$2.7
ResNet50	multi-class mask	8x8	6.7 $\pm$ 2.5
ResNet50	multi-class mask	8x8, 4x4, 2x2, 1x1	5.4 $\pm$ 1.7
CDCN	binary mask	32x32	7.5 $\pm$ 4.7
CDCN	binary mask	32x32, 16x16	6.8 $\pm$ 4.6
CDCN	binary mask	32x32, 16x16, 8x8	6.6 $\pm$ 5.0
CDCN	binary mask	32x32, 16x16, 8x8, 4x4	6.7 $\pm$ 4.8
CDCN	depth map	32x32	6.9 $\pm$ 2.9
CDCN	depth map	32x32, 16x16	4.8$\pm$1.8
CDCN	depth map	32x32, 16x16, 8x8	5.4 $\pm$ 2.5
CDCN	depth map	32x32, 8x8, 4x4	6.7 $\pm$ 3.7
CDCN	depth map	32x32, 16x16, 8x8, 4x4	7.5 $\pm$ 2.9

fine-grained depth guidance. One more interesting finding is that when replacing binary mask with multi-class mask labels (3 classes, i.e., live, print and replay are used in OULU-NPU), there is no performance gains achieved. With more detailed spoofing cues, supervision with multi-class pixel-wise label is worth rethinking and exploring in the future. In terms of the CDCN, it is reasonable that supervision with depth map performs better than that with binary mask. In the following experiments, we adopt the configurations, i.e., ResNet50 with four-scale (8×8 , 4×4 , 2×2 , 1×1) binary mask supervision and CDCN with two-scale (32×32 , 16×16) depth map supervision.

Impact of Scales of Pyramid Supervision: It can be seen from Table II that, 1) pyramid binary mask supervision with four scales (8×8 , 4×4 , 2×2 , 1×1) has the largest gain ($-1.9 \pm 1.0\%$ ACER) for ResNet50 compared with the vanilla one scale (8×8 only) setting; and 2) pyramid depth map supervision with two scales (32×32 , 16×16) significantly boosts the vanilla CDCN with only 32×32 depth supervision for $-2.1 \pm 1.1\%$ ACER. As a result, these two kinds of scales are utilized as the default settings for pyramid binary mask and depth map supervision, respectively. It is interesting to find that more scales always benefit the ResNet with pyramid binary mask supervision but not the CDCN with pyramid depth map supervision. This is possible because 1) pyramid binary mask supervision is likely to provide multi-scale fine-grained signals for ResNet50 to learn reasonable semantics; and 2) lower-resolution depth map labels might disturb CDCN to learn the local detailed representation.

We also study the impacts of pyramid depth map supervision for ResNet50 (the 2nd block) and pyramid binary mask supervision for CDCN (in the second last block) in Table II.

Fig. 5. Performance of various architectures w/ and w/o pyramid supervision on Protocol-4 of OULU-NPU. The pyramid supervision for ‘DepthNet’ is with two-scale (32×32 , 16×16) depth maps while the other architectures are with four-scale (8×8 , 4×4 , 2×2 , 1×1) binary masks.

On one hand, compared with single-scale 8×8 supervision, ResNet50 supervised with multi-scale 8×8 , 4×4 and 8×8 , 4×4 , 2×2 perform slightly better (with $<1\%$ ACER decline). The low resolutions 4×4 and 2×2 of the downsampled depth maps cannot bring very satisfied improvement. On the other hand, as for the CDCN, multi-scale binary mask supervision helps to further performance boosts compared with single-scale 32×32 mask. However, there is no obvious improvement when using more scales such as 4×4 and 2×2 .

Pyramid Supervision vs. Spatial Pyramid Pooling (SPP) [57]: As pyramid supervision and SPP share the similar insight of multi-scale features, here we also study the performance of SPP. Specifically, SPP module is cascaded with the deep features from ResNet50, and then binary score label is utilized for supervision. Despite with gained performance in generic object detection and classification tasks, SPP could not benefit the FAS task. It is interesting to find that when assembled SPP with more multiple scales, the model performs even worse. To be specific, ResNet50 with two-scale SPP 2×2 , 1×1 brings slight performance improvement while that with three scales 4×4 , 2×2 , 1×1 or four scales 8×8 , 4×4 , 2×2 , 1×1 causes ACER decline. It is possibly caused by the overfitting from SPP module with high-dimensional features. Note that compared with vanilla ResNet50, dimensions of the embedding features from the ones with two-, three-, and four-scale SPPs are 4, 20, 84 times larger. In contrast, the proposed pyramid binary mask supervision with scales 2×2 , 1×1 , 4×4 , 2×2 , 1×1 , and 8×8 , 4×4 , 2×2 , 1×1 outperforms ResNet50 with SPP by large margins (1.0 ± 2.7 , 2.4 ± 1.7 , and $4.5 \pm 1.8\%$ ACER, respectively), indicating the efficiency of the pyramid supervision for the predicted compact multi-scale binary masks.

Generalization to Architectures: Besides ResNet50 and CDCN, we also study the generalization ability on more architectures. As illustrated in Fig. 5, pyramid binary mask supervision improves both large models (VGG16 [66] and DenseNet121 [56]) and mobile models (ShuffleNetV2 [67] and MobileNetV2 [68]) consistently, indicating the promising application of the proposed pyramid supervision for further

TABLE III
THE RESULTS OF INTRA TESTING ON OULU-NPU [58]. ‘ 8×8 ’ AND ‘-PS’ DENOTE WITH 8×8 BINARY MASK AND PYRAMID SUPERVISION, RESPECTIVELY

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
1	DRL-FAS	5.4	4.0	4.7
	STASN [11]	1.2	2.5	1.9
	Auxiliary [10]	1.6	1.6	1.6
	FaceDs [34]	1.2	1.7	1.5
	FAS-SGTD [45]	2.0	0.0	1.0
	SpoofTrace [50]	0.8	1.3	1.1
	Disentangled [52]	1.7	0.8	1.3
	BCN [9]	0.0	1.6	0.8
	DeepPixBiS [26]	0.8	0.0	0.4
	ResNet50_8x8 [55]	2.3	6.7	4.5
	ResNet50-PS (Ours)	1.9	6.4	4.1
	CDCN [8]	0.4	1.7	1.0
	CDCN-PS (Ours)	0.4	1.2	0.8
2	DeepPixBiS [26]	11.4	0.6	6.0
	FaceDs [34]	4.2	4.4	4.3
	Auxiliary [10]	2.7	2.7	2.7
	STASN [11]	4.2	0.3	2.2
	FAS-SGTD [45]	2.5	1.3	1.9
	Disentangled [52]	1.1	3.6	2.4
	DRL-FAS	3.7	0.1	1.9
	BCN [9]	2.6	0.8	1.7
	SpoofTrace [50]	2.3	1.6	1.9
	ResNet50_8x8 [55]	0.6	10	5.3
	ResNet50-PS (Ours)	1.9	7.5	4.7
	CDCN [8]	1.5	1.4	1.5
	CDCN-PS (Ours)	1.4	1.4	1.4
3	DeepPixBiS [26]	11.7±19.6	10.6±14.1	11.1±9.4
	FAS-SGTD [45]	3.2±2.0	2.2±1.4	2.7±0.6
	FaceDs [34]	4.0±1.8	3.8±1.2	3.6±1.6
	DRL-FAS	4.6±3.6	1.3±1.8	3.0±1.5
	Auxiliary [10]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [11]	4.7±3.9	0.9±1.2	2.8±1.6
	Disentangled [52]	2.8±2.2	1.7±2.6	2.2±2.2
	SpoofTrace [50]	1.6±1.6	4.0±5.4	2.8±3.3
	BCN [9]	2.8±2.4	2.3±2.8	2.5±1.1
	ResNet50_8x8 [55]	2.9±1.2	4.2±1.9	3.6±1.4
	ResNet50-PS (Ours)	2.2±1.4	4.0±1.7	3.1±1.6
	CDCN [8]	2.4±1.3	2.2±2.0	2.3±1.4
	CDCN-PS (Ours)	1.9±1.7	2.0±1.8	2.0±1.7
4	DeepPixBiS [26]	36.7±29.7	13.3±14.1	25.0±12.7
	Auxiliary [10]	9.3±5.6	10.4±6.0	9.5±6.0
	FAS-SGTD [45]	6.7±7.5	3.3±4.1	5.0±2.2
	STASN [11]	6.7±10.6	8.3±8.4	7.5±4.7
	DRL-FAS [69]	8.1±2.7	6.9±5.8	7.2±3.9
	FaceDs [34]	1.2±6.3	6.1±5.1	5.6±5.7
	Disentangled [52]	5.4±2.9	3.3±6.0	4.4±3.0
	SpoofTrace [50]	2.3±3.6	5.2±5.4	3.8±4.2
	BCN [9]	2.9±4.0	7.5±6.9	5.2±3.7
	ResNet50_8x8 [55]	5.8±3.2	6.8±4.2	6.3±3.7
	ResNet50-PS (Ours)	2.9±4.0	5.8±4.9	4.4±2.7
	CDCN [8]	4.6±4.6	9.2±8.0	6.9±2.9
	CDCN-PS (Ours)	2.9±4.0	5.8±4.9	4.8±1.8

vision models. In addition, with pyramid depth map supervision, both the mean and standard deviation ACER are decreased on DepthNet, which shows the excellent generalization of pyramid supervision.

D. Intra-Dataset Intra-Type Testing on OULU-NPU

We strictly follow the four protocols on OULU-NPU for the evaluation. As shown in Table III, with our proposed pyramid supervision, the ACER are decreased on four protocols consistently for both ResNet50 (-0.4% , -0.6% , -0.5% , and -1.9%) and CDCN (-0.2% , -0.1% , -0.2% , and -2.1%), respectively. It indicates that pyramid supervision performs well at the generalization of the external environment, attack mediums and input camera variation.

In terms of architectures, ‘CDCN-PS’ performs better than or on par with the state-of-the-art methods in all four protocols. As for ‘ResNet50-PS’, despite with bad performance

TABLE IV
RESULTS OF CROSS-DATASET INTRA-TYPE TESTING ON OULU-NPU, CASIA-MFSD, REPLAY-ATTACK, AND MSU-MFSD

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Color Texture [70]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
Auxiliary (Depth) [10]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	66.61
MMD-AAE [71]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [39]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
DR-MD-Net [41]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
BCN [9]	19.81	88.23	25.12	82.68	22.75	83.85	21.24	86.25
OCA-FAS [72]	15.51	91.92	18.45	89.33	19.51	88.29	19.25	88.55
RFMeta [40]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
ResNet50_8x8 [55]	18.80	90.48	18.26	86.42	19.83	88.24	26.47	80.72
ResNet50-PS (Ours)	<u>14.32</u>	94.51	18.23	89.75	18.86	89.63	21.44	87.56
CDCN [8]	22.90	85.45	22.46	86.64	19.98	84.75	16.92	90.46
CDCN-PS (Ours)	20.42	87.43	18.25	86.76	19.55	86.38	15.76	92.43

for the first three protocols, it surprisingly achieves better performance than ‘CDCN-PS’ in the most challenging Protocol-4, which indicates the efficacy of pyramid binary mask supervision even with limited training samples. In the future, it is promising to plug pyramid supervision into more advanced architectures as well as pixel-wise labels.

E. Cross-Dataset Intra-Type Testing

Here we conduct experiments on two kinds of cross-dataset intra-type testing protocols to verify the effectiveness of pyramid supervision.

Firstly, we follow the well-established domain generalization cross-dataset intra-type testing protocols [39], [40], [41], [72] including four datasets OULU-NPU (O), CASIA-MFSD (C), Idiap Replay-Attack (I) and MSU-MFSD (M) for experiments. Specifically, the training sets randomly selected from three datasets and the testing set is from the remained one. For example, when doing ‘O&C&I’ to M, the training sets from ‘O&C&I’ are used for training while the testing set is used from M for testing. The HTER threshold is calculated based on the EER threshold from the targeted testing set directly. It can be seen from Table IV that with pyramid binary mask supervision, the performances of ‘ResNet50-PS’ are improved remarkably by -4.48% and -5.03% HTER on protocols ‘O&C&I to M’ and ‘I&C&M to O’, respectively. Similarly, CDCN with pyramid depth map supervision outperforms the vanilla CDCN by -2.48% , -4.21% and -1.16% HTER on protocols ‘O&C&I to M’, ‘O&M&I to C’ and ‘I&C&M to O’, respectively. The experimental results indicate that the proposed pyramid supervision is also helpful for providing rich multi-scale guidance on multiple source domains.

Secondly, we conduct extra cross-dataset testings to further testify the generalization ability of our models there is only one single training dataset (domain). There are two sub-protocols: first is to train on the training set of CASIA-MFSD and test on the testing set of Replay-Attack, i.e., protocol CR; the second one is exchanging the training dataset and the testing dataset, i.e., protocol RC. The HTER threshold is also determined by the EER threshold achieved on testing dataset. As shown in Table V, our proposed pyramid supervision could improve the performance of both vanilla ResNet50 (-3.1% and -2.9%

TABLE V
THE HTER(%) RESULTS OF CROSS-DATASET TESTING BETWEEN CASIA-MFSD AND REPLAY-ATTACK

Method	Train	Test	Train	Test
	CASIA-MFSD	Replay-Attack	Replay-Attack	CASIA-MFSD
ColorTexture [73]		30.3		37.7
FaceDs [34]		28.5		41.1
STASN [11]		31.5		30.9
DRL-FAS [69]		28.4		33.2
Auxiliary [10]		27.6		28.4
Disentangled [52]		22.4		30.3
FAS-SGTD [45]		17.0		22.8
BCN [9]		16.6		36.4
ResNet50_8x8 [55]		29.6		35.2
ResNet50-PS (Ours)		26.5		32.3
CDCN [8]		15.5		32.6
CDCN-PS (Ours)		13.8		31.3

HTER) and CDCN (-1.7% and -1.3% HTER) on the protocols CR and RC, respectively, which further demonstrates the generalization ability of the proposed methods.

F. Intra-Dataset Cross-Type Testing on SiW-M

Following the same cross-type testing protocol (13 attacks leave-one-out) on SiW-M, we compare our proposed methods with several recent FAS methods [9], [10], [27], [50], [58] to validate the generalization capacity of unseen attacks. As shown in Table VI, compared with the vanilla pixel-wise supervision, ‘ResNet50-PS’ and ‘CDCN-PS’ achieve an overall better EER with the improvement by 17% and 12% respectively. Specifically, benefitted from the pyramid supervision, our method can perform more robustly in the two challenging attacks (‘Transparent Mask’ and ‘Partial Paper’) while the vanilla supervision performs poorly on these two attacks. Note that the proposed methods only with one specific pixel-wise label achieve comparable performance with the state-of-the-art method BCN [9] supervised by three kinds of pixel-wise labels (i.e., binary mask, depth map and reflection map). In the future, extending pyramid supervision on multiple pixel-wise labels simultaneously for exploiting richer and more fine-grained cues is favorable.

G. Visualization and Analysis

Fig. 6 illustrates the predicted binary maps of live and spoof faces on the SiW-M dataset under cross-type testing protocol.

TABLE VI
RESULTS OF THE CROSS-TYPE TESTING ON SiW-M [27]

Method	Metrics(%)	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Average
				Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Im.	Cos.	Fun.	Glasses	Partial	
SVM+LBP [58]	ACER	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	26.9±14.5
	EER	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	24.5±12.9
Auxiliary [10]	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 [27]	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
SpoofTrace [50]	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 [9]	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
ResNet50_8x8 [55]	ACER	12.8	12.4	19.5	17.1	13.6	4.3	9.2	26.2	2.8	18.7	22.5	25.1	12.4	15.1±7.3
	EER	12.4	10.4	18.1	18.5	10.2	0.0	7.5	26.1	1.6	18.0	20.1	24.6	11.6	13.8± 8.0
ResNet50-PS (Ours)	ACER	10.6	10.6	20.2	15.9	10.0	1.2	10.0	23.3	0.4	14.0	20.5	24.3	5.8	12.8±7.8
	EER	11.3	9.6	19.4	14.8	6.8	0.0	7.5	21.7	0.0	12.4	19.5	21.4	5.0	11.5± 7.6
CDCN [8]	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
CDCN-PS (Ours)	ACER	12.1	7.4	9.9	9.1	14.8	5.3	5.9	43.1	0.4	13.8	24.4	18.1	3.5	12.9±11.1
	EER	10.3	7.8	8.3	7.4	10.2	5.9	5.0	43.4	0.0	12.0	23.9	15.9	0.0	11.5±11.4

Fig. 6. Visualization of the predicted binary masks from ‘ResNet50’ and ‘ResNet50-PS’ on SiW-M. Higher predicted scores (yellow) indicates higher probability of liveness.

On one hand, in terms of the predictions from ‘Live’, ‘Replay’ and ‘Transparent Mask’, both ‘ResNet50_8j’ and ‘ResNet50-PS’ perform well and keep high confidence for discrimination. On the other hand, when fed with the unseen and challenging attack types (e.g., ‘Partial Print’ and ‘Half Mask’), the predictions become less reliable and chaotic. To be specific, it can be seen from the ‘ResNet50_8 × 8’ results in 3rd row, 2nd column that besides the print parts in the eye regions, other facial regions are predicted as low liveness. The similar result also occurs in 3rd row, 2nd column with high liveness confidence in the mask region. In contrast, assembled with pyramid supervision, the interpretability about the spoof localization improves significantly. From the predicted 8 × 8 and 4 × 4 maps We can find that the high scores (liveness) and low

activations (spoofing) position in the facial skin regions and spoof mediums, respectively. With the evolution of spoofing attacks, we believe that network interpretability becomes more and more important in spoof localization and understanding.

V. CONCLUSION

In this paper, we give a elaborate review about face anti-spoofing (FAS) with pixel-wise supervision and conduct extensive experiments to study corresponding factors. Moreover, we propose a novel pyramid supervision, intending to provide richer multi-scale spatial context for fine-grained supervision. Experimental results demonstrate the effectiveness of our proposed pyramid supervision on both performance improvement and interpretability enhancement.

The possible future directions include: 1) discovering the optimal pixel-wise labels with suitable supervision strategies (e.g., pyramid supervision) automatically is worth exploring; and 2) large-scale FAS benchmark with fine-grained (i.e., multiple classes) pixel/patch-level annotated labels should be established, which also helps to quantitatively evaluate the interpretability.

ACKNOWLEDGMENT

The authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

REFERENCES

- [1] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, “Learning meta face recognition in unseen domains,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 6163–6172.
- [2] S. Jia, G. Guo, and Z. Xu, “A survey on 3D mask presentation attack detection and countermeasures,” *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107032.
- [3] G. Pan, L. Sun, Z. Wu, and S. Lao, “Eyeblink-based anti-spoofing in face recognition from a generic webcam,” in *Proc. Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.
- [4] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, “Generalized face anti-spoofing by detecting pulse from face videos,” in *Proc. IEEE 23rd Int. Conf. Assoc. Pattern Recognit.*, Cancun, Mexico, 2016, pp. 4244–4249.
- [5] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, “LBP—Top based countermeasure against face spoofing attacks,” in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 121–132.

- [6] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *Proc. IEEE Biometrics Theory Appl. Syst.*, Arlington, VA, USA, 2013, pp. 1–8.
- [7] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, pp. 2268–2283, 2016.
- [8] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 5295–5305.
- [9] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 557–575.
- [10] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 389–398.
- [11] X. Yang *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3507–3516.
- [12] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics*, Denver, CO, USA, 2017, pp. 319–328.
- [13] Z. Yu *et al.*, "Multi-modal face anti-spoofing based on central difference networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Seattle, WA, USA, 2020, pp. 650–651.
- [14] S. Zhang *et al.*, "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 2, no. 2, pp. 182–193, Apr. 2020.
- [15] H.-K. Jee, S.-U. Jung, and J.-H. Yoo, "Liveness detection for embedded face recognition system," *Int. J. Biol. Med. Sci.*, vol. 1, no. 4, pp. 235–238, 2006.
- [16] J.-W. Li, "Eye blink detection based on multiple gabor response waves," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, vol. 5, Kunming, China, 2008, pp. 2852–2856.
- [17] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *Tsinghua Sci. Technol.*, vol. 14, no. 6, pp. 685–690, Dec. 2009.
- [18] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Linhai, China, 2009, pp. 233–236.
- [19] J. Bigun, M. Fronthaler, and K. Kollreider, "Assuring liveness in biometric identity authentication by real-time face tracking," in *Proc. IEEE Int. Conf. Comput. Intell. Homeland Security Pers. Safety (CIHSPS)*, Venice, Italy, 2004, pp. 104–111.
- [20] A. Ali, F. Deravi, and S. Hoque, "Liveness detection using gaze collinearity," in *Proc. IEEE 3rd Int. Conf. Emerg. Security Technol.*, Lisbon, Portugal, 2012, pp. 62–65.
- [21] B. Lin, X. Li, Z. Yu, and G. Zhao, "Face liveness detection by rPPG features and contextual patch-based CNN," in *Proc. Int. Conf. Biomed. Eng. Appl.*, 2019, pp. 61–68.
- [22] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 151–160.
- [23] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, 2015, pp. 2636–2640.
- [24] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 504–517.
- [25] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
- [26] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. Int. Conf. Biometrics*, Crete, Greece, 2019, pp. 1–8.
- [27] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 4680–4689.
- [28] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, "Face spoofing detection based on local ternary label supervision in fully convolutional networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3181–3196, 2020.
- [29] T. Kim, Y. Kim, I. Kim, and D. Kim, "BASN: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing," in *Proc. Int. Conf. Comput. Vis. Workshops*, Seoul, South Korea, 2019, pp. 494–503.
- [30] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo, and S. Z. Li, "3DPC-Net: 3D point cloud network for face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics*, Houston, TX, USA, 2020, pp. 1–8.
- [31] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014. [Online]. Available: arXiv:1408.5601.
- [32] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. Int. Conf. Image Process. Theory Tools Appl.*, Oulu, Finland, 2016, pp. 1–6.
- [33] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. Chin. Conf. Biometric Recognit.*, 2016, pp. 611–619.
- [34] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 290–306.
- [35] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3D face anti-spoofing with factorized bilinear coding," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 15, 2020, doi: [10.1109/TCSVT.2020.3044986](https://doi.org/10.1109/TCSVT.2020.3044986).
- [36] L. Li, Z. Xia, X. Jiang, F. Roli, and X. Feng, "CompactNet: Learning a compact space for face presentation attack detection," *Neurocomputing*, vol. 409, pp. 191–207, Oct. 2020.
- [37] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 9, 2020, doi: [10.1109/TPAMI.2020.3036338](https://doi.org/10.1109/TPAMI.2020.3036338).
- [38] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2639–2652, 2018.
- [39] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 10023–10031.
- [40] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proc. Int. Conf. Assoc. Adv. Artif. Intell.*, 2020, pp. 11974–11981.
- [41] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 6678–6687.
- [42] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.
- [43] Y. Qin *et al.*, "Learning meta model for zero-and few-shot face anti-spoofing," in *Proc. Int. Conf. Assoc. Adv. Artif. Intell.*, 2020, pp. 11916–11923.
- [44] D. Peng, J. Xiao, R. Zhu, and G. Gao, "Ts-Fen: Probing feature selection strategy for face anti-spoofing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 2942–2946.
- [45] Z. Wang *et al.*, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 5042–5051.
- [46] Y. Chen, T. Wang, J. Wang, P. Shi, G. Shan, and H. Snoussi, "Towards good practices in face anti-spoofing: An image reconstruction based method," in *Proc. IEEE Chin. Autom. Congr.*, Hangzhou, China, 2019, pp. 4700–4705.
- [47] M. S. Hossain, L. Rupty, K. Roy, M. Hasan, S. Sengupta, and N. Mohammed, "A-DeepPixBis: Attentional angular margin for face anti-spoofing," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl.*, Melbourne, VIC, Australia, 2020, pp. 1–8.
- [48] Z. Yu *et al.*, "Auto-FAS: Searching lightweight networks for face anti-spoofing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 996–1000.
- [49] Y. Ma, L. Wu, Z. Li, and F. Liu, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," *Pattern Recognit. Lett.*, vol. 131, pp. 261–267, Mar. 2020.
- [50] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 406–422.
- [51] W. Sun, Y. Song, H. Zhao, and Z. Jin, "A face spoofing detection method based on domain adaptation and lossless size adaptation," *IEEE Access*, vol. 8, pp. 66553–66563, 2020.
- [52] K.-Y. Zhang *et al.*, "Face anti-spoofing via disentangled representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 641–657.
- [53] Y. Zhang *et al.*, "CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 70–85.
- [54] H. Feng *et al.*, "Learning generalized spoof cues for face anti-spoofing," 2020. [Online]. Available: arXiv:2005.03922.

- [55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
- [56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4700–4708.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
- [58] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, 2017, pp. 612–618.
- [59] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *Proc. IEEE 5th IAPR Int. Conf. Biometrics*, New Delhi, India, 2012, pp. 26–31.
- [60] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Spec. Interest Group*, Darmstadt, Germany, 2012, pp. 1–7.
- [61] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, pp. 746–761, 2015.
- [62] *ISO/IEC JTC 1/SC 37 Biometrics: Information Technology—Biometric Presentation Attack Detection—Part 1: Framework*, Standard ISO/IEC 30107-1:2016, 2016. [Online]. Available: <https://www.iso.org/obp/ui/iso>
- [63] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 152–168.
- [64] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [65] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017. [Online]. Available: [arXiv:1708.04552](https://arxiv.org/abs/1708.04552).
- [66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
- [67] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.
- [68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.
- [69] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 937–951, 2020.
- [70] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, pp. 1818–1830, 2016.
- [71] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5400–5409.
- [72] Y. Qin, W. Zhang, J. Shi, Z. Wang, and L. Yan, "One-class adaptation face anti-spoofing with loss function search," *Neurocomputing*, vol. 417, pp. 384–395, Dec. 2020.
- [73] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017.
- [74] A. Liu *et al.*, "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometri.*, vol. 10, no. 1, pp. 24–43, 2021.

Xiaobai Li (Member, IEEE) received the B.Sc. degree in psychology from Peking University, the M.Sc. degree in biophysics from the Chinese Academy of Science, and the Ph.D. degree in computer science from the University of Oulu, where she is currently an Assistant Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. Her research of interests include spontaneous versus posed facial expression comparison, micro-expression and deceitful behaviors, and heart rate measurement from facial videos.

Jingang Shi received the B.S. and Ph.D. degrees from the Department of Electronics and Information Engineering, Xi'an Jiaotong University, China. From 2017 to 2020, he was a Postdoctoral Researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Since 2020, he has been an Associate Professor with the School of Software, Xi'an Jiaotong University. His current research interests mainly include image restoration, face analysis, and biomedical signal processing.

Zhaoqiang Xia (Member, IEEE) received the B.E. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2014, respectively. He has been a Visiting Scholar with the University of North Carolina at Charlotte from 2011 to 2013, and a Senior Researcher with the University of Oulu in 2019. He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. He has authored or coauthored more than 60 papers in international journals and conferences. His current research interests include image processing, visual search and recognition, and statistical machine learning. He has served as a reviewer for mainstream journals and a session co-chairs for international conferences.

Zitong Yu (Graduate Student Member, IEEE) received the M.S. degree from the University of Nantes, France, in 2016. He is currently pursuing the Ph.D. degree with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His research interests focus on remote physiological measurement, face anti-spoofing, and video understanding.

Guoying Zhao (Senior Member, IEEE) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where she has been a Senior Researcher since 2005, and an Associate Professor since 2014. She has authored or coauthored more than 240 papers in journals and conferences. Her papers have currently over 14 500 citations in Google Scholar (H-index 55). She has served as a area chairs for several conferences and is an Associate Editor for *Pattern Recognition*, *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, and *Image and Vision Computing* journals. She is a Fellow of IAPR.