Learnable Descriptive Convolutional Network for Face Anti-Spoofing

Pei-Kai Huang[1]
alwayswithme@gapp.nthu.edu.tw
Hui-Yu Ni[1]
huiyu8794@gapp.nthu.edu.tw
Yan-Qin Ni[2]
niyanqin1022@gmail.com
Chiou-Ting Hsu[1]
cthsu@cs.nthu.edu.tw

[1] Department of Computer Science, National Tsing Hua University, Taiwan
[2] Department of Computer Science and Information Engineering, National Central University, Taiwan

Abstract

Face anti-spoofing aims to counter facial presentation attacks and heavily relies on identifying live/spoof discriminative features. In this paper, we propose a novel Learnable Descriptive Convolution (LDC) to expand the representation capacity of vanilla convolution and especially focus on learning intrinsic textural features of live and spoof faces. In terms of LDC, we develop a convolutional network LDCNet for face anti-spoofing. In addition, to facilitate cross-domain detection, we introduce two strategies, including triplet mining and dual-attention supervision, to constrain the model training. We adopt triple mining to encourage LDCNet to learn to narrow the domain gap, and adopt the dual-attention supervision to guide LDCNet on learning discriminative features from regional live and spoof attentions. With the collaborative supervision of the two strategies, we conduct extensive experiments and show that LDCNet achieves promising results on many benchmark datasets. The codes are available at https://github.com/huiyu8794/LDCNet.

1 Introduction

Facial recognition and identification systems have penetrated our daily lives to quickly facilitate many applications, such as cellphone unlock, authentication of on-line banking, and identification of criminals. However, extensive applications of facial recognition also incur potential security risks and require specific techniques to support the application security. Therefore, many face anti-spoofing methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 30, 31, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47] have been developed to counter facial presentation attacks, such as Print Attack (i.e., printing a face on a paper) and Replay Attack (i.e., replaying a face video on digital devices).

Earlier methods [3, 9, 17, 24, 26, 40] developed their live/spoof classifiers in terms of handcrafted feature descriptors, such as local binary pattern (LBP) [3, 9, 24], histogram of gradient (HoG) [17, 40], and scale-invariant feature transform (SIFT) [26]. Recent deep

learning-based methods [1, 4, 5, 7, 8, 12, 13, 14, 16, 18, 19, 20, 21, 22, 23, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 46, 47] directly learned the features and/or the classifier from labeled training data and have achieved significant improvement over earlier methods. Convolutional neural networks (CNNs), with shared-weight architecture and vanilla convolutions, have achieved great success in many computer vision tasks. By stacking locally smoothing filters, CNNs successfully suppress image noises in different spatial scales and encode task-representative features for different applications. However, face anti-spoofing, unlike other computer vision tasks, deals with highly similar characteristics between live and spoof faces and requires a more delicate representation on characterizing the intrinsic features relating to face presentation attacks. For example, grid artifacts and moiré patterns are considered as distinctive textural features for detecting print attacks and replay attacks, respectively. Therefore, several extensions of vanilla CNNs, such as Sobel Convolution [37], Local Binary Convolution [15], and Central Difference Convolution [42], have been proposed to learn the textural features of face presentation attacks. These extensions, by including pre-defined local descriptors into vanilla convolution, have been shown to improve the representation capacity of vanilla convolution.

Another challenge in face anti-spoofing relates to the cross-domain issue. Because different benchmark datasets are independently collected and have various distributions, the model trained on one dataset often fails to perform well on the others. Several methods [6, 12, 13, 19, 30, 31, 34, 35, 38] have been proposed to address this cross-domain issue in face anti-spoofing problem. Moreover, because most benchmark datasets only provide binary live/spoof ground-truth labels to indicate whether an image is live or spoof, there is a lack of fine-grained supervision to guide the model learning. Therefore, many methods included external supervision, e.g., facial depth [6, 12, 21, 30, 31, 35, 42], rPPG signal [12, 18, 20, 21], and reflection [16, 41, 45, 47], as an auxiliary guide to help the models on learning discriminative features. These external auxiliary supervisions, though effective on specific scenarios, heavily rely on the availability and quality of the adopted information.

To address the above issues, we propose a Learnable Descriptive Convolutional Network (LDCNet) for face anti-spoofing and design a novel Learnable Descriptive Convolution (LDC) through a learnable local descriptor to increase the representation capacity of vanilla convolutions. The proposed LDC not only enables CNNs to adaptively learn the intrinsic textural features of live and spoof faces but also exhibits a generalized formulation of the other vanilla extensions [15, 37, 42]. To tackle the issues of cross-domain gap and the lack of fine-grained supervision, we further include two strategies to collaboratively supervise LDCNet, including triple mining and dual-attention supervision. We adopt the idea of triple mining to narrow the domain gap and to encourage LDCNet to learn domain-invariant features. Moreover, we introduce a dual-attention supervision, including a live attention and a spoof attention, to constrain LDCNet to focus on regional attentions in learning live/spoof discriminative features.

Our contributions are summarized as follows:

- We design a novel Learnable Descriptive Convolution (LDC) to adaptively learn the delicate textural features in face anti-spoofing and show that LDC is a generalized operation of other vanilla extensions.
- We incorporate the strategies of triplet mining and dual-attention supervision to collaboratively supervise LDCNet to learn domain-invariant and live/spoof discriminative features.
- Extensive experiments demonstrate effectiveness of the proposed method.

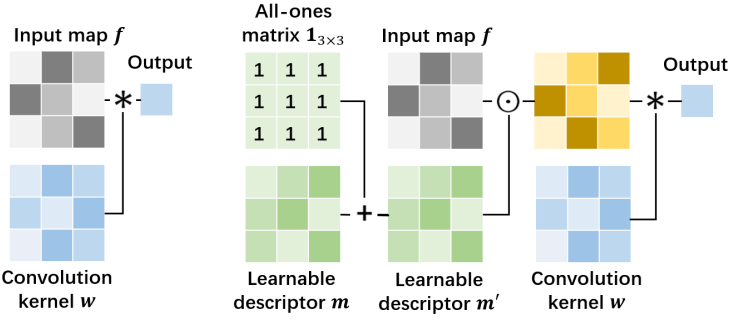


Figure 1: Illustration of the difference between (a) the vanilla convolution, and (b) the proposed Learnable Descriptive convolution (LDC). $\odot$ and $*$ denote the element-wise multiplication and the convolution operation, respectively.

2 Proposed Method

In Section 2.1, we first present the proposed Learnable Descriptive Convolution (LDC) on learning intrinsic textural features of face presentation attacks and then develop a LDC-based convolutional network (LDCNet) to address the face anti-spoofing problem. In Section 2.2 and Section 2.3, we introduce using triplet mining and dual-attention supervision to collaboratively supervise LDCNet towards learning domain-invariant and live/spoof discriminative features. Section 2.4 summarizes the total loss of LDCNet and describes the detection score for live/spoof classification in the inference stage.

2.1 Learnable Descriptive Convolutional Network

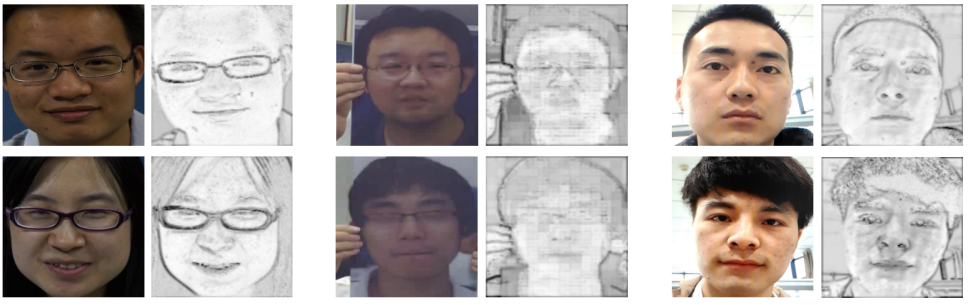
2.1.1 Review of Vanilla Convolution and Central Difference Convolution

Vanilla 2D spatial convolution is essential to CNNs for learning representative features for different tasks. The 2D convolution is a linear operation involving the multiplication of the filter w (i.e., a set of weights) with the input f in a local neighborhood $\mathcal{R}$ by,

$$g(p) = w(p) * f(p) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot f(p + p_n), \tag{1}$$

where $*$ denotes the convolution operation, p is the pixel of current location, p_n is the location of neighboring pixels in $\mathcal{R}$, and g is the output feature map. Figure 1 (a) shows an example of a 3×3 convolution operation in $\mathcal{R} = \{(-1,-1),(-1,0),...,(0,1),(1,1)\}$.

However, the representation capacity of vanilla convolution is not equally effective for all the computer vision tasks. In particular, as noted in [42], the weighted summation in vanilla convolution tends to overly smooth sharp details and diminish the discriminative textural features in face anti-spoofing. Therefore, several variations, such as Sobel Convolution [37], Local Binary Convolution [15], and Central Difference Convolution [42], have been proposed by involving local descriptors into vanilla convolution to better focus on edge or textural details. In [42], the authors proposed a Central Difference Convolution, by combin-



(a) (b) (c)

Figure 2: Examples of facial images and their **low-level** feature maps obtained by the proposed Learnable Descriptive Convolution. (a) Live images, (b) spoof images of print attacks, and (c) spoof images of replay attacks.

ing vanilla convolution with a weighted summation of central difference in $\mathcal{R}$ by,

$$g(p) = (1-\theta)\underbrace{\sum_{p_n\in\mathcal{R}} w(p_n)\cdot f(p+p_n)}_{\text{vanilla convolution}} + \theta\underbrace{\sum_{p_n\in\mathcal{R}} w(p_n)\cdot (f(p+p_n)-f(p))}_{\text{central difference convolution}}, \tag{2}$$

where $\theta \in [0,1]$ is a hyperparameter. When $\theta = 0$, Equation (2) degenerates to the vanilla convolution in Equation (1).

2.1.2 Learnable Descriptive Convolution

Although the methods [15, 37, 42] incorporated different local descriptors to extend the vanilla convolution, they all adopted predefined local descriptors and still kept all the learning capabilities in the convolution kernel w. That is, the local descriptors in [15, 37, 42] are fixed and not updated along with the model training. We argue that, the predefined and unlearnable descriptors are inflexible to capture various textural features and that their applicability in face anti-spoof is limited. Therefore, instead of predetermining the descriptor, we propose a novel Learnable Descriptive Convolution (LDC) by incorporating a learnable descriptor m' into vanilla convolution by,

$$g(p) = w(p) * (f(p) \odot m'(p)), \tag{3}$$

where $\odot$ denotes the element-wise multiplication. To simplify the description, we assume the local neighborhood $\mathcal{R}$, the convolution kernel w, and the learnable descriptor m' are of size 3×3. The descriptor m' is composed of a base matrix $\mathbf{1}_{3\times 3}$ and a learnable matrix m as,

$$m' = (1-\varepsilon)\cdot \mathbf{1}_{3\times 3} + \varepsilon \cdot m, \tag{4}$$

where ε is a learnable parameter and is used to balance the contribution of the two matrices. As shown in Figure 1 (b), the base matrix $\mathbf{1}_{3\times 3}$ is an all-ones matrix and is included to preserve the vanilla convolution function in LDC; and the 3×3 matrix m is initialized with $\mathbf{1}_{3\times 3}$ and then is jointly updated with ε during the training stage. By substituting Equation (4) into Equation (3), we rewrite the proposed LDC as,

$$\begin{aligned} g(p) &= w(p) * (f(p) \odot ((1-\varepsilon)\cdot \mathbf{1}_{3\times 3}(p) + \varepsilon \cdot m(p))) \\ &= (1-\varepsilon)\cdot (w(p) * f(p)) + \varepsilon \cdot (w(p) * (f(p) \odot m(p))) \\ &= (1-\varepsilon)\underbrace{\sum_{p_n\in\mathcal{R}} w(p_n)\cdot f(p+p_n)}_{\text{vanilla convolution}} + \varepsilon\underbrace{\sum_{p_n\in\mathcal{R}} w(p_n)\cdot (f(p+p_n)\cdot m(p_n))}_{\text{learnable descriptive convolution}}. \end{aligned} \tag{5}$$

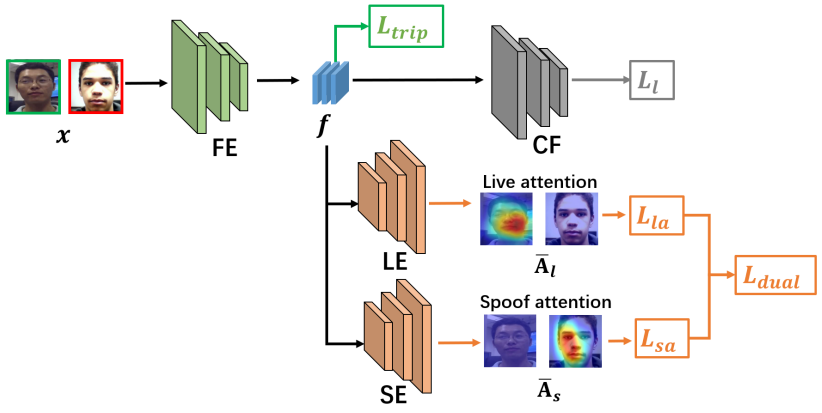


Figure 3: The proposed LDCNet, which consists of a feature extractor **FE**, a classifier **CF**, a live attention estimator **LE** and a spoof attention estimator **SE**.

Equation (5) shows that the proposed LDC is a weighted combination of vanilla convolution and the learnable descriptive convolution with m, and that the two convolutions share the same kernel w. Note that, the proposed LDC exhibits a good generalization of other convolutions [15, 37, 42]. When $\varepsilon = 0$, LDC apparently becomes vanilla convolution. In addition, by comparing Equation (5) with Equation (2), we show that Central Difference Convolution [42] is a special case of LDC when the matrix m in Equation (5) is

$$m = \varepsilon \cdot \mathbf{1}_{3\times 3} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\frac{1}{w(p)} \sum_{p_n \in \mathcal{R}} w(p_n) & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{6}$$

Figure 2 gives several examples of live and spoof faces and their feature maps obtained by LDC. These features maps show that LDC not only well preserves the textural details of facial images but also highlights the distinctive characteristics of spoof faces. As shown in Figures 2 (b) and (c), the grid artifacts and moiré patterns of print attacks and replay attacks are clearly visible in the corresponding feature maps.

2.1.3 LDCNet

With the proposed LDC, we develop a Learnable Descriptive Convolutional Network (LDCNet) to address the face anti-spoofing problem. As shown in Figure 3, LDCNet consists of a feature extractor **FE**, a live/spoof classifier **CF**, and two attention estimators, including a live attention estimator **LE** and a spoof attention estimator **SE**. Note that, instead of vanilla convolution, we adopt the proposed LDC in all the convolutional layers of **FE**, **LE** and **SE**.

We define the liveness loss $\mathcal{L}_l$ for classifying live and spoof faces by,

$$\mathcal{L}_l = -\sum_{\forall x} y log(\mathbf{CF}(\mathbf{FE}(x))) + (1-y) log(1-\mathbf{CF}(\mathbf{FE}(x))), \tag{7}$$

where x is an input image, and y is its binary liveness label, i.e., $y = 1$ for live images and $y = 0$ for spoof images.

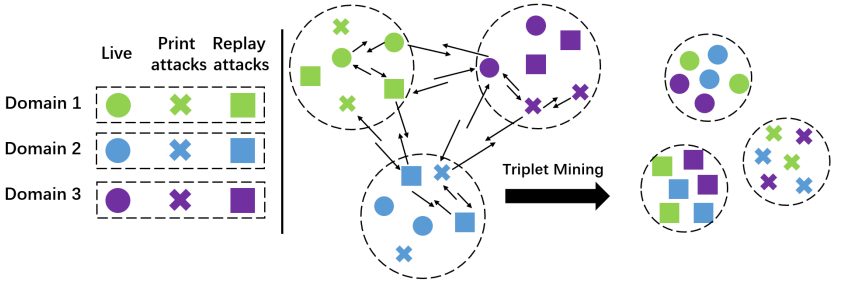


Figure 4: Triplet mining for learning domain-invariant features in LDCNet. Note that, although this illustration was inspired from [13], which focuses on learning robust live/spoof decision boundary, our goal is different from [13] and aims to learn domain-invariant features. Hence, we devise the triplet loss to separate data of the three classes (i.e., live, print attack, and replay attack) while aggregating the data from different domains but having the same class label together.

2.2 Triplet Mining

Next, we adopt triplet mining in LDCNet to constrain the feature extractor **FE** to learn domain-invariant features. We assume each benchmark dataset indicates one domain. Note that, because images under different presentation attacks have different characteristics, here we do not assume all the spoof images belong to one class. Instead, we assign spoof images of different attacks into different classes. For example, in Figure 4, the training data are labeled with 3 classes, i.e., live, print attack, and replay attack.

We define the triplet loss $\mathcal{L}_{trip}$ to enforce the inter-class pairs to be distant from intra-class pairs by least a margin α by,

$$\mathcal{L}_{trip} = \sum_{\forall x_i^a} (\|\mathbf{FE}(x_i^a) - \mathbf{FE}(x_i^p)\|_2^2 - \|\mathbf{FE}(x_i^a) - \mathbf{FE}(x_i^n)\|_2^2 + \alpha), \tag{8}$$

where x_i^a is an anchor input, x_i^p is a positive input of the same class as x_i^a, and x_i^n is a negative input of a different class from x_i^a. By minimizing the triplet loss, we encourage LDCNet to narrow the distance between different domains during the model training so as to learn domain-invariant features.

2.3 Dual Attention Supervision

As mentioned in Sec. 1, the binary ground labels only indicate whether an image is live or spoof but give no regional indication about where the attacked regions locate. To tackle this issue, we propose a dual-attention supervision, including a live attention and a spoof attention, to offer LDCNet additional guidance with fine-grained information.

Therefore, we include two attention estimators **LE** and **SE** in LDCNet to further encourage the feature extractor **FE** to learn from regional live and spoof attentions. The two attention estimators **LE** and **SE** are jointly trained with LDCNet but need further indications to constrain their learning. Instead of including external auxiliary information, we propose using the well-known Class Activation Map [29, 49] to generate the quasi-ground truth for **LE** and **SE**. We pre-train the feature extractor **FE** and the live/spoof classifier **CF** to obtain

Total loss $\mathcal{L}_T$				[O,C,I]→ M		[O,M,I]→ C		[O,C,M]→ I		[I,C,M]→ O	
$\mathcal{L}_l$	$\mathcal{L}_{trip}$	$\mathcal{L}_{dual}$		HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
		$\mathcal{L}_{A_l}$	$\mathcal{L}_{A_s}$								
✓				15.24	90.43	18.33	89.45	17.14	89.84	18.22	88.37
✓	✓			13.10	91.98	16.67	90.98	13.87	92.84	16.19	90.56
✓	✓	✓		11.43	93.68	13.07	93.89	10.64	94.92	14.32	92.54
✓	✓		✓	12.38	93.14	14.72	92.57	12.71	94.13	14.91	91.39
✓	✓	✓	✓	**9.29**	**96.86**	**12.00**	**95.67**	**9.43**	**95.02**	**13.51**	**93.68**

Table 1: Ablation study on all the cross-domain protocols, using different combinations of loss terms. The evaluation metrics are HTER(%) ↓ and AUC(%) ↑.

Method	HTER	AUC
Vanilla Convolution [23]	16.65	84.19
Sobel Convolution [30] (*CVPR 20*)	14.96	90.00
Local Binary Convolution [15] (*CVPR 17*)	15.10	90.50
Central Difference Convolution [30] (*TPAMI 20*)	14.94	91.33
Ours: LDC	**13.51**	**93.68**

Table 2: Ablation study on [I,C,M]→ O using different convolution kernels. The evaluation metrics are HTER(%) ↓ and AUC(%) ↑.

the live activation map $\mathbf{A}_l$ and the spoof activation map $\mathbf{A}_s$ of an input image x by,

$$\begin{aligned} \mathbf{A}_l &= \textbf{Grad-CAM}(\textbf{CF}(\textbf{FE}(x)); y=1), \text{and} \\ \mathbf{A}_s &= \textbf{Grad-CAM}(\textbf{CF}(\textbf{FE}(x)); y=0), \end{aligned} \tag{9}$$

where **Grad-CAM** indicates the class activation operation [29].

Next, by referring to $\mathbf{A}_l$ and $\mathbf{A}_s$ as the live and spoof attentions, we define the dual attention loss $\mathcal{L}_{dual}$ by,

$$\mathcal{L}_{dual} = \mathcal{L}_{A_l} + \mathcal{L}_{A_s}, \tag{10}$$

which incorporates the live attention loss $\mathcal{L}_{A_l}$ and the spoof attention loss $\mathcal{L}_{A_s}$ by,

$$\mathcal{L}_{A_l} = ||\mathbf{A}_l - \bar{\mathbf{A}}_l||_2^2; \mathcal{L}_{A_s} = ||\mathbf{A}_s - \bar{\mathbf{A}}_s||_2^2, \tag{11}$$

where $\bar{\mathbf{A}}_l$ and $\bar{\mathbf{A}}_s$ are the estimated attentions by **LE** and **SE**, respectively. Note that, we only use the original activation maps $\mathbf{A}_l$ of live images and $\mathbf{A}_s$ of spoof images in the model training, but set all the values of $\mathbf{A}_l$ of spoof image and $\mathbf{A}_s$ of live images into zeros.

2.4 Total Loss and Live/Spoof Classification

Finally, we include the liveness loss L_l, the triplet loss L_{trip}, and the dual attention loss $\mathcal{L}_{dual}$ to define the total loss by:

$$\mathcal{L}_T = \mathcal{L}_l + \beta \mathcal{L}_{trip} + \gamma \mathcal{L}_{dual}, \tag{12}$$

where β and γ are the weight factors. In all our experiments, we empirically set $\beta = 0.1$ and $\gamma = 0.004$.

In the inference stage, we follow previous face anti-spoofing methods [12, 21, 22, 31, 37, 42] to measure the detection score in terms of the classification output. For each test image x, we define the detection score s_{ls} by,

$$s_{ls} = \textbf{CF}(\textbf{FE}(x)). \tag{13}$$

3 Experiments

3.1 Experimental Setting

3.1.1 Datasets and Evaluation Metrics

We evaluate our method on the face anti-spoofing databases: **OULU-NPU** [4] (denoted as O), **MSU-MFSD** [39] (denoted as M), **CASIA-MFSD** [48] (denoted as C), and **Idiap Replay-Attack** [6] (denoted as I). To have a fair comparison with previous methods [5, 12, 13, 19, 30, 31, 34, 35, 38], we follow the same settings in [30] to conduct cross-testing on these datasets and report the results using the same evaluation metrics, including Attack Presentation Classification Error Rate (APCER) [32], Bona Fide Presentation Classification Error Rate (BPCER) [32], Average Classification Error Rate (ACER) [32], Half Total Error Rate (HTER) [7], and Area Under Curve (AUC).

3.1.2 Network Architecture and Implementation Details

We develop LDCNet by using Res-18 [10] as the network backbone for the feature extractor **FE**, and using three convolutional blocks to build the two estimators **LE** and **SE**. Each convolutional block consists of convolutional layers, batch normalization layers and ReLU activation functions. The feature maps of the last three layers of **FE** are resized to 32×32 and are then concatenated together for **LE** and **SE**. Note that, we replace all the vanilla convolutions in **FE**, **LE**, and **SE** with the proposed LDC.

We implement the proposed method by Pytorch. To pre-train **FE** and **CF**, we set a constant learning rate of $5\mathbf{e}{-4}$ with Adam optimizer up to 100 epochs. As to LDCNet, we set a constant learning rate of $1\mathbf{e}{-4}$ with Adam optimizer to train **FE**, **LE**, **SE** and **CF** up to 200 epochs.

3.2 Ablation Study

3.2.1 Comparison between Different Losses

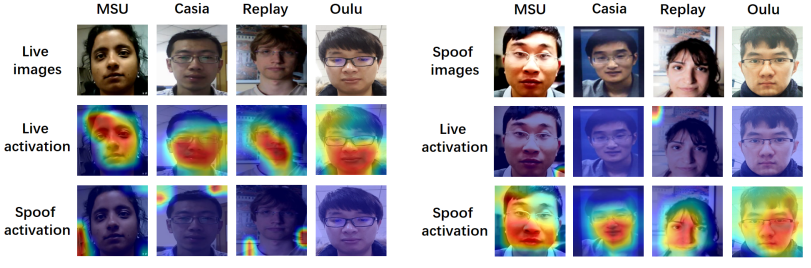
In Table 1, we compare using different combinations of loss terms to train LDCNet and test on all the cross-domain protocols. When including $\mathcal{L}_{trip}$ with $\mathcal{L}_{l}$, we show that the triplet

Method	[O,C,I]→ M		[O,M,I]→ C		[O,C,M]→ I		[I,C,M]→ O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
MADDG [30] (*CVPR 19*)	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
DR-MD-Net [29] (*CVPR 20*)	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
SSDG-M [13] (*CVPR 20*)	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83
RFM [31] (*AAAI 20*)	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
RAEDFL [14] (*ACPR 21*)	16.67	87.93	17.78	86.11	14.64	85.64	18.06	90.04
ANRL [19] (*ACM MM 21*)	10.83	96.75	17.83	89.26	16.03	91.04	15.67	91.90
D^2AM [8] (*AAAI 21*)	15.43	91.22	12.70	95.66	20.98	85.58	15.27	90.87
SDA [35] (*AAAI 21*)	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30
CDCN-PS [43] (*TBBIS 21*)	20.42	87.43	18.25	86.76	19.55	86.38	15.76	92.43
FAS-DR-BC(MT) [24] (*TPAMI 22*)	11.67	93.09	18.44	89.67	11.93	94.95	16.23	91.18
LMFD-PAD [9] (*WACV 22*)	10.48	94.55	12.50	94.17	18.49	84.72	**12.41**	**94.95**
SSAN-M [38] (*CVPR 22*)	10.42	94.76	16.47	90.81	14.00	94.58	19.51	88.17
SSAN-R [38] (*CVPR 22*)	**6.67**	**98.75**	**10.00**	**96.67**	**8.88**	**96.79**	13.72	93.63
Ours	**9.29**	**96.86**	**12.00**	**95.67**	**9.43**	**95.02**	**13.51**	**93.68**

Table 3: Comparison of cross-domain face presentation attack detection. The evaluation metrics are HTER(%) ↓ and AUC(%) ↑.

Method	[M,I]→ C		[M,I]→ O	
	HTER	AUC	HTER	AUC
MADDG [40] (*CVPR 19*)	41.02	64.33	39.35	65.10
DR-MD-Net [34] (*CVPR 20*)	31.67	75.23	34.02	72.65
SSDG-M [12] (*CVPR 20*)	31.89	71.29	36.01	66.88
RFM [31] (*AAAI 20*)	36.34	67.52	29.12	72.61
RAEDFL [12] (*ACPR 21*)	31.11	72.63	29.23	74.62
ANRL [15] (*ACM MM 21*)	31.06	72.12	30.73	74.10
D^2AM [8] (*AAAI 21*)	32.65	72.04	**27.70**	75.36
SDA [33] (*AAAI 21*)	32.17	72.79	28.90	73.33
SSAN-M [38] (*CVPR 22*)	**30.00**	**76.20**	29.44	**76.62**
Ours	**22.22**	**82.87**	**21.54**	**86.06**

Table 4: Comparison of limited cross-domain testing on **[M, I]→ C** and **[M, I]→ O**. The evaluation metrics are HTER(%) ↓ and AUC(%) ↑.



(a) (b)

Figure 5: Activation maps of (a) live and (b) spoof images from **high-level** features on different datasets.

mining effectively encourages LDCNet to learn domain-invariant features and improve the performance over the case of $\mathcal{L}_l$. When including $\mathcal{L}_l + \mathcal{L}_{trip}$ with either $\mathcal{L}_{A_l}$ or $\mathcal{L}_{A_s}$, we do have improved performance over the case of $\mathcal{L}_l + \mathcal{L}_{trip}$ and verify the effectiveness of each single attention. When further including $\mathcal{L}_{dual}$, we show that the two attentions indeed offer LDCNet a fine-grained supervision to learn discriminative features and achieve the best performance. These results verify that each of the proposed components steadily contributes to the overall performance.

3.2.2 Comparison between Different Convolutions

In Table 2, we compare using Sobel Convolution [37], Local Binary Convolution [15], Central Difference Convolution [42], and the proposed LDC to replace the vanilla convolution in **FE**, **LE** and **SE** and use the same total loss $\mathcal{L}_T$ to train the model and test on the protocol **[I,C,M]→ O**. The results show that LDC outperforms the others and verify that the proposed learnable descriptor better adapts to various textural details than the pre-defined descriptors. We believe this learnable characteristic of LDC indeed facilitates the model to learn intrinsic features for face anti-spoofing.

3.3 Experimental Comparisons on Cross- and Intra-Domain Testing

First, we follow the setting of [30] to conduct cross-domain testing, i.e., using the model trained on training domains to detect face presentation attacks on unseen domain. Table 3 shows the detection performance of four cross-domain testing protocols on the datasets

Method	P.	ACPER	BPCER	ACER	P.	ACPER	BPCER	ACER
SGTD [30] (*CVPR 20*)		2.0	0.0	1.0		3.2±2.0	2.2±1.4	2.7±0.6
BCN [41] (*ECCV 20*)		0.0	1.6	0.8		2.8±2.4	2.3±2.8	2.5±1.1
Disentangle [32] (*CVPR 20*)	1	1.7	0.8	1.3	3	2.8±2.2	1.7±2.6	2.2±2.2
RAEDFL [10] (*ACPR 21*)		1.67	0.00	0.83		1.38±1.78	0.28±0.68	**0.83±0.86**
Structure [15] (*IJCB 21*)		1.3	0.0	**0.6**		2.3±2.7	1.4±2.6	**1.9±1.6**
Ours		0.0	0.0	**0.0**		4.55±4.55	0.58±0.91	2.57±2.67
SGTD [30] (*CVPR 20*)		2.5	1.3	1.9		6.7±7.5	3.3±4.1	5.0±2.2
BCN [41] (*ECCV 20*)		2.6	0.8	1.7		2.9±4.0	7.5±6.9	5.2±3.7
Disentangle [32] (*CVPR 20*)	2	2.7	2.7	2.4	4	5.4±2.9	3.3±6.0	4.4±3.0
RAEDFL [10] (*ACPR 21*)		0.69	1.67	**1.18**		5.41±6.40	2.50±2.74	3.96±3.90
Structure [15] (*IJCB 21*)		2.2	2.2	2.2		6.7±6.8	0.0±0.0	**3.3±3.4**
Ours		0.8	1.0	**0.9**		4.50±1.48	3.17±3.49	**3.83±2.12**

Table 5: Comparison of intra-domain face presentation attack detection on **OULU-NPU**. The evaluation metrics are APCER(%) ↓, BPCER(%) ↓, and ACER(%) ↓.

OULU-NPU, **MSU-MFSD**, **CASIA-MFSD**, and **Idiap Replay-Attack**. The results in Table 3 show that LDCNet achieves promising performances on the four protocols in both metrics HTER and AUC. This performance improvement demonstrates the efficacy of LDC and also shows that collaborative supervision of triplet mining and dual-attention effectively promotes LDCNet to learn domain-invariant and live/spoof discriminative features.

In Figure 5, we further use the high-level features extracted from the last layer of **FE** to generate the live and spoof activation maps. We see that: (1) the live activation maps of live images concentrate mainly on facial regions, (2) the spoof activation maps of live images exhibit nearly no responses, and vice versa for the spoof images. These visualization results demonstrate excellent ability of the proposed LDCNet on learning live/spoof discriminative features.

Next, we conduct the limited cross-domain testing by using the model trained on only two source domains to evaluate the domain generalization ability. Among the four datasets, as mentioned in [30], there exists a significant domain gap between **MSU-MFSD** and **Idiap Replay-Attack**. Therefore, we use these two datasets as training domains and then conduct the cross-domain testing on the two protocols **[M, I]→ C** and **[M, I]→ O**. The results in Table 4 show that the proposed method significantly outperforms all the other methods and achieves 26.38% improvement in HTER and 9.50% improvement in AUC over the state-of-the-art method SSAN-M [38]. These results demonstrate the superior generalization ability of LDCNet even when training on a small number of source domains.

Finally, we further conduct the intra-domain testing on **OULU-NPU**. As shown in Table 5, the results again demonstrate the effectiveness of the proposed method.

4 Conclusion

This paper proposes a novel Learnable Descriptive Convolution (LDC) to adaptively learn the intrinsic textural features for face anti-spoofing and to extend the representation capacity of vanilla convolution. In terms of LDC, we develop the Learnable Descriptive Convolutional Network (LDCNet) and incorporate the idea of triplet mining and dual-attention supervision to collaboratively guide LDCNet on learning domain-invariant as well as discriminative textural features. Extensive experiments are conducted to verify the effectiveness of the proposed method and also show significant improvement over previous methods on the limited cross-domain testing.

References

[1] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. Boosting face presentation attack detection in multi-spectral videos through score fusion of wavelet partition images. *Frontiers in big Data*, page 53, 2022.

[2] André Anjos and Sébastien Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[3] 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.

[4] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017.

[5] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1132–1139, 2021.

[6] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[7] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3722–3731, 2022.

[8] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing. *arXiv preprint arXiv:2005.03922*, 2020.

[9] 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. Springer, 2012.

[10] 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.

[11] Pei-Kai Huang, Chu-Ling Chang, Hui-Yu Ni, and Chiou-Ting Hsu. Learning to augment face presentation attack dataset via disentangled feature learning from limited spoof data. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[12] Pei-Kai Huang, Ming-Chieh Chin, and Chiou-Ting Hsu. Face anti-spoofing via robust auxiliary estimation and discriminative feature learning. In *Asian Conference on Pattern Recognition*, pages 443–458. Springer, 2022.

[13] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020.

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

[15] 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.

[16] Taewook Kim, YongHyun Kim, Inhan Kim, and Daijin Kim. Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

[18] 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.

[19] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021.

[20] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 558–573, 2018.

[21] 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.

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

[23] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *European Conference on Computer Vision*, pages 406–422. Springer, 2020.

[24] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[25] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[26] 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. doi: 10.1109/TIFS.2016.2578288.

[27] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[28] Nilay Sanghvi, Sushant Kumar Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Mixnet for generalized face presentation attack detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5511–5518. IEEE, 2021.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[30] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019.

[31] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11974–11981, 2020.

[32] I Standard. Information technology—biometric presentation attack detection—part 1: Framework. *ISO: Geneva, Switzerland*, 2016.

[33] Xiaoguang Tu, Zheng Ma, Jian Zhao, Guodong Du, Mei Xie, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–19, 2020.

[34] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020.

[35] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2746–2754, 2021.

[36] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1955–1964, 2022.

[37] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5042–5051, 2020.

[38] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022.

[39] 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.

[40] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013.

[41] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *European Conference on Computer Vision*, pages 557–575. Springer, 2020.

[42] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020.

[43] Zitong Yu, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):285–295, 2021.

[44] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *European Conference on Computer Vision*, pages 641–657. Springer, 2020.

[45] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Shice Liu, Bangjie Yin, Shouhong Ding, and Jilin Li. Structure destruction and content combination for face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021.

[46] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020.

[47] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European Conference on Computer Vision*, pages 70–85. Springer, 2020.

[48] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face anti-spoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012.

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.