

LEARNING MULTIPLE EXPLAINABLE AND GENERALIZABLE CUES FOR FACE ANTI-SPOOFING

Ying Bian^{*}, Peng Zhang^{*}, Jingjing Wang, Chunmao Wang, Shiliang Pu[†]

Hikvision Research Institute, China

ABSTRACT

Although previous CNN based face anti-spoofing methods have achieved promising performance under intra-dataset testing, they suffer from poor generalization under cross-dataset testing. The main reason is that they learn the network with only binary supervision, which may learn arbitrary cues overfitting on the training dataset. To make the learned feature explainable and more generalizable, some researchers introduce facial depth and reflection map as the auxiliary supervision. However, many other generalizable cues are unexplored for face anti-spoofing, which limits their performance under cross-dataset testing. To this end, we propose a novel framework to learn multiple explainable and generalizable cues (MEGC) for face anti-spoofing. Specifically, inspired by the process of human decision, four mainly used cues by humans are introduced as auxiliary supervision including the boundary of spoof medium, moiré pattern, reflection artifacts and facial depth in addition to the binary supervision. To avoid extra labelling cost, corresponding synthetic methods are proposed to generate these auxiliary supervision maps. Extensive experiments on public datasets validate the effectiveness of these cues, and state-of-the-art performances are achieved by our proposed method.

Index Terms— Face Anti-spoofing, Explainable Cue Learning, Generalizable Cue Learning

1. INTRODUCTION

Currently, face anti-spoofing has become a crucial part to guarantee the security of face recognition systems and drawn increasing attention in the face recognition community. Previous methods mainly extract handcrafted features such as color [1], texture and distortion cues [2] for face anti-spoofing. However these methods are vulnerable to illumination variations and scene changes.

As deep learning has proven to be effective in many computer vision problems, many researchers turn to employ CNNs to extract more discriminative features [3–5], and show significant improvement over the conventional ones. These methods treat face anti-spoofing as a binary classification

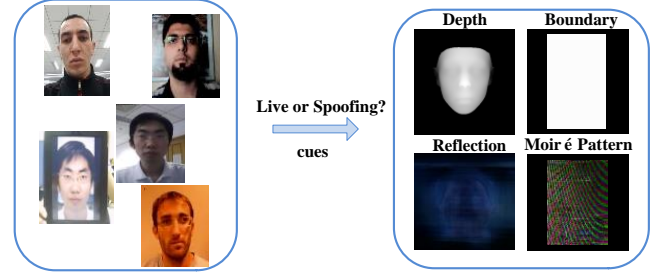


Fig. 1. Face anti-spoofing can be regarded as a binary classification (live or spoofing) problem, which relies on the intrinsic cues such as depth, reflection, boundary and moiré pattern.

problem and train the network with only softmax loss. A CNN with binary supervision might discover arbitrary cues to separate the two classes without explanation, which causes overfitting on the training dataset. When the learned cues change or even disappear during testing, these models would fail to distinguish spoof vs. live faces and achieve poor generalization performance under cross-dataset testing. Therefore, it is desirable to learn explainable and generalizable cues for face anti-spoofing.

To achieve this goal, Liu et al. [6] regard live faces have face-like depth, while faces in print or replay attacks have flat or planar depth. Therefore they utilize depth as auxiliary information to supervise both live and spoof faces. Considering light rays that are reflected from a surface of spoof medium may cause the reflection artifacts in recaptured images, the reflection map [7] is used as additional auxiliary supervision for more robust feature learning. However, only limited cues are leveraged in these approaches, and many other generalizable cues (such as moiré pattern, boundary of spoof medium, etc.) are discarded for face anti-spoofing, which limits their performance under cross-dataset testing. Therefore, more generalizable cues are desirable to be explored to improve the robustness under severe variations.

When a human is distinguishing spoof vs. live faces, the following four main artifacts are usually leveraged. Firstly, the boundary of the spoof medium, such as the screen border of the phone and computer, or the boundary of the printed photographs is easily spotted. Secondly, there exists obvious moiré pattern under replay-attack due to the aliasing caused

^{*}Equal contribution.

[†]Corresponding author: Shiliang Pu (pushiliang.hri@hikvision.com).

by different frequencies of capture devices. Thirdly, reflection artifacts may be caused by the reflection from a surface of spoof medium. Finally, facial depth difference between live and spoofing faces is also a cue as most spoofing faces are broadcasted in plane presentation attack instruments.

Inspired by the way humans distinguish spoof vs. live faces, we propose a novel framework to learn multiple explainable and generalizable cues for face anti-spoofing. Specifically, the network is trained end-to-end with boundary of spoof medium, moiré pattern, reflection artifacts and facial depth as auxiliary supervision in addition to the binary supervision. These extracted cues are visualized in Fig.1. Due to the expensive cost for labelling these cues, we propose synthetic methods to generate the corresponding maps.

2. PROPOSED METHOD

In order to learn the proposed explainable and generalizable cues for face anti-spoofing, we need to get the auxiliary supervision maps including the ones of the boundary of spoof medium, moiré pattern, reflection artifacts and facial depth. The reflection and depth maps are extracted as [7], while the boundary and moiré maps are generated using our proposed synthetic methods. In the section, we first introduce the methods to generate the boundary and moiré maps, and then elaborate the proposed MEGC framework as illustrated in Fig.2.

2.1. Extracting Moiré Map

When a fine pattern on the subject meshes with the pattern on the imaging chip of the shooting camera, the moiré pattern occurs. It is inevitable to get moiré pattern in the relay-attack, since a screen is the subject photographed. Therefore, moiré pattern is a strong generalizable cue under relay-attack.

Directly labelling the moiré pattern is intractable, we need algorithms to extract it automatically. At present, there is no model to directly estimate the moiré map of an input image. As the process of generating moiré pattern is known, we can use different interference fringes with similar frequency to generate moiré pattern physically, and add it into an image without moiré pattern to get a corresponding pair of input image and its moiré map. Another way is to leverage the existing mature demoiréing methods to get an output image without moiré pattern given the input image with moiré pattern. Then subtracting the output image from the input image, we can obtain the corresponding moiré map. However the first method only uses the live images without moiré pattern and discards the various spoofing images with moiré pattern, which limits its performance. While the second method focuses on removing the moiré pattern to make the output image have no moiré pattern visually, which often leads to the residual image contains some image content due to over-removing as shown in Fig.3 (c). This noising moiré map hinders the feature learning and leads to performance degradation.

In order to solve this problem, we propose a network to estimate the moiré map of an input image as shown in Fig.4. In the training phase, we use the above mentioned first method to get corresponding pairs of input images and their moiré maps. The images are as input for the network, and the corresponding moiré maps are as supervision. To reduce the difficulty of learning, the final moiré map learning is based on the above mentioned second method. That is, we first use the image demoiréing method to get the residual image and then refine it to get the final moiré map. The image demoiréing part is based on the SOTA demoiréing method MRGAN [8]. We use a trained MRGAN model [8] to initialize the parameters of the upper half branch of our network and its parameters are fixed during training, while it is followed by two learnable 3×3 convolution layers for adaptation. The image demoiréing part is also followed by two 3×3 convolution layers to refine the moiré map. During testing, we use the trained network to extract the moiré maps of replay-attack images which are used as moiré cues to train the network in our MEGC framework. Moiré maps are set to all zeros for live faces, and we don't perform gradient back propagation on spoofing samples which don't belong to replay spoofing types when training the auxiliary moiré part in our MEGC framework.

2.2. Generating Boundary Map

During recapturing, the boundary of the spoof medium, such as the screen border of the phone and computer, or the boundary of the printed photograph is often captured by the camera due to the larger field of view. Therefore, the boundary of the spoof medium is another strong generalizable cue for face anti-spoofing.

To label the boundary is labor-consuming, we propose a synthetic method to generate pairs of spoofing images with boundaries and their corresponding boundary maps. Firstly, we randomly select a live sample and a spoofing sample. Then the face in the spoofing sample is cut out and pasted to the live sample to replace the live face. In this way, we can get spoofing samples with known boundaries as shown in Fig.3 (e). To generate the corresponding boundary map, the values inside the boundary are set to one, the ones outside the boundary are set to zero as shown in Fig.1. As for live faces, boundary maps are set to all zeros. This boundary map will act as an auxiliary supervision to help the network in our MEGC framework to learn the boundary cues. For the original spoofing images, as we don't know their boundary information, these samples are not used to train the boundary part of the network.

2.3. MEGC Framework

The proposed MEGC framework consists of four main modules, i.e., common feature extraction (backbone network), multi-auxiliary feature extraction (MAFE), multi-feature enrichment (MFE), and classifier as shown in Fig.2. To make a

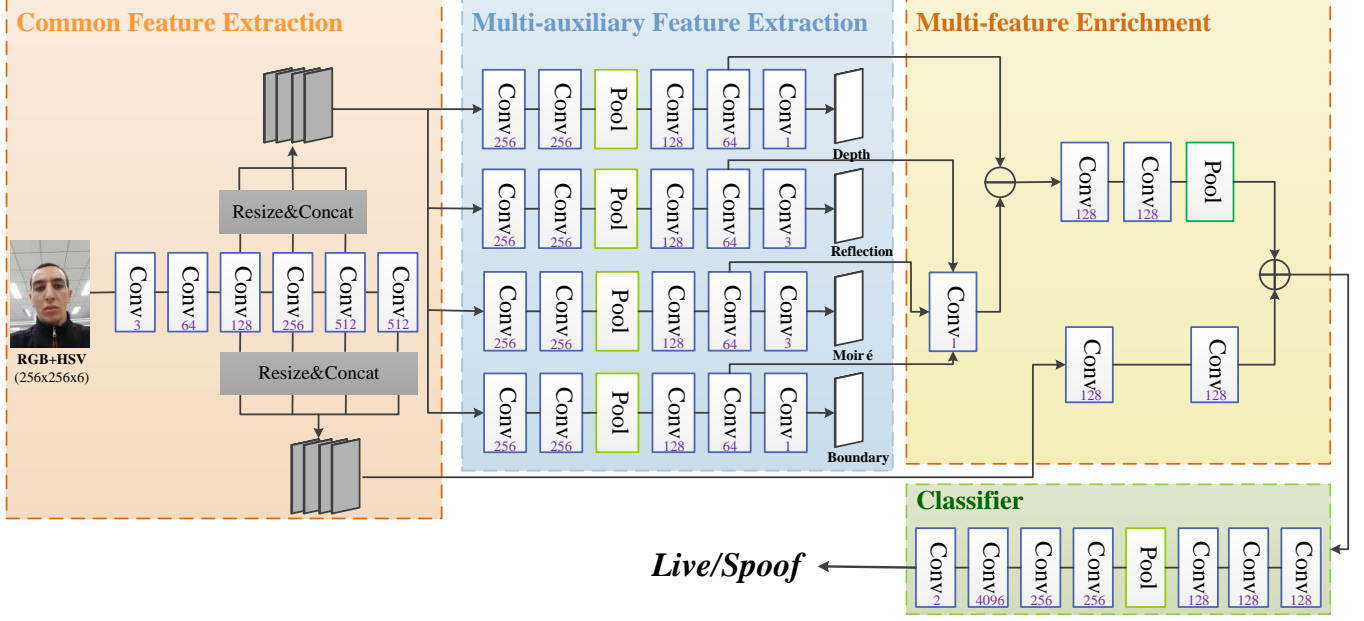


Fig. 2. The architecture of our MEGC framework. Individual numbers indicate the channel numbers of feature maps.



Fig. 3. (a) generated moiré pattern, (b) generated moiré image, (c) the residual moiré map learned by demoiré method MRGAN [8], (d) moiré map estimated by our network, (e) synthetic image with boundary.

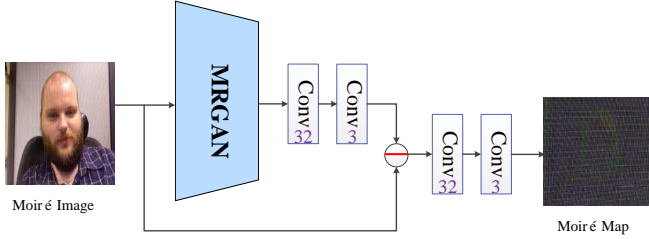


Fig. 4. The network architecture for moiré map extraction.

fair comparison with BASN [7], we use the same backbone network and classifier as BASN. Feature maps of conv3, conv4, conv5 from the backbone network are resized to the fixed size of 64×64 , and are then concatenated to be passed to the MAFE. Feature maps of conv3, conv4, conv5, and conv6 of are resized to the size of 16×16 and are concatenated to be passed to the MFE.

Multi-auxiliary Feature Extractor. MAFE consists of four auxiliary feature extractors, including depth feature extractor, reflection feature extractor, moiré feature extractor

and boundary feature extractor. The depth and reflection feature extractors are the same as the ones in BASN [7]. As for the moiré and boundary feature extractors, we get the ground truth maps using the methods proposed in section 2.1 and 2.2 respectively with size of 32×32 . Given an input face image I , the MAFE predicts the depth map D_{pre} , reflection map R_{pre} , moiré map M_{pre} and boundary map B_{pre} . The loss functions can be formulated as:

$$\mathcal{L}_D = \frac{1}{N} \sum_{i \in N} \|D_{pre}(i) - D_{gt}(i)\|_2^2 \quad (1)$$

$$\mathcal{L}_R = \frac{1}{N} \sum_{i \in N} \|R_{pre}(i) - R_{gt}(i)\|_2^2 \quad (2)$$

$$\mathcal{L}_M = \frac{1}{N} \sum_{i \in N} \|M_{pre}(i) - M_{gt}(i)\|_2^2 \quad (3)$$

$$\mathcal{L}_B = \frac{1}{N} \sum_{i \in N} \|B_{pre}(i) - B_{gt}(i)\|_2^2 \quad (4)$$

where, D_{gt} , R_{gt} , M_{gt} and B_{gt} denote ground truth depth map, reflection map, moiré map and boundary map respectively. N is the batch size. Finally, the overall loss function is $\mathcal{L}_{overall} = \mu * \mathcal{L}_{cls} + \lambda * (\mathcal{L}_D + \mathcal{L}_R + \mathcal{L}_B + \mathcal{L}_M)$, where μ and λ denote the weight of each loss functions.

Multi-feature Enrichment. MFE enriches the feature representations by fusing feature maps from MAFE and the backbone network. Finally, the fused feature map will go through the binary classifier. Different from BASN [7], we first fuse the reflection, moiré and boundary features from MAFE as the spoofing feature map. Then the spoofing feature map is subtracted from the depth feature map.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. Two public face anti-spoofing datasets are utilized to evaluate the effectiveness of our method: Replay-Attack [9] (denoted as R) and CASIA-MFSD [10] (denoted as C). We select one dataset as source domain for training and the remaining one as target domain for cross-testing. Thus, we have two cross-testing tasks in total. Following [11], the Half Total Error Rate (HTER) is used as the evaluation metric.

Implementation Details. The size of face image is $256 \times 256 \times 6$ with both the RGB and HSV channels. The face boxes are detected with open source toolbox Dlib. To expose the boundary, the face boxes are expanded by 1 times in size. Other hyperparameters μ , λ are set to 10, 0.1 respectively. For every training epoch, the ratio of positive and negative images is 1:1.

3.2. Ablation Study

To verify the effectiveness of the learned spoofing cues, we discard one of the spoofing cues in turn to get our methods without reflection, moiré and boundary cues, which are denoted as Ours_wo/R, Ours_wo/M and Ours_wo/B respectively. The results of Tab.1 show that the performances of these methods decrease in different degrees which validates the effectiveness of each one of the proposed spoofing cues. It is also worth noting that different cues have different impact on different datasets. The performance of Ours_wo/M is the worst on the Replay-Attack which shows the moiré cue is the most robust cue for Replay-Attack. Similarly, we can get the boundary cue is the most robust cue for CASIA-MFSD. Therefore, learning multiple generalizable cues can improve the robustness of the model under cross-testing.

To verify the effectiveness of our moiré extracting method, we compare our method with the two mentioned extracting methods in subsection 2.1. The first one adds synthetic moiré (as shown in Fig.3 (a)) into the live images to get the training images with moiré, which is denoted as Ours_w/moiré1. The second one leverages the demoiré method to get the residual moiré map, which is denoted as Ours_w/moiré2. The results of Tab.1 show the effectiveness of our proposed moiré extracting method. Ours_w/moiré1 discards the real spoofing images with moiré pattern, while the residual moiré map learned by Ours_w/moiré2 contains some image content as shown in Fig.3 (c), which hinders their performances. Our method can learn clean moiré map as shown in Fig.3 (d).

3.3. Comparison with State-of-the-Art Methods

In this subsection, we compare the proposed MEGC with previous state-of-the-art methods. The competitive approaches include LBP-TOP [12], Spectral cubes [13], LBP [14], Color Texture [1], CNN [3], STASN [15], FaceDe-S [16], Auxiliary

Table 1. Ablation study on cross-dataset testing.

Methods	Train	Test	Train	Test
	C	R	R	C
Ours	20.2		27.9	
Ours_wo/R	25.7		35.2	
Ours_wo/M	29.0		34.1	
Ours_wo/B	23.7		37.2	
Ours_w/moiré1	27.9		39.6	
Ours_w/moiré2	30.8		39.8	

Table 2. Comparison to SOTA methods.

Methods	Train	Test	Train	Test
	C	R	R	C
LBPTOP [12]		49.7		60.6
Spectral cubes [13]		34.4		50.0
LBP [14]		47.0		39.6
Color Texture [1]		30.3		37.7
CNN [3]		48.5		45.5
STASN [15]		31.5		30.9
FaceDe-S [16]		28.5		41.1
Auxiliary [6]		27.6		28.4
BASN [7]		23.6		29.9
BCN [17]		16.6		36.4
Ours		<u>20.2</u>		27.9

[6], BASN [7], BCN [17]. As shown in Tab.2, the bold type indicates the best performance, and the under-line type indicates the second best performance. Our method outperforms the baseline method BASN [7], which verifies the effectiveness of the two extra introduced cues: moiré and boundary. Our approach achieves the best overall performance, which verifies the effectiveness of proposed multiple generalizable cues learning. It is notable that our method is slightly worse than BCN [17] on the Replay-Attack. However, BCN uses more sophisticated network and introduces other cues, such as surface texture cues. These cues are compatible with our method, and combination of them can further improve the performance. We leave it in the future work.

4. CONCLUSION

In this paper, we propose a novel framework to learn multiple explainable and generalizable cues for face anti-spoofing. Moiré pattern and boundary of spoof medium are introduced to improve the generalization capacity. Two synthetic methods are proposed to generate the corresponding maps to avoid the expensive cost for labelling. Extensive experiments show the effectiveness of these cues, and state-of-the-art performances are achieved.

5. REFERENCES

- [1] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid, "Face spoofing detection using colour texture analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, 2016.
- [2] Wen, D., Han, H., Jain, and A.K., "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.
- [3] Jianwei Yang, Zhen Lei, and Stan Z Li, "Learn convolutional neural network for face anti-spoofing," in *arXiv preprint arXiv:1408.5601*, 2014.
- [4] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu, "Face anti-spoofing using patch and depth-based cnns," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2017.
- [5] Jourabloo, A., Liu, and X., "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [6] 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 (CVPR)*, 2018.
- [7] Kim, T., Kim, Y., Kim, I., Kim, and D., "Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing," *International Conference on Computer Vision Workshops (ICCVW)*, 2019.
- [8] Huanjing Yue, Yijia Cheng, Fanglong Liu, and Jingyu Yang, "Unsupervised moiré pattern removal for recaptured screen images," *Neurocomputing*, vol. 456, pp. 352–363, 2021.
- [9] Ivana Chingovska, André Anjos, and Sebastien Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, 2012.
- [10] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li, "A face antispoofing database with diverse attacks," in *Proceedings of the IEEE International Conference on Biometrics (ICB)*, 2012.
- [11] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [12] Freitas Pereira, T., Komulainen, J., Anjos, A., De Martino, J., Hadid, A., Pietikäinen, M., Marcel, and S., "Face liveness detection using dynamic texture," *Eurasip Journal on Image and Video Processing*, vol. 1, no. 2, 2014.
- [13] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Transactions on Image Processing*, vol. 24, pp. 12, 2015.
- [14] Boulkenafet, Z., Komulainen, J., Hadid, and A., "Face anti-spoofing based on color texture analysis," in *International Conference on Image Processing (ICIP)*, 2015, pp. 2636–2640.
- [15] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu, "Face antispoofing: Model matters, so does data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [16] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu, "Face despoofing: Anti-spoofing via noise modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [17] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao, "Face anti-spoofing with human material perception," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.