Multi-Frames Temporal Abnormal Clues Learning Method for Face Anti-Spoofing

Heng Cong*, Rongyu Zhang, Jiarong He, Jin Gao

Interactive Entertainment Group of Netease Inc, Guangzhou, China

{congheng, zhangrongyu, gzhejiarong, jgao}@corp.netease.com

Abstract—Face anti-spoofing researches are widely used in face recognition and has received more attention from industry and academics. In this paper, we propose the EulerNet, a new temporal feature fusion network in which the differential filter and residual pyramid are used to extract and amplify abnormal clues from continuous frames, respectively. A lightweight sample labeling method based on face landmarks is designed to label large-scale samples at a lower cost and has better results than other methods such as 3D camera. Finally, we collect 30,000 live and spoofing samples using various mobile ends to create a dataset that replicates various forms of attacks in a real-world setting. Extensive experiments on public OULU-NPU show that our algorithm is superior to the state of art and our solution has already been deployed in real-world systems servicing millions of users.

I. Introduction

Face anti-spoofing (FAS) plays an important role in interactive AI systems and is also a challenging task without specific hardware equipped in the industry. The existing methods based on multi-modal information (e.g. infrared light, structured light, and light field) are widely used in scenarios, where there are specialized hardware devices. Although these methods [1]–[3] perform well in classification, they cannot be used on mobile devices on a broad scale.

Initially, the traditional manual feature is used to recognize face spoofing [4]–[6]. With the development of convolution neural network (CNN), classification techniques are excelled in the FAS task [7]–[9]. Recently, the application of face depth information has further improved the FAS effectiveness. The previous method performed excellently, but there are many shortcomings. In [10]–[13], single-frame information is used to estimate the face depth for video prediction, but the inter-frame information of the video is discarded. [14]–[17] achieved better effects by considering the inter-frame information, whereas the feature of face abnormal clues is not be utilized. [18] label face depth with the 3D camera equipped to focus on face detail, which is difficult to promote. By contrast, [19]–[21] convert 2D faces to 3D faces using PRNet, which has a lower label cost but a higher error rate. Especially in the prediction of large-angle side faces, [19]–[21] is not effective. The binary mask label is obtained by the middle part of the face, in [10], [22]–[25], where live and attack is marked as 1 and 0, respectively. However, [10], [22]–[25] ignores the face depth and edge texture information, which degrades the effect of the model. Moreover, many datasets such as NUAA [26], CASIA-MFSD [27], Replay-Attack [28], MSU-USSA [29], and OULU-NPU [30] released several FAS data that are collected in the laboratory state. However, because of the uncertainty of lighting, scenarios, device, etc., the samples in the laboratory varies greatly from the samples in the real world.

To address the above problems, we propose a temporal feature fusion network named EulerNet. The eulerian video magnification is introduced to extract temporal clues and fuse the inter-frame feature from the video. Residual Pyramid is designed in EulerNet to fuse features of different resolutions. We propose a feature-compressed attention module (FCAM) to process features in the temporal dimension. To balance the cost and accuracy, we propose a lightweight face labeling method based on face landmarks, which is faster than [18] and more accurate than [19]–[21]. The face position map obtained by the proposed labeling method is easier to model at fine-grain than depth map. Compared with binary mask, more gradient information of the face edges is retained by face position map. Finally, a large and comprehensive FAS dataset is built to simulate the real usage background under the industry environment. According to the real proportion of cell phone models and application scenarios in real-world, we collect live and attack by different types of mobile ends in various scenarios for our dataset.

In summary, the main contributions of this work as follows:

- We propose a novel network architecture to extract abnormal clues and process multi-frame temporal information. The feature-compressed attention module (FCAM) and Residual Pyramid are designed to improve the computational efficiency, focus on prominent face information, and amplify the weak signal.
- A lightweight face labeling method based on face landmarks is presented to balance the labeling cost and accuracy. Compared with binary mask and depth map, face position map is accurate and fast.
- We build a large and comprehensive dataset, in which live and attack samples are collected by different types of mobile ends in various scenarios.
- The effectiveness, performance, and advantages of the proposed method are numerically and experimentally verified.

---

* denotes corresponding author

II. THE PROPOSED METHOD

In this section, we first introduce a temporal feature fusion network based on eulerian magnification (EulerNet), including feature-compressed attention module (FCAM) and Residual Pyramid. Following EulerNet, a lightweight face labeling method based on face landmarks is described and the supervision method based on face location map is presented.

A. EulerNet

We propose a temporal feature fusion network based on eulerian magnification, namely EulerNet, to extract the abnormal clues from live and spoofing. By applying eulerian video magnification [31] to live and spoofing faces, the import clues for face anti-spoofing are discovered. As illustrated in Fig. 1, the result of print attack sample, lack a natural motion transition has obvious distinguishability with other attacks. Although replay attack and live sample have the same motion transition, the pattern characteristics are different, which provides reliable information for the network. Fig. 2 shows EulerNet architecture includes a Residual Pyramid and a series of feature-compressed attention modules (FCAM) and max pooling. Instead of processing the whole video, we extract a sequence (length 4 and frame interval 3) from the video as input to avoid the problems of excessive memory usage and high derivation time overhead. The attention structure was designed in FCAM to process the temporal features. Using differential infinite impulse response filtering, FCAM amplify the subtle changes in faces between different frames. In Fig. 2, the feature level is divided from 256×256 into 128×128, 64×64, and 32×32 in EulerNet, which is fused by Residual Pyramid. Different levels of features (128×128, 64×64, and 32×32) are sampled for residuals to obtain effective frequency for face anti-spoofing.
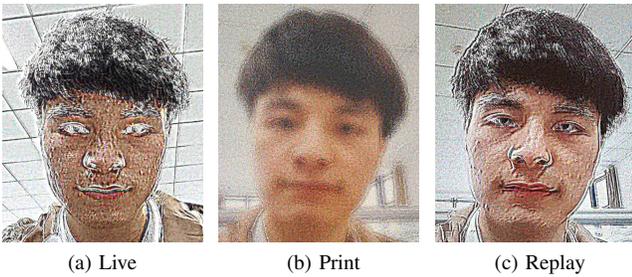
Fig. 1: Results of Applying Eulerian Video Magnification

(a) Live   (b) Print   (c) Replay

Feature-compressed Attention Module. In FCAM, the feature map channels are adjusted to 1 by the 3×3 convolution (Conv3×3), which synthesizes information from each channel. Inspired by the realization of speech signal [32], we transform the implementation of an infinite impulse response (IIR) filter on the image feature map to construct a differential infinite impulse response filter (DIIRF). By initializing a state matrix at the feature map size with 0, the output and the update for state in time sequence can be described as
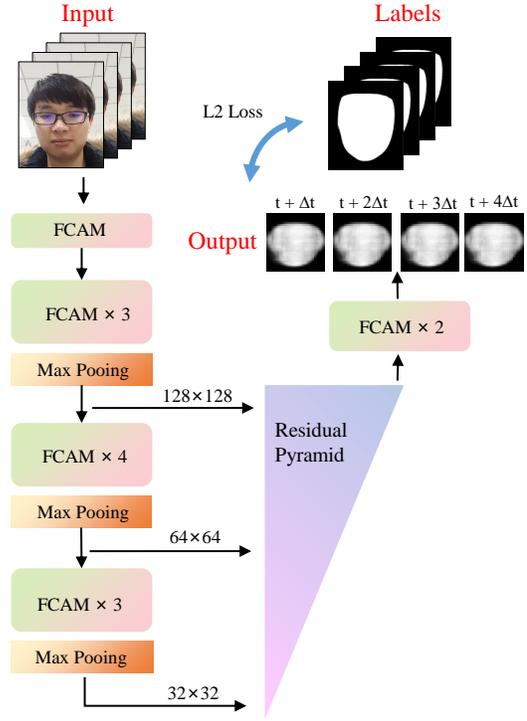
Fig. 2: The proposed EulerNet.

$$y[n] = b_0 x[n] + h_1[n-1] \tag{1}$$

$$h_1[n] = b_1 x[n] + h_2[n-1] - a_1 y[n] \tag{2}$$

$$h_2[n] = b_2 x[n] - a_2 y[n] \tag{3}$$

where $y[n]$ and $x[n]$ present output and input at nth timestamp, respectively. h_i is the state matrix with 0. b_i and a_j is the training parameters of the filter layer, $i \in [0, 2]$, $j \in [1, 2]$. The sequential feature is utilized to filter video signals in DIIRF and the effective frequency for face anti-spoofing is reserved. The sigmoid function as the gate control unit process the feature from DIIRF. Attentional structure is built by multiplying the feature map obtained by sigmoid back to the original input. FCAM architecture is shown in Fig. 3.
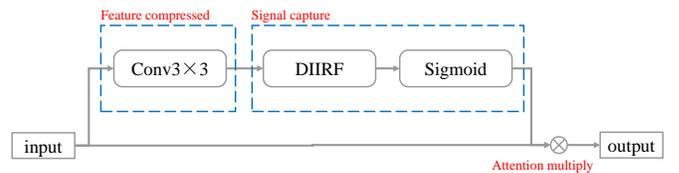
Fig. 3: Feature-compressed attention module.

Residual Pyramid. To amplify the weak signal for face anti-spoofing, we design Residual Pyramid to fuse feature. As shown in Fig. 2, 3-level outputs from backbone network are collected as the input of Residual Pyramid. In Residual Pyramid, the 32×32 ($F_{32\times32}$) and 64×64 ($F_{64\times64}$) features are upsampled to generate 64×64 ($F'_{64\times64}$) and 128×128

$F'_{128\times128}$) feature. Then, the $F'_{64\times64}$ and $F'_{128\times128}$ feature are subtracted by $F_{64\times64}$ and $F_{128\times128}$, respectively. After Conv3×3 extracting, the output $S_{64\times64}$ and $S_{128\times128}$ are obtained.

$$S_{128\times128} = Conv\left(F_{128\times128} - Upsample\left(F_{64\times64}\right)\right) \quad (4)$$

$$S_{64\times64} = Conv\left(F_{64\times64} - Upsample\left(F_{32\times32}\right)\right) \quad (5)$$

$$Ouput_{32\times32} = Concat(Donwsample(S_{128\times128}),$$
$$Donwsample(S_{64\times64}), \quad (6)$$
$$F_{32\times32})$$

The Residual Pyramid can help the network directly learn useful information from features of different depths. In particular, upsampling and downsampling are introduced to calculate the residual between different resolutions.
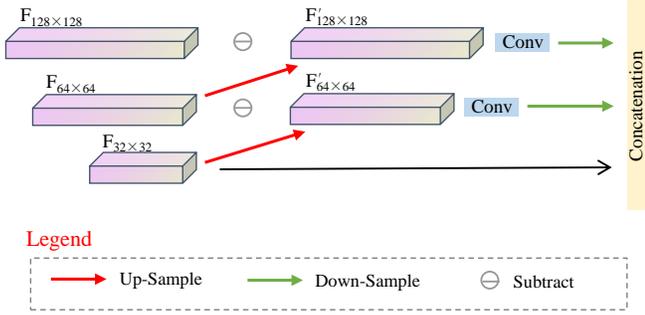
Fig. 4: Residual Pyramid for feature fusing of different depths.

B. A lightweight face labeling method based on face landmarks

Binary mask-based and Depth map-based labeling methods are effective, but both have drawbacks. Binary mask [22] does not distinguish between the face and background, and discards the depth information, which reduces the network efficiency. For depth map-based labeling methods, additional labeling resources are required. Compared with binary mask-based and Depth map-based labeling methods, we connect the key points of the face contour to form a closed face region, namely face position map. In particular, the pixel points within the face region are labeled as 1 and 0 on live face and spoofing, respectively. Face position map is used to reduce the labeling time, preserve the gradient information of the face edges, and improve accuracy. Fig. 5 shows the binary mask, depth map, and face position map.
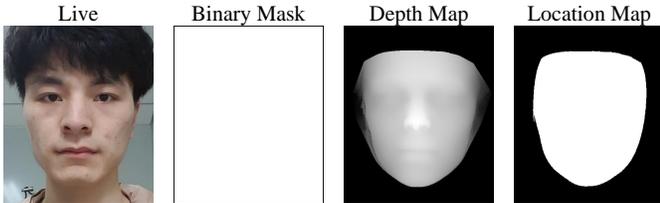
Fig. 5: Comparison of different supervision for FAS.

III. DATASET COLLECTION

The difference of data distribution in different application scenarios determines the model performance in a specific scenario. To address this problem, we simulate the real-world scenario to establish the face anti-spoofing dataset, where each data sample is captured by the mobile end camera. In our dataset, the majority of mobile ends currently on the market are used for data collection. The spoofing type is divided into three parts: complex attack, certificate attack, and image dynamic generation attack. In particular, it is difficult to create and conduct 3D attacks in practical scenarios, so we mainly collect 2D attacks used in [33].

In our dataset, the complex attack includes various paper attacks, projection attacks, high-definition quality shots, and low-quality shots. Certificate attack is common in real-world scenarios and different from other attacks, which consists of ID card attacks, passport document attacks, etc. Image dynamic generation attack is an AI attack method, where dynamic images are generated from single images for face spoofing.

Our dataset outperforms previous public datasets [27], [28], [34] in three points: 1) Our dataset is the largest one that includes 30,000 live and spoofing videos (average duration to be 2s), collected from 10000 subjects, compared to 12,000 videos from 200 subjects in [18]. 2) As shown in Fig. 6, our dataset covers dozens of commonly used mobile ends. 3) The data distribution of our dataset matches the real-world mobile verification scenarios.
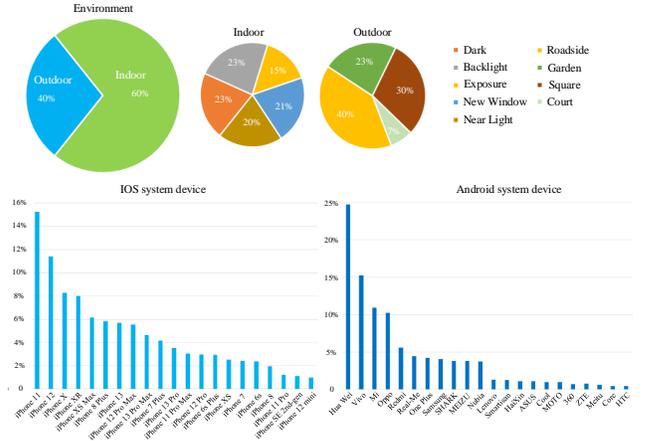
Fig. 6: Statistics of electronic devices and scenes of our dataset.

IV. EXPERIMENTS

A. Implementation Details

To verify the effectiveness of our proposed network architecture and supervision method, we conduct experiments on the collected dataset, dividing the training, development, and test sets according to the ratio of 3:1:1. The learning rate is initialized as 1e-4 and the batchsize is set to 16 for a synergistic optimizer Ranger [35]. The input to the model are

sequences of 4 consecutive frames taken from the same video at intervals of 3. Thus we compute 64 images in one gradient update. The model is trained with Nvidia Tesla V100s. Pytorch is utilized as the backend for the network architecture. L2 loss is utilized to guide the network predict the location map.

Average Classification Error Rate (ACER) is the mean value of Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER), given by

$$APCER = \frac{FP}{TN + FP}, \quad BPCER = \frac{FN}{TP + FN} \quad (7)$$

$$ACER = \frac{APCER + BPCER}{2} \quad (8)$$

In the next subsections, ACER is uniformly utilized for evaluation. For each video, we extracted multiple non-overlapping sequences, then calculated the live score individually and finally average scores for video.

B. Ablation Study

To investigate the behavior of EulerNet, we conducted several ablation studies. Table I shows the ablation study on our dataset. The baseline model is the proposed method using all improvements. Compare 1 (C1) and Compare 5 (C5) discuss the influence of different labels, including binary mask and depth map. In Compare 3 (C3), the Residual Pyramid is replaced by channels concatenation. Compare 4 (C4) removes the FCAM. Compare 2 (C2) discards all structure improvements and uses the common depth map for global comparison.

Effectiveness of FCAM and Residual Pyramid. After adding FCAM and d Residual Pyramid, ACER decreased by 0.34 percent and 0.38 percent, respectively, indicating that they were beneficial in FAS. The results suggest that extracting the effective frequency information of the input and fusing different resolution features together in a residual manner is meaningful for face anti-spoofing. Table I demonstrate that we must guide the network to mine as many anomalous patterns as possible in the attack video. The ACER of the proposed model is reduced by 0.7% concerning Compare 2 in Table I. We optimize the combination of models and labels to maximize the benefits and obtain excellent accuracy on our dataset.

Effectiveness of Face Location Map Supervision. The optimal model on the development set is utilized to check the test set. Our proposed model yields the best ACER, achieving 0.18% lower than the model supervised with depth map and 0.96% lower than the model supervised with binary mask. This indicates that the features with binary mask supervision are not robust and discriminative, which are learned from labeling all pixels of living images as 1. The practice of regression on face location map reduces the difficulty of pixel modeling task and makes the network focus more on the color texture of face region and edge gradient information.

As shown in the Fig. 7, we plot the training performance curves of different models. At the beginning, the curve without FCAM drops the fastest because the model does not extract frequency information. Although the early iteration is fast, the

TABLE I: Ablation study on our dataset.

| Tag | Structure | | | ACER(%)↓ | |
	Label	FCAM	Residual Pyramid	Dev	Test
Compare 1	Binary Mask	✓	✓	3.95	2.84
Compare 2	Depth Map	✗	✗	3.62	2.57
Compare 3	Face Location Map	✓	✗	2.85	2.26
Compare 4	Face Location Map	✗	✓	3.13	2.22
Compare 5	Depth Map	✓	✓	2.74	2.06
Baseline	Face Location Map	✓	✓	**2.48**	**1.88**
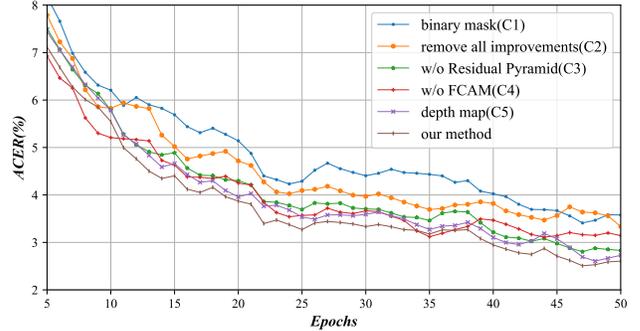
Fig. 7: Performance comparison curve on the development set in training process.

accuracy in the later stage is low, which suggests that it is significant to mine the specific frequency information. In the late training phase (epochs 45-50), the fluctuating curves are sorted by mean value size as C2>C1>C4>C3>C5>our method. Especially, the proposed method curve shows a smoother decreasing trend during training with less fluctuation.

C. Visualization

The Grad-cam [38] is used to visualize the output of the neurons in the model to compare the changes brought by changing the structure or the supervision method. Fig. 8 details
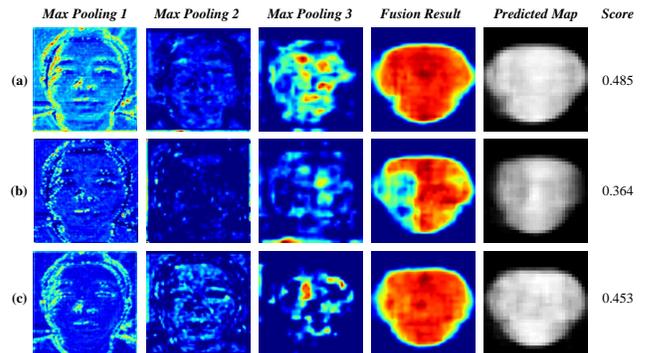
Fig. 8: Visualization study of neuron outputs at different depths: (a) our method, (b) w/o FCAM, (c) w/o Residual Pyramid. From left to right are the results with increasing downsampling multiplier as well as the multi-resolution fusion results and the final prediction maps. Their resolutions are 128×128, 64×64, 32×32, 32×32, 32×32 in order.
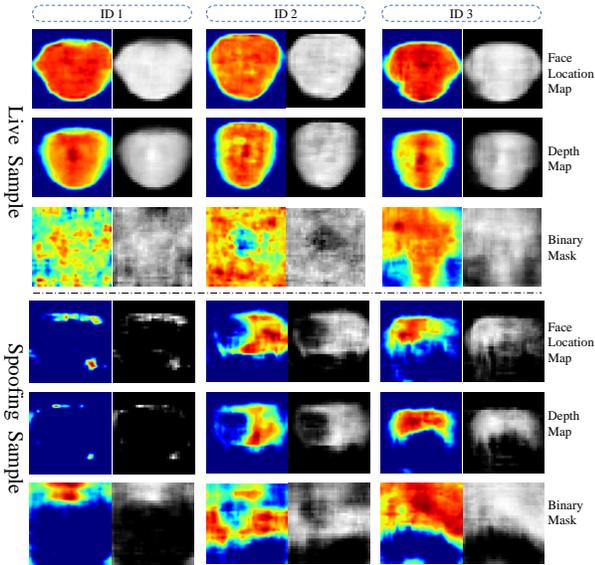
Fig. 9: Model output and heatmap visualization under different supervision methods. The same column indicates the results of the same face, marked with ID 1-3.

TABLE II: The results of intra testing on four protocols of OULU-NPU compared with deep learning methods

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Disentangled [36]	1.7	0.8	1.3
	FAS-SGTD [14]	2.0	**0.0**	1.0
	DeepPixBiS [22]	0.8	**0.0**	0.4
	CDCN [37]	**0.4**	1.7	1.0
	EulerNet(Ours)	**0.4**	3.3	1.9
2	DeepPixBiS [22]	11.4	**0.6**	6.0
	Disentangled [36]	**1.1**	3.6	2.4
	FAS-SGTD [14]	2.5	1.3	1.9
	CDCN [37]	1.5	1.4	**1.5**
	EulerNet(Ours)	2.1	1.4	1.7
3	DeepPixBiS [22]	11.7±19.6	10.6±14.1	11.1±9.4
	FAS-SGTD [14]	3.2±2.0	2.2±1.4	2.7±0.6
	CDCN [37]	**2.4±1.3**	2.2±2.0	2.3±1.4
	Disentangled [36]	2.8±2.2	1.7±2.6	2.2±2.2
	EulerNet(Ours)	2.6±1.3	**1.6±0.8**	**2.1±0.5**
4	DeepPixBiS [22]	36.7±29.7	13.3±14.1	25.0±12.7
	CDCN [37]	4.6±4.6	9.2±8.0	6.9±2.9
	FAS-SGTD [14]	6.7±7.5	**3.3±4.1**	5.0±2.2
	Disentangled [36]	5.4±2.9	3.3±6.0	4.4±3.0
	EulerNet(Ours)	**1.8±1.9**	4.3±2.4	**3.1±0.9**

0.485, which is better than the model removed FCAM and Residual Pyramid. In addition, as the network deepened, the extracted features mined more abnormal clues and produced more accurate face location maps.

Fig. 9 shows the visualization results for the three different labels (face location map, binary map, and depth map) are studied in this paper. For the live face, the heat map supervised by the face localization map is sharper and finer at the edges of the face region. As shown in Fig. 9, the binary mask-based supervised method ignores the gradient information of face edges. Compared with the depth map, the prediction map based on the face location map has higher contrast in distinguishing faces and backgrounds. When the input is spoofing face, the predictions obtained by the face location map-based and depth map-based methods are similar and still bias the face region, which demonstrates that the network does not need to simulate depth information in practice. As shown in Fig. 9, the results of the binary mask have less distinct semantic features. The data distribution on the heat map is more uniform and scattered with the binary mask supervision. Compared with the above three supervision methods, the loss of generalization of the features extracted by the network without the drive of depth and location information.

D. Comparison on Public Dataset

We compare the proposed EulerNet with recent deep learning methods on OULU-NPU [30]. Protocol 1-3 respectively evaluate the generalization of the algorithm under previously unseen environmental conditions, attacks created with different printers or displays, and input camera variation. Protocol 4 considers all the above three factors. Generally, protocols 3 and 4 are more difficult than other protocols.

The Comparison results on OULU-NPU are shown in Table II. The best performance obtained by the proposed method in protocols 3 and 4 demonstrates that our method can maintain accuracy under complex conditions. Table II proves that the structures we have designed are capable of extracting abnormal clues, which has a stronger generalization in facing changes of the shooting device. The complexity of protocols 3 and 4 is similar to the realistic scenario where electronic products are changing rapidly, there are demonstrates the practicality of our proposed method.

V. CONCLUSION

In this paper, we propose a novel face anti-spoofing method, which effectively recognize the subtle differences between real face and spoofing in the video. The novel network architecture, namely EulerNet, is designed to fuse temporal information and extract abnormal clues. We propose a lightweight labeling method based on face landmarks to reduce the labeling cost and improve the labeling speed. Extensive experimental results on our datasets and public OULU-NPU validate the effectiveness of our method.

the effect of FCAM and Residual Pyramid on the features extracted by the network. Max Pooling 1-3 indicates the results using different downsampling multipliers. Fusion result is the multi-resolution fusion results. The final prediction are shown in Predicted Map. Scores are presented in the last column of Fig. 8. The higher the score on the far right, the higher the likelihood of being judged as live faces. The comparison between Max Pooling 1 and Max Pooling 2 shows that the visualization features are clearly distinct under different downsampling scales. The model with FCAM pays more attention to the parts where the action occurs, so there are higher activation values at pixels of the contour, eyes, and nose. Finally, the face score obtained by our method is

REFERENCES

[1] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[2] X. Xie, Y. Gao, W.-S. Zheng, J. Lai, and J. Zhu, "One-snapshot face anti-spoofing using a light field camera," in *Chinese Conference on Biometric Recognition*. Springer, 2017, pp. 108–117.

[3] D. Yi, Z. Lei, Z. Zhang, and S. Z. Li, "Face anti-spoofing: Multi-spectral approach," in *Handbook of Biometric Anti-Spoofing*. Springer, 2014, pp. 83–102.

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 2636–2640.

[5] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha, "Face anti-spoofing with multifeature videolet aggregation," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 1035–1040.

[6] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 4244–4249.

[7] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise modeling, synthesis and classification for generic object anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7294–7303.

[8] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *International conference image analysis and recognition*. Springer, 2017, pp. 27–34.

[9] Y. A. U. Rehman, L. M. Po, and M. Liu, "Deep learning for face anti-spoofing: An end-to-end approach," in *2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2017, pp. 195–200.

[10] Z. Yu, Y. Qin, X. Xu, C. Zhao, Z. Wang, Z. Lei, and G. Zhao, "Auto-fas: Searching lightweight networks for face anti-spoofing," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 996–1000.

[11] 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 transactions on pattern analysis and machine intelligence*, vol. 43, no. 9, pp. 3005–3023, 2020.

[12] 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, 2020.

[13] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, "Meta-teacher for face anti-spoofing," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[14] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. 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*, 2020, pp. 5042–5051.

[15] M. Asim, Z. Ming, and M. Y. Javed, "Cnn based spatio-temporal feature extraction for face anti-spoofing," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017, pp. 234–238.

[16] Z. Yu, X. Li, P. Wang, and G. Zhao, "Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1290–1294, 2021.

[17] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3507–3516.

[18] Y. Liu, Y. Tai, J. Li, S. Ding, C. Wang, F. Huang, D. Li, W. Qi, and R. Ji, "Aurora guard: Real-time face anti-spoofing via light reflection," *arXiv preprint arXiv:1902.10311*, 2019.

[19] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *European Conference on Computer Vision*. Springer, 2020, pp. 557–575.

[20] T. Kim, Y. Kim, I. Kim, and D. 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*, 2019, pp. 0–0.

[21] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, G. Qi, J. Wan, and Z. Lei, "Exploiting temporal and depth information for multi-frame face anti-spoofing," *arXiv preprint arXiv:1811.05118*, 2018.

[22] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[23] M. S. Hossain, L. Rupty, K. Roy, M. Hasan, S. Sengupta, and N. Mohammed, "A-deeppixbis: Attentional angular margin for face anti-spoofing," in *2020 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2020, pp. 1–8.

[24] Y. Ma, L. Wu, Z. Li *et al.*, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," *Pattern Recognition Letters*, vol. 131, pp. 261–267, 2020.

[25] 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. 66 553–66 563, 2020.

[26] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *European Conference on Computer Vision*. Springer, 2010, pp. 504–517.

[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*. IEEE, 2012, pp. 26–31.

[28] I. Chingovska, A. Anjos, and S. 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)*. IEEE, 2012, pp. 1–7.

[29] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE transactions on information forensics and security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[30] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. 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)*. IEEE, 2017, pp. 612–618.

[31] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[32] B. Kuznetsov, J. D. Parker, and F. Esqueda, "Differentiable iir filters for machine learning applications," in *Proc. Int. Conf. Digital Audio Effects (eDAFx-20)*, 2020, pp. 297–303.

[33] L. Li and X. Feng, "Face anti-spoofing via deep local binary pattern," in *Deep Learning in Object Detection and Recognition*. Springer, 2019, pp. 91–111.

[34] P. P. Chan, W. Liu, D. Chen, D. S. Yeung, F. Zhang, X. Wang, and C.-C. Hsu, "Face liveness detection using a flash against 2d spoofing attack," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 521–534, 2017.

[35] L. Wright, "Ranger - a synergistic optimizer." https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019.

[36] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 641–657.

[37] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.