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.

DOI reference number: 10.18293/SEKE2022-076
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 in EulerNet, which is fused by Residual Pyramid) are collected as the input of Residual Pyramid. In Residual Pyramid, 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

\[ y[n] = b_0x[n] + h_1[n - 1] \]  
\[ h_1[n] = b_1x[n] + h_2[n - 1] - a_1y[n] \]  
\[ h_2[n] = b_2x[n] - a_2y[n] \]

where \( y[n] \) and \( x[n] \) present output and input at nth timestamp, respectively. \( h_i \) is the state matrix with 0. \( b_j \) 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.

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

Residual Pyramid. To amplify the weak signal for face anti-spoofing, we design Residual Pyramid to fuse feature. As shown in Fig. 4 3-level outputs from backbone network are collected as the input of Residual Pyramid. In Residual Pyramid, the 32×32 \( F'_{32×32} \) and 64×64 \( F_{64×64} \) features are upsampled to generate 64×64 \( F'_{64×64} \) and 128×128 \( F_{128×128} \) features.
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.

![Residual Pyramid](image)

**Legend**

- **Up-Sample**
- **Down-Sample**
- **Subtract**

**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.

![Comparison of different supervision for FAS](image)

**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.

![Statistics of electronic devices and scenes of our dataset](image)

**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 (BFCER), given by

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

\[
\text{ACER} = \frac{\text{APCER} + \text{BPCER}}{2} \quad \text{(8)}
\]

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 Table I, 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 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 the changes between the model with depth map, model without Residual Pyramid, and model with only FCAM.
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.

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.
