SEKE 2022

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

Heng Cong, Rongyu Zhang, Jiarong He, Jin Gao

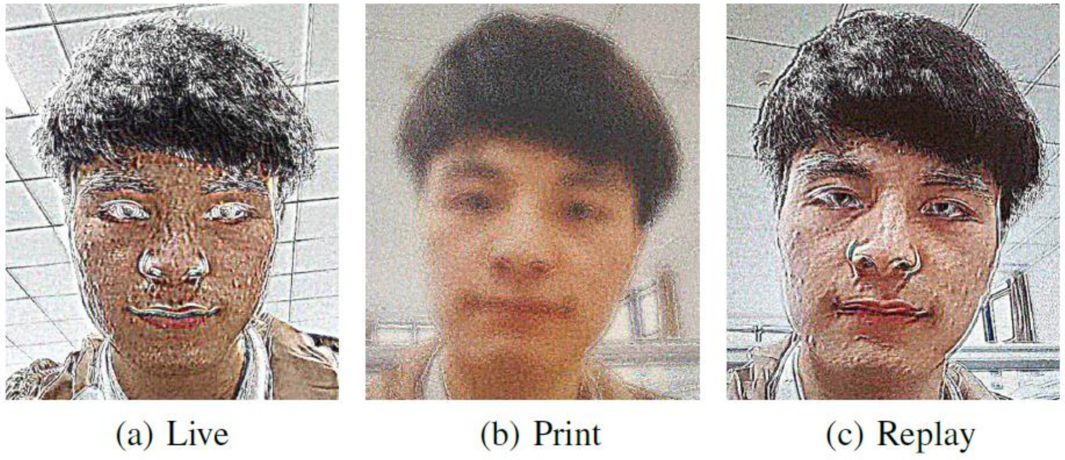
網易 NetEase

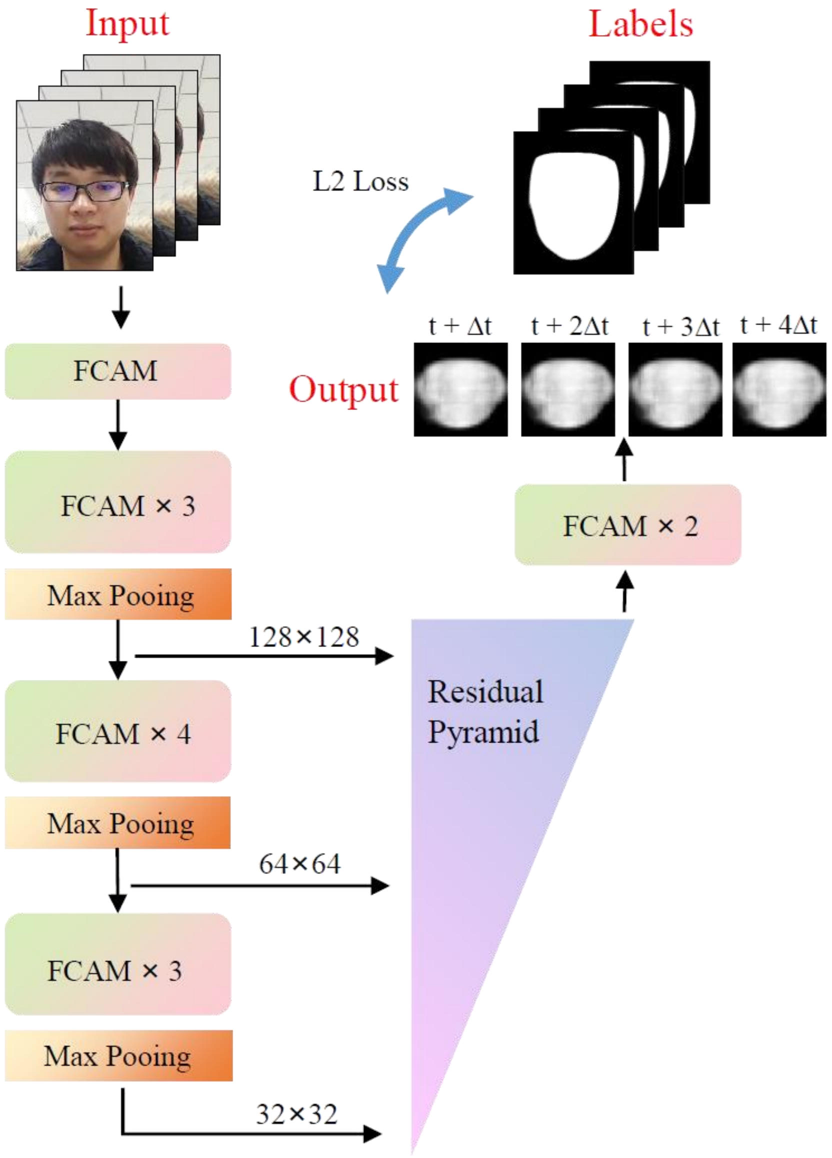
Challenge for Face Anti-Spoofing

- The existing methods are mostly based on multi-modal information (e.g. infrared light, structured light, and light field), which cannot be used on mobile devices on a broad scale.
- The single-frame-based CNN methods discard inter-frame information of the video. The potential of the multi-frame-based methods remains to be explored.
- Face information supervision is an important part of the face anti-spoofing task. Depth camera requires specific hardware equipment and is difficult to promote.
- Datasets collected in the laboratory vary greatly from the samples in the real world.

The Proposed EulerNet

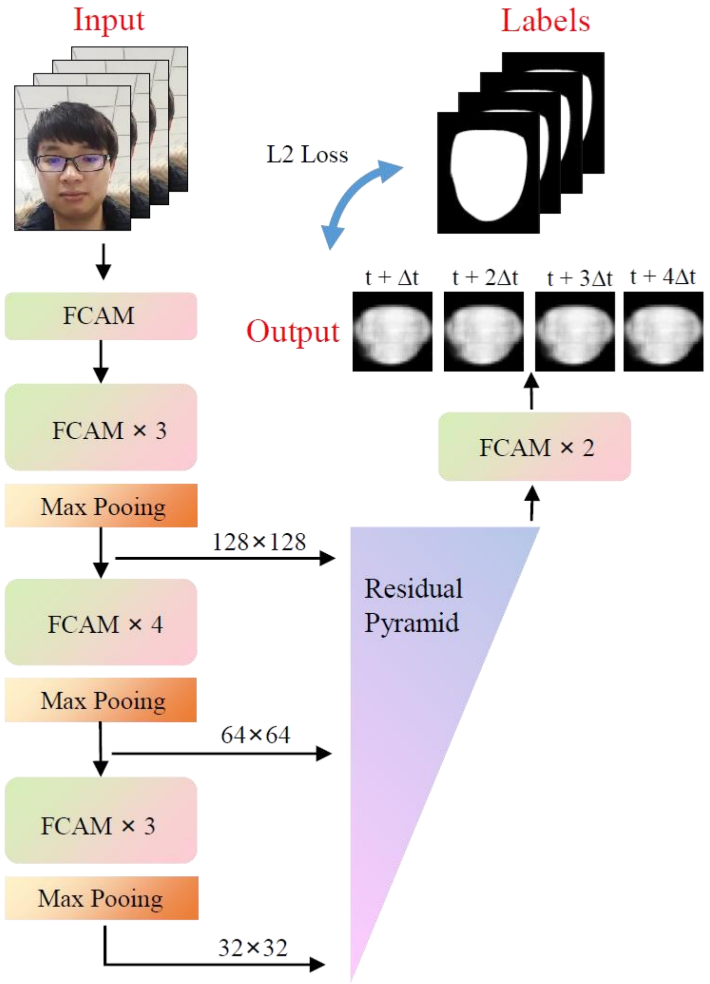
- By applying **eulerian video magnification** to live and spoofing faces, the import clues for face anti-spoofing are discovered.





網易 NetEase

The Proposed EulerNet



- **Input**: a sequence (length 4 and frame interval 3) from the video
- **Feature-compressed attention modules (FCAM)**: Using differential infinite impulse response filtering, FCAM amplify the subtle changes in faces between different frames.
- **Residual Pyramid**: Fusing features from different depths.
- **Face position map**: lightweight labeling, balance the labeling cost and accuracy.

FCAM

Feature-compressed Attention Module

Feature compressed: synthesizes information from each channel.

DIIRF: differential infinite impulse response filter

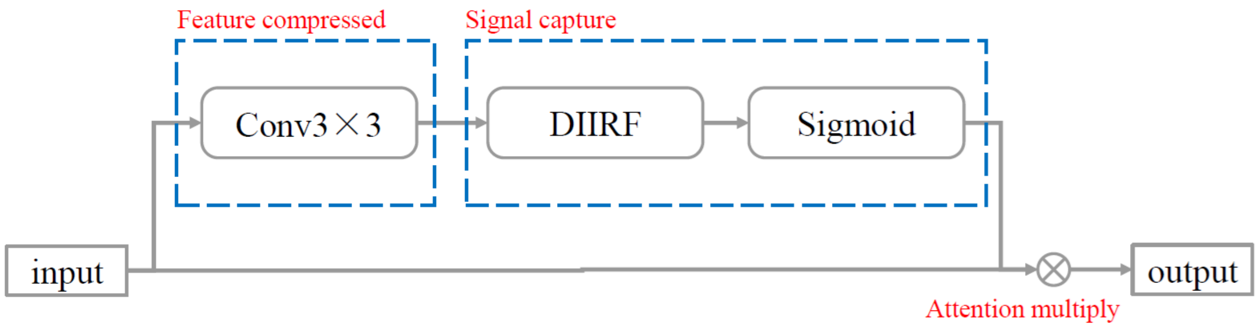
$$y[n] = b_0 x[n] + h_1[n-1] \quad (1)$$

$$h_1[n] = b_1 x[n] + h_2[n-1] - a_1 y[n] \quad (2)$$

$$h_2[n] = b_2 x[n] - a_2 y[n] \quad (3)$$

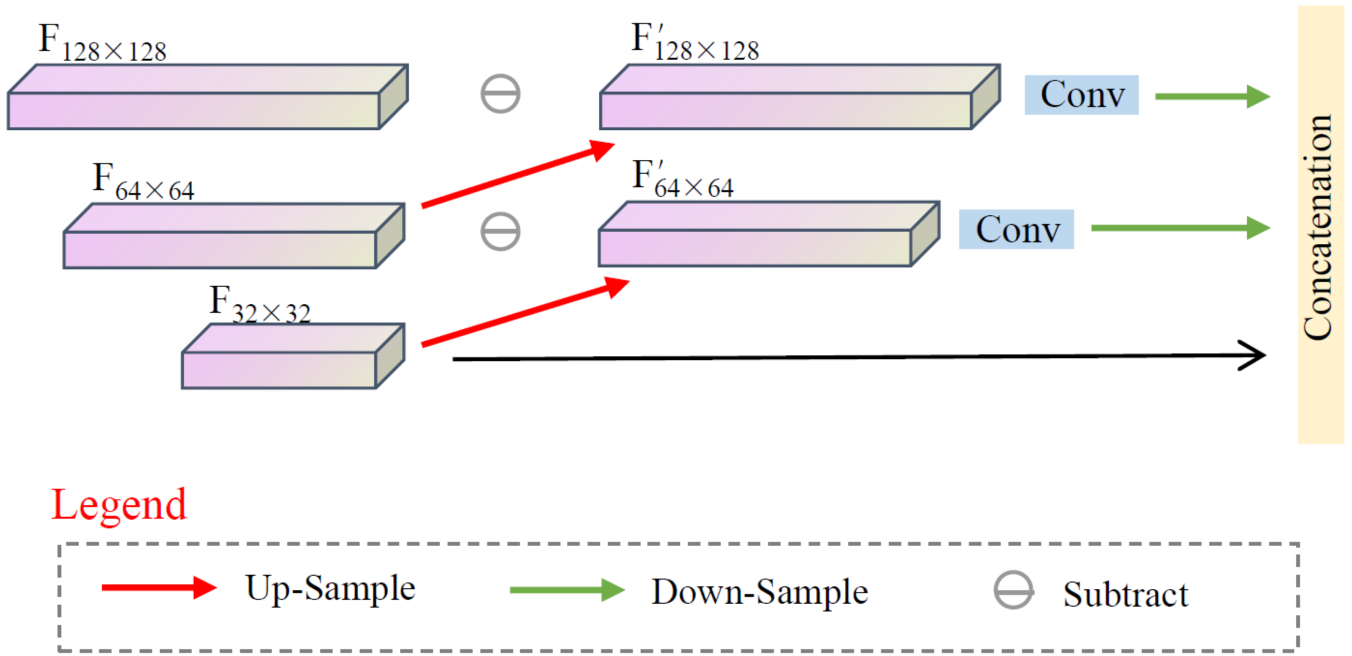
- y[n] is the output at nth timestamp
- x[n] is the input at nth timestamp
- h_i is the parameter of state matrix
- a_i and b_i are the training parameters of the filter layer

Attention: multiplying the feature map obtained by sigmoid back to the original input.



網易 NetEase

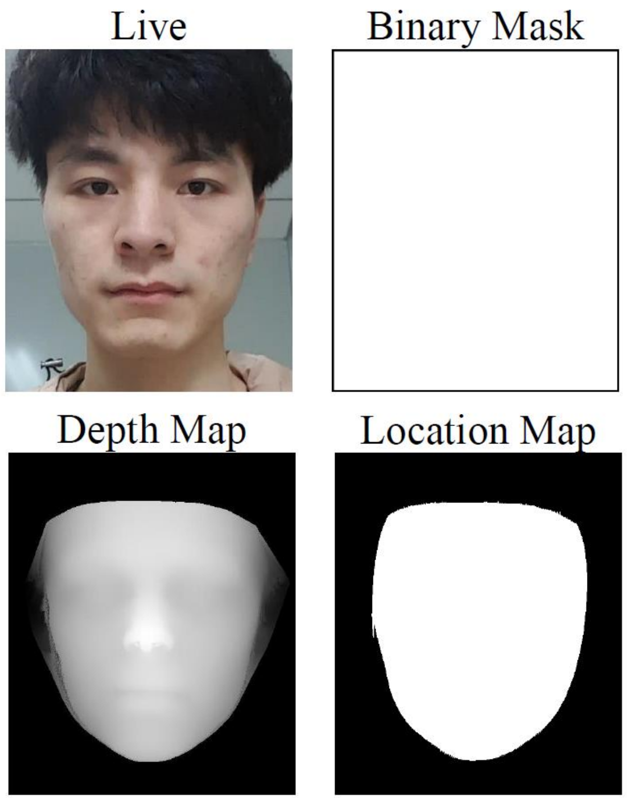
Residual Pyramid



Advantages

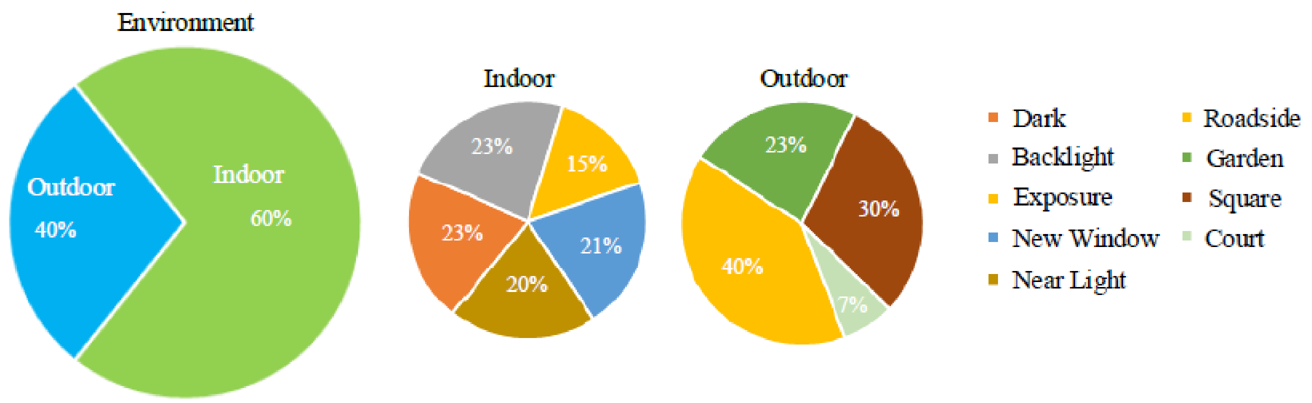
- **Weak signal amplification**
- **Different depths aggregation**
- **Multi-resolution residual utilization**

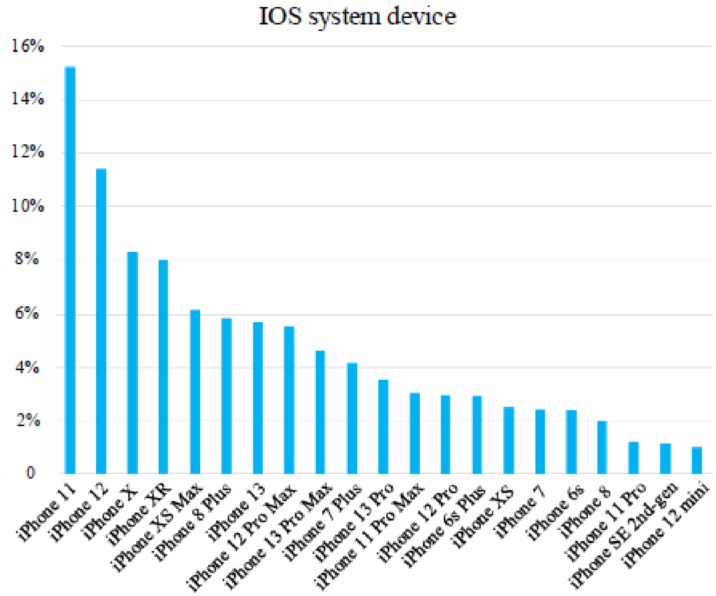
Face Location Map

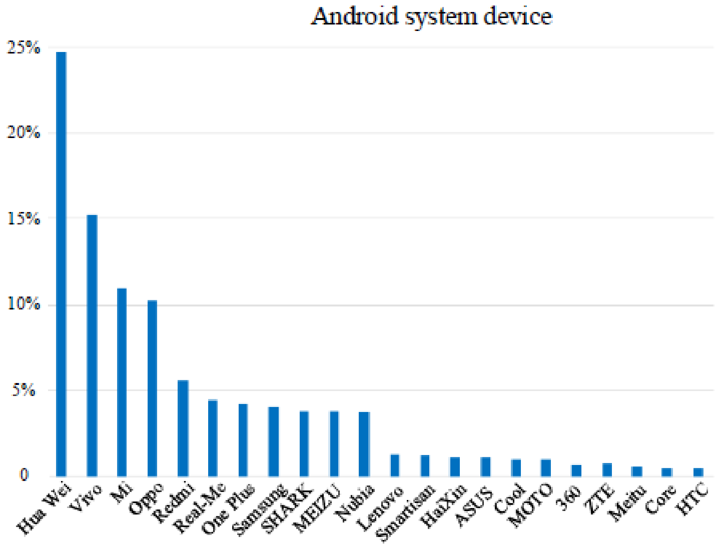


- **binary mask**: fast ✔ lost information ✖
- **depth map**: slow ✖ abundant ✔ difficult to learn ✖
- **location map**: fast ✔ abundant ✔ easy to learn ✔

Dataset Collection



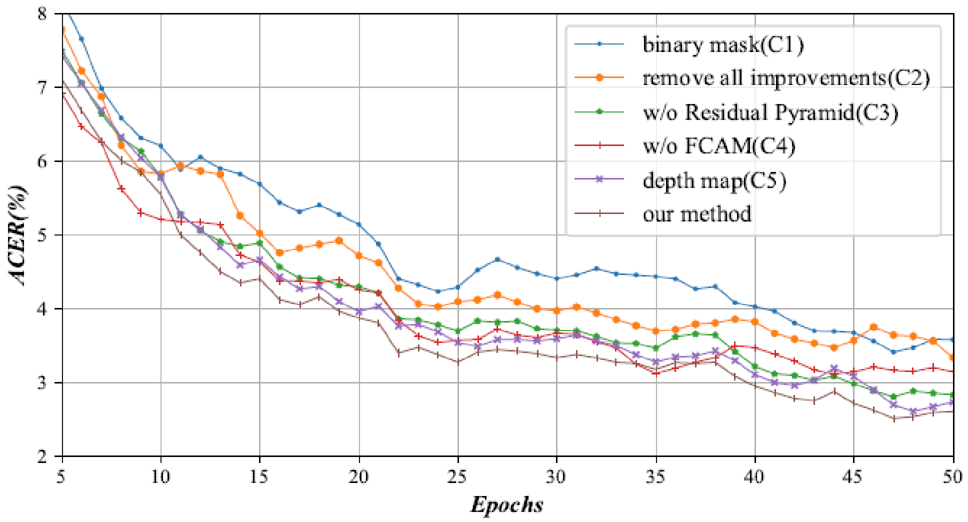




網易 NetEase

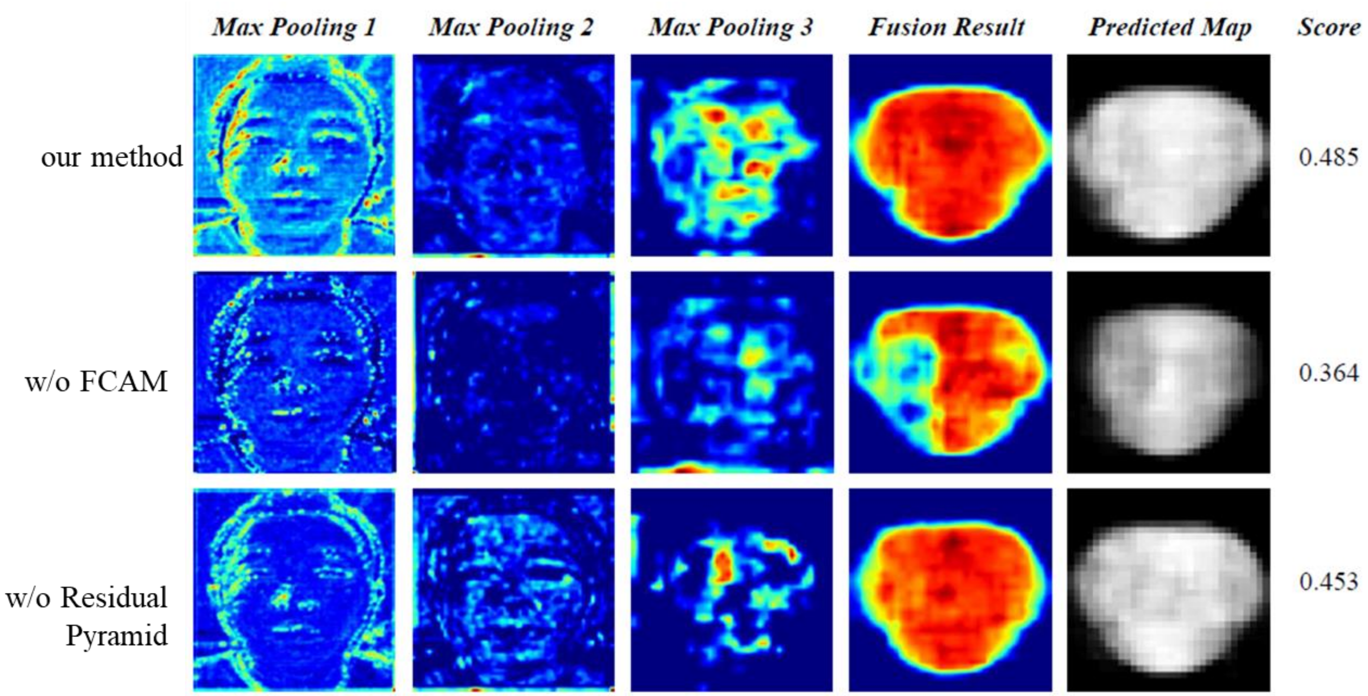
Ablation Study

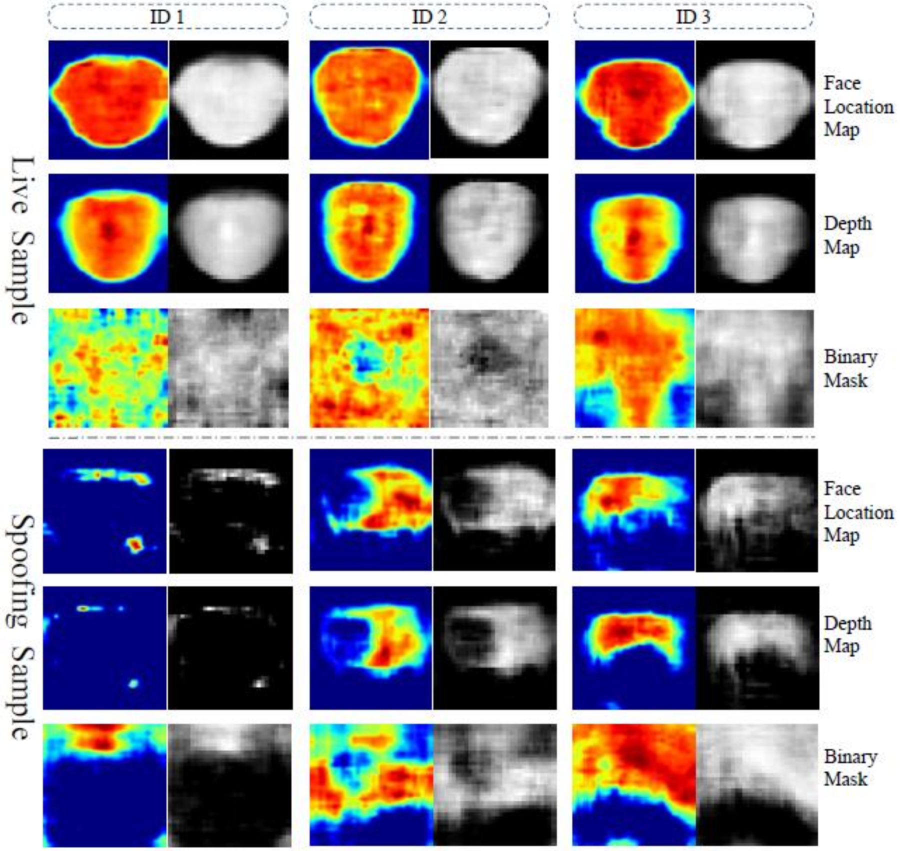
Tag	Structure			ACER(%)↓	
	Label	FCAM	Residual Pyramid	Dev	Test
Compare 1	Binary Mask	√	√	3.95	2.84
Compare 2	Depth Map	×	×	3.62	2.57
Compare 3	Face Location Map	√	×	2.85	2.26
Compare 4	Face Location Map	×	√	3.13	2.22
Compare 5	Depth Map	√	√	2.74	2.06
Baseline	Face Location Map	√	√	**2.48**	**1.88**



- After adding FCAM and Residual Pyramid, ACER decreased by **0.34%** and **0.38%**, respectively.
- Location map supervision 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.
- The proposed method curve shows a smoother decreasing trend during training with less fluctuation.

Visualization





- The model with **FCAM** pays more attention to the parts where the action occurs, so there are higher activation values at pixels.
- The prediction map based on the **face location map** has higher contrast in distinguishing faces and backgrounds.

網易 NETEASE

Comparison on OULU-NPU

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**

Prot.1 environmental conditions
Prot.2 attack means
Prot.3 input camera variation
→ **Prot.4**

- The complexity of protocols 3 and 4 is similar to the realistic scenario where electronic products are changing rapidly.
- The best performance obtained by the proposed method in protocols 3 and 4 demonstrates that our method can maintain accuracy **under complex conditions**.

Conclusion

- 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**.
- 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.

Thank you

SEKE 2022

KSIR Virtual Conference Center, Pittsburgh, USA

July 1 - July 10, 2022

網易 NetEase