University of Zurich[UZH]

Department of Informatics

FAME: Feature Activation Map Explanation on Image Classification and Face Recognition

Xinyi Zhang, Manuel Günther

The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop 2026

1. Introduction

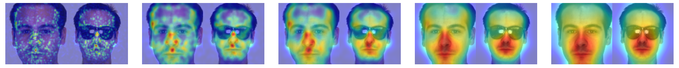
- Image processing networks extract information in highly non-linear and obscured ways, making them **difficult to interpret** — the black-box problem.
- Existing XAI streams are variants of **Grad-CAM** [1] and **Perturbation-based** [2].
- We propose **Feature Activation Map Explanation** (**FAME**): network gradients used at the **pixel level**, combining both worlds without patches or thresholds.
- FAME is applied to visualize **Image Classification** and **Face Recognition**.
- We show that CAM's spatial assumption does not hold for deep networks.

2. Approach

- FAME reinterprets **LOTS** adversarial images as attribution technique [3]:

$$\bar{x} \leftarrow \bar{x} - \eta \frac{\nabla_{\bar{x}}}{\max|\nabla_{\bar{x}}|}, \quad \nabla_{\bar{x}} = \frac{\partial \mathcal{L}(f(\bar{x}), t)}{\partial \bar{x}}, \quad \Delta x = |\bar{x} - x| \in \mathbb{R}^{H \times W}$$

- **Task-specific loss functions** $\mathcal{L}$ are explored below.
- The **attribution map** Δx is Gaussian-smoothed and max-normalized to $[0, 1]$.
- We explore the **Effect of Gaussian Blur in Figure 1:** Small σ is noisy; medium σ balances smoothness and localization; large σ becomes overly diffuse.
- We smooth as in Figure 1(d) for comparability with CAM-based methods.



(a) $\sigma = 0.8$ (b) $\sigma = 2.5$ (c) $\sigma = 4$ (d) $\sigma = 7.7$ (e) $\sigma = 13$

Figure 1. Effect of Gaussian Smoothing on FAME

3. Experiments

- We **analyze feature maps** via $\mathcal{L}_a = \|a[k] - 0\|_1$ in Figure 2.
- **How Reliable is CAM?** In **shallow** networks, each feature map cell corresponds to a local input area; in **deep** networks, receptive fields expand and shift.

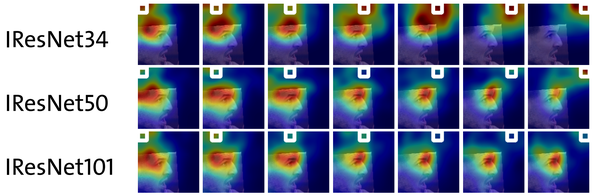


Figure 2. Feature Map Cell Receptive Fields across Network Depths

- We visualize **Facial Embedding Similarity** via $\mathcal{L}_+ = \cos(\varphi_g, \varphi_p)$ in Figure 3.
- FAME produces consistent attributions across datasets, focusing on **identity-relevant regions** under occlusion, resolution degradation, and pose variation.
- FAME consistently **outperforms all baselines** in Insert and achieves competitive or lower Delete across all backbones, see Table 1.

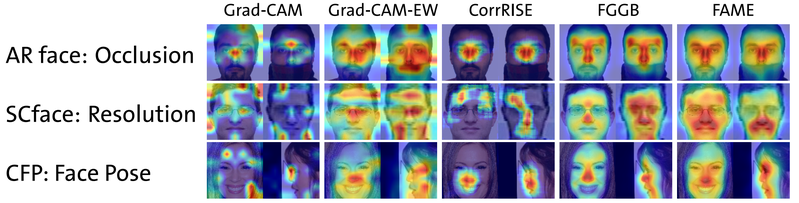


Figure 3. Comparison of Face Recognition Explanations across Face Variations

- We quantify **Image Classification** via class logit: $\mathcal{L}_{\text{cls}} = z_o$ in Table 2.
- FAME **leads in ROAD-Delete** at lower P, and is **competitive in IoU**.

Table 1. IResNet101 Delete ↓, Insert ↑

	AR-glass		SCface-med		CFP-FP	
Method	Del	Ins	Del	Ins	Del	Ins
Grad-CAM	72.6	77.9	67.6	69.9	74.9	85.2
Grad-CAM-EW	60.4	90.1	59.1	80.6	59.3	94.4
CorrRISE	58.4	91.2	57.4	80.4	56.3	96.1
FGGB	58.1	83.7	58.5	75.6	58.7	90.9
FAME	**56.1**	**92.1**	**57.2**	**82.0**	**55.8**	**96.2**

Table 2. ResNet50 on ImageNet

	IoU (%) ↑			ROAD-Delete ↑		
Method	0.3	0.5	0.7	10%	30%	50%
Grad-CAM	38.8	23.4	10.2	0.267	1.035	2.256
Grad-CAM-EW	40.2	24.4	10.7	0.252	0.975	2.124
HiResCAM	38.8	23.4	10.2	0.267	1.035	2.256
FullGradCAM	**48.8**	**33.1**	**13.1**	0.253	1.020	**2.289**
FAME	46.1	29.1	10.8	**0.425**	**1.150**	2.167

4. Conclusions

- CAM's upsampling assumption does not hold for deep networks.
- FAME provides a **generalized**, **unified**, and **thresholdless** framework.
- FAME outperforms Grad-CAM-EW, CorrRISE, and FGGB on ImageNet and three FR datasets under occlusion, pose, and resolution variations.
- Future work extends FAME to Transformer-based architectures.

[1] Selvaraju and others: *Grad-CAM: Visual explanations from deep networks via gradient-based localization.* ICCV, 2017.

[2] Ivanovs, Kadikis, and Ozols: *Perturbation-based methods for explaining deep neural networks: A survey.* Pattern Recognition Letters, 2021.

[3] Rozsa, Günther, and Boult: *LOTS about attacking deep features.* IJCB, 2017.

Code on GitHub