Faces in Focus: Evaluating Generalizability and Transferability of Facial Recognition Explainability Techniques for CNN-based Models

Paweł Borsukiewicz[1]; El-hacen Diallo[1]; Abdoul A. Bonkoungou[1]; Wendkûuni C. Ouédraogo[1]; Jacques Klein[1]; Charles Beumier[2] ; Tegawendé F. Bissyandé[1]
[1]University of Luxembourg, Luxembourg; [2]Royal Military Academy, Brussels

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

Most explainability (XAI) methods for facial recognition provide pairwise explanations, highlighting why two specific images match.

Privacy-oriented applications, such as face de-identification, however, need to identify regions critical to an individual's identity across multiple images. This calls for XAI methods that **generalize** and **transfer**.

We select salient regions for each individual and measure their impact on verification across multiple images of that person, rather than on a specific image pair. We optimize hyperparameters: layers and loss functions for gradient-based techniques.

Generalizability

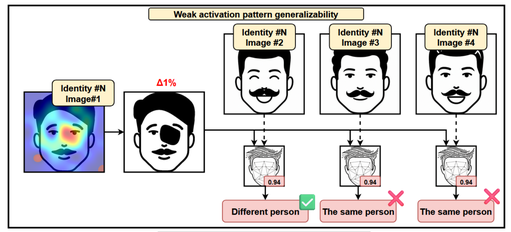


Figure 1. Generalizability refers to the identified facial regions being relevant to a broader range of images.

Transferability

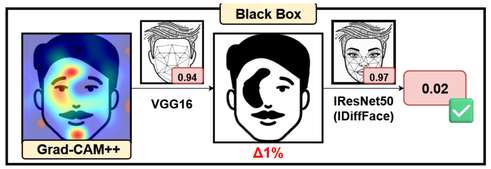
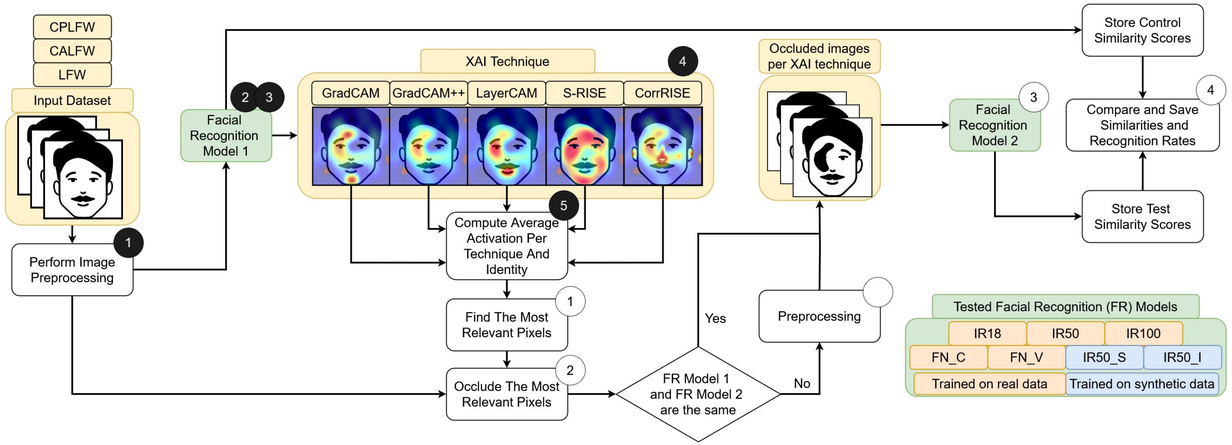


Figure 2. Transferability refers to the successful reuse of key facial regions from a source model to influence a target model.

Experimental Setup

Models: 7 pretrained CNNs (variants of IResNet [1] and FaceNet [2])
Evaluation datasets: LFW, CALFW, CPLFW
XAI techniques: gradient-based (LayerCAM [3], Grad-CAM [4], Grad-CAM++ [5]) and perturbation-based (S-RISE [6], CorrRISE [7])
Pixel occlusion types: black, white, random, mean, and blur



Results - generalizability

Technique	LFW	CPLFW	CALFW
GradCAM++	0.622	0.375	0.485
GradCAM	0.851	0.661	0.748
S-RISE	0.945	0.797	0.853
CorrRISE	0.954	0.817	0.859
LayerCAM	0.958	0.812	0.847

Table 1. Average recognition accuracy per technique across datasets using 1% black pixel occlusions

.

Results - transferability

Generation model	Evaluation model						
	IR50_I	FN_C	FN_V	IR100	IR18	IR50	IR50_S
IR50_I	0.759	0.898	0.936	0.935	0.935	0.765	0.852
FN_C	**0.284**	**0.337**	**0.332**	**0.392**	**0.351**	**0.284**	**0.326**
FN_V	**0.042**	**0.075**	**0.051**	**0.147**	**0.081**	**0.052**	**0.077**
IR100	**0.112**	**0.161**	**0.146**	**0.232**	**0.173**	**0.120**	**0.158**
IR18	**0.142**	**0.195**	**0.181**	**0.261**	**0.209**	**0.150**	**0.189**
IR50	**0.122**	**0.176**	**0.162**	**0.244**	**0.188**	**0.132**	**0.170**
IR50_S	0.761	0.898	0.936	0.934	0.934	0.766	0.852

Table 2. Activation mapping transferability of Grad-CAM++ against black pixel occlusion in CPLFW. Lower values indicate stronger transferability. Values below 0.5 are in bold.

Key Findings

Generalizability:
- Grad-CAM++ is the only method that reliably highlights identity-critical regions.
- LayerCAM, S-RISE, and CorrRISE mainly target areas with negligible impact on recognition.
- XAI techniques are proving less effective for facial recognition models trained on synthetic data.
- Occlusion type has a minor impact on the generalizability evaluation.

Transferability:
- Strong cross-model transferability of the most relevant areas.
- Black-box attacks can be more successful than white-box settings.
- Models trained on synthetic data are equally exposed to the transferability as their traditional counterparts.

Hyperparameter optimization:
- Grad-CAM and Grad-CAM++ are typically the most effective when the second-to-last layer is used. The last layer, the most common choice for these techniques, is a close second.
- Layer choice has a negligible influence on LayerCAM.
- We identify cosine similarity as the best loss function for LayerCAM. In most cases, ArcFace with a 0.2-0.3 margin worked the best for Grad-CAM and Grad-CAM++.

Conclusions

The majority of methods designed for pairwise explanations cannot capture the identity-level consistency required for generalizable explanations, whereas Grad-CAM++ is the only method whose highlighted regions remain predictive of recognition performance across identities and models.

We observe **strong cross-model transferability** of the identified regions, a property particularly **valuable for models trained on synthetic data, where XAI methods tend to underperform**.

Research with financial support from the Luxembourg Armed Forces.

Contact

Paweł Borsukiewicz
University of Luxembourg, SnT, TruX
Email: pawel.borsukiewicz@uni.lu

LinkedIn

References

1. Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
2. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
3. Jiang, Peng-Tao, et al. "Layercam: Exploring hierarchical class activation maps for localization." IEEE transactions on image processing 30 (2021): 5875-5888.
4. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
5. Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.
6. Lu, Yuhang, and Touradj Ebrahimi. "Explanation of face recognition via saliency maps." Applications of Digital Image Processing XLVI. Vol. 12674. SPIE, 2023.
7. Lu, Yuhang, Zewei Xu, and Touradj Ebrahimi. "Towards visual saliency explanations of face verification." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.