Explainable Artificial Intelligence for Face Presentation Attack Detection

Wilson Silva, João Ribeiro Pinto, Tiago Gonçalves, Ana F. Sequeira, and Jaime S. Cardoso

INESC TEC and Faculty of Engineering, University of Porto, Porto, Portugal – wilson.j.silva@inesctec.pt

Objectives

- **Assess** the **robustness** of face PAD models.
- Define **interpretability**-related **properties** of a **robust** face **PAD model**.

Introduction

- **Deep learning** algorithms are **excelling** in **most** of the artificial intelligence **fields**.
- Sometimes deep learning incredible performances are obtained by a **focus** in **wrong/biased** dataset-related **information** instead of domain significant information [1].
- An **evaluation** performed based on only the **traditional metrics** may be **misleading**.
- We propose the use of **interpretability methods** to further **assess** model **robustness**.

Methodology

- A **PAD method** receives as input a **biometric trait measurement** and returns as output a **prediction**: living individual (***bona fide***) or spoof attempt to intrude the system (***attack***).

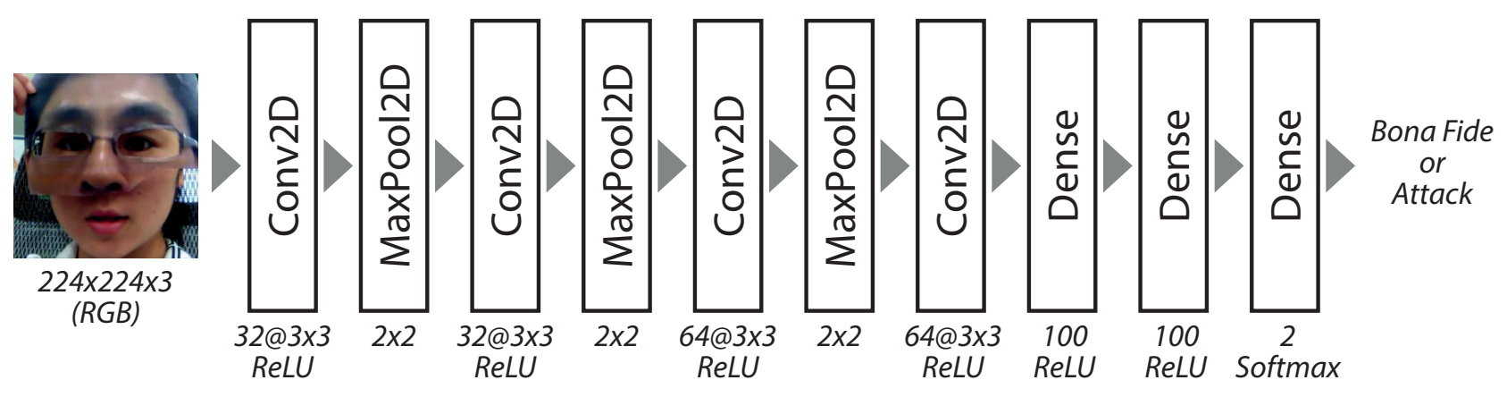

Figure 1:Architecture of the implemented PAD model.

- With regards to the interpretability method, we selected **Grad-CAM** [2], as it has the flexibility to generate explanations for any layer of the network, and also allow us to obtain class-specific explanations.
- The **experiments** were performed with the **ROSE-Youtu Face Liveness Detection** Dataset [3].

Table 1:Characteristics of the presentation attack instruments in the ROSE Youtu dataset [3].

Attack	Type of presentation attack instruments	N.I.
-	Genuine (bona fide)	2794
#1	Still printed paper	1136
#2	Quivering printed paper	1188
#3	Video of a Lenovo LCD display	923
#4	Video of a Mac LCD display	1113
#5	Paper mask without cropping	1194
#6	Paper mask with two eyes and mouth cropped out	608
#7	Paper mask with the upper part cut in the middle	1162

Experimental Assessment

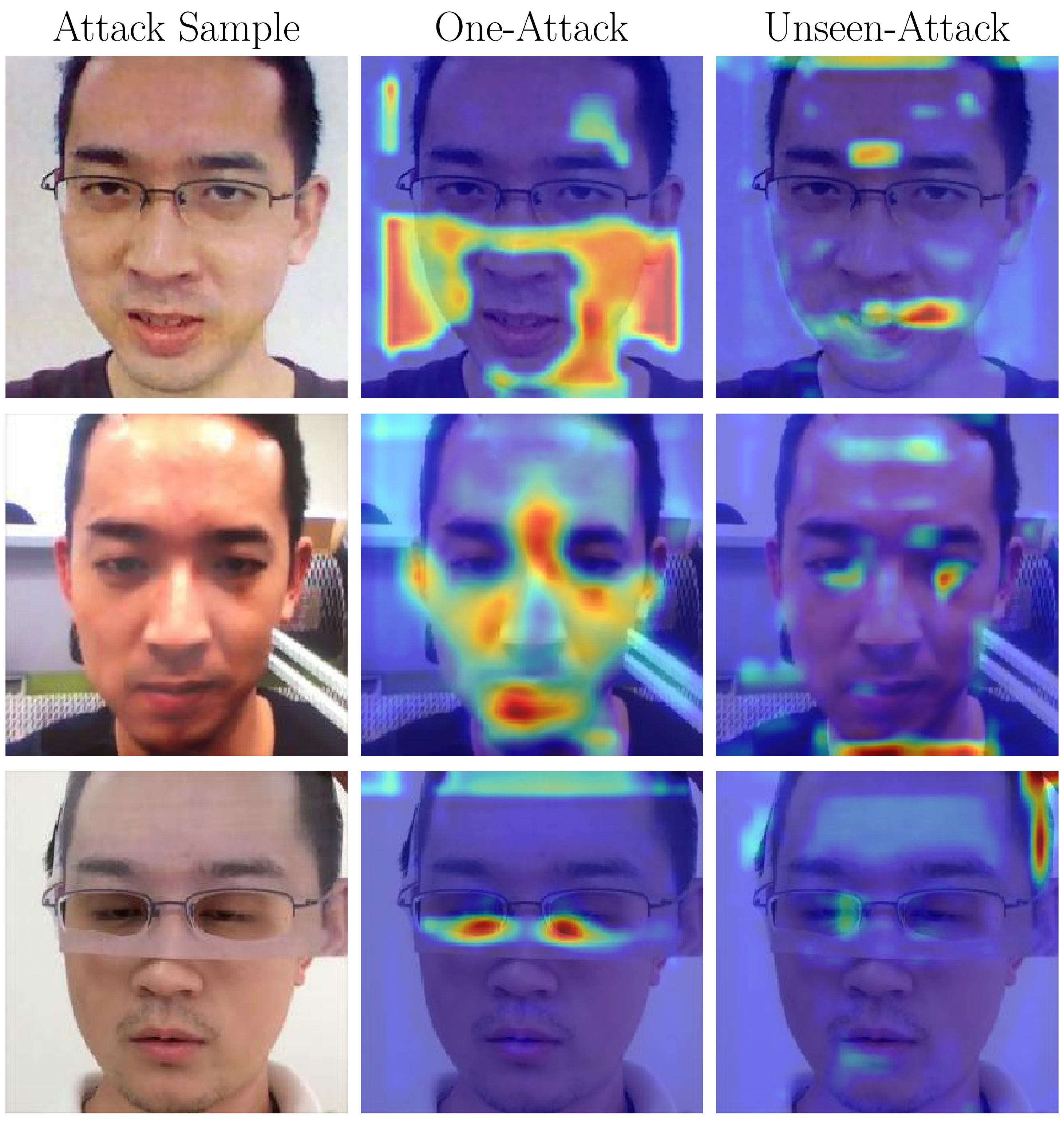

Figure 2:Explanations for correctly classified attack samples (TP) in the One-Attack (2^{nd} column) or Unseen-Attack (3^{rd} column) frameworks. Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

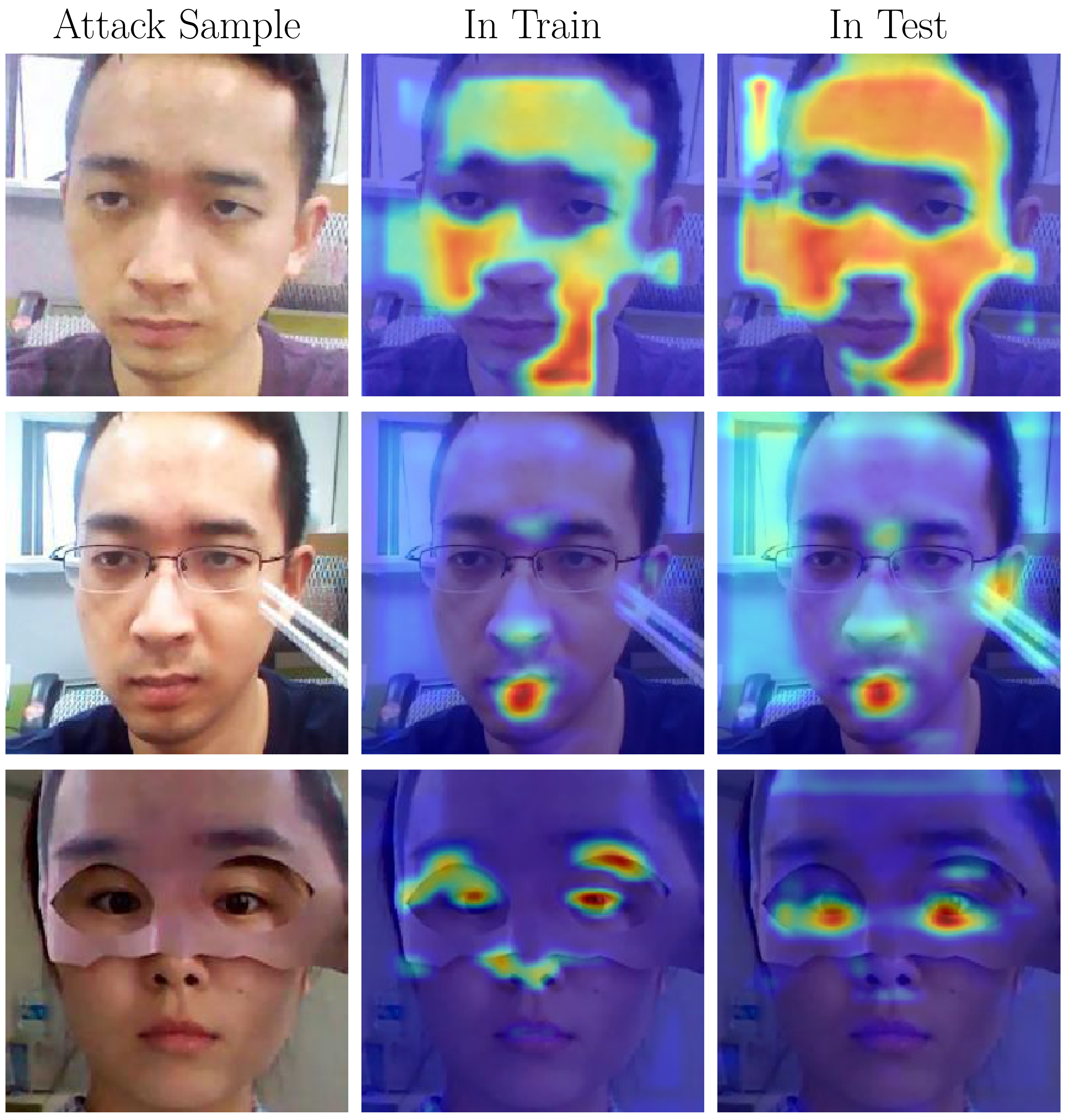

Figure 3:Grad-CAM Explanations for correctly classified attack samples when a subject is in the train set (2^{nd} column) or in the test set (3^{rd} column). Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

Desirable Properties

- **Explanations** for the **same sample** should be **similar whether or not** it is **seen** during **training** (data swap).
- **Explanations** for the **same sample** should be **similar whether or not** the **model** is **trained** to detect that **specific attack** (One-Attack vs. Unseen-Attack).

Findings and Conclusions

- **Interpretability** was explored to further **assess** the **robustness** of face PAD models.
- We defined **desirable properties** for a face PAD model to fulfill that are **verifiable** through an **interpretability** analysis of the models.
- This **interpretability evaluation** can only be done **qualitatively**, therefore, **lacking objectivity**.
- **Future work** will focus on finding ways of **quantifying** the information obtained with the **interpretability** analysis.

References

[1] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek.
Analyzing classifiers: Fisher vectors and deep neural networks.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2912–2920, 2016.

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[3] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot.
Unsupervised domain adaptation for face anti-spoofing.
IEEE Transactions on Information Forensics and Security, 13(7):1794–1809, 2018.

[4] Ana F Sequeira, Wilson Silva, João Ribeiro Pinto, Tiago Gonçalves, and Jaime S Cardoso.
Interpretable biometrics: Should we rethink how presentation attack detection is evaluated?
In 2020 8th International Workshop on Biometrics and Forensics (IWBF), pages 1–6. IEEE, 2020.

Acknowledgements

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707", and within the PhD grants "SFRH/BD/137720/2018", "SFRH/BD/139468/2018" and "SFRH/BD/06434/2020"