Supplementary Documents for Detecting Presentation Attacks on ID cards Using Feature Refinement

Raghavendra Mudgalgundurao, Patrick Schuch, Raghavendra Ramachandra, Kiran Raja

I. Introduction

This document serves as supplementary material for the paper titled "Detecting Presentation Attacks on ID cards Using Feature Refinement" It includes results from ablation studies, benchmarking our models using various data mixup strategies with different α values, and a detailed analysis of the impact of each attack on the proposed model. Additionally, we present an explainability analysis of the proposed model using class activation maps. This supplementary material ensures that reviewers and readers can effectively reference and evaluate the visual and tabulated data supporting our research findings.

II. Dataset Description and Splits

The experiments were conducted using two primary datasets: MIDV-500 and KID34K which were partitioned into training (70%), testing (20%), and validation(10%) subsets to ensure a rigorous evaluation of the proposed model. The datasets include bona fide samples and four types of presentation attacks: screen, replay, diffusion generated attacks, and crafted texture transfer. These attack types simulate real-world scenarios where fraudulent attempts are made to deceive document verification systems. A detailed breakdown of the training, testing, and validation numbers are provided in the Table I.

1) Dataset Composition:

- Training Set:
  - Bona fide Samples: The training set consists of 67,011 bona fide images. From these, additional attack samples were generated:
  - Print Attacks: Created by printing and capturing the bona fide images.
  - Screen Attacks: Generated by displaying the bona fide images on a screen and capturing them.
  - Crafted Texture Transfer: Synthesized by applying texture manipulation techniques to mimic genuine documents.
  - Diffusion-Based Images: Using the ULD network, 10,500 diffusion-based images were generated to enhance the diversity of the training data.
  - Remaining Attack Samples: The remaining 56,511 images were distributed across screen, print, and texture transfer attacks, ensuring a balanced representation of attack types.
- Testing and Validation Sets:
  - The testing and validation sets were constructed using unseen data to evaluate the model's generalization capabilities. These sets include bona fide samples and the same types of attacks (screen, replay, texture transfer, and ULD generated) as the training set, ensuring a consistent evaluation framework.

2) Dataset Splits: The datasets were partitioned as follows:

- MIDV-500:
  - Training: Folders 1–35 (35 folders, 55,668 images).
  - Testing: Folders 36–45 (10 folders, 16,323 images).
  - Validation: Folders 46–50 (5 folders, 7,783 images).
- KID34K:
  - DL (Driving License) Subset:
    - Training: Folders 1–31 (31 folders, 7,727 images).
    - Testing: Folders 32–40 (9 folders, 865 images).
    - Validation: Folders 41–45 (5 folders, 479 images).
  - ID Cards Subset:
    - Training: Folders 101–126 (26 folders 3,614 images).
    - Testing: Folders 127–133 (7 folders, 674 images).
    - Validation: Folders 134–137 (4 folders, 382 images).

3) Mixed Data Augmentation: To enhance the model's discriminative capability and robustness, we employ mixup augmentation during training. Mixup generates synthetic training examples by blending pairs of data samples and their corresponding labels. Specifically, we apply mixup between bona fide and attack classes, creating interpolated samples that combine features of bona fide and attacks. This approach exposes the model to a diverse range of blended samples, improving its ability to distinguish subtle differences between bona fide and adversarial inputs. By training on these interpolated samples, the model achieves better generalization and increased robustness against adversarial inputs.

Mixup augmentation is applied exclusively to the training set, ensuring that the testing and validation sets remain untouched and representative of real-world scenarios. This strategy aligns with our goal of maintaining a rigorous evaluation framework while leveraging data augmentation to improve model performance.

4) Attack Types: All subsets (training, testing, and validation) include the following attack types:

- Screen: Attacks where the document is displayed on a screen and captured.

Dataset	Training	Testing	Validation
MIDV-500			
Folders	1–35	36–45	46–50
Images	55,668	16,323	7,783
KID34K DL			
Folders	1–31	32–40	41–45
Images	7,727	865	479
KID34K ID			
Folders	101–126	127–133	134–137
Images	3,614	674	382
Diffusion Images			
Images	10,500	3,000	1,500
Screen Images			
Images	27,786	7,938	3,969
Print Images			
Images	27,785	7,938	3,969
Crafted Template Transfer Images			
Images	938	268	134

Table STI: Dataset split for training, testing, and validation

- Replay: Attacks where the document is replayed using a secondary device.
- Crafted Texture Transfer: Attacks involving synthetic texture manipulation to mimic genuine documents.
- Diffusion Generated images: Attacks created from the ULD network, based on the training data.

III. Results and Discussions

The results presented in Table STIII provide a detailed comparative analysis of various approaches in terms of Equal Error Rate (EER%), Bona Fide Presentation Classification Error Rate (BPCER) at Attack Presentation Classification Error Rate (APCER) = 5%, and APCER = 10%. The table encompasses state-of-the-art benchmarks, deep learning models, Vision Transformers (ViTs), and the proposed EfficientNet-Transformer architecture, all evaluated under mixed data augmentation settings.

A. State-of-the-Art Benchmarks

Among the baseline methods, Gonzalez [1] achieves the lowest EER of 7.61%, with a BPCER@APCER=5% of 10.41% and BPCER@APCER=10% of 5.95%. In contrast, the PixelWise Model [2] and Gonzalez and Tapia [3] exhibit significantly higher error rates, with EERs of 21.02% and 23.46%, respectively. These results highlight the limitations of the existing SOTA in handling complex presentation attack scenarios.

B. Deep Models with Mixed Data Augmentation

The EfficientNet-B4 architecture demonstrates robust performance across varying augmentation intensities (α). At $\alpha = 0.2$, it achieves an EER of 7.09%, BPCER@APCER=5% of 12.97%, and BPCER@APCER=10% of 3.51%. Notably, at $\alpha = 0.7$, EfficientNet-B4 achieves its lowest EER of 4.83%, suggesting that moderate augmentation levels enhance model generalization. However, at $\alpha = 1.0$, the EER increases to 6.88%, indicating that the model's performance degrades when trained without augmentation, suggesting the importance of data augmentation for optimal results.

MobileNet V2 and MobileNetV3-Large also exhibit competitive performance. MobileNet V2 ($\alpha = 0.2$) records an EER of 4.89%, BPCER@APCER=5% of 4.81%, and BPCER@APCER=10% of 3.37%, while MobileNetV3-Large ($\alpha = 0.2$) achieves an EER of 7.74%, BPCER@APCER=5% of 9.84%, and BPCER@APCER=10% of 6.64%. Both models show a trade-off between EER and BPCER, with MobileNet V2 achieving lower EERs but higher BPCERs compared to EfficientNet-B4.

C. Vision Transformers (ViTs) with Mixed Data Augmentation

Vision Transformers, while promising in other domains, exhibit suboptimal performance in this task. For instance, ViT ($\alpha = 0.2$) achieves an EER of 21.68%, BPCER@APCER=5% of 56.11%, and BPCER@APCER=10% of 38.98%. The performance deteriorates further at higher α values, with ViT ($\alpha = 1.0$) reaching an EER of 25.18%, BPCER@APCER=5% of 61.71%, and BPCER@APCER=10% of 50.41%. This suggests that ViTs may require additional architectural adaptations or training strategies to effectively handle presentation attack detection.

D. Proposed EfficientNet-Transformer Model

The proposed EfficientNet-Transformer hybrid architecture achieves state-of-the-art performance, particularly at $\alpha = 0.2$, with an EER of 3.14%, BPCER@APCER=5% of 2.42%, and BPCER@APCER=10% of 1.33%. This represents a significant improvement over both traditional benchmarks and standalone deep learning models. The model's ability to maintain low error rates across varying α values (e.g., EER of 4.63% at $\alpha = 0.7$ and 5.92% at $\alpha = 1.0$) underscores its robustness and generalization capabilities. The optimal performance at $\alpha = 0.2$ suggests that a balanced augmentation strategy is

critical for maximizing detection accuracy while minimizing false positives.

E. Discussion

The results demonstrate that the proposed EfficientNet-Transformer architecture, combined with mixed data augmentation, effectively addresses the challenges of presentation attack detection. Its superior performance, particularly in terms of BPCER metrics, highlights its potential for deployment in high-security applications. Future work could explore further architectural refinements and training strategies to enhance the performance of Vision Transformers in this domain.

IV. Analysis for explainability of proposed EfficientNet-Transformer Network

We provide insights into the proposed EfficientNet-Transformer network by utilizing Class Activation Maps (CAMs). CAM analysis is conducted to identify regions of interest across various areas of ID cards. This is achieved by computing linear combinations of the activations from the final convolutional layer with the output weights corresponding to the target class. The resulting CAMs are overlaid on the ID cards, visually highlighting the regions that influence the model's decision-making process in distinguishing between bona fide and attack presentations.

To take a closer look at the suggested method, we use three Class Activation Mapping (CAM) techniques: GradCAM[4], HiResCAM[5], and ScoreCAM[6]. GradCAM creates class-specific localization maps by using the gradients of the target class with the final convolutional layer. It combines activation maps with weights and then applies a ReLU function, to highlight the areas that matter most for the prediction. HiResCAM improves on GradCAM by not relying on gradient backpropagation. Instead, it scales feature map activations using their global average pooling values, which results in high-resolution visual explanations and cuts down on noise and instability. ScoreCAM however, doesn't use gradient information at all and checks how much each activation map contributes through a series of forward passes. By combining activation maps based on how important they are and applying ReLU, ScoreCAM provides clear and understandable visualizations that help solve problems like sensitivity to noise and issues related to gradients.

A. CAM Analysis with Different α Values

CAM visualizations correspond to varying values of the alpha parameter in the mixup strategy: 0.2, 0.5, 0.7, 0.8, and 1.0. Each visualization highlights the focus regions for the proposed deep learning model when detecting different presentation attacks (e.g., print attack, replay attack, diffusion) and distinguishing them from bona fide samples. We provide a summary of the analysis in Table STII.

B. CAM with $\alpha = 0.2$

GradCAM Figure SF8, the model focuses intensely on narrow and specific regions of the ID cards. The activation maps show highly localized attention, particularly on edges and distinct textual regions (e.g., photo areas and machine read-able zone (MRZ)). This suggests that the model relies on specific fine-grained details to distinguish attacks from bona fide samples.

ScoreCAM Figure SF14 ScoreCAM produces smooth and consistent heatmaps with sharp activations around critical areas, such as faces, text regions, and edges. Manipulated regions in Print and Replay attacks receive strong attention, especially around printed textures and boundaries.

HighResCAM Figure SF20 HiResCAM at this α value provides the most localized and fine-grained activations. The heatmaps demonstrate a precise focus on facial features and text while also clearly highlighting attack-specific artifacts, such as fine edges and distortions. This indicates the model is learning crucial information to distinguish between bona fide versus presentation attacks.

C. CAM with $\alpha = 0.5$

GradCAM Figure SF9 The model continues to highlight critical features such as facial photos and MRZ regions but with broader coverage across the ID card. This balance indicates the model is capturing more global patterns while still focusing on key areas.

ScoreCAM Figure SF15 ScoreCAM maintains smooth and sharp activations but starts to exhibit slightly broader heatmaps compared to lower α. Facial and text regions remain dominant, with a good emphasis on attack-specific anomalies. The boundaries and high-texture regions in replay attacks still receive significant attention.

HiResCAM Figure SF21 The heatmaps remain highly focused on critical regions, such as faces and text, with enhanced visibility of attack artifacts. This indicates model is focussing more on the general and global aspects of the ID cards, which results in less focus on the finer details and the same can be corroborated with Table III - Proposed EfficientNet-Transformer - With Mixed Data Augmentation.

D. CAM with $\alpha = 0.7$

GradCAM Figure SF10 For bona fide samples, the focus remains on the facial photo and card details. However, for presentation attacks like print and replay, the model starts identifying broader areas of artifacts (e.g., smudging, color inconsistencies).

ScoreCAM SF16 ScoreCAM shows the model focuses on broader activations, with slightly less focus on fine details. While the key regions, such as faces and text, remain prominent, the precision in highlighting specific manipulated areas reduces slightly compared to lower α values.

HiResCAM Figure SF22 The heatmaps remain focused and provide clear insights into attack artifacts. Fine-textured details, particularly in Print and Replay attacks, are still well-emphasized.

E. CAM with $\alpha = 0.8$ and $\alpha = 1$

GradCAM Figure SF11, Figure SF12 The visualizations highlight large areas across the ID card for all attack types,

α	**GradCAM**	**ScoreCAM**	**HiResCAM**
0.2	Localized but less smooth	Smooth and sharp	Best fine-grained focus
0.5	Slight loss of sharpness	Smooth with moderate spread	Retains sharp focus
0.7	Spread-out, less clarity	Smooth but diluted focus	Maintains better precision
0.8 and 1.0	Highly diffused, unclear	Smooth but general focus	Best clarity among methods

Table STII: Comparison of GradCAM, ScoreCAM, and HiResCAM across different α values.

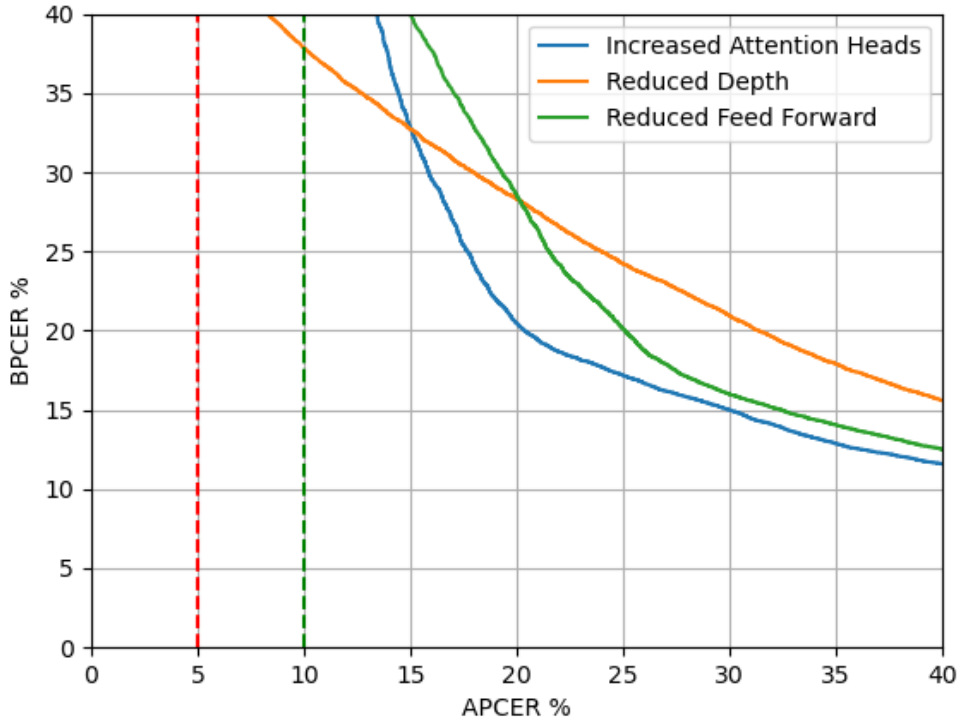

Fig. SF1: DET Plots for ablation studies.

including bona fide samples. While key regions like the photo and textual areas remain prominent, the broader coverage indicates that the model is learning more generalized patterns. This may result in slightly reduced precision in distinguishing subtle attack artifacts.

ScoreCAM Figure SF17, Figure SF18 Visualizations still highlight the general regions of interest, but the emphasis on attack-specific artifacts and fine details becomes less distinct.

HiResCAM Figure SF23, Figure SF24 HiResCAM retains its fine-grained focus on facial regions, text, and attack-specific anomalies, demonstrating its resilience to variations in α.

In summary, α = 0.2 appears to strike a balance between local and global attention, suggesting it may provide optimal performance for distinguishing bona fide samples from presentation attacks. Whereas higher alpha values (e.g., 0.8 and 1.0) result in broader attention across the ID card, promoting generalization but potentially reducing precision for detecting localized attack-specific cues.

Approach	EER%	BPCER @ APCER@5%	BPCER @ APCER@10%
State of the Art Benchmark			
Gonzalez[1]	7.61	10.41	5.95
PixelWise Model[2]	21.02	56.13	36.35
Gonzalez and Tapia[7]	23.46	70.42	47.59
Deep models and Vision Transformer - With Mixed Data Augmentation			
Efficientnet-B4 ($\alpha = 0.1$)	5.12	5.36	1.80
Efficientnet-B4 ($\alpha = 0.2$)	7.09	12.97	3.51
Efficientnet-B4 ($\alpha = 0.5$)	6.22	4.48	3.5
Efficientnet-B4 ($\alpha = 0.7$)	4.83	5.40	1.87
Efficientnet-B4 ($\alpha = 0.8$)	5.20	5.88	1.02
Efficientnet-B4 ($\alpha = 1$)	6.88	9.57	5.11
MobileNet V2 ($\alpha = 0.1$)	4.49	4.23	3.12
MobileNet V2 ($\alpha = 0.2$)	4.89	4.81	3.37
MobileNet V2 ($\alpha = 0.5$)	5.02	5.38	2.29
MobileNet V2 ($\alpha = 0.7$)	4.75	4.42	1.67
MobileNet V2 ($\alpha = 0.8$)	4.92	5.21	2.46
MobileNet V2 ($\alpha = 1$)	4.81	6.16	1.92
MobileNetV3-Large ($\alpha = 0.1$)	4.39	8.23	5.27
MobileNetV3-Large ($\alpha = 0.2$)	7.74	9.84	6.64
MobileNetV3-Large ($\alpha = 0.5$)	5.93	6.54	2.26
MobileNetV3-Large ($\alpha = 0.7$)	5.04	5.24	2.64
MobileNetV3-Large ($\alpha = 0.8$)	5.31	6.65	3.79
MobileNetV3-Large ($\alpha = 1$)	4.99	5.95	3.43
ViT ($\alpha = 0.2$)	21.68	56.11	38.98
ViT ($\alpha = 0.5$)	22.57	55.21	40.97
ViT ($\alpha = 0.7$)	23.46	63.52	47.26
ViT ($\alpha = 0.8$)	22.14	54.21	41.75
ViT ($\alpha = 1$)	25.18	61.71	50.41
Proposed EfficientNet-Transformer - With Mixed Data Augmentation			
Proposed ($\alpha = 0.2$)	**3.14**	**2.42**	**1.33**
Proposed ($\alpha = 0.5$)	10.88	11.14	10.56
Proposed ($\alpha = 0.7$)	4.63	6.05	3.19
Proposed ($\alpha = 0.8$)	7.37	9.71	5.44
Proposed ($\alpha = 1.0$)	5.92	5. 64	2.18

Table STIII: The results from our proposed approach compared to baseline methods show that our approach achieves a lower Equal Error Rate (EER) than the other methods evaluated. **Bar Graphs for the above table are presented in Figures* SF2 - SF6

Approach	EER%	BPCER@5%	BPCER@10%
Proposed Network - With Bona fide vs Diffusion Generated Images			
Proposed ($\alpha = 0.2$)	0.42	0.64	0.31
Proposed ($\alpha = 0.5$)	1.96	6.31	2.14
Proposed ($\alpha = 0.7$)	0.42	0.64	0.31
Proposed ($\alpha = 0.8$)	0.69	2.28	0.96
Proposed ($\alpha = 1.0$)	0.64	1.38	0.18

Table STIV: Ablation studies for our proposed model, evaluating bona fide vs diffusion attacks.

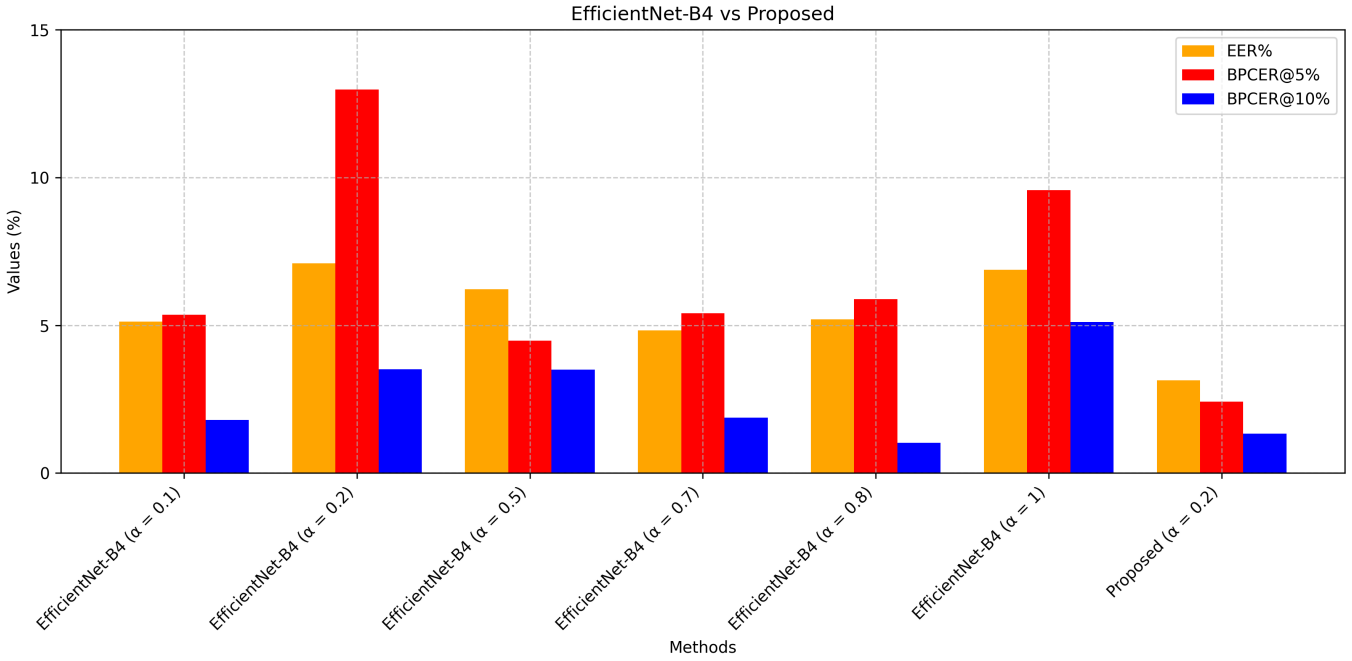

Fig. SF2: Graphical representation of EER, BPCER@5% and 10% for EfficientNet-B4 with different α values vs Proposed.

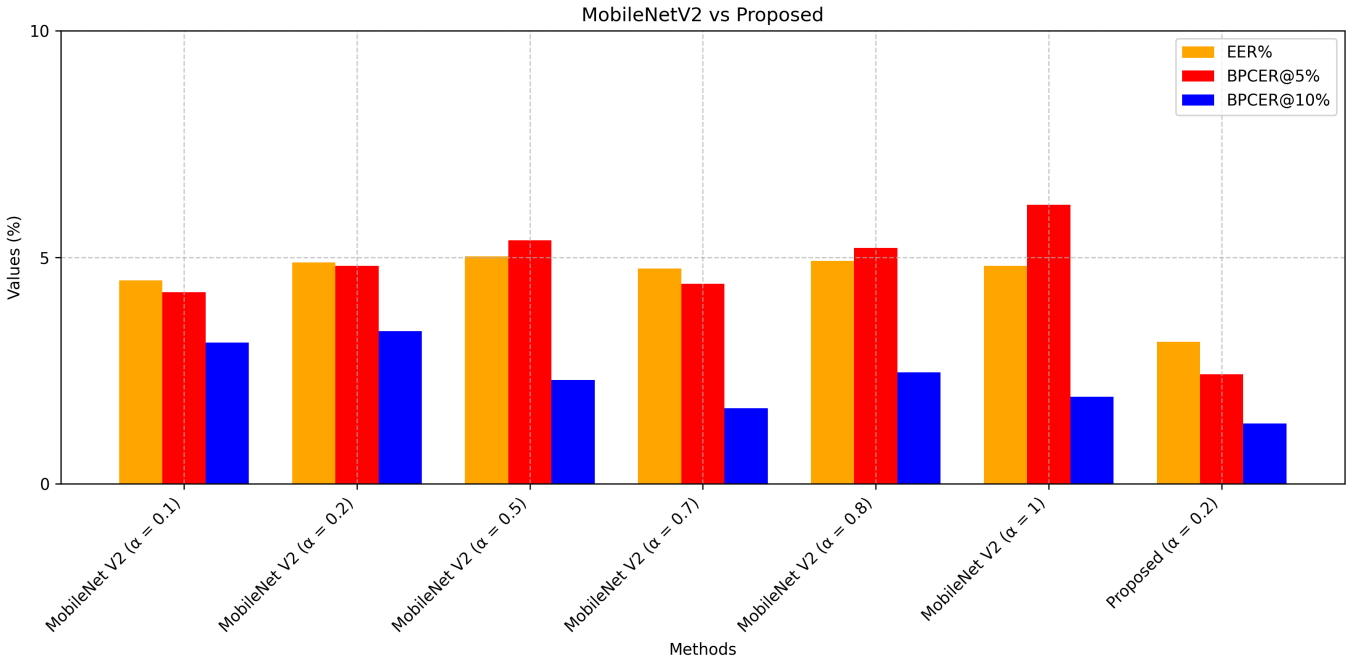

Fig. SF3: Graphical representation of EER, BPCER@5% and 10% for MobileNetV2 with different α values vs Proposed.

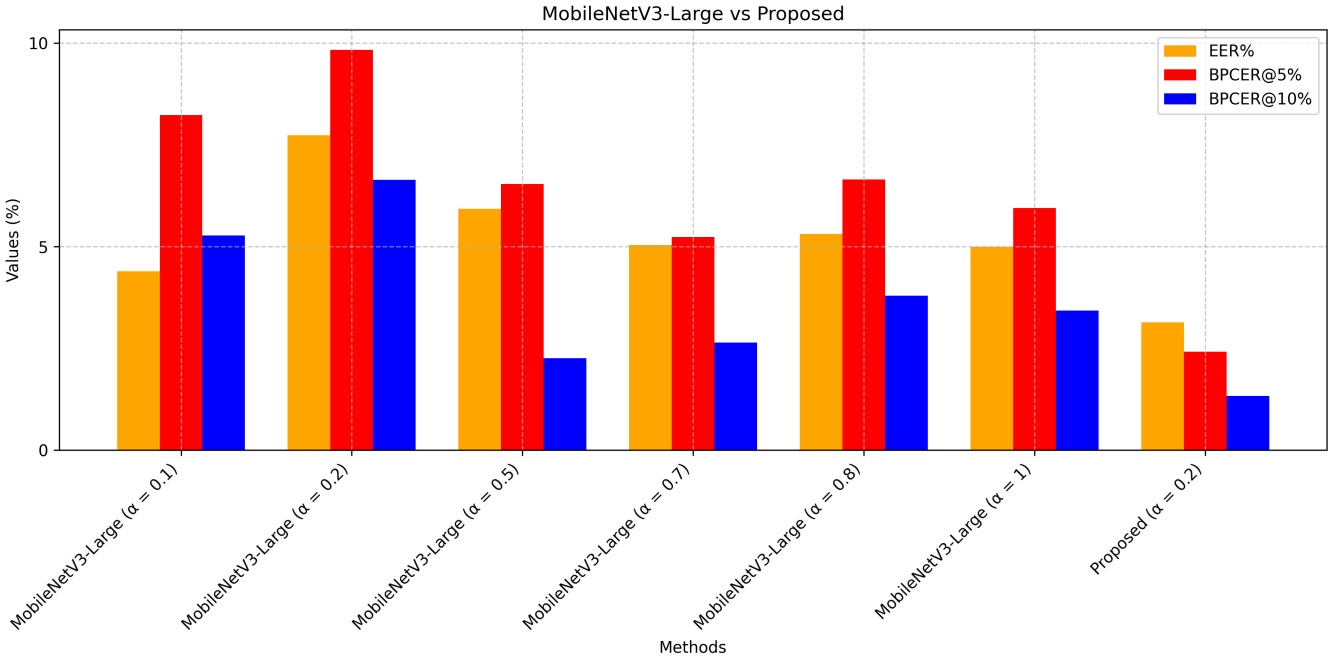

Fig. SF4: Graphical representation of EER, BPCER@5% and 10% for MobileNetV3 with different α values vs Proposed.

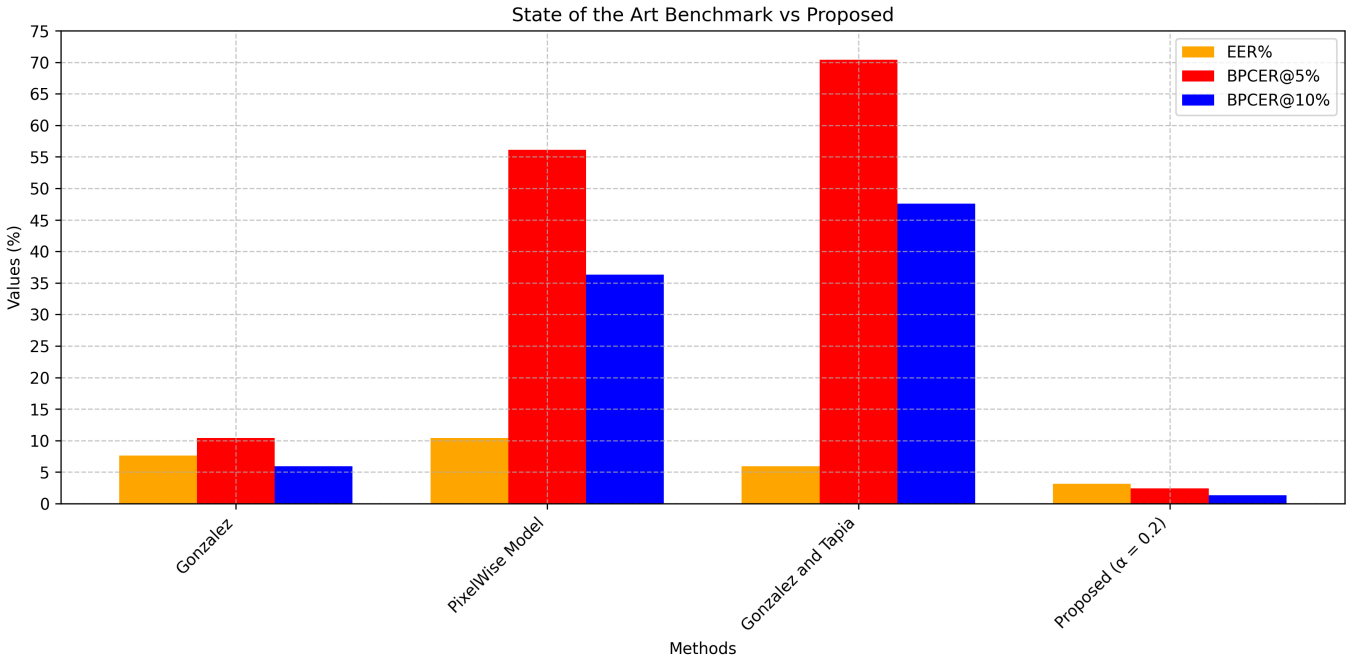

Fig. SF5: Graphical representation of EER, BPCER@5% and 10% for SOTA with different α values vs Proposed.

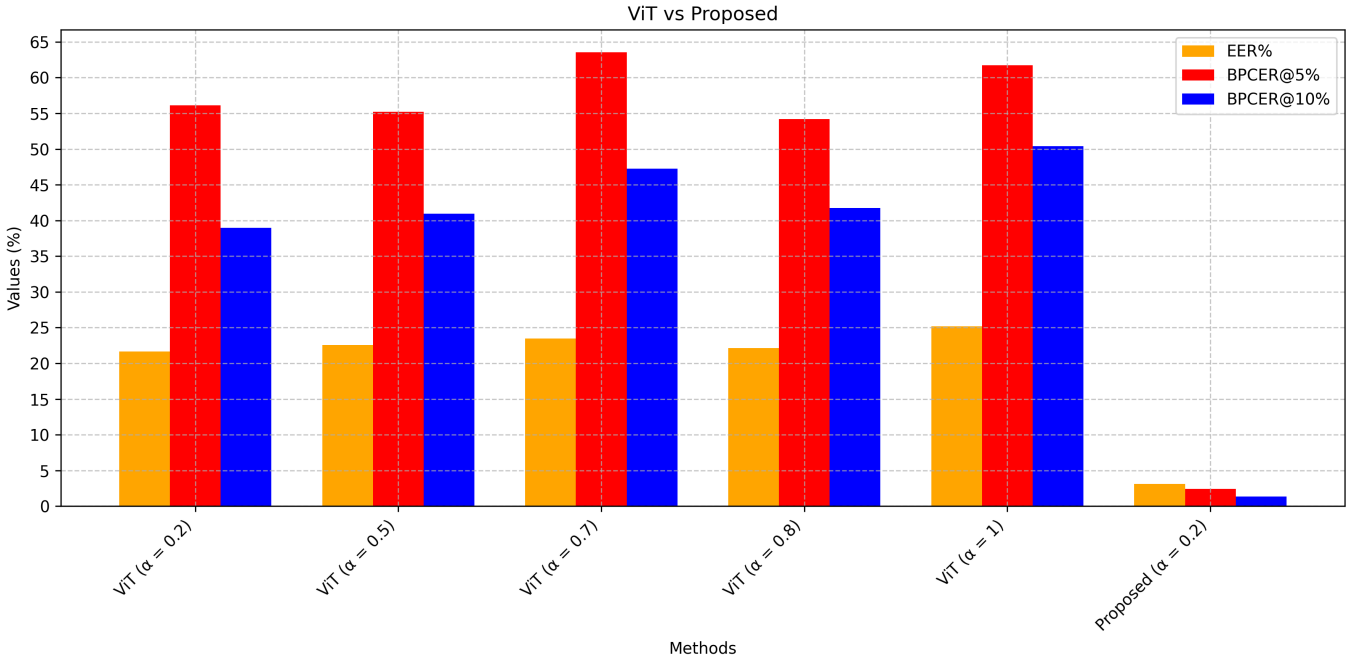

Fig. SF6: Graphical representation of EER, BPCER@5% and 10% for ViT with different α values vs Proposed.

Approach	**EER%**	**BPCER@5%**	**BPCER@10%**
Proposed Network - With Bona fide vs Print Attacks			
Proposed ($\alpha = 0.2$)	1.32	5.5	2.78
Proposed ($\alpha = 0.5$)	2.63	12.97	8.54
Proposed ($\alpha = 0.7$)	2.22	13.27	10.27
Proposed ($\alpha = 0.8$)	1.47	12.62	7.83
Proposed ($\alpha = 1.0$)	1.58	7.16	3.54

Table STV: Ablation studies for our proposed model, evaluating bona fide vs print attacks.

Approach	**EER%**	**BPCER@5%**	**BPCER@10%**
Proposed Network - With Bona fide vs Screen Attacks			
Proposed ($\alpha = 0.2$)	3.82	10.21	4.51
Proposed ($\alpha = 0.5$)	2.83	8.72	5.07
Proposed ($\alpha = 0.7$)	5.01	8.98	3.27
Proposed ($\alpha = 0.8$)	4.35	12.69	6.19
Proposed ($\alpha = 1.0$)	3.83	12.03	8.45

Table STVI: Ablation studies for our proposed model, evaluating bona fide vs screen attacks.

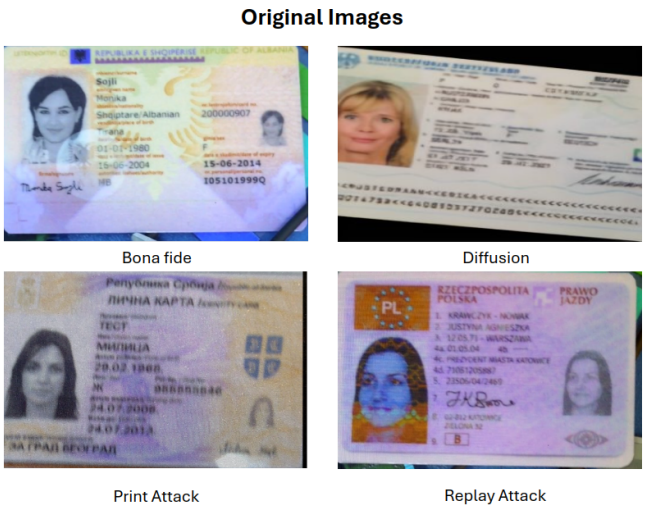

Fig. SF7: Original Images

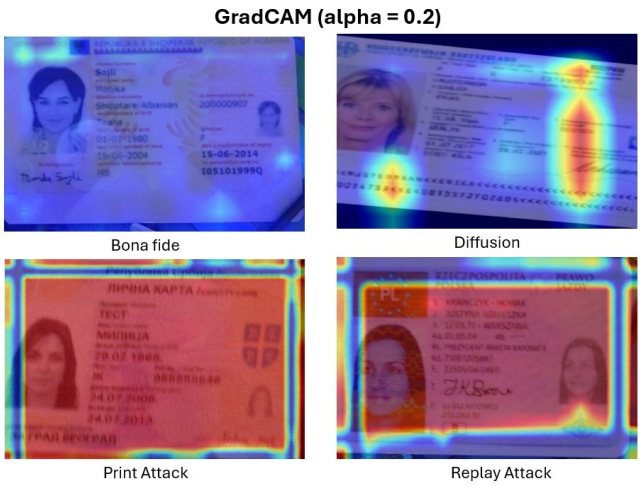

Fig. SF8: GradCAM (α=0.2)

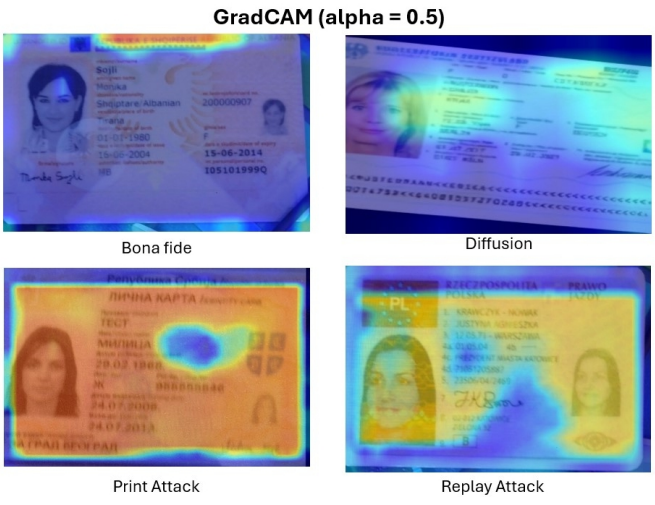

Fig. SF9: GradCAM (α=0.5)

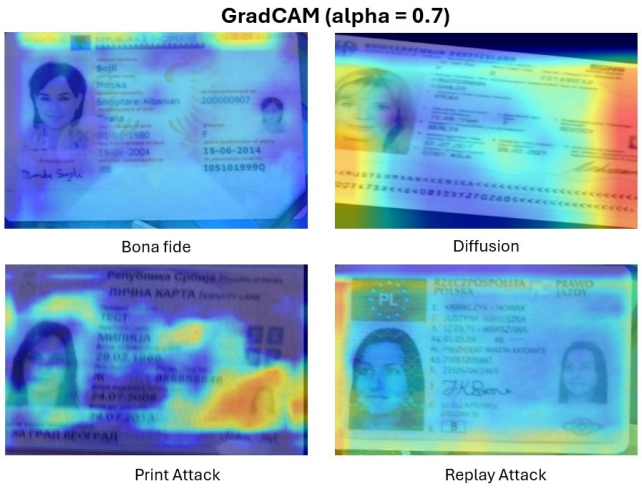

Fig. SF10: GradCAM (α=0.7)

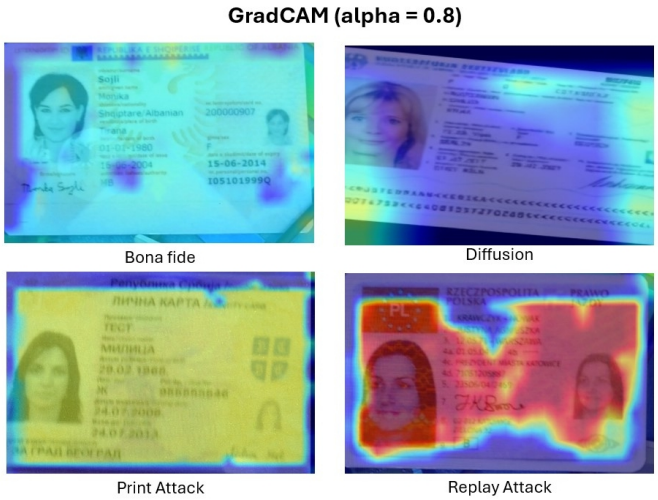

Fig. SF11: GradCAM (α=0.8)

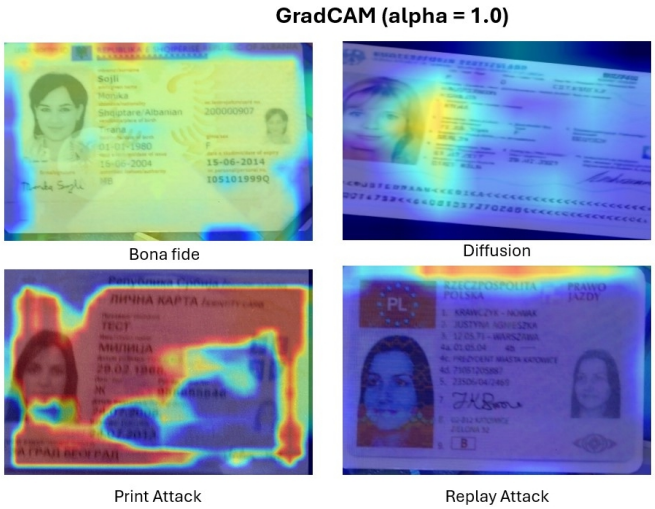

Fig. SF12: GradCAM (α=1.0)

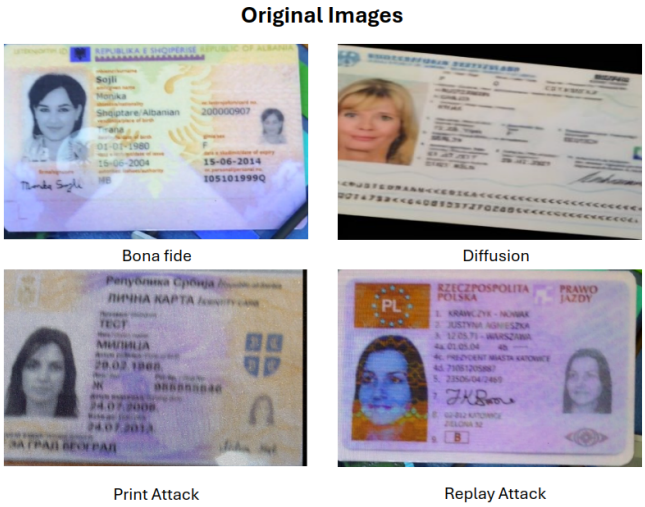

Fig. SF13: Original Images

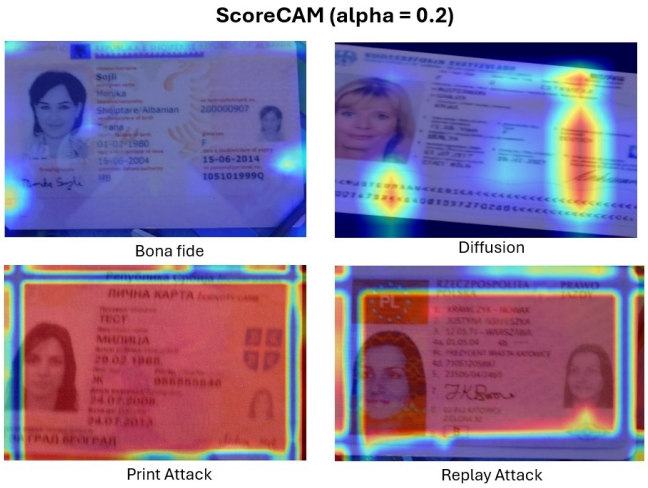

Fig. SF14: ScoreCAM (α=0.2)

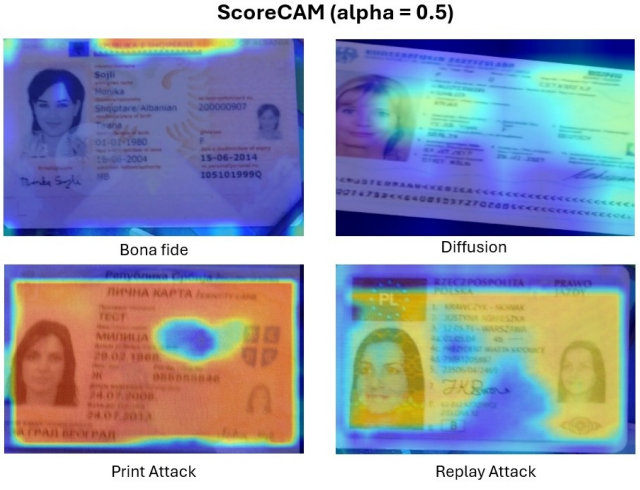

Fig. SF15: ScoreCAM (α=0.5)

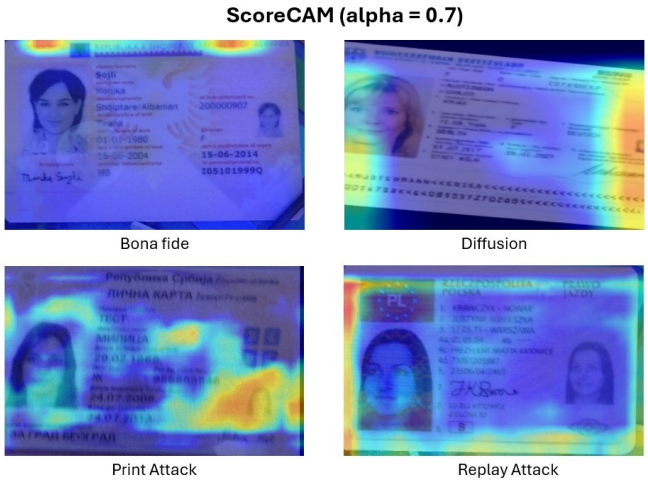

Fig. SF16: ScoreCAM (α=0.7)

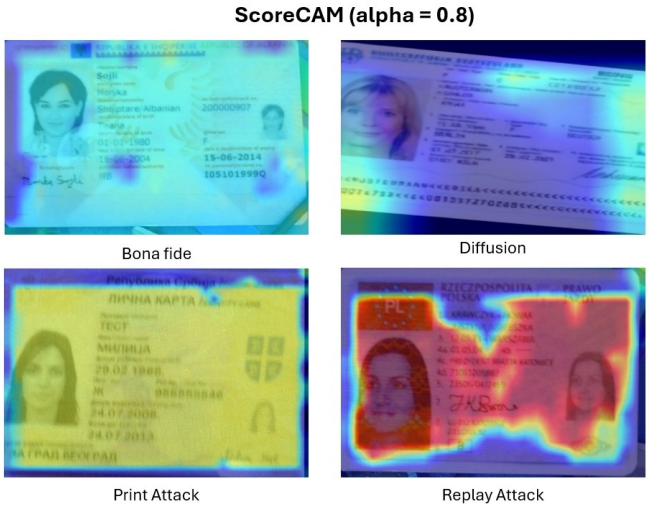

Fig. SF17: ScoreCAM (α=0.8)

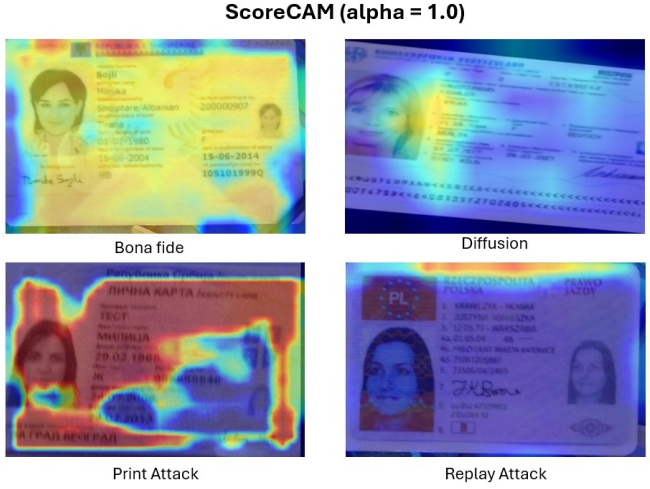

Fig. SF18: ScoreCAM (α=1.0)

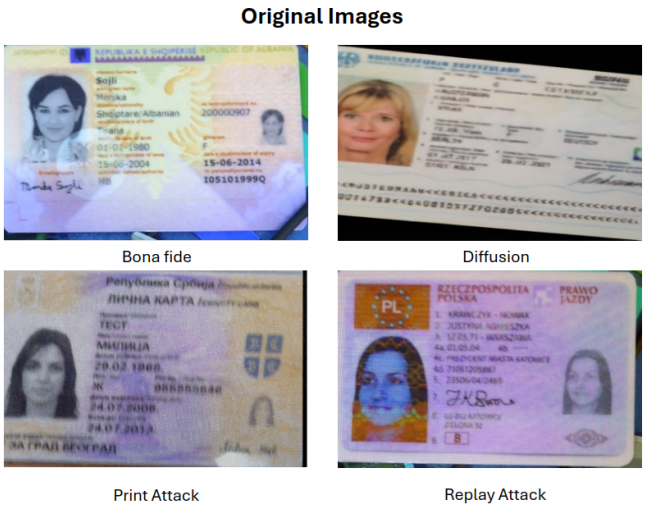

Fig. SF19: Original Images

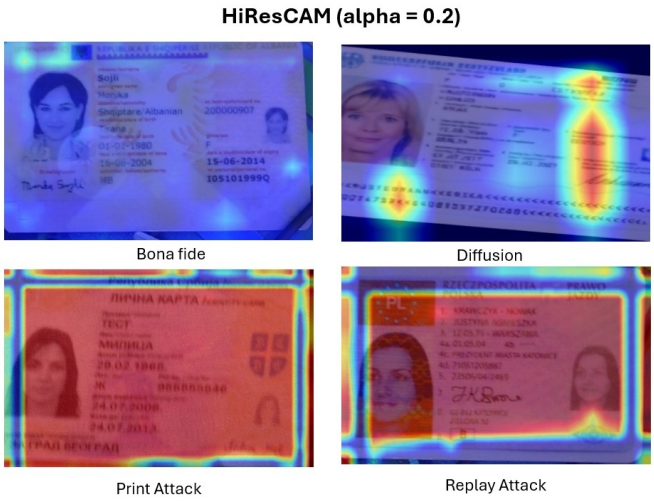

Fig. SF20: HighResCAM (α=0.2)

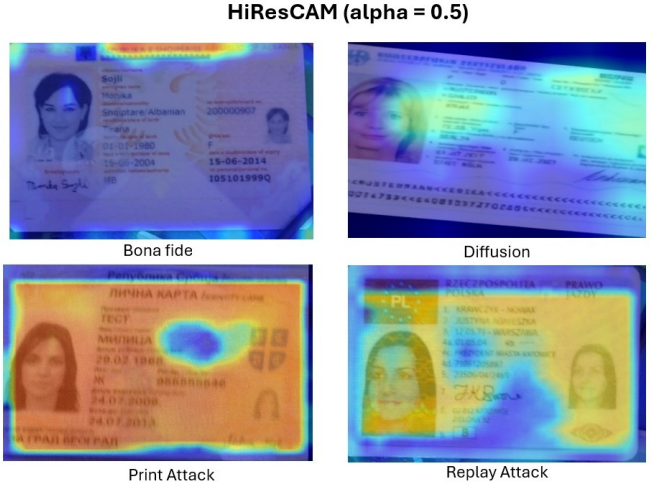

Fig. SF21: HighResCAM (α=0.5)

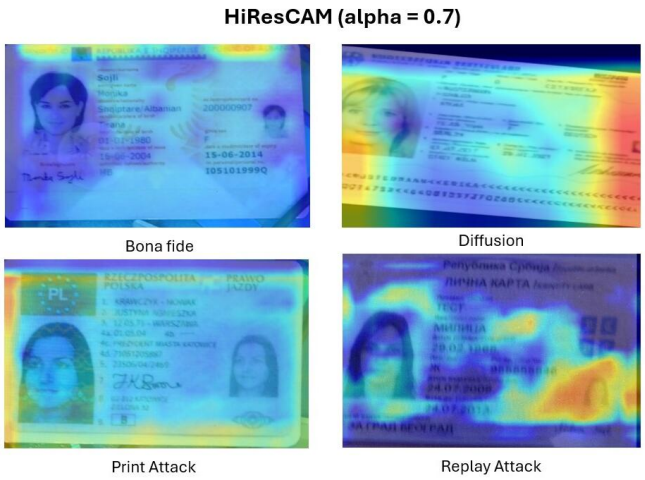

Fig. SF22: HighResCAM (α=0.7)

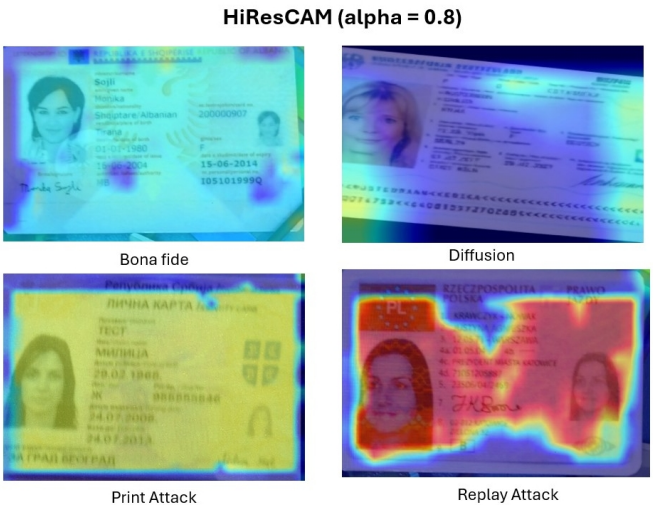

Fig. SF23: HighResCAM (α=0.8)

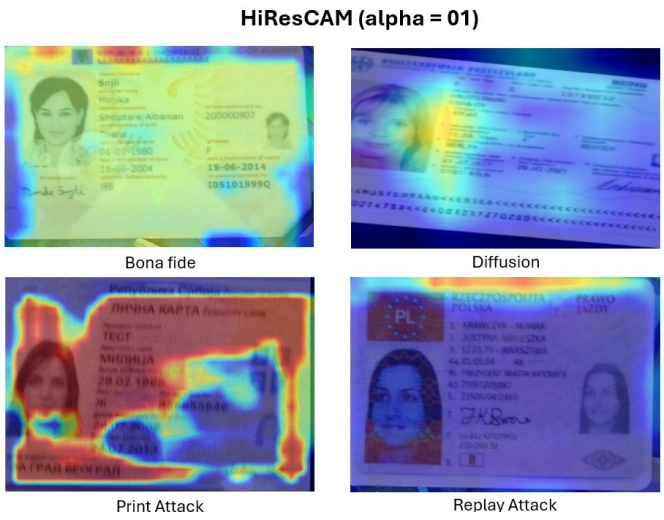

Fig. SF24: HighResCAM (α=1)

REFERENCES

[1] S. González and J. Tapia, "Towards refining id cards presentation attack detection systems using face quality index," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1027–1031, IEEE, 2022.

[2] R. Mudgalgundurao, P. Schuch, K. Raja, R. Ramachandra, and N. Damer, "Pixel-wise supervision for presentation attack detection on identity document cards," *IET Biometrics*, vol. 11, no. 5, pp. 383–395, 2022.

[3] S. Gonzalez and J. Tapia, "Improving presentation attack detection for id cards on remote verification systems," *arXiv preprint arXiv:2301.09542*, 2023.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[5] R. L. Draelos and L. Carin, "Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks," *arXiv preprint arXiv:2011.08891*, 2020.

[6] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

[7] S. Gonzalez and J. E. Tapia, "Forged presentation attack detection for id cards on remote verification systems," *Pattern Recognition*, p. 111352, 2025.