On the Robustness of Iris Presentation Attack Detectors

Aditya Sneh
adityas19@iiserb.ac.in
Akshay Agarwal
akagarwal@iiserb.ac.in

Trustworthy BiometraVision Lab,
IISER Bhopal, India

Abstract

Iris recognition is a predominantly used medium for person authentication; however, it is severely impacted by several presentation attacks. One such stealthy presentation attack is the use of contact lenses, but the relief is that several research efforts ensure that these contact lenses can be effectively detected. The prime worry is that it is observed that deep learning-based algorithms are susceptible to image corruption, and iris presentation attack detection (IPAD) algorithms are not adequately evaluated against image corruption. *Therefore, through this research for the first time, we comprehensively investigate the robustness of IPAD algorithms using several novel input corruptions*. The extensive experiment performed using multiple datasets reveals the sensitivity of state-of-the-art IPAD algorithms and demands the development of a robust algorithm. It is demonstrated that blind corruption that does not involve a classifier network, while its learning and perturbation of a few critical pixels can fool multiple IPAD algorithms. For example, on the LivDet-17 dataset, the proposed input corruption increases the equal error rate of ViT from 7.78% to 58.36%.

1 Introduction

Contact lens presentation attack instrument (PAI) is one of the most complex attacks to deceive the iris recognition algorithms [34, 59]. Several IPAD algorithms are proposed in the literature to tackle it, where the current defenses are heavily biased toward deep neural networks [35, 36]. Surprisingly, it is seen that deep neural networks are susceptible to input modification [37]. Therefore, we assert that developed deep iris presentation attack detection algorithms must also be adequately evaluated to check whether these algorithms are also sensitive to input corruptions [38]. The prime reason for such evaluation is that iris systems are highly deployed for security-related applications, including border control and verification of users on restricted devices with online transaction control as well [65]. Therefore, one can safely assume that deploying non-robust defense algorithms can be dangerous, as it can lead to the illegal entry of an intruder into the biometrics system, which is now used for mobile unlocking, border access, and digital payments. Therefore, we have carefully evaluated the sensitivity of state-of-the-art (SOTA) deep iris presentation attack detection (IPAD)

algorithms to advance the research in this critical direction and develop robust defense algorithms. Our investigation extends beyond conventional image manipulation, exploring adversarial attacks at multiple levels, including modifying texture and raw pixels using gradient and attention mechanisms. Recognizing the broader implications of compromised IPAD algorithms, we aim to contribute to developing robust defenses against emerging presentation attack methodologies. In brief, the contributions of this research are multifold: (i) Several novel input level corruptions are proposed which does not utilize the network information in learning its noise structure; (ii) An extensive experimental evaluation has been performed to evaluate the robustness of deep IPAD algorithms; and (iii) Comprehensive comparison with existing corruption and adversarial attacks demonstrates the effectiveness of the proposed attacks.

2 Related Work

Since the problem of iris presentation attacks is well known and has existed for more than a decade, several defense algorithms have been developed, ranging from using handcrafted features to current deep learning architectures. The prominent handcrafted and/or combination with deep learning features codebook [53], Haralick texture [54], local binary pattern [56], spatial pyramidal matching [57]. The handcrafted features are simple to compute but have a limited capacity and generalizability. The recent success of deep learning architectures, including convolutional networks and transformer architectures, has seen a tremendous jump in their utilization for IPAD [59, 60, 61]. Later, several research efforts have started utilizing the capacity of deep learning to develop an effective and generalized architecture. Agarwal et al. [58] proposed a Siamese architecture by combining the original and enhanced images to capture the contact lenses' texture features effectively. Agarwal et al. [60] proposed the deep contraction expansion network to improve the IPAD performance on multiple datasets. Recently, Jaswal et al. [62] proposed a network to learn global and local iris features through feature calibration, convolution, and residual learning.

While deep learning architectures have proven effective in detecting iris presentation attacks, their susceptibility to corruption and adversarial noise remains a significant issue [37, 38]. The vulnerability of deep IPAD algorithms to adversarial perturbations and common image corruption is still underexplored. Soleymani et al. [63] developed an adversarial perturbation to deceive iris recognition systems, yet its impact on presentation attack detection remains unclear. Recently, Sharma et al. examined weight perturbation in IPAD algorithms, though input perturbations were not addressed. Jain et al. [64] proposed a feature alteration approach to enhance IPAD robustness, though it requires deep network access and lacks generalizability to unseen networks. While the generalizability issue in IPAD has received attention [70, 71], unlike IPAD, vulnerability studies [69] of PAD classifiers in face and fingerprint modalities are more advanced.

We propose several input corruption algorithms to address gaps in IPAD vulnerability research, leveraging local and global texture features relevant to contact lens detection. Unlike weight perturbations, our attacks are classifier-agnostic and require no classifier-specific information, enabling broader application.

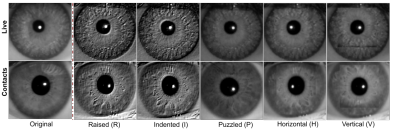

Figure 1: Showcasing the impact of different texture embossing alterations proposed in this research to fool the IPAD algorithms.

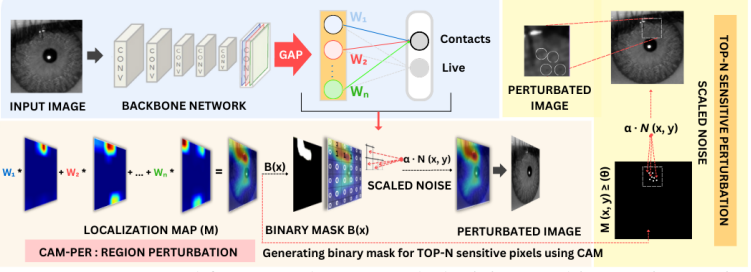

Figure 2: CAM-PER: A novel framework to perturb decision-making regions using Gaussian noise from the region level to top-N sensitive pixels. Perturbations are guided by Class Activation Maps (CAM) through Global Average Pooling and binary mask generation.

3 Proposed Iris Perturbations

The iris is rich in textural features, which are practical for iris recognition and for detecting presentation attacks. Therefore, one primary feature we have explored for generating iris corruption is the perturbation of textural features. In this section, we first defined the perturbation generated by modifying texture regions. It is observed that the texture attacks modify an entire area of the iris that the human examiners might notice. Henceforth, to increase the stealthy nature and make the attack imperceptible, we identify the critical region by using the attention mechanism and modifying either the highlighted region or a few essential pixels to generate an adversarial iris image.

3.1 Unlearnable Perturbations: Texture Manipulation

Texture embossing [14] creates a three-dimensional effect and exploits the system's sensitivity to textural details by mimicking natural iris variations. It involves creating a 3D effect by highlighting the textured surface's variations in intensity or color. It can be achieved by stimulating how light and shadow interact with the surface, creating a sense of depth. To alter the iris, different forms of texture embossing methods have been explored and are outlined as follows: (i) **Raised (R):** Following the detection of the iris and exclusion of the pupil region, an embossing kernel is applied to emphasize changes in intensity and simulate the impact of light and shadow on a three-dimensional surface. The convolution process enhances edges, producing a 3D-like raised texture effect on the iris. (ii) **Indented (I):** The kernel applied for this introduces an indentation effect by emphasizing specific pixel values based on their neighborhood relationships. Positive coefficients contribute to a brightening effect, while negative coefficients contribute to a darkening effect. The resulting indented iris image exhibits alterations in texture, with regions appearing to be pushed inward or indented, creating a unique visual impact. (iii) **Adaptive Embossing Kernel:** To find the optimal embossing kernel for achieving a desired texture effect while optimizing the structural similarity index measure (SSIM), an iterative optimization process is employed. This process involves ad-

Algorithm 1 CAM-PER: CAM Guided Perturbation and TOP-N Sensitive Pixels

Require: Original image I, CAM model *cam*, threshold value θ, top N value = top_N
Apply CAM to generate color map M: $M = \sum_k w_k \cdot A_k$
Threshold M to create binary mask binary_mask using θ:
$$\text{binary\_mask}(x,y) = \begin{cases} 1, & \text{if } M(x,y) \geq \theta \\ 0, & \text{otherwise} \end{cases}$$
Generate Gaussian noise for each pixel: $N(x,y)$
Scale the noise: scaled_noise $= N(x,y) \times \alpha$
Initialize noisy image noisy_image as a copy of I
for each pixel (x,y) in I **do**
    if binary_mask(x,y) is 1 **then**
        Add scaled_noise(x,y) to noisy_image(x,y): noisy_image$(x,y) = I(x,y) +$ scaled_noise(x,y)
    end if
end for
return noisy_image
— TOP-N Sensitive Region Selection —
Sort CAM values in descending order: M_{sorted}
Select top N sensitive regions: top_N_indices $=$ argsort$(M_{\text{sorted}})[:N]$
return binary_mask

justing the kernel's coefficients and evaluating the impact on SSIM between the embossed and the original image. Using strategies like grid search, the goal is to balance the enhancement of visual texture with structural fidelity to the original. The outcome is a finely tuned kernel that effectively meets both the quantitative SSIM standards and the qualitative aesthetic requirements of the embossing task. (iv) **Puzzle (P):** The puzzling effect is introduced by dividing the iris into small pieces, which are then randomly shuffled, generating a disordered arrangement of pixel values. The final step combines the original iris image with the puzzle effect, utilizing bitwise operations to preserve the pupil region and background. (v) **Geometric Variator (H & V):** Since the iris textures are random patterns, it is difficult for a human to identify a correct arrangement of an entire iris image. Therefore, it gives the flexibility to apply Horizontal (H) and Vertical (V) flips to this iris region without being noticed. Once the flipping operations are executed, the altered iris region is composited back onto the original iris image using the pupil center and radius. Gaussian smoothing is applied to the entire iris image to enhance the visual coherence of the altered image, aiming to reduce noise and create a more visually appealing result. Figure 1 shows the impact of the different texture embossing alterations performed to generate attack images. It can be seen that due to effective pre- and post-processing, the proposed alterations are imperceptible. On top of that, the proposed alterations are based on essential properties of iris images and do not involve any network to find the external perturbations.

3.2 Proposed Learnable Perturbations

The class activation map (CAM) is one of the strongest mediums for identifying the critical region in an image, based on which a deep network makes a particular decision. Inspired by this understanding, to further boost the strength of the proposed texture manipulation attacks, we first identify the sensitive regions in an iris image, and later these regions are modified using Gaussian noise. The CAM is generated using an unseen network; in our case, we have used the ResNet-50 architecture [19]. First, a coarse localization map (M) from an input image (I) is generated to highlight significant areas for classification, followed by applying Global Average Pooling (GAP) to feature maps (A_K) for targeted noise addition. These are

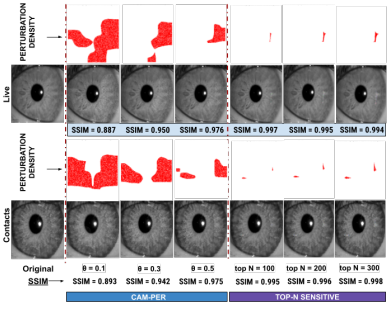

Figure 3: Comparison of perturbed iris generated using the proposed CAM-PER and TOP-N sensitive methods alongside its perturbation density.

expressed as :

$$\text{GAP}(A_k) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} A_k(i,j), \quad M = \sum_k w_k \cdot A_k$$

Feature maps (A_k) are pulled from the last layer of the model, capturing essential details from the input image. The Class Activation Mapping (CAM) uses these maps, considering their height (H) and width (W), and applies weights (w_k) to highlight crucial areas for classification. By weighting and summing these maps ($w_k \cdot A_k$), CAM pinpoints the most critical regions for the model's decisions. The binary mask extraction step is crucial in selectively perturbing sensitive regions highlighted by CAM. Pixels with CAM values surpassing a predefined threshold ($\theta = 0.03$) are labeled as sensitive (1), while those below the threshold are considered non-sensitive (0). Mathematically, the binary mask is derived as:

$$B(x)_{i,j} = \begin{cases} 1, & \text{if } M(x)_{i,j} \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

Gaussian noise (N(x,y)) is generated for each pixel location in the input image. The noise is scaled by a factor $(\alpha) = 0.001$ based on the sensitivity of each pixel in the binary mask. Noise is only added to pixels labeled as sensitive in the binary mask. Mathematically, the perturbed image is formulated as:

$$I'(x,y) = \begin{cases} I(x,y) + \alpha \cdot N(x,y), & \text{if } B(x,y) = 1 \\ I(x,y), & \text{otherwise} \end{cases}$$

where (x,y) denotes pixel coordinates, M(x,y) represents the CAM value at pixel (x,y), α controls the magnitude of perturbation, and θ is a predefined threshold. Figure 2 shows the schematic diagram, and Algorithm 1 shows the pseudocode of the proposed attack based on the knowledge of identifying critical iris regions using CAM. This approach effectively balances preserving image integrity with targeted adjustments to improve defense against adversarial attacks.

3.2.1 TOP-N Sensitive Pixels

To further make the proposed CAM-based perturbation effective and to perturb every critical region, we have proposed an adaptive variant where we have identified the top-N pixels in the influential regions. To prioritize the top-N-sensitive regions, we rank pixels based on their CAM values and perturb only the top-sensitive pixels. Figure 3 compares the proposed

CAM-PER and TOP-N sensitive methods, analyzing perturbation density and SSIM scores on original contact and live iris images. Then, the perturbed image I'(x,y) is formulated as follows:

$$I'(x,y) = \begin{cases} I(x,y) + \alpha \cdot \mathcal{N}(x,y), & \text{if pixel} \in \text{top-N} \\ I(x,y) & \text{otherwise} \end{cases}$$

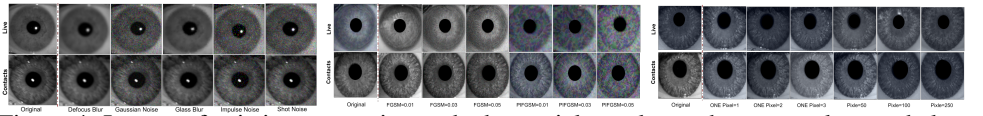

Figure 4: Impact of existing corruption and adversarial attacks on the contact lens and clean (original) images.

3.3 Existing Adversarial Attacks

3.3.1 Image Corruptions

Apart from the proposed attack, to effectively identify the sensitivity of the IPAD algorithms, we have utilized the existing image corruptions [38] and adversarial perturbation algorithms [37]. We have used several image noises, which have recently been explored to benchmark the robustness of deep neural networks, namely Gaussian Noise, Shot Noise, Impulse Noise, Defocus Blur, and Glass Blur. Gaussian, Impulse, and Shot are grouped as NOISE; on the other hand, Defocus and Glass are grouped as BLUR.

3.3.2 Adversarial Perturbations

Adversarial perturbations are imperceptible noises learned from the network information, such as the gradient or parameter of the network, to fool the DNNs. In this research, we have used state-of-the-art (SOTA) adversarial perturbation algorithms, namely Fast Gradient Sign Method (FGSM) [10], Patch-wise Iterative Fast Gradient Sign Method (PIFGSM) [20], Pixle [21], and one-pixel attack [23]. The FGSM and PIFGSM utilize the network's gradient information to manipulate the iris images; whereas, Pixle is a black-box attack that rearranges the pixel information, and a one-pixel attack restricts the modification of pixels to a few pixels, which can lead to the misclassification of an image. As shown in Figure 4, while these corruptions are natural, they can drastically alter the appearance of an image, and if applied with less severity, they are found less effective in fooling deep neural networks [38]. Further, the adversarial perturbations are minute but need access to the DNNs.

4 Experiments Results and Analysis

In this section, we first define the IPAD datasets used to generate the perturbations and evaluate the robustness of SOTA IPAD models. Later, the SOTA models used for extensive experiments are described in detail, followed by the results and analysis reflecting the possible vulnerability of each IPAD model.

4.1 IPAD Datatsets

The experiments utilize several benchmark datasets, including MUIPAD [24], LivDet-17 IIT-WVU [25], and IIITD-CLI [26]. This study focuses exclusively on images depicting

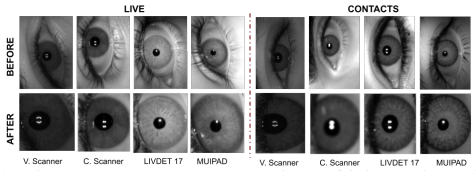

Figure 5: These examples illustrate the segmentation of iris regions before and after processing for each dataset, encompassing both Live and Contacts classes.

textured contact lenses and live iris captures. The IIITD-CLI dataset employs two scanners, namely the Cogent and Vista. The dataset distribution is as follows: IIITD-CLI (Cogent Scanner) comprises 1163 live iris and 1160 contact lens images, while IIITD-CLI (Vista Scanner) includes 1000 live iris and 1065 contact lens iris images. Additionally, LivDet-17 IIT-WVU features 1725 live iris and 1625 contact lens images, and MUIPAD contains 1466 live iris and 1681 contact lens images.

Three primary methods are utilized sequentially to segment the iris region accurately so that only the texture region of the iris gets perturbed. First, a Gaussian blur is applied to grayscale iris images to reduce noise and enhance smoothness. Later, the Hough Circle Transform [27] is employed to detect the iris by identifying circular shapes within the image, guided by parameters like the minimum and maximum radius. If the Hough Circle Transform cannot precisely localize the iris, Haar Cascades [28] are employed as an alternative method for eye detection. The examples in Figure 5 show iris region segmentation for live and contact lens classes using Vista and Cogent scanners from the IIITD CLI dataset.

4.2 IPAD Models

In this research, we have used state-of-the-art (SOTA) convolutional neural networks (CNNs) along with specific deep IPAD models: (i) D-Net PAD [32] (D-Net), (ii) Vision Transformer ViT-B/16 [33], (iii) HDA-IDVC [66], (iv) IPAD CNN [67], and (v) BUCEA Algo1 [66]. The use of CNNs in iris presentation attack detection is prominent and has shown tremendous success in identifying presentation attack instruments (PAIs) [61]. Therefore, their evaluation and another SOTA image classification model, ViT, are critical. The robustness of the models and success of the proposed and existing attacks are reported using equal error rate (EER%) and average classification error rate (ACER%)% which is the average of attack presentation error classification rate (APCER%) and bona fide (original) classification error rate (BPCER%).

4.3 Results and Analysis

Examining the equal error rate (EER%) on clean images in Table 1 across various models and datasets reveals distinct performance differences. Notably, models like D-Net PAD, ViT, and BUCEA Algo1 consistently demonstrate high clean accuracy, achieving 0.00% EER across multiple datasets such as IIITD-CLI (Cogent and Vista) and MUIPAD, indicating the effectiveness of the models in detecting contact lenses when the images are not perturbed. In contrast, models like HDA-IDVC, IPAD CNN, and BUCEA Algo1 show a broader range of EERs, reflecting their varying accuracy. For example, IPAD CNN achieves a 1.56% EER on the IIITD-CLI dataset, while HDA-IDVC EER peaks at 59.08% on LivDet-17, highlighting the importance of model selection for specific applications due to performance variability. The one question we asked is, while the CNNs show tremendous accuracy in

EER (%)															
Dataset ↓	**Model ↓**	**Proposed Methods**								**Existing Methods**					
		Clean	**H**	**V**	**I**	**P**	**R**	**CP**	**TN**	**BLUR**	**NOISE**	**FGSM**	**PIFGSM**	**Pixel**	**OnePixel**
Parameter →								$t = 0.3$	$t = 100$			$e = 0.01$	$e = 0.01$	$i = 250$	$p = 3$
IIITD-CLI (Cogent)	D-Net	0.00	4.15	5.19	1.72	0.00	0.29	3.45	0.00	14.08 ± 18.69	10.25 ± 8.72	0.00	50.57	50.00	**51.44**
	VIT	0.00	8.01	15.08	22.13	**81.03**	46.55	43.68	54.60	54.74 ± 9.95	61.40 ± 12.45	0.86	35.63	50.29	52.01
	HDA-IDVC	31.32	51.12	64.55	20.98	**53.74**	47.13	50.72	48.56	52.01 ± 1.22	47.79 ± 9.13	40.52	49.43	49.14	51.72
	IPAD CNN	2.87	11.19	17.54	9.77	4.89	9.20	30.95	9.48	11.63 ± 1.83	24.33 ± 7.85	3.16	**49.14**	47.99	47.13
	BUCEA Algo1	0.00	0.00	1.87	7.47	1.72	46.55	**50.43**	47.99	6.61 ± 2.43	23.75 ± 11.87	0.00	50.29	49.14	47.13
IIITD-CLI (Vista)	D-Net	0.00	7.28	8.75	0.00	1.80	0.31	52.81	0.00	**55.16 ± 4.63**	43.33 ± 7.41	0.00	50.62	46.88	53.20
	VIT	0.00	17.50	11.00	51.56	45.00	53.75	53.12	53.12	**56.72 ± 18.78**	48.02 ± 10.15	0.00	47.50	47.19	50.31
	HDA-IDVC	59.38	54.92	50.82	42.81	60.31	49.06	51.25	50.94	45.00 ± 3.97	55.53 ± 5.30	51.61	48.39	27.50	**60.94**
	IPAD CNN	1.56	18.85	16.39	0.00	2.81	10.94	35.94	11.25	17.34 ± 5.52	0.21 ± 0.36	4.06	49.69	48.75	**50.00**
	BUCEA Algo1	0.00	32.79	36.89	2.50	0.63	14.06	54.37	43.13	**56.72 ± 5.97**	44.06 ± 3.25	0.31	52.19	47.50	45.31
LivDet-17	D-Net	5.19	6.31	7.34	5.76	4.61	6.63	0.00	4.90	7.42 ± 0.50	11.41 ± 9.63	6.92	50.72	48.66	**51.44**
	VIT	7.78	6.23	7.32	15.56	38.90	34.73	**58.36**	37.03	37.60 ± 0.20	38.38 ± 1.71	5.33	48.27	48.87	50.58
	HDA-IDVC	59.08	**60.13**	52.29	56.34	39.48	54.32	53.88	25.22	53.74 ± 1.63	54.27 ± 20.43	51.15	48.27	51.59	56.92
	IPAD CNN	15.13	20.70	20.92	7.78	17.00	13.98	20.75	18.16	19.81 ± 1.73	24.30 ± 0.80	14.84	49.71	49.14	**50.58**
	BUCEA Algo1	5.04	10.46	10.89	10.23	18.59	41.93	**100.00**	47.69	9.58 ± 0.30	40.05 ± 11.43	8.21	49.14	50.14	48.70
MUIPAD	D-Net	0.00	11.26	12.90	5.35	2.38	4.36	48.12	0.00	3.36 ± 3.35	27.53 ± 11.46	0.40	**51.88**	51.49	47.52
	VIT	0.00	23.73	12.11	42.18	39.01	39.41	43.17	42.57	40.10 ± 4.62	35.71 ± 6.70	0.59	**50.30**	50.10	49.11
	HDA-IDVC	38.81	58.28	53.50	15.25	45.15	52.48	49.31	60.83	55.84 ± 4.76	52.80 ± 4.54	49.50	52.08	45.74	**63.37**
	IPAD CNN	2.77	8.28	13.38	6.73	4.36	3.76	13.07	6.53	7.82 ± 3.50	11.15 ± 9.32	3.17	**52.87**	50.69	47.92
	BUCEA Algo1	0.00	0.64	0.96	**63.96**	30.30	35.64	26.53	47.72	4.65 ± 3.78	51.55 ± 11.20	3.17	50.69	51.29	49.11

Table 1: A detailed EER(%) comparison between proposed attacks such as Horizontal (H), Vertical (V), Indented (I), Puzzle (P), Raised (R), CAM-PER (CP), and TOP-N (TN) to existing attacks like BLUR, NOISE, FGSM, PIFGSM, Pixle, and OnePixel.

detecting contact lens attacks, is it their actual effectiveness? For that, we have applied several proposed and existing perturbations to both the clean and contact lens iris images of each dataset. It is observed that the proposed region-sensitive induced perturbation, namely CAM-PER (CP), drastically reduces the performance of each model on each dataset. For example, it is observed that the BUCEA Algo1 yields a significant jump in EER from 5.04% to 100.00% on the LivDet-17 dataset. This vulnerability suggests iris textures may be more prone to manipulations impacting feature clarity. Similarly, on the IIIT-CLI (Cogent) dataset, the ViT model's EER spikes to 81.03% under CAM-PER (CP), highlighting the potential model-agnostic capacity of the proposed attack, which has been trained on an unseen model (ResNet-50). Interestingly, the IPAD algorithms are found to be less susceptible to flip attacks (both horizontal and vertical) than the texture embossing attacks. It highlights that even if the features are flipped without modification, they can be effectively used to detect presentation attacks. In contrast, D-Net shows exceptional resilience, maintaining a 0.00% EER under several attacks, including a minimal rise to 3.45% under CAM-PER on Cogent, suggesting inherent robustness or limitations in attack effectiveness against it. CAM-PER (CP) and TOP-N (TN) attacks notably degrade model performance, exemplified by CAM-PER's increase in EER to 100.00% for BUCEA Algo1 on LivDet-17 and 58.36% for ViT on LivDet-17, underlining the urgent need for effective defense mechanisms. Among existing methods, NOISE proves particularly potent, elevating ViT (EER to 61.40% on IIITD-CLI), surpassing the disruption caused by BLUR. **Apart from the proposed attack, the PIFGSM attack is highly stealthy and increases the EER by a significant margin**. It is to be noted here that the attack drastically distorts the image quality and hence is a primary factor for its high attack success.

4.4 Comparison with existing perturbations

The box plot shown in Figure 6 (left) shows the equal error rates (EER%) for proposed and existing methods, revealing key insights into their performance against various adversarial attacks. IPAD algorithms show resilience against flip attacks but are notably vulnerable to CAM-PER (CP) and TOP-N (TN) attacks, highlighted by high median EER values. Proposed Indented (I) and Puzzle (P) attacks outperform existing attack methods in effec-

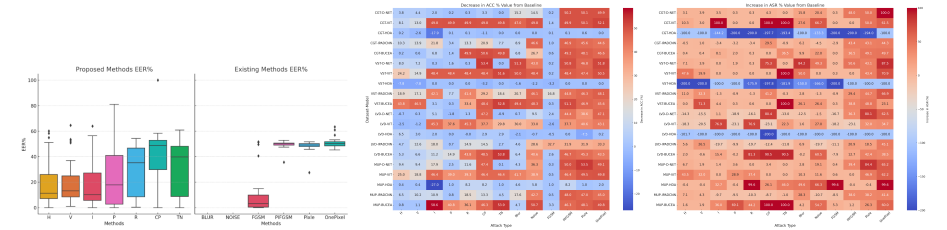

Figure 6: (left) Box-Plot analysis of EER% for proposed and existing attacks. (right) The detailed heatmap analysis of ASR (%) increment and ACC(%) decrement relative to baseline (clean ASR (%) & ACC (%)). The datasets Cogent, Vista, LivDet-17, and MUIPAD are abbreviated as CGT, VST, LVD, and MUP.

tiveness, achieving higher median EERs, and display wider EER spreads in the box plot analysis. Similarly, the IPAD models are significantly vulnerable to learning-based adversarial perturbations, including OnePixel attacks. The proposed Top-N (TN) attack perfectly misled the ViT model when used to classify presentation attack and clean images, and all the images of both these classes are swapped into each other, leading to a perfect ACER value of 100%. Conversely, the Puzzle (P) attack's significant impact, reaching a 50% ACER on the ViT model, shows that even if the model bifurcates the images into patches, their puzzled nature in the original image can mislead the classifier. In contrast, existing attacks such as PIFGSM showcase their robustness and consistently high impact across various models and datasets. For instance, the PIFGSM attack manifests a potent disruption with an ACER reaching up to 70.53% on the D-Net PAD (D-Net) model in the MUIPAD dataset, underscoring the enduring effectiveness of gradient-based adversarial methods. However, it is noted that the PIFGSM attack drastically reduces the visual quality of iris images compared to the proposed attacks. In contrast, the D-NET model showcases resilience with a 0.00% ACER against multiple attacks on the IIITD-CLI dataset, though its ACER spikes to 52.65% under the proposed CAM-PER (CM) attack on the Vista dataset.

Figure 6 (right) shows the high attack success rate (ASR) of each of the proposed and existing attacks, highlighting the critical vulnerability of each deep IPAD algorithm. Starting with the heatmap analysis (increase in ASR from clean baseline), we observe a significant rise in ASR for specific models and attack combinations, indicating a high susceptibility to those particular attacks. For example, Cogent-ViT (CGT_VIT), Vista-ViT (VST_VIT), and MUP-BUCEA show a 100.0% increase in ASR when subjected to the CAMPER (CP) attack, implying a complete breakdown against the attack. On the other hand, some models are relatively robust against specific attacks. For instance, Cogent HDA-IDVC (CGT_HDA) only shows a 0.00% increase in ASR when faced with the Horizontal ('H') attack, indicating better resilience. *'We assert that such a missing understanding of the vulnerability of IPAD algorithms against a wide range of corruptions and adversarial perturbation can lead to a false sense of security and might limit the development of a robust solution'*. The proposed study can help understand the limitations of current deep IPAD algorithms and aim to bring robustness in its mechanism [64].

5 Conclusion

Iris recognition is one of the most popular and effective mediums for secure access to sophisticated mediums like mobile phones, borders, and digital payments. However, it can be

hacked by simply wearing contact lenses. Several iris presentation attack detection (IPAD) algorithms have been developed to secure such systems, but their resiliency against input perturbations is not adequately addressed. Henceforth, we comprehensively evaluated the vulnerability of several IPAD networks through extensive experiments. Our findings reveal that while some models exhibit foundational robustness against input corruptions, no model is wholly impervious to nuanced and sophisticated attacks. This vulnerability underscores the pressing need for developing resilient defenses to counter adversarial tactics. In the future, we aim to build an accurate yet robust iris presentation algorithm to handle a variety of adversarial and noise perturbations.

References

[1] Jung, Hyungsik, and Youngrock Oh. "Towards better explanations of class activation mapping." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[2] Yambay, David, et al. "LivDet iris 2017—Iris liveness detection competition 2017." 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017.

[3] Spinoulas, Leonidas, et al. "Multispectral biometrics system framework: Application to presentation attack detection." IEEE Sensors Journal 21.13 (2021): 15022-15041.

[4] Bowyer, Kevin W., and Patrick J. Flynn. "The ND-IRIS-0405 iris image dataset." arXiv preprint arXiv:1606.04853 (2016).

[5] He, Z., Tan, T., Sun, Z., & Qiu, X. (2019). IrisConvNet: Deep learning for iris recognition. IEEE International Joint Conference on Biometrics (IJCB), 1-8.

[6] Zuo, J., Schmid, N. A., & Bowyer, K. W. (2020). Multispectral iris recognition: A preliminary study. IEEE Transactions on Information Forensics and Security, 15, 1268-1282.

[7] Zhuo, Wenqi, et al. "Irisguard: image forgery detection for iris anti-spoofing." Chinese Conference on Biometric Recognition. Cham: Springer Nature Switzerland, 2022.

[8] Gangwar, Abhishek & Joshi, Akanksha. (2016). DeepIrisNet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. 10.1109/ICIP.2016.7532769.

[9] Zola, Francesco, et al. "Verification system based on long-range iris and Graph Siamese Neural Networks." Proceedings of the 2022 European Symposium on Software Engineering. 2022.

[10] Dillak, Rocky Yefrenes, and Martini Ganantowe Bintiri. "A novel approach for iris recognition." 2016 IEEE Region 10 Symposium (TENSYMP). IEEE, 2016.

[11] Xu, Jin, Zhendong Cai, and Wei Shen. "Using FGSM targeted attack to improve the transferability of adversarial example."international conference on electronics and communication engineering (ICECE) 2019.

[12] Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

[13] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). Ieee, 2017.

[14] Combey, Théo, et al. "Probabilistic jacobian-based saliency maps attacks." Machine learning and knowledge extraction 2.4 (2020): 558-578.

[15] Kim, J., Kwang-Eui, L., & Kwon, K. (2010). A New Embossing Method for Color Images. International Journal of Computer Science and Network Security (IJCSNS), 10(2), 144.

[16] Hore, Alain, and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM." 2010 20th international conference on pattern recognition. IEEE, 2010.

[17] Liashchynskyi, Petro, and Pavlo Liashchynskyi. "Grid search, random search, genetic algorithm: a big comparison for NAS." arXiv preprint arXiv:1912.06059 (2019).

[18] Feissel, Martine, and Wlodzimierz Lewandowski. "A comparative analysis of Vondrak and Gaussian smoothing techniques." Bulletin géodésique 58 (1984): 464-474.

[19] Siddiqui, Shoaib Ahmed, and Thomas Breuel. "Identifying Layers Susceptible to Adversarial Attacks." arXiv preprint arXiv:2107.04827 (2021).

[20] Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).

[21] Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).

[22] Pomponi, Jary, et al. "Rearranging Pixels is a Powerful Black-Box Attack for RGB and Infrared Deep Learning Models." IEEE Access 11 (2023): 11298-11306.

[23] Pomponi, Jary, Simone Scardapane, and Aurelio Uncini. "Pixle: a fast and effective black-box attack based on rearranging pixels." 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022.

[24] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5), 828-841.

[25] Yadav, Daksha, et al. "Iris presentation attack via textured contact lens in unconstrained environment." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

[26] Yambay, David, et al. "LivDet iris 2017—Iris liveness detection competition 2017." 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017.

[27] Yadav, Daksha, et al. "Unraveling the effect of textured contact lenses on iris recognition." IEEE Transactions on Information Forensics and Security 9.5 (2014): 851-862.

[28] Hassanein, Allam Shehata, et al. "A survey on Hough transform, theory, techniques and applications." arXiv preprint arXiv:1502.02160 (2015).

[29] Yustiawati, Ratna, et al. "Analyzing of different features using Haar cascade classifier." 2018 International Conference on Electrical Engineering and Computer Science (ICECOS). IEEE, 2018.

[30] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[31] Sani Zulkarnaen, Aulia Chusnyriani, et al. "Application of Convolutional Neural Network Method with MobileNet V1 and ResNet-152 V2 Architecture in Batik Motif Classification." International Conference on Broadband and Wireless Computing, Communication and Applications. Cham: Springer Nature Switzerland, 2023.

[32] Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[33] Sharma, Renu, and Arun Ross. "D-netpad: An explainable and interpretable iris presentation attack detector." 2020 IEEE international joint conference on biometrics (IJCB). IEEE, 2020.

[34] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.

[35] Boyd A, Fang Z, Czajka A, Bowyer KW. Iris presentation attack detection: Where are we now?. Pattern Recognition Letters. 2020 Oct 1;138:483-9.

[36] Safaa El-Din Y, Moustafa MN, Mahdi H. Deep convolutional neural networks for face and iris presentation attack detection: Survey and case study. IET Biometrics. 2020 Sep;9(5):179-93.

[37] Pooshideh, Matineh, et al. "Presentation Attack Detection: A Systematic Literature Review." ACM Computing Surveys 57.1 (2024): 1-32.

[38] Han S, Lin C, Shen C, Wang Q, Guan X. Interpreting adversarial examples in deep learning: A review. ACM Computing Surveys. 2023 Jul 17;55(14s):1-38.

[39] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261. 2019 Mar 28.

[40] Yadav, Daksha, et al. "Unraveling the effect of textured contact lenses on iris recognition." IEEE Transactions on Information Forensics and Security 9.5 (2014): 851-862.

[41] El-Naggar, Susan, and Arun Ross. "Which dataset is this iris image from?." 2015 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2015.

[42] Suvarchala, P. V. L., and S. Srinivas Kumar. "Feature Set Fusion for Spoof Iris Detection." Engineering, Technology & Applied Science Research 8.2 (2018).

[43] Sequeira, Ana F., et al. "Mobilive 2014-mobile iris liveness detection competition." IEEE international joint conference on biometrics. IEEE, 2014.

[44] Czajka, Adam. "Database of iris printouts and its application: Development of liveness detection method for iris recognition." 2013 18th International Conference on Methods & Models in Automation & Robotics (MMAR). IEEE, 2013.

[45] Sequeira, Ana, et al. "Cross-eyed-cross-spectral iris/periocular recognition database and competition." 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2016.

[46] Boyd, Aidan, et al. "State Of The Art In Open-Set Iris Presentation Attack Detection." arXiv preprint arXiv:2208.10564 (2022).

[47] Boyd, Aidan, et al. "Iris presentation attack detection: Where are we now?." Pattern Recognition Letters 138 (2020): 483-489.

[48] Li, Chengcheng, Weidong Zhou, and Shasha Yuan. "Iris recognition based on a novel variation of local binary pattern." the visual computer 31 (2015): 1419-1429.

[49] Kulkarni, Shrinivasrao B., et al. "GLCM-based multiclass iris recognition using FKNN and KNN." International Journal of Image and Graphics 14.03 (2014): 1450010.

[50] Jenadeleh, M.; Pedersen, M.; Saupe, D. Blind Quality Assessment of Iris Images Acquired in Visible Light for Biometric Recognition. Sensors 2020, 20, 1308. https://doi.org/10.3390/s20051308

[51] Gupta, Priyanshu, et al. "On iris spoofing using print attack." 2014 22nd international conference on pattern recognition. IEEE, 2014.

[52] Morales, Aythami, et al. "Introduction to Presentation Attack Detection in Iris Biometrics and Recent Advances." arXiv preprint arXiv:2111.12465 (2021).

[53] Czajka, Adam, and Andrzej Pacut. "Replay attack prevention for iris biometrics." 2008 42nd Annual IEEE International Carnahan Conference on Security Technology. IEEE, 2008.

[54] Gomez-Barrero, Marta, Javier Galbally, and Julian Fierrez. "Efficient software attack to multimodal biometric systems and its application to face and iris fusion." Pattern Recognition Letters 36 (2014): 243-253.

[55] Löfstedt, Tommy, et al. "Gray-level invariant Haralick texture features." PloS one 14.2 (2019): e0212110.

[56] Sun, Zhenan, et al. "Iris image classification based on hierarchical visual codebook." IEEE Transactions on pattern analysis and machine intelligence 36.6 (2013): 1120-1133.

[57] Z. He et al., “Efficient Iris Spoof Detection via Boosted Local Binary Patterns,” in Proc. International Conference on Biometrics (ICB), 2009, pp. 1080–1090.

[58] Y. Hu, K. Sirlantzis, and G. Howells, “Iris Liveness Detection using Regional Features,” Pattern Recognition Letters, vol. 82, pp. 242–250, 2016.

[59] Agarwal A, Noore A, Vatsa M, Singh R. Generalized contact lens iris presentation attack detection. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2022 May 24;4(3):373-85.

[60] Agarwal A, Noore A, Vatsa M, Singh R. Enhanced iris presentation attack detection via contraction-expansion CNN. Pattern Recognition Letters. 2022 Jul 1;159:61-9.

[61] C. Chen and A. Ross, “An Explainable Attention-Guided Iris Presentation Attack Detector,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 97–106.

[62] M. Choudhary, V. Tiwari, and U. Venkanna, “Iris Anti-Spoofing through Score-Level Fusion of Handcrafted and Data-Driven Features,” Applied Soft Computing, vol. 91, p. 106206, 2020.

[63] M. Fang et al., “Deep Learning Multi-Layer Fusion for an Accurate Iris Presentation Attack Detection,” in Proc. International Conference on Information Fusion (FUSION), 2020, pp. 1–8.

[64] Jaswal G, Verma A, Roy SD, Ramachandra R. Learning Joint Local-Global Iris Representations via Spatial Calibration for Generalized Presentation Attack Detection. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2024 Jan 17.

[65] Soleymani, Sobhan, et al. "Adversarial examples to fool iris recognition systems." 2019 International Conference on Biometrics (ICB). IEEE, 2019.

[66] Wu, Dongxian, Shu-Tao Xia, and Yisen Wang. "Adversarial weight perturbation helps robust generalization." Advances in neural information processing systems 33 (2020): 2958-2969.

[67] Agarwal, Akshay, et al. "Deceiving face presentation attack detection via image transforms." 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, 2019.

[68] Fei J, Xia Z, Yu P, Xiao F. Adversarial attacks on fingerprint liveness detection. EURASIP Journal on Image and Video Processing. 2020 Jan 13;2020(1):1.

[69] V. Jain, A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Robust IRIS Presentation Attack Detection Through Stochastic Filter Noise," International Conference on Pattern Recognition (ICPR), 2022, pp. 1134-1140

[70] Daksha Yadav, Naman Kohli, Akshay Agarwal, Mayank Vatsa, Richa Singh, Afzel Noore; "Fusion of Handcrafted and Deep Learning Features for Large-Scale Multiple Iris Presentation Attack Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 572-579

[71] M. Gupta, V. Singh, A. Agarwal, M. Vatsa, and R. Singh, "Generalized Iris Presentation Attack Detection Algorithm under Cross-Database Settings," International Conference on Pattern Recognition (ICPR), 2021, pp. 5318-5325