An Explainable Attention-Guided Iris Presentation Attack Detector

Cunjian Chen
Michigan State University
cunjian@msu.edu

Arun Ross
Michigan State University
rossarun@msu.edu

Abstract

Convolutional Neural Networks (CNNs) are being increasingly used to address the problem of iris presentation attack detection. In this work, we propose an explainable attention-guided iris presentation attack detector (AG-PAD) to augment CNNs with attention mechanisms and to provide visual explanations of model predictions. Two types of attention modules are independently placed on top of the last convolutional layer of the backbone network. Specifically, the channel attention module is used to model the inter-channel relationship between features, while the position attention module is used to model inter-spatial relationship between features. An element-wise sum is employed to fuse these two attention modules. Further, a novel hierarchical attention mechanism is introduced. Experiments involving both a JHU-APL proprietary dataset and the benchmark LivDet-Iris-2017 dataset suggest that the proposed method achieves promising detection results while explaining occurrences of salient regions for discriminative feature learning. To the best of our knowledge, this is the first work that exploits the use of attention mechanisms in iris presentation attack detection.

1. Introduction

Iris recognition systems are vulnerable to various types of presentation attacks (PAs), where an adversary presents a fabricated artifact or an altered biometric trait to the iris sensor in order to circumvent the system [5]. Commonly discussed attacks include cosmetic contacts, printed eyes, and artificial eye models [3]. To address these challenges, convolutional neural networks (CNNs) are being increasingly used for addressing the problem of iris presentation attack detection (PAD) [10, 19, 31, 30, 17, 14, 22, 8], which is often formulated as a binary-classification task. The output of the network is a presentation attack (PA) score indicating whether the input image should be classified as "live" or "PA".[1] Existing work in the literature have shown remarkable performance on *known* or *seen* presentation attacks, where attacks encountered in the test set are observed in the train set [16]. Detecting *unknown* or *unseen* presentation attacks remains a very challenging problem [34, 6]. Further, degraded iris image quality can significantly affect detection accuracy [31].

Attention networks [26, 12] model the interdependencies between channel-wise features and/or spatial-wise features on CNN feature maps. Feature maps are obtained when a series of convolution filters are applied to outputs from a previous layer in a CNN. The dimensionality of feature maps is channel×height×width. The channel corresponds to the number of convolution filters. The height×width defines the spatial dimension. Channel-wise features are derived from channel-wise convolution that operates along the direction of the channel of feature maps, whereas spatial-wise features are derived from spatial convolution that operates along the direction of the width and height of feature maps. Attention networks have been appropriated in the context of the face modality [25, 4], but have not yet been exploited by other biometric modalities. Wang et al. [25] proposed a multi-modal fusion approach by sequentially combining the spatial and channel attention modules to improve the generalization capability of face PAD systems. Chen et al. [4] developed an attention-based fusion scheme that can effectively capture the feature complementarity from the outputs of two-stream face PAD networks. However, attention mechanisms have not been leveraged for use in *iris* presentation attack detection. Furthermore, effectively integrating such attention modules within the CNN architecture is yet to be systematically studied for presentation attack detection as well as model explanation.

In this paper, we present an explainable attention-guided iris presentation attack detector (AG-PAD) that improves both the generalization and explanation capability of existing PAD networks. This is due to the capability of attention mechanisms to model long-range pixel dependencies such that they can refine the feature maps to focus on regions of interests. Here, long-range dependencies are modeled via the receptive fields formed by a series of sequential convolutional operations. Given a set of convo-

[1] Sometimes, the term "bonafide" is used instead of "live", and the term "spoof" is used instead of "PA", in the biometrics literature.

lutional feature maps obtained from a backbone network, a *channel-attention* module and a *position-attention* module are independently used to capture the inter-channel and inter-spatial feature dependencies, respectively. After that, the refined feature maps are input to convolutional blocks to extract more compact features. Finally, the outputs from these two attention modules are fused using an element-wise sum to capture complementary attention features from both channel and spatial dimensions. This is followed by global average pooling and softmax operations to compute the class probabilities ("live" or "PA"). The flowchart of the AG-PAD network is depicted in Figure 1. In order to provide visual explanations for decisions made by AG-PAD models, we utilize gradient-weighted class activation mapping (Grad-CAM) to flow the gradients back from the class output to the convolutional layer to visualize class-discriminative information. This offers insights on explaining model decisions by highlighting important regions of the image that correspond to such predictions. Compared with other explainable techniques such as saliency maps [23], feature visualization [35], or inverted image representations [15], Grad-CAM provides a simple way to visualize the importance of different input regions for the predicted class.

The main contributions of this work are summarized here:

- We propose an attention-guided PAD approach that can enhance the generalization and explanation capability of iris presentation attack detection.
- We extend the proposed PAD network by introducing hierarchical attention to attend to lower-layer feature representations.
- We evaluate the proposed method on challenging datasets, involving both unseen and unknown PAs.

Although attention modules have been exploited in the face PAD literature [25], our proposed method differs in several different ways: (a) we propose a parallel combination of these two different modules via an element-wise sum, instead of sequentially combining the attention modules; (b) we construct a hierarchical attention method to better attend to the low-layer feature representations; and (c) we provide better visualizations of the discriminative regions after integrating the attention modules.

The rest of the paper is organized as follows. Section 2 presents a literature survey on recent developments in iris presentation attack detection and attention mechanisms. Section 3 describes the proposed attention-guided network used in this work, including the hierarchical attention network. Section 4 discusses the results of the proposed method on both benchmark datasets and challenging proprietary datasets, and compares it with the state-of-the-art methods. Conclusions are reported in Section 5.

2. Related Work

In this section, we provide a brief discussion on (a) existing iris PAD techniques that utilize convolutional neural networks and (b) the applications of attention mechanisms in various computer vision tasks.

CNN-based Iris PAD: Development of Iris PAD approaches using CNN typically operate on either geometrically normalized [10, 19, 30] or un-normalized iris images [17, 3, 29, 11, 14, 32]. This requires the use of an iris segmentation or an iris detection method as a preprocessing step [2]. He et al. [10] proposed a multi-patch convolutional neural network (MCNN) approach that densely samples iris patches from the normalized iris image. Each patch was independently fed into a convolutional neural network and a final decision layer was used to fuse the outputs. On the other hand, Chen and Ross [3] directly used the un-normalized iris image. Their proposed method simultaneously performs iris localization and presentation attack detection. Kuehlkamp et al. [14] computed multiple representations of binarized statistical image features (BSIF) of un-normalized iris images by exploiting different filter sizes, and used these as inputs to the CNNs. An ensemble model was then used to combine the outputs from the different BSIF representations. To handle the problem of unseen presentation attacks, Yadav et al. [32] used relativistic average standard generative adversarial network (RaSGAN) to synthesize iris images in order to train a PAD system that can generalize well to "unseen" attacks. The relativistic discriminator (RD) component of the trained RaSGAN was later extended to design a one-class classifier [33]. However, all the aforementioned methods do not use attention modules to enhance presentation attack detection performance.

Attention Mechanism: The use of an attention mechanism has been adopted in a variety of tasks such as image captioning [28], segmentation [9, 24], classification, and detection [26, 12, 27, 1]. In addition, attention modules have also been used by generative adversarial networks (GANs) to allow for long-range pixel-dependency modeling for the image generation task [36, 7]. Generally, attention mechanisms can be coarsely divided into two types: the generation of channel attention module (CAM) and position attention module (PAM).[2] Hu et al. [12] proposed a novel architecture, termed the "Squeeze-and-Excitation" (SE) block, to explicitly model the interdependencies between channels. Their proposed SENet won the ILSVRC 2017 image classification competition and generalized well across challenging datasets. Though CAM and PAM can

[2] PAM is sometimes also referred to as spatial attention module.

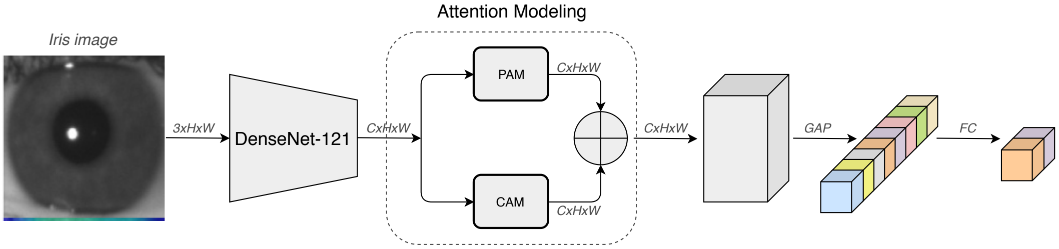


Figure 1: Flowchart depicting the proposed AG-PAD developed in this work. GAP and FC refer to the global average pooling and fully connected layers, respectively. Attention modeling involves both the position attention module (PAM) and the channel attention module (CAM). $\oplus$ denotes element-wise sum.

be independently integrated into existing network architectures, they can also be combined together to provide complementary attention information. Woo et al. [18] proposed a bottleneck attention module (BAM) to first compute the channel and spatial attentions along two separate network branches. Then, an element-wise sum was used to combine these two attention branches. The BAM was demonstrated to show consistent improvement in both classification and detection tasks. There are, of course, other ways to compute the channel and spatial attention modules. Fu et al. [9] utilized the self-attention mechanism to compute the PAM and CAM. The PAM and CAM were combined via an element-wise sum. Their proposed dual attention network (DANet) achieved state-of-the-art performance in image segmentation. Inspired by the works of [27, 9], we observe that PAM and CAM are able to capture long-range pixel dependencies along the spatial and channel dimensions, thereby refining the feature maps to focus on salient iris regions, which can improve the generalization capability of iris PAD solutions.

3. Proposed Algorithm

Our proposed AG-PAD network aims to automatically learn discriminative features from the cropped iris regions that are relevant to presentation attack detection. However, there is a need for using attention mechanisms to further enhance the feature discrimination along both the spatial and channel dimensions. This will ensure that the network will focus more on salient regions during backpropagation learning. The attention-based CNN model is designed by leveraging knowledge via transfer learning. Given a cropped iris image, I, from a detection network [3], the feature maps are extracted by a backbone network f, which is formulated as:

$$A = f(I|\theta). \tag{1}$$

Here, $A \in \mathbb{R}^{C\times H\times W}$ denotes the feature maps of the last convolution layer, where C, H and W are the number of channels, height and width of the feature maps, respectively. θ is a set of parameters associated with the network. Without loss of generality, DenseNet121 [13] is chosen as the backbone network in this study.

3.1. Position Attention Module

Given output feature maps, $A \in \mathbb{R}^{C\times H\times W}$, obtained from a backbone network, they are first fed into two different convolution layers to produce feature maps $B \in \mathbb{R}^{C/r\times H\times W}$ and $C \in \mathbb{R}^{C/r\times H\times W}$, respectively. Here, r is the reduction ratio. Then, these two feature maps are reshaped to $\mathbb{R}^{C/r\times N}$, where $N = H \times W$. After that, C is transposed to $\mathbb{R}^{N\times C/r}$, and multiplied with B, to obtain a feature map of size $\mathbb{R}^{N\times N}$. Finally, a softmax layer is applied to compute the position attention map, $P \in \mathbb{R}^{N\times N}$, as follows:

$$P_{ij} = \frac{exp(C_i \cdot B_j)}{\sum_{i=1}^{N} exp(C_i \cdot B_j)}, \tag{2}$$

where, C_i denotes the i-th row of C and B_j denotes the j-th column of B. P_{ij} is a probability value measuring the position dependency between C_i and B_j, meaning that it can be considered as a weight to refine a pixel value in the spatial position of a feature map.

In addition, the feature map A is fed into another convolution layer to obtain a feature map $D \in \mathbb{R}^{C\times H\times W}$, which is later reshaped to $\mathbb{R}^{C\times N}$. After that, a matrix multiplication is performed between the reshaped D and P to obtain $\mathbb{R}^{C\times N}$. This can be simply reshaped back to the dimensions of the original feature map: $\mathbb{R}^{C\times H\times W}$. The final refined output is obtained as:

$$M_{ij} = \alpha \sum_{k=1}^{N} (D_{ik} P_{kj}) + A_{ij}. \tag{3}$$

Visually, the procedure to compute the position attention map can be seen in Figure 2.

3.2. Channel Attention Module

Given an output feature map, $A \in \mathbb{R}^{C\times H\times W}$, obtained from a backbone network, it is first reshaped to $\mathbb{R}^{C\times N}$. Then, matrix multiplication is performed between A and the transpose of A, resulting in $\mathbb{R}^{C\times C}$. The channel attention map, Q, can be obtained as:

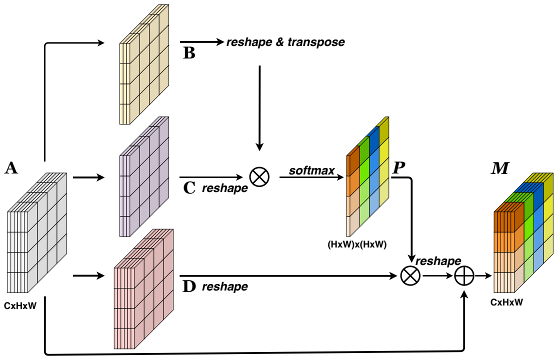


Figure 2: Flowchart depicting the computation of PAM in this work. $\oplus$ denotes element-wise sum and $\otimes$ denotes matrix multiplication.

$$Q_{ij} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)}, \tag{4}$$

where, A_i and A_j are used to denote the i-th row and j-th column, respectively. Q_{ij} is a probability value measuring the channel dependency, meaning that it can be considered as a weight to refine a pixel value in the channel position of a feature map. After that, the channel attention map Q is multiplied with A, where the ensuing feature map is reshaped back to $\mathbb{R}^{C \times H \times W}$. The final refined output from the channel attention map is computed by rescaling with β and an element-wise summation with A:

$$M_{ij} = \beta \sum_{k=1}^{C} (Q_{ik} A_{kj}) + A_{ij}. \tag{5}$$

Visually, the procedure to compute the channel attention map can be seen in Figure 3.

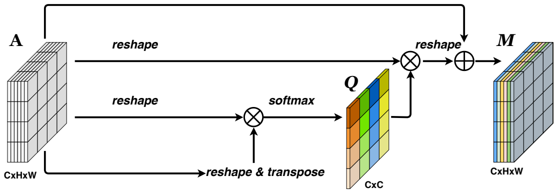


Figure 3: Flowchart depicting the computation of CAM in this work. $\oplus$ denotes element-wise sum and $\otimes$ denotes matrix multiplication.

3.3. Attention-Guided PAD

After utilizing the position attention module and the channel attention module to generate the refined outputs from the input feature maps A, the remaining task is to effectively fuse the attention information from the complementary modules.

First, we consider performing an element-wise sum to fuse the two attention modules.

$$A' = M_p(A) \oplus M_q(A). \tag{6}$$

Here, $M_p(A)$ and $M_q(A)$ are the refined feature maps after applying PAM and CAM, respectively, to the input feature maps A.

Then, we also consider sequentially combining these two attention modules.

$$A' = M_q(A) \tag{7}$$

$$A'' = M_p(A'). \tag{8}$$

Here, A' is the output after applying the channel attention module to the input feature maps A. A'' is the output after applying the position attention module to the input feature maps A.

In addition, a novel hierarchy-attention architecture is proposed, which applies PAM and CAM to mid-level feature maps. Specifically, we have considered feature maps extracted from the *conv3* block and *conv4* block, that have dimensions of $28 \times 28 \times 512$ and $14 \times 14 \times 1024$, respectively. In addition, CAM is applied to the last convolutional block *conv5* of size $7 \times 7 \times 1024$. Since the feature dimensions from individual convolutional blocks are different, max pooling is applied first to reduce both dimensions of $28 \times 28 \times 512$ and $14 \times 14 \times 1024$ to $7 \times 7 \times 512$ and $7 \times 7 \times 1024$, respectively, prior to feature concatenation. We demonstrate later that the hierarchy-attention architecture is more superior for iris presentation attack detection on low-quality samples due to its multi-scale representation.

3.4. Implementation Details

Due to a limited number of training samples in existing iris presentation attack datasets, our PAD models are pre-trained on ImageNet [20]. We fine-tune our models using the feature maps of the last convolutional layer. Unless specified otherwise, DenseNet121 [13] is used as the backbone network in this work. The input spatial size is 224×224 pixels. Our models are trained for 50 epochs in total using Adam optimizer with a learning rate of 0.0001. A mini-batch size of 32 is used during the training. We add two convolutional layers right after the attention layers, prior to the element-wise sum of the refined feature maps.

To better generalize to unseen attacks, extensive data augmentation is applied to populate the training dataset. A number of operations, including horizontal flipping, rotation, zooming, and translation are applied. It must be noted that sensor interoperability is implicitly addressed by gleaning iris samples from different datasets in the training phase. The iris detection module is implemented using the Darknet framework and the iris PAD is implemented using the Keras framework.

4. Experimental Result

4.1. Datasets and Metrics

The proposed AG-PAD is evaluated on the datasets collected by Johns Hopkins University Applied Physics Laboratory (JHU-APL). The datasets are collected across two different sessions, which are termed as JHU-1 and JHU-2. To facilitate the comparison against state-of-the-art methods, the proposed method is also evaluated on the benchmark LivDet-Iris-2017 datasets [34].

JHU-APL: The training set consists of 7,191 live samples and 7,214 PA samples that are assembled from a variety of datasets. JHU-1 consists of 1,378 live samples and 160 PA samples. Types of PAs in JHU-1 include colored contact lenses and Van Dyke/Doll fake eyes. JHU-2 consists of 1,368 live samples and 227 PA samples. Types of PAs in JHU-2 include colored contact lenses and Van Dyke/Doll fake eyes. The colored contact lens in JHU-2 can be further divided into Air Optix colored contact lenses, Acuvue Define colored contact lenses, and Intrigue partial coverage lenses (some examples are shown in Figure 10). The image quality between JHU-1 and JHU-2 were observed to be different. Samples collected in JHU-2 have improved capture quality by minimizing variations such as blur and reflection (see Figure 4). The experiments were conducted in a cross-dataset setting, where training and test subsets are from different datasets.

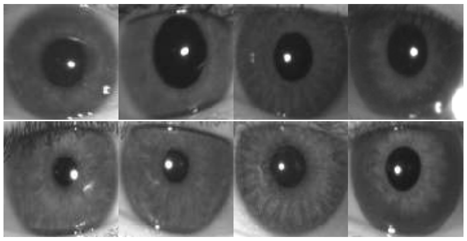


Figure 4: Examples of differences in image capture quality between JHU-1 and JHU-2 datasets. The images in the top and bottom rows are from JHU-1 and JHU-2, respectively. The left two columns denote live samples, whereas the right two columns denote PA samples.

LivDet-Iris-2017: For LivDet-Iris-2017-Warsaw, the training subset consists of 2,669 print samples and 1,844 live samples. The test subset contains both known spoofs and unknown spoofs. The known-spoofs subset includes 2,016 print samples and 974 live samples, while the unknown-spoofs subset includes 2,160 print samples and 2,350 live samples. For LivDet-Iris-2017-ND, the training subset consists of 600 textured contact lenses and 600 live samples. The testing subset is split into known spoofs and unknown spoofs. The known-spoofs subset includes 900 textured contact lenses and 900 live samples. The unknown-spoofs subset includes 900 contact lens PAs, where the types of contact lenses are not represented in the training set, and 900 live samples.

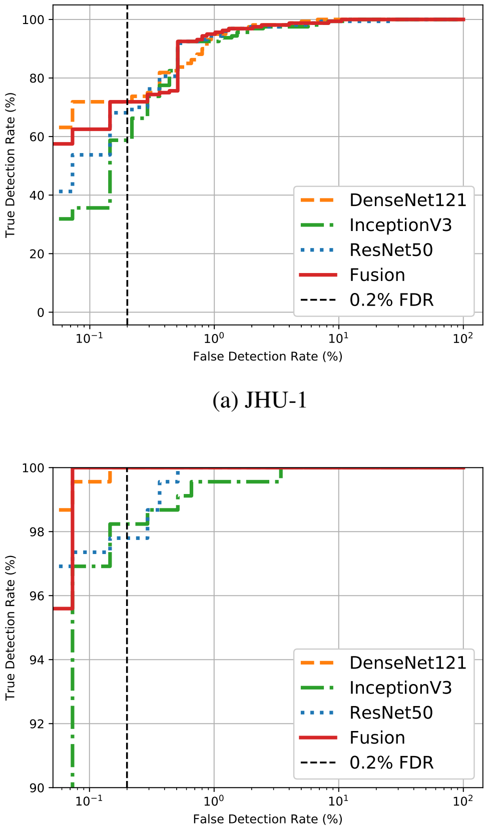
Evaluation Metrics: To report results on the JHU-APL datasets, we use True Detection Rate (TDR: proportion of PAs that are correctly classified) and False Detection Rate (FDR: proportion of live images that are misclassified as PAs),[3] along with the Receiver Operating Characteristic (ROC) curve. To report results on LivDet-Iris-2017, we use the same evaluation metrics as outlined in the LivDet-Iris-2017 competition: (a) BPCER is the rate of misclassified live images ("live" classified as "PA"); and (b) APCER is the rate of misclassified PA images ("PA" classified as "live"). Note that FDR is the same as BPCER and TDR equals (1-APCER). Evaluations on the JHU-APL and LivDet-Iris-2017 datasets follow the same training protocol described in Section 3.4. The difference lies in what datasets are used for training and testing.

4.2. Evaluation on JHU-APL Datasets

The purpose of JHU-APL datasets is to evaluate the generalizability of the iris PAD solutions in practical applications. In addition to the DenseNet121 backbone network, we also investigated the use of other backbones, viz., InceptionV3 and ResNet50. The reason for testing various different backbone networks is to showcase the effectiveness of attention modules regardless of the choice of the backbone network. Moreover, an ensemble of multiple backbone networks may further boost PAD performance. As can be seen from Figure 5, the obtained TDR accuracy at 0.2% FDR on the JHU-1 dataset is not high. This is due to the degraded iris image quality originating from blur, reflections and glasses, to name a few. The proposed method achieved significantly better performance on the JHU-2 dataset, where image quality is much better (see Figure 5). Among all the three evaluated backbone networks, DenseNet121 obtains the best performance (see Table 1). Though the average-score fusion of all three backbone networks does not improve performance on the JHU-1 dataset, it shows improved performance on the JHU-2 dataset at a lower FDR (e.g., 0.1% FDR).

The iris PAD performance was observed to vary with the nature of the presentation attack. While a presentation attack with patterned contact lenses results in only a subtle texture change to the iris region, those based on artificial eye models can extend beyond the iris region and change the iris appearance significantly. Hence, the latter is much easier to detect, as evidenced by a higher true detection rate at 0.2% FDR (see Figure 6). This highlights the necessity of developing more effective PAD solutions for the cosmetic

[3] FDR defines how many 'live" images are misclassified as "PA". The TDR at 0.2% FDR was used to demonstrate the performance of this algorithm in practice.



(a) JHU-1

(b) JHU-2

Figure 5: Evaluation of the proposed AG-PAD method with different backbone networks on the JHU-APL datasets.

Table 1: Evaluation of the proposed AG-PAD method with different backbone networks on the JHU-APL datasets. TDR at 0.2% FDR is used to report the PAD accuracy.

	JHU-1	**JHU-2**
DenseNet121	71.87	100.0
InceptionV3	58.75	98.23
ResNet50	68.12	97.79
Fusion	71.87	100.0

contact PA.

4.3. Architecture Design

Previously, we mentioned three different ways to combine the PAM and CAM. In this section, we show the evaluation results for all the three architectures on the JHU-APL datasets as well. As can be seen from Figure 7, the parallel combination (AG-PAD) obtains better performance than the sequential combination of CAM and PAM on both JHU-1 and JHU-2 datasets. The parallel combination obtains TDRs of 71.87% and 100.0% at 0.2% FDR on JHU-1 and JHU-2 datasets, respectively. The sequential combination obtains TDRs of 53.75% and 98.23% at 0.2% FDR on JHU-1 and JHU-2 datasets, respectively. The hierarchy architecture, on the other hand, shows more promising results on the challenging JHU-1 dataset that has more image quality degradations.

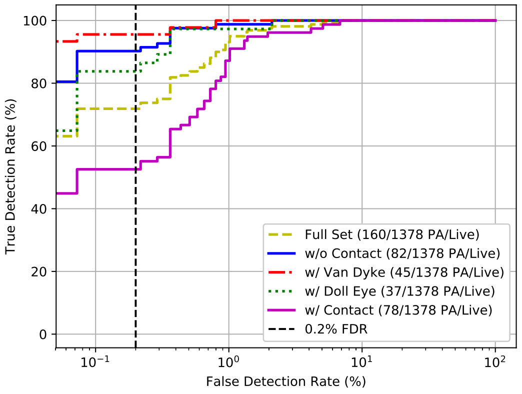
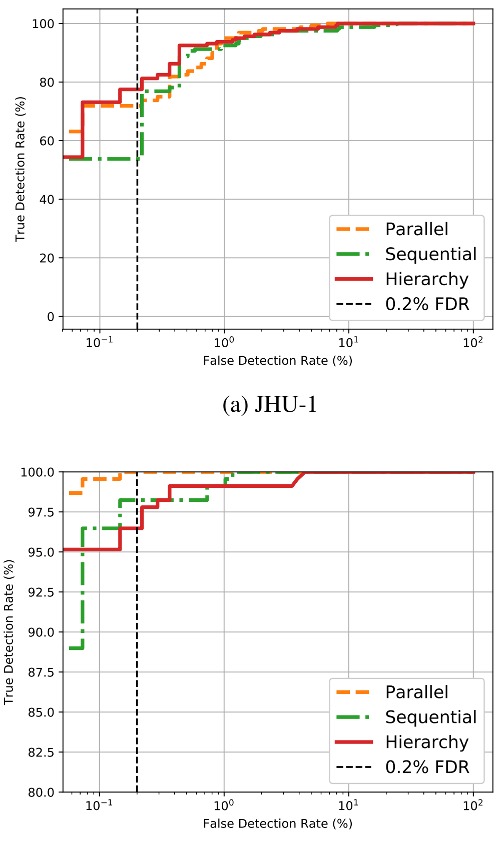


Figure 6: Evaluation of the proposed AG-PAD method on JHU-1 dataset with respect to different presentation attacks. The backbone network used is DenseNet121.

4.4. Ablation Study

To demonstrate the effectiveness of different attention modules, an ablation study was conducted using both the JHU-1 and JHU-2 datasets. In particular, four different variants are considered: *w/o Attention, w/ PAM, w/ CAM,* and *w/ PAM-CAM.* Baseline *w/o Attention* refers to the results obtained by retraining the PAD network without any attention module. *w/ PAM* and *w/ CAM* refer to the results obtained by retraining the PAD network with appended PAM and CAM, respectively. Finally, *w/ PAM and CAM* refers to the results obtained by augmenting the PAD network with both PAM and CAM.

As can be seen in Figure 8, the use of attention modules significantly improves the accuracy over the baseline. Without the attention modules, the baseline only gives 58.75% TDR at 0.2% FDR on JHU-1. After integrating both attention modules, the results improved to 71.87% TDR (see Table 2). A similar trend was observed in the JHU-2 dataset. Further, comparing with other state-of-the-art attention modules [18, 27, 1], the combination of PAM and CAM achieves the best performance. This justifies the significance of using the PAM and CAM attention modules in iris presentation attack detection.



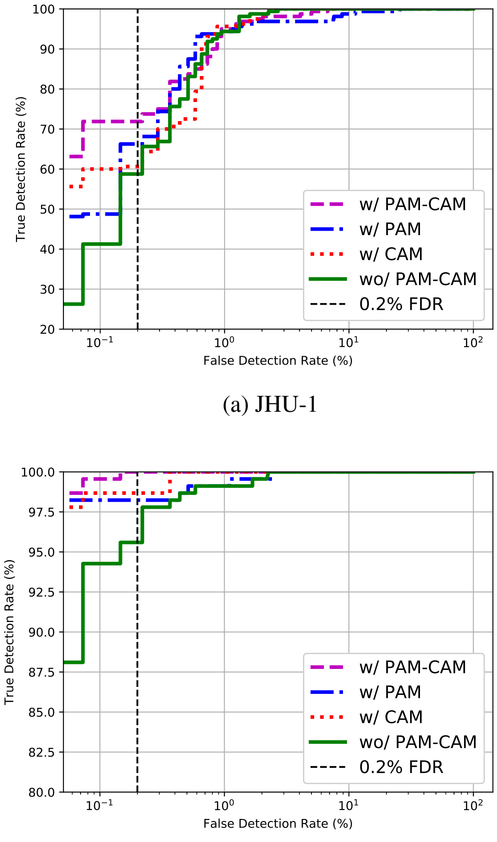
(a) JHU-1

(b) JHU-2

Figure 7: Evaluation of the proposed methods with different architectures on the JHU-APL datasets.

Table 2: The TDR at 0.2% FDR when using the PAM and CAM attention modules with DenseNet121 as the backbone network on the JHU-1 and JHU-2 datasets. The ablation study involving the attention module used in this work and its comparison against other attention modules can be seen here.

	JHU-1	**JHU-2**
w/o Attention	58.75	95.59
w/ PAM	66.25	98.23
w/ CAM	60.62	98.67
w/ PAM and CAM	**71.87**	**100.0**
BAM [18]	68.75	97.35
CBAM [27]	70.0	97.79
GC [1]	66.87	92.95



(a) JHU-1

(b) JHU-2

Figure 8: Ablation study with different attention mechanisms on the JHU-APL datasets.

4.5. Explainable Visualization

To further identify the important regions of an iris image that is used to render a PA decision, Grad-CAM [21] is utilized to generate the visualizations before and after the application of the attention modules in Figure 9. Here, Grad-CAM is used to calculate the gradient of the presentation attack detection score with respect to the feature maps to measure pixel importance. **It is evident that the use of attention modules has enabled the network to shift the focus on to the annular iris region.** By observing the activation maps generated for both live and cosmetic contact samples (see Figure 9), the application of attention modules has forced the network to attend to the annular iris region in order to make the final decision. This is consistent with our intuition that iris texture positioned beyond the pupil region plays a much more significant role for presentation attack detection.

In addition, we also visualize the PAD results for both known PAs and unknown PAs in Figure 10. The bound-

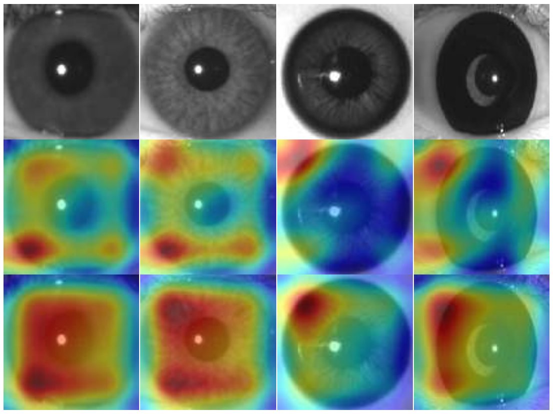


Figure 9: Visualization using Grad-CAM before and after the integration of attention maps for live iris (column 1), contact lens (columns 2 and 4) and fake eye images (column 3). The second row is the result before the use of attention module and the third row is the result after the use of attention module.

ing boxes are obtained from a pre-trained iris detection network. As can be seen, it is much more challenging to perform PAD on unknown PAs, such as Acuvue Define colored contact lenses and Intrigue partial coverage lenses (see Figure 10). The PA scores for unknown attacks are observed to be much lower than the PA scores for known attacks.

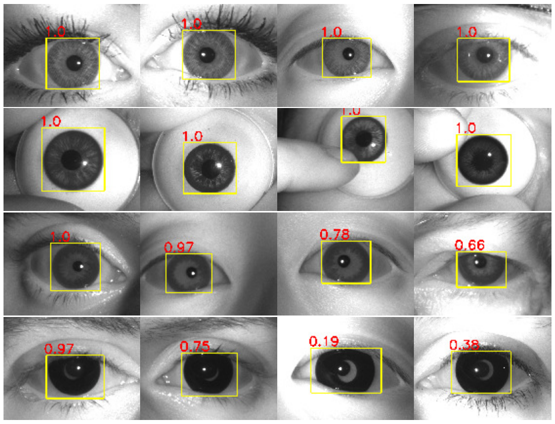


Figure 10: Evaluation of the proposed AG-PAD method on JHU-2 dataset with known and unknown presentation attacks. The first and second rows show the PA results for Air Optix colored contact lenses and Van Dyke fake eye, respectively (known attack). The third row shows the PA results for Acuvue Define colored contact lenses (unknown attack). The last row shows the PA results for Intrigue partial coverage lenses (unknown attack) with both successful and failure cases when using a threshold of 0.5.

4.6. Comparison with State-of-the-art Methods

We also compare the proposed method against state-of-the-art methods evaluated on the LivDet-Iris-2017 [34]. Results of three algorithms that participated in the competition, under both known presentation attacks and unknown presentation attacks, were used. According to the evaluation protocol, a threshold of 0.5 was used to calculate the APCER and BPCER.

As indicated in Table 3, error rates are much higher for unknown presentation attacks on both Warsaw and Notre Dame datasets. **Nevertheless, the proposed method achieves excellent performance for both known and unknown presentation attacks.** The proposed AG-PAD method achieves 1.34% APCER and 0% BPCER for unknown PAs in the Warsaw dataset. This is expected, since this dataset is limited to print attacks only.

Table 3: Evaluation of the proposed method on the LivDet-Iris-2017 dataset. Both known/unknown (K/U) attack results (%) are reported whenever available. Otherwise, a combined error rate is reported.

	Warsaw		**Notre Dame**	
Algorithm	**APCER (K/U)**	**BPCER (K/U)**	**APCER (K/U)**	**BPCER**
CASIA [34]	0.15/6.43	5.74/9.78	1.56/21.11	7.56
Anon1 [34]	0.4/11.44	2.77/6.64	**0**/15.56	0.28
UNINA [34]	0.1/**0**	0.62/20.64	0.89/50	0.33
Proposed	**0.09**/1.34	**0/0**	0.11/**8.33**	**0.22**

5. Conclusions

This paper proposes a novel attention-guided CNN framework for iris presentation attack detection that enables better generalization and explainability. The proposed AG-PAD method utilizes attention-guided feature maps extracted by a channel attention module and a position attention module to regularize the network to focus on salient iris regions, thereby improving the generalization capability of iris PAD solutions. Experiments on several datasets indicate that the proposed method is effective in detecting both known and unknown presentation attacks. Further, a visualization scheme was used to explain the performance of the proposed network.

Acknowledgment

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA RD Contract No. 2017 - 17020200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE International Conference on Computer Vision Workshops*, pages 1971–1980, 2019.

[2] Cunjian Chen and Arun Ross. Exploring the use of iriscodes for presentation attack detection. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–9, 2018.

[3] Cunjian Chen and Arun Ross. A multi-task convolutional neural network for joint iris detection and presentation attack detection. In *IEEE Winter Applications of Computer Vision Workshops*, pages 44–51, 2018.

[4] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil Martin Robertson, and Stan Z. Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2020.

[5] Adam Czajka and Kevin W. Bowyer. Presentation attack detection for iris recognition: An assessment of the state-of-the-art. *ACM Computing Surveys*, 51(4):86:1–86:35, 2018.

[6] Priyanka Das, Joseph McGrath, Zhaoyuan Fang, Aidan Boyd, Ganghee Jang, Amir Mohammadi, Sandip Purnapatra, David Yambay, Sbastien Marcel, Mateusz Trokielewicz, Piotr Maciejewicz, Kevin Bowyer, Adam Czajka, Stephanie Schuckers, Juan Tapia, Sebastian Gonzalez, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Renu Sharma, Cunjian Chen, and Arun Ross. Iris liveness detection competition (LivDet-Iris) – the 2020 edition. In *IEEE International Joint Conference on Biometrics*, 2020.

[7] Xing Di, He Zhang, and Vishal M. Patel. Polarimetric thermal to visible face verification via attribute preserved synthesis. In *IAPR International Conference on Biometrics*, 2019.

[8] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Deep learning multi-layer fusion for an accurate iris presentation attack detection. In *IEEE International Conference on Information Fusion*, 2020.

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[10] Lingxiao He, Haiqing Li, Fei Liu, Nianfeng Liu, Zhenan Sun, and Zhaofeng He. Multi-patch convolution neural network for iris liveness detection. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2016.

[11] Steven Hoffman, Renu Sharma, and Arun Ross. Convolutional neural networks for iris presentation attack detection: Toward cross-dataset and cross-sensor generalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[14] Andrey Kuehlkamp, Allan da Silva Pinto, Anderson Rocha, Kevin W. Bowyer, and Adam Czajka. Ensemble of multi-view learning classifiers for cross-domain iris presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 14(6):1419–1431, 2019.

[15] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015.

[16] David Menotti, Giovani Chiachia, Allan da Silva Pinto, William Robson Schwartz, Hélio Pedrini, Alexandre Xavier Falcão, and Anderson Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Trans. Information Forensics and Security*, 10(4):864–879, 2015.

[17] Federico Pala and Bir Bhanu. Iris liveness detection by relative distance comparisons. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[18] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. In *British Machine Vision Conference*, 2018.

[19] Ramachandra Raghavendra, Kiran B. Raja, and Christoph Busch. ContlensNet: Robust iris contact lens detection using deep convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2017.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[21] 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 *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[22] Renu Sharma and Arun Ross. D-NetPAD: An explainable and interpretable iris presentation attack detector. In *IEEE International Joint Conference on Biometrics*, 2020.

[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshops*, 2014.

[24] Vishwanath A. Sindagi and Vishal M. Patel. HA-CCN: hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2020.

[25] Guoqing Wang, Chuanxin Lan, Hu Han, Shiguang Shan, and Xilin Chen. Multi-modal face presentation attack detection via spatial and channel attentions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-Local Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19, 2018.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[29] Daksha Yadav, Naman Kohli, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Afzel Noore. Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[30] Daksha Yadav, Naman Kohli, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting textured contact lens in uncontrolled environment using DensePAD. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[31] Daksha Yadav, Naman Kohli, Shivangi Yadav, Mayank Vatsa, Richa Singh, and Afzel Noore. Iris presentation attack via textured contact lens in unconstrained environment. In *IEEE Winter Conference on Applications of Computer Vision*, pages 503–511, 2018.

[32] Shivangi Yadav, Cunjian Chen, and Arun Ross. Synthesizing iris images using RaSGAN with application in presentation attack detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[33] Shivangi Yadav, Cunjian Chen, and Arun Ross. Relativistic discriminator: A one-class classifier for generalized iris presentation attack detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2624–2633, 2020.

[34] David Yambay, Benedict Becker, Naman Kohli, Daksha Yadav, Adam Czajka, Kevin W. Bowyer, Stephanie Schuckers, Richa Singh, Mayank Vatsa, Afzel Noore, Diego Gragnaniello, Carlo Sansone, Luisa Verdoliva, Lingxiao He, Yiwei Ru, Haiqing Li, Nianfeng Liu, Zhenan Sun, and Tieniu Tan. LivDet Iris 2017 - iris liveness detection competition 2017. In *International Joint Conference on Biometrics*, 2017.

[35] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[36] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.