Machine Vision and Applications (2022) 33:8
https://doi.org/10.1007/s00138-021-01256-9

SPECIAL ISSUE PAPER

The overlapping effect and fusion protocols of data augmentation techniques in iris PAD

Meiling Fang[1,2] · **Naser Damer**[1,2] · **Fadi Boutros**[1,2] · **Florian Kirchbuchner**[1] · **Arjan Kuijper**[1,2]

Received: 26 February 2021 / Revised: 23 September 2021 / Accepted: 6 October 2021 / Published online: 26 November 2021

Abstract
Iris Presentation Attack Detection (PAD) algorithms address the vulnerability of iris recognition systems to presentation attacks. With the great success of deep learning methods in various computer vision fields, neural network-based iris PAD algorithms emerged. However, most PAD networks suffer from overfitting due to insufficient iris data variability. Therefore, we explore the impact of various data augmentation techniques on performance and the generalizability of iris PAD. We apply several data augmentation methods to generate variability, such as shift, rotation, and brightness. We provide in-depth analyses of the overlapping effect of these methods on performance. In addition to these widely used augmentation techniques, we also propose an augmentation selection protocol based on the assumption that various augmentation techniques contribute differently to the PAD performance. Moreover, two fusion methods are performed for more comparisons: the strategy-level and the score-level combination. We demonstrate experiments on two fine-tuned models and one trained from the scratch network and perform on the datasets in the Iris-LivDet-2017 competition designed for generalizability evaluation. Our experimental results show that augmentation methods improve iris PAD performance in many cases. Our least overlap-based augmentation selection protocol achieves the lower error rates for two networks. Besides, the shift augmentation strategy also exceeds state-of-the-art (SoTA) algorithms on the Clarkson and IIITD-WVU datasets.

Keywords Iris presentation attack detection · Data augmentation · Deep learning

1 Introduction

Iris recognition systems are vulnerable to presentation attacks (PAs). An imposter can use a printed image or replay an iris video to impersonate an enrolled user or wear textured contact lenses to escape recognition. Therefore, developing a reliable iris PAD algorithm is still a challenging task. Considering that neural networks successfully improve the performance in many computer vision fields, deep learning-based algorithms are further applied for iris PAD [6,15,17,30,39]. However, most neural networks suffer from overfitting, where the network does not generalize very well on an unseen test set. Several strategies are therefore proposed to improve the generalizability of networks, e.g., Dropout [34], Batch normalization [25]. In contrast to such methods, the data augmentation technique targets the root problem and insufficient training data variability. Most iris PAD datasets are limited to a small-scale compared to the datasets used for general purposes, for privacy security. Data augmentation can be categorized into *data warping* and *oversampling*. Data warping creates more images based on affine transformation like rotation or translation. Oversampling generates synthetic images, such as using Generative Adversarial Networks (GAN) [18]. Data augmentation techniques improve the performance of modern image classifiers without doubts [10,23,32]. In the iris PAD field, several studies also showed the improvement of performance by augmentation techniques. Gragnaniello *et al.* [19] utilized a data augmentation to generate more training data by rotating the original images for iris PAD task. Their results are slightly improved when applying data augmentation. Raghavendra

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

✉ Meiling Fang
meiling.fang@igd.fraunhofer.de

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany

[2] Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany

et al. [30], Chen *et al.* [6] and Choudhart *et al.* [7] also utilized the augmentation techniques to avoid the overfitting in training phase (see Table 1). However, the contribution of augmentation techniques is not clear because no analysis or experimental comparison is provided as summarized in Table 1. It is worth noting that iris images generated by GAN [18] cannot be used as augmented data in our application to improve the performance, as done for general computer vision tasks. This is because the generated iris images are considered another type of presentation attack for impersonation [37]. As a result, we chose to explore the effect of the data warping technique on iris PAD performance due to the restricted condition of augmentation techniques in the PAD field.

Furthermore, the detailed effect of the data augmentation on iris PAD performance is relatively understudied. In this regard, this work provides answers to the following questions: (1) What is the relative effect of various data augmentation techniques on the performance of iris PAD? (2) Does the combination of all augmentation techniques at various design levels always lead to superior performance, or can there be a formal approach to augmentation methods selection? (3) Do different augmentation strategies improve PAD performance by bringing the "same" misclassified samples to the correct classes? Or do they have a less overlapping effect?

To answer these questions, we explore the impact of different augmentation techniques, specifically data warping techniques, on the generalization of deep learning-based iris PAD. The main contributions of the work are as follows: (1) provide a first in-depth analysis of data augmentation techniques role on the performance and reliability for iris PAD, (2) propose a classification error overlap-based augmentation selection protocol, (3) demonstrate the experiments in terms of fine-tuned and trained from scratch networks with various augmentations on multiple cross-validation scenarios datasets, (4) visualize and discuss the overlapping effect of different augmentation techniques to provide a better explanation of the generalizability induced by augmentation techniques.

2 Related work

Iris recognition systems have been widely applied in different recognition scenarios due to the uniqueness and high accuracy of iris features [2–5]. However, the operational security of the iris recognition has raised many concerns. This section provides a brief review of deep learning-based iris PAD algorithms and general data augmentation techniques. The Iris-LivDet-2017 [42] is the most recent published competition. The used competition datasets and protocols indicated that improving the generalizability of iris PAD is a major challenge. Some recent iris PAD competitions, such as Iris-LivDet-2017 [42] or Iris-LivDet-2020 [11], are organized to evaluate the generalizability of iris PAD algorithms. In contrast to Iris-LivDet-2017 [42], the 2020 edition competition [11] did not offer any official training data and the test data are not yet publicly available, the experiments and analysis in this work are still based on the protocols designed in Iris-LivDet-2017 competition [42]. Hence, we focus here on the algorithms and results in Iris-LivDet-2017. The protocols in this competition are designed under cross-dataset and cross-PA scenarios to reflect the real-world situation. In this competition [42], CASIA proposed to train two SpoofNets to detect printouts and textured contact lenses separately, while UNINA relied on the Scale Invariant Descriptor (SID) and Bag of Words (BoW) to classify the attacks. Afterward, Kuehlkamp *et al.* [27] proposed to combine 61 CNN lightweight CNNs via meta-fusion to classify multiple Binarized Statistical Image Features (BSIF) views of the iris image to overcome such generalization problems. Their results outperformed the winners of the competition. Furthermore, Sharma *et al.* [31] proposed a DenseNet network-based iris PA detector, D-NetPAD, to demonstrate the experiments on a proprietary dataset and four public competition datasets. They trained a D-NetPAD model on their private dataset, including 12,772 training data. Then, this pre-trained model was used in three ways to examine the generalizability on the competition datasets: 1) the pre-trained D-NetPAD is used directly on the test sets in the competition, 2) train a D-NetPAD model from scratch on the competition training sets, 3) fine-tune this pre-trained model on the competition training sets. As expected, the fine-tuned model performed the best. They achieved the lowest error rate (0.30% ACER values) on the Notre Dame dataset in the competition, whereas the second-lowest error is 3.28% from the previous Meta-Fusion method. However, there is a slight problem that their proprietary training data include the testing data of Notre Dame. To fairly compare the results using the same data, we only report the D-NetPAD trained from scratch, and also the Meta-Fusion results later on in Table 12. Besides, we compare our results with the multi-layer fusion (MLF) method achieving the 2.31% ACER in Notre Dame and the recently published micro-stripe analysis (MSA) method ([14,17]) obtaining good performance (11.13% ACER value) in the IIITD-WVU dataset.

Even though such neural network-based algorithms obtain good performance, they still suffer from overfitting. One reason is that training data are insufficient, in quantity and variation. For example, there are only 1200 training iris images in the Notre Dame dataset in the competition [42], which is quite limited compared to datasets designed for generic computer vision tasks. Moreover, not only iris PAD algorithms have this problem, and most networks suffer from overfitting leading to low generalization. Under this condi-

Table 1 Algorithm properties including the used augmentation techniques and relation study

Author	Year	Algorithm	Augmentation	Ablation study
Gragnaniello *et al.* [19]	2016	Domain-aware CNN	Rotate images of multiples of 90 degrees	Yes
Raghavendra *et al.* [30]	2017	Three-class classification based ContensNet	Rotate images through four angles	No
Kohli *et al.* [26]	2017	Synthesize iris images as attacks	Synthetic oversampling by iGCGAN	No
Chen and Ross [6]	2018	Multi-task CNN	Random flipping, cropping, color jittering	No
Yadav *et al.* [39]	2018	Fuse multi-level haralick and VGG features	No	No
Kuehlkamp *et al.* [27]	2019	Ensemble multi CNNs fed with mBSIF features	No	No
Sharma and Ross [31]	2019	DenseNet based PA detector (D-NetPAD)	No	No
Yadav *et al.* [41]	2019	Synthesize iris images as attacks	Synthetic oversampling by RaSGAN	No
Choudhary *et al.* [7]	2020	Fuse top-k features selected thorough Friedman test	Various transformation (rotation, flip, shear, etc.)	No
Fang *et al.* [13]	2020	Fuse features from multiple layers of CNN	No	No
Fang *et al.* [17]	2020	Train MobileNet based on Micro stripes	No	No
Fang *et al.* [14]	2020	Cross-dataset scenarios investigation in iris PAD	No	No

Only [19] compared the PAD performance of their proposed method without and with augmentation.

tion, data augmentation can help to reduce overfitting and enhance the generalizability of networks by virtually generating more training images (more variations) from the original data. The data augmentation techniques can be categorized into data warping and synthetic oversampling [38]. The term data warping can be traced back to the distortion of handwriting in [1]. The warped data are created by applying geometric and color augmentations, such as rotation, shift, flipping, and changing the contrast. In addition to data warping applied in data-space, synthetic oversampling creates images in feature-space by using GANs. The recent iris PAD studies and their used augmentation techniques are presented in Table 1. It is noticed that many works did not mention applying data augmentation, and those who did, did not study the effect of that augmentation in an ablation study. Only [19] did measure this effect, however, as all other works, did not study multiple augmentation methods nor provided a formal selection protocol for augmentation selection. It should be noticed that such generated synthetic images [26,41] are classified as a type of presentation attack in the PAD field, i.e., only increase the number of attack samples without bona fide samples. Such synthetically generated iris images are exploited by an adversary to impersonate someone else's identity. For example, Yadav *et al.* [41] studied the impact of the synthetic data on PAD algorithms when used as a presentation attack. Hence, we explore the impact of augmentation techniques on the performance of iris PAD algorithms. Nevertheless, we only perform data warping augmentation methods due to the imbalance generation of synthetic oversampling techniques.

As summarized in Table 1, the augmentation techniques used in most iris PAD works are rotation, flip, and shear. However, the exact impact of these transformations on PAD performance is unspecified in these works. Moreover, our experimental results (in Sect. 5) show that not all single or combined augmentations increase the iris PAD performance. Therefore, it is essential to find out the most contribute augmentations by considering the unique characteristic of iris data, e.g., NIR illumination, specific sensors, and no noise background. Furthermore, as shown later in Sect. 5, these individual data augmentations that can improve the performance and generalizability of networks help understand the nature of the variations in the attacks. Consequently, studying the specific role of augmentations inspired us to fuse them by sorting overlap classification rates.

3 Methodology

In this section, we will introduce the investigated data augmentation techniques along with the augmentation selection

and fusion protocols, as well as the three CNNs used in our iris PAD study.

3.1 Data augmentation techniques

The collection of large-scale iris datasets is challenging for iris research because of various factors, e.g., privacy concerns and high demand for acquisition environment specifications. Deep learning-based iris PAD studies are thus limited by inadequate datasets. Compared to datasets designed for general purposes like ImageNet dataset [12], most iris PAD datasets have only a dozen to a hundred distinct irises (distinct subjects) as summarized in [8]. The problem of training on small-scale datasets is overfitting, which refers to the phenomenon that a trained network can not generalize well on unseen data. Besides, the Iris-LivDet-2017 competition results suggested that cross-PA and cross-dataset scenarios can be considered the major challenges of current iris PAD fields. To simultaneously validate against insufficient data resources and cross scenarios, we explore the impact of data augmentation methods on iris PAD generalization ability.

To observe the respective impact of data augmentation strategies, we perform six geometric transformation-based augmentation techniques. Notably, the oversampling augmentation technique is neglected in this work because the iris data generated by the GAN [18] are considered *fake* iris [26,41], i.e., an attack. The explored six basic augmentations in this study are: horizontal shift, vertical shift, brightness adjustment, zoom in/out, and horizontal flipping. Such augmentation techniques are widely used in the computer vision field with proved positive effect [10,23,32] and also in the iris PAD field [6,7,19,30]. More reasons that lead us to choose these augmentations are: (1) even under a controlled environment, the irises are not in the same position and same viewpoint. There is still a small geometric variation between iris images. (2) the capture light condition varies between the different datasets when performing the cross-dataset evaluation. (3) the size of the captured irises varies slightly depending on the collectors. (4) iris textures are distinct between the left and right eyes of the same person [8]. However, in some cases, only a single eye of a person is contained in PAD datasets [12]. Hence, it is interesting to explore if horizontal flipping of iris images can improve the performance of PAD algorithms. Considering that the position, direction, size, and illumination differences of iris images are small, we augment the images in a relatively small degree to avoid inducing unwanted noise. The detailed augmentation parameters are listed in Table 4, and the corresponding explanation is in Sect. 4.2. Most interestingly, we look at the effect of each of these augmentation in respect to the other methods.

3.2 Fusion and augmentation protocol

Furthermore, we investigate two methods to fuse the above individual augmentation strategies: strategy-level and score-level fusion. For the former category, the training data are generated by using a combination of several augmentation strategies. For example, an iris image can be rotated, shifted, zoomed, and other operations simultaneously. For the latter category, the prediction scores by each network (trained with one of the single augmentation methods) are fused to calculate a final prediction.

On the other hand, we investigate an augmentation selection protocol. This protocol is based on the overlapping ratio of misclassified samples caused by the different augmentations (as explained later) and thus their relative effect on the performance. This selection step is based on two assumptions: (1) different augmentation techniques contribute to different aspects of the PAD performance, (2) selecting augmentations with the lower overlap of misclassified samples to fuse may improve the results as they focus on the different types of variability in the images.

Let $A = \{A_1, ..., A_n\}$ define a set of augmentation techniques. $I^a_{A_n} = \{I^{a_1}_{A_n}, ..., I^{a_m}_{A_n}\}$ presents a set of misclassified attack images with augmentation A_n and $I^{bf}_{A_n} = \{I^{bf_1}_{A_n}, ..., I^{bf_k}_{A_n}\}$ is a set of misclassified bona fide images with augmentation A_n. The misclassified attacks overlap ratio $O^a_{A_{pq}}$ denotes the ratio of attack samples classified incorrectly with augmentation technique A_p that are also classified wrongly with augmentation A_q. Similarly, the misclassified bona fides overlap ratio $O^{bf}_{A_{pq}}$ denotes the ratio of bona fide samples misclassified with augmentation technique A_p that are also misclassified with augmentation A_q. The ratios can be computed as followed equations:

$$O^a_{A_{pq}} = \frac{\#(I^a_{A_p} \bigcap I^a_{A_q})}{\#I^a_{A_p}} \tag{1a}$$

$$O^{bf}_{A_{pq}} = \frac{\#(I^{bf}_{A_p} \bigcap I^{bf}_{A_q})}{\#I^{bf}_{A_p}} \tag{1b}$$

where $p, q \in \{1, ..., n\}$. Then, the overall overlap ratio $O_{A_{pq}}$ between augmentation techniques A_p and A_q is:

$$O_{A_{pq}} = (O^a_{A_{pq}} + O^{bf}_{A_{pq}})/2 \tag{2}$$

The detailed pseudo-code of the selection protocol can be found in Algorithm 1. We set $k = 3$ in our experiment and select the A_b with the minimum Equal Error Rate (EER) values.

Algorithm 1 Augmentation Selection Protocol

$A = \{A_1, ...A_n\} \leftarrow$ a set of augmentation techniques
$O = \{O_{A_{11}}, ..., O_{A_{ij}}\}$ for $i, j \in \{1, ..., n\} \leftarrow$ a set of overall overlap ratio values per pair of augmentations
$A_b, b \in \{1, ..., n\} \leftarrow$ an augmentation achieved the best performance
$k, k <= n \leftarrow$ desired number of augmentations
procedure FINDLEASTOVERLAP(A, O, A_b, k)
    $S = \{\}$ ▷ Initialize an empty set of selected augmentations
    $S \leftarrow S \cup A_b$ ▷ Start with the best augmentation
    $count \leftarrow 1$
    while $count < k$ **do**
        $temp_{aug} \leftarrow None$
        $temp_o \leftarrow 1$ ▷ The largest overlap ratio is 1
        for A_p in S **do**
            for A_q in $(A \setminus S)$ **do**
                if $O_{A_{pq}} < temp_o$ **then**
                $temp_{aug} \leftarrow A_j$
                $temp_o \leftarrow O_{A_{pq}}$
        $S \leftarrow S \cup temp_{aug}$
        $count \leftarrow count + 1$
    return S

3.3 Neural networks

To evaluate the effect of data augmentation on iris PAD more generally, we train three neural networks: (1) fine-tuning ResNet50, (2) fine-tuning VGG16, (3) training from scratch MobileNetV3-small. On the one hand, ResNet and VGG networks are used widely either as feature extractor or end-to-end architectures in biometric research fields [29,36,39]. For example, Nguyen *et al.* [35] used ResNet [21], VGGNet [33], etc., to extract image features for iris recognition. Yadav *et al.* [39] fused features extracted from off-the-shelf VGG16 model and handcrafted haralick features to detect iris presentation attacks. Therefore, we fine-tune the pre-trained ResNet50 [21] and VGG16 [33] to perform iris PAD. On the other hand, most generic models trained on ImageNet datasets [12] have different patterns compared to iris images. Therefore, we train a lightweight network architecture, MobileNet V3 Small [22] from scratch to target iris PAD issues additionally. MobileNet V3 small has only 2.25M parameters, which is suitable to deploy on mobile devices and to be trained on limited iris data, while ResNet50 has 25.64M parameters and VGG16 has 138M parameters. MobileNet V3 [22] uses the depth-wise convolution and squeeze-and-excitation to reduce parameters and preserve the accuracy at the same time. The training hyperparameters are listed in Table 3. In this work, we focus on the impact of various augmentation techniques and aim to discover the consistency of data augmentation effects, the augmentation selection protocols, and the fusion protocols under diverse network architectures and training strategies. Therefore, we opted to intentionally select a diverse set of networks and training protocols that have shown good performances on iris PAD in previous works [14,16,17,39]. Hence, we fine-tune the ResNet50 and VGG16 networks and train from scratch MobileNetV3 following the experimental settings adopted in [14,16,17,39].

4 Experimental setup

This section describes the datasets, the used parameters in the neural networks and data augmentation techniques, and the evaluation metrics.

4.1 Datasets

The experiments are demonstrated on publicly available benchmark datasets used in the Iris-LivDet-2017 competition [42] to explore the impacts of different data augmentation techniques on PAD performance. The Iris-LivDet-2017 competition [42] contains four datasets: Clarkson, Warsaw, Notre Dame, and IIITD-WVU. Because the Warsaw dataset is no longer publicly available, we use the remaining three datasets in our experiments. Furthermore, the Iris-LivDet-2017 are designed for cross-PA, cross-sensor, and cross-dataset evaluation. Figure 1 presents iris samples from the training and test sets of each of the used datasets. The varying appearance between different datasets indicates the challenging task of cross-dataset PAD. Table 2 summarizes the description of the used datasets, including the number of images in the training and test sets and sensors.

Clarkson dataset The Clarkson dataset is designed as a cross-PA evaluation. The test set consists of additional unknown attack image types that are not present in the training set. The unknown data include visible-light image printouts attack and the extra pattern contact lenses produced by different manufactures. The bona fide visible-light images are presented neither in the training set or the test set.

Notre dame dataset The Notre Dame dataset contains bona fide iris images (without lenses) and textured contact lens attacks. The test set is a combination set of the known subset and unknown subset, corresponding to the cross-PA scenario. The unknown subset includes iris images with textured lenses produced by different manufacturers (different patterns) and not represented in the training data. Another difficulty of this dataset is the limited training data.

IIITD-WVU dataset The IIITD-WVU dataset is an amalgamation of two datasets: the IIITD dataset used for training and the WVU dataset for testing. The experiments performed on the IIITD-WVU dataset correspond to the cross-dataset evaluation because the sensors, data acquisition environments, subject population, and PA generation procedures for the training and testing are different. The training set (IIITD set) was selected from the IIIT-Delhi Contact Lens Iris (CLI) dataset [40] and IIITD Iris Spoofing (IIS) dataset [20], where the images were captured by multiple sensors under a con-

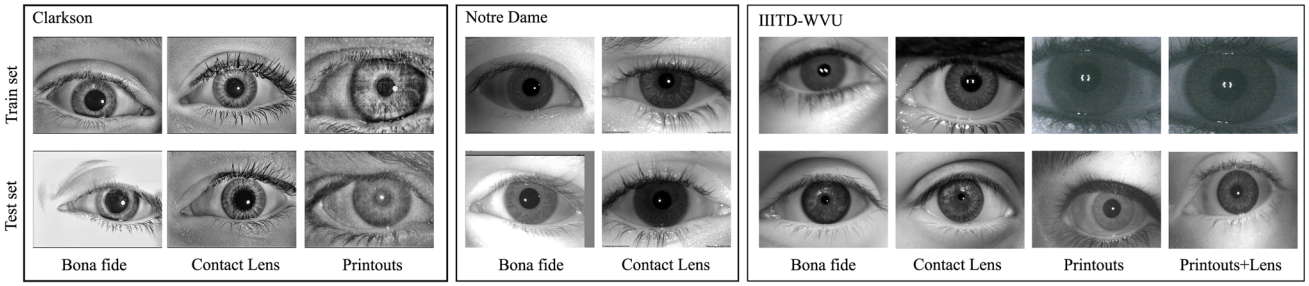

Fig. 1 Samples of iris images from the used datasets. It can be seen that the bona fide and attack samples of different datasets have distinctive appearance, which are affected by capture sensors, light conditions, and printer types, among other factors. This variation indicates the challenging task of cross-dataset PAD

Table 2 Summarized information of the used datasets

Dataset	Clarkson		Notre Dame		IIITD-WVU	
	Train	Test	Train	Test	Train	Test
Bona Fide	2469	1485	600	1800	2250	702
Printous	1346	908	–	–	3000	2806
Contact Lens	1122	765	600	1800	1000	701
Overall	4937	3158	1200	3600	6250	4209
Type	Cross-PA		Cross-PA		Cross-Dataset	
Sensor	IrisAccess EOU2200		IrisGuard AD100, IrisAccess LG4000		Cognet, CIS 202, VistaFA2E	IrisShield MK2120U

trolled environment. The test set (WVU set) was captured using a mobile iris sensor under both controlled (indoor) and uncontrolled (outdoor) environments. The Iris-LivDet-2017 competition results [42] indicated that the cross-dataset evaluation was considered the most challenging task on account of the significant variations.

4.2 Parameters setting

To make our experimental setting compliant with the Iris-LivDet-2017 competition [42], we use the pre-defined training and the test set as described in Sect. 4.1. Additionally, 20% of the images are selected randomly from each training set to serve as a validation set during the training procedure. The training hyperparameters listed in Table 3 are used to fine-tune the ResNet50 [21] and VGG16 [33] networks, and train the MobileNetV3-small [22] from scratch. The input size of the three networks is $480 \times 640 \times 3$, where the grey-scale iris images are converted to three-channel images filled with the same pixel values. The number of actual training epochs is controlled by the early stopping method. The training stops if the validation loss does not decrease after ten epochs or the training reaches its maximum training epochs in our experiments.

The parameters of augmentation techniques are listed in Table 4. An image can be shifted horizontally or vertically by a specific ratio of the image width or height. The range of the shift is 0 to 100%. In our case, the specific ratio sets to 10%. A rotation augmentation randomly rotates the image clockwise between 0 and 360 degrees. We limited the maximum rotation degree to 15 degrees. Also, the brightness of the image can be augmented by either randomly darkening or brightening. The range of the brightness argument is from 0 to 200%. The brightness is not changed when the value is 100. The values less than 100 darken the image, whereas values larger than 100 brighten the image. Furthermore, the iris image can be zoomed in/out with a specific ratio. The range of zoom arguments is 0 to 200%. The image is not changed when the zoom argument is 100%. In the experiment, we zoom the images between 85% and 115% randomly. Finally, more iris images can be produced by horizontally flipping. The code is implemented based on the Keras library. [1]

Table 3 The training hyperparameters

Parameter	Value
No. epochs (max)	35
Earlystopping patience	10
Learning rate	0.001
Optimizer	Adam
Batch Size	32

[1] Keras: A high-level NNs API (https://keras.io/).

Table 4 The parameters of different data augmentations

	$Shift_h$	$Shift_v$	Rotation	Brightness	Zoom	$Flip_h$
Unit	%	%	Degree	%	%	Bool
Range	[0,100]	[0,100]	[0,360]	[0, 200]	[0, 200]	T, F
Used	10	10	15	[70, 120]	[85, 115]	F

The subscripts h and v represent horizontal and vertical, respectively. The value 100 in brightness and zoom refers to an unchanged images

4.3 Evaluation metrics

The following metrics are used to measure the PAD algorithm performance:

- **Attack Presentation Classification Error Rate (APCER)**: The proportion of attack images incorrectly classified as bona fide samples.
- **Bona Fide Presentation Classification Error Rate (BPCER)**: The proportion of bona fide images incorrectly classified as attack samples.
- **Average Classification Error Rate (ACER)**: corresponds to the average of BPCER and APCER.

The APCER, BPCER, and ACER follow the standard definition presented in the ISO/IEC 30107-3 [24]. The threshold used to decide an iris image is bona fide is 0.5, as defined in the Iris-LivDet-2017 protocol [42]. Moreover, the Detection Equal Error Rate (D-EER) and the BPCER value by fixing the APCER value at 1% are reported for more analysis.

Furthermore, we use the Fisher Discriminant Ratio (FDR) to examine the achieved class separability (attack and bona fide) induced by different augmentation settings to indicate classification generalizability. The FDR is described in [28] and [9] as the measurement of separability between genuine and imposter scores. In our work, the high separation between bona fide and attack scores indicates higher reliability of the applied augmentation technique in the iris PAD system. The FDR is described in Equ. 3:

$$FDR = \frac{(\mu^{bf} - \mu^{a})^2}{(\sigma^{bf})^2 + \sigma^{a})^2} \tag{3}$$

where μ^{bf} and μ^{a} are the respective standard deviation of bona fide and attack scores, and σ^{bf} and σ^{a} are their mean values. We also analyze the differences in the augmentation-induced enhancement of different augmentation strategies with the help of the confusion matrix plotted based on the overlap misclassified samples as mentioned in Sect. 3.1. The details of this confusion matrix are described in Sect. 5.2.

5 Experiments evaluation

This section evaluates the several augmentation techniques using the three models on three datasets in terms of the different metrics. In addition to the individual augmentation methods (see Tables 5, 6, 7, 8, 9, 10), we also report the results of strategy-level and score-level combination in Table 11. We also draw the ROC curves of either single augmentation technique (as appended in Fig. 2) or multiple fusion methods (as appended in Fig. 3). Furthermore, we analyze the overlapping misclassified images by employing the confusion matrix (as shown Fig. 6, 7 and 8).

5.1 Results

In this subsection, we first analyze the results in terms of individual datasets per specific augmentation technique. Then, for further study, the fusion-based results are discussed. Finally, we compare our results with the SoTA algorithms for an overall analysis.

Clarkson Results Table 5 reports the iris PAD performance in terms of D-EER, the BPCER value at 1% APCER value, and FDR. It can be observed that (1) translation, brightness, and horizontal flip augmentation produce better results in some cases, e.g., applying the MobileNetV3 model, (2) however, not all augmentations can improve the PAD performance, (3) the higher FDR values mostly coincide with the lower D-EER value. By looking at Table 5 and Table 6 together, we can find that the FDR value has a greater potential to suggest a lower ACER value relative to the D-EER metric.

Notre Dame Results Table 7 and Table 8 describe the iris PAD performance on Notre Dame dataset. As shown in Table 7, the model fine-tuned without augmentation (ResNet50 and Vgg16) outperforms than most other augmentations. In contrast, the performance of the MobileNetV3 (scratch) is mostly improved compared to training without any augmentation techniques. Moreover, unlike the lowest ACER acquired by MobileNetV3 on the Clarkson dataset, ResNet50 achieved the best result (9.56% ACER) in Table 8 by using brightness augmentation on the Notre Dame dataset. The Clarkson and Notre Dame datasets both correspond to the cross-PA scenarios that include unseen cosmetic lens patterns. However, the same network architectures show a significant difference. As

Table 5 Iris PAD performance (%) reported in terms of D-EER, the BPCER value at 1% the APCER value on Clarkson dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR
No	7.44	57.54	7.02	**6.42**	55.62	**8.03**	7.31	34.75	5.87
Shift_h	**6.74**	**48.15**	**8.59**	6.78	**50.77**	7.72	**0.28**	**0.00**	**75.09**
Shift_v	**7.35**	53.06	**8.75**	6.49	**56.09**	7.47	8.01	27.41	4.81
Rotation	**6.93**	61.75	**8.19**	6.87	62.30	7.13	3.26	6.53	**8.36**
Brightness	7.92	63.03	6.34	**6.37**	58.72	**8.35**	**1.27**	**1.75**	5.17
Zoom	8.52	**24.85**	4.69	6.74	**50.98**	7.77	6.81	29.70	6.54
Flip_h	8.36	**52.39**	7.16	**6.36**	57.31	**8.49**	**0.35**	**0.00**	**21.86**

The bold values in *D-EER* and *BPCER @ 1%* column are the Top-3 lowest error rate, and the bold values in the *FDR* column indicate the Top-3 highest separability

Table 6 Iris PAD performance (%) reported in terms of APCER, BPCER and ACER on Clarkson dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)
No	11.78	2.56	7.17	10.58	1.75	**6.17**	12.91	2.83	7.87
Shift_h	**9.15**	2.02	**5.58**	11.78	**1.08**	6.43	**1.67**	**0.00**	**0.84**
Shift_v	**9.62**	**1.14**	**5.38**	11.12	**1.68**	6.40	19.67	0.94	10.30
Rotation	**10.58**	**1.21**	**5.90**	11.18	1.82	6.50	**11.42**	**0.00**	**5.71**
Brightness	12.97	1.62	7.29	**10.52**	**1.08**	**5.80**	17.27	**0.00**	8.64
Zoom	18.23	**0.61**	9.42	**10.52**	1.89	6.21	14.76	0.27	7.52
Flip_h	10.94	2.02	6.48	**9.50**	2.22	**5.86**	**5.20**	**0.00**	**2.60**

The bold numbers in columns are the Top-3 lowest error rates.

Table 7 Iris PAD performance (%) reported in terms of D-EER, the BPCER value at 1% the APCER value on Notre Dame dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR
No	**3.72**	12.56	2.35	**6.00**	20.56	**4.39**	14.11	56.78	1.69
Shift_h	8.94	**12.28**	1.95	6.78	19.28	**5.16**	14.22	59.39	2.08
Shift_v	**5.83**	20.06	**4.88**	**6.44**	**18.61**	3.83	**9.22**	**45.22**	**3.05**
Rotation	7.83	40.11	2.00	7.05	27.67	3.13	12.78	77.50	**3.28**
Brightness	**3.56**	**8.78**	**5.04**	**6.44**	**18.67**	**4.21**	**9.67**	**43.00**	2.23
Zoom	100	**0.67**	**2.50**	7.61	26.17	3.80	**9.94**	50.17	**3.03**
Flip_h	12.33	19.50	2.33	**5.61**	**17.94**	3.67	11.72	**23.00**	3.00

The bold values in *D-EER* and *BPCER @ 1%* column are the Top-3 lowest error rate, and the bold values in the *FDR* column indicate the Top-3 highest separability

shown in Table 6 and Table 8, ResNet50 performed worst on the Clarkson and best on the Notre Dame dataset, whereas MobileNetV3 performed best on the Clarkson and worst on the Notre Dame. One possible reason for this opposite variation is insufficient training data for the Notre Dame dataset (4937 training data in Clarkson and 1200 training in Notre Dame). Another possibility is the differences in the ratio of their unknown PA in the test set (21.03% unknown attack in the attack test subset in Clarkson and 50% in Notre Dame). Considering these two reasons, we argue that models pre-trained on large-scale datasets may perform better on unseen pattern data with insufficient training data for fine-tuning. Besides, similar to the third finding in the Clarkson dataset, the augmentation technique obtained with the higher FDR values also achieves the lower ACER values determined by a pre-defined threshold.

IIITD-WVU Results Table 9 and Table 10 denote the results of the IIITD-WVU dataset, which corresponds to a challenging cross-dataset scenario. It can be observed that D-EER values and ACER values of the IIITD-WVU are higher than

Table 8 Iris PAD performance (%) reported in terms of APCER, BPCER and ACER on Notre Dame dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)
No	35.72	0.06	17.89	**25.72**	**0.56**	**13.14**	37.78	**0.06**	18.92
Shift_h	34.67	**0.00**	17.33	**22.28**	1.00	**11.64**	38.83	**0.00**	19.42
Shift_v	**18.78**	0.66	**9.72**	27.39	0.78	14.09	27.50	2.94	15.22
Rotation	35.33	**0.00**	17.67	31.77	0.61	16.19	**21.39**	3.39	**12.39**
Brightness	**19.06**	0.06	**9.56**	**27.11**	**0.38**	**13.75**	33.44	**0.06**	16.75
Zoom	33.78	**0.00**	16.89	27.88	**0.50**	14.19	**26.00**	1.44	**13.72**
Flip_h	**31.16**	**0.00**	**15.58**	29.89	**0.56**	15.23	**22.72**	2.11	**12.42**

The bold numbers in columns are the Top-3 lowest error rates.

Table 9 Iris PAD performance (%) reported in terms of D-EER, the BPCER value at 1% the APCER value on IIITD-WVU dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR	D-EER (%)	BPCER @ 1%	FDR
No	13.24	46.72	1.90	21.93	**50.43**	1.41	**13.10**	**42.59**	2.67
Shift_h	**12.71**	**35.61**	**3.07**	**18.85**	**48.86**	**1.76**	**12.81**	**35.47**	**3.28**
Shift_v	**9.26**	**40.31**	**4.81**	21.83	58.83	1.29	22.36	**44.44**	1.70
Rotation	**12.53**	**41.60**	**3.17**	21.93	59.54	1.19	13.67	45.30	**3.18**
Brightness	16.66	53.98	1.85	**20.69**	**50.71**	**1.44**	18.51	46.87	2.43
Zoom	14.38	46.72	2.70	**18.37**	52.85	**1.70**	**12.95**	56.70	**3.48**
Flip_h	19.65	77.78	1.83	21.80	64.96	1.34	19.08	52.99	1.84

The bold values in *D-EER* and *BPCER @ 1%* column are the Top-3 lowest error rate, and the bold values in the *FDR* column indicate the Top-3 highest separability

Table 10 Iris PAD performance (%) reported in terms of APCER, BPCER and ACER on IIITD-WVU dataset

Augmentation	ResNet50			VGG16			MobileNetV3		
	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)
No	**1.88**	40.17	21.03	**13.43**	26.78	**20.11**	**3.96**	24.07	14.02
Shift_h	17.54	8.55	**13.05**	**16.97**	**20.09**	**18.53**	4.25	**20.37**	**12.31**
Shift_v	14.68	**5.41**	**10.05**	22.70	21.08	21.89	6.62	25.78	16.20
Rotation	**14.14**	11.39	**12.77**	25.86	**19.80**	22.83	**3.51**	21.51	**12.51**
Brightness	29.31	**7.21**	18.22	19.82	20.94	20.38	**3.91**	24.50	14.20
Zoom	**8.50**	20.37	14.43	23.01	**15.53**	**19.27**	4.73	**18.09**	**11.41**
Flip_h	28.37	**7.83**	18.10	**14.80**	26.35	20.58	15.08	**20.09**	17.58

The bold numbers in columns are the Top-3 lowest error rates.

the Clarkson and Notre Dame datasets. As shown in Fig. 2, when fixing the APCER values (x-axis), the ROC curves indicate that the IIITD-WVU dataset has higher BPCER values (the y-axis coordinate is 1-BPCER) than the Clarkson and Notre Dame datasets. Moreover, the variation between individual augmentation techniques is more pronounced on the IIITD-WVU dataset. In addition to such variations on different datasets, the variations of augmentation techniques are slightly different across methods. For example, ResNet50 and VGG16 achieve better results with vertical shift on all datasets; however, the MobileNetV3 model performs worse when using vertical shift (See referable AUC values). Looking at Table 5, horizontal shift and zoom yields better results with VGG16 and MobileNetV3 networks. The lowest D-EER (9.26%) and the lowest ACER (10.05%) are achieved by the vertical shift when fine-tunning the ResNet50 model. Consistent with the observations in the Clarkson and Notre Dame datasets, the higher FDR value potentially points to a lower ACER value in most cases. Therefore, we can conclude that the FDR metric is more suitable than the D-EER metric

Table 11 Fusion-based PAD performance (%) reported in terms of D-EER, ACER and FDR

Dataset	Method	ResNet50			VGG16			MobileNetV3		
		D-EER (%)	ACER (%)	FDR	D-EER (%)	ACER (%)	FDR	D-EER (%)	ACER (%)	FDR
Clarkson	BS	**6.74**	**5.58**	**8.59**	**6.36**	**5.86**	**8.49**	**0.28**	**0.84**	**75.09**
	ST	**6.02**	**5.12**	**9.59**	6.87	6.43	7.71	5.73	7.49	6.47
	SC	7.28	8.07	7.28	6.46	5.98	8.43	**0.28**	5.08	28.36
	LO_{ST}	7.44	6.96	6.56	**6.17**	**5.81**	**8.49**	1.04	**1.49**	**38.97**
	LO_{SC}	7.09	6.55	7.74	6.46	5.96	8.11	0.50	7.53	14.84
Notre Dame	BS	**3.56**	**9.56**	5.04	**5.61**	15.23	**5.16**	**9.22**	12.39	3.28
	ST	**3.44**	**7.47**	**6.25**	**5.06**	**13.89**	4.34	11.11	17.75	1.98
	SC	6.56	16.72	2.89	6.06	14.50	3.90	**9.33**	13.42	3.61
	LO_{ST}	10.56	14.44	3.28	5.72	**11.64**	**5.24**	10.83	**10.28**	**4.29**
	LO_{SC}	3.56	10.58	**5.14**	5.88	**13.89**	4.35	10.17	**12.36**	**3.96**
IIITD-WVU	BS	**9.26**	**10.05**	**4.81**	**18.37**	**19.27**	**1.70**	**12.81**	12.31	3.28
	ST	**12.11**	**13.12**	**3.05**	**18.79**	20.23	**1.63**	12.95	13.11	**3.29**
	SC	13.31	13.41	2.71	20.07	**20.08**	1.52	14.38	**11.20**	2.80
	LO_{ST}	15.96	16.04	2.35	20.12	20.38	1.42	**10.96**	**10.79**	**3.69**
	LO_{SC}	12.54	15.68	2.94	21.01	21.19	1.40	16.37	13.95	2.71

The bold values in *D-EER* and *ACER* column are the Top-2 lowest error rate, and the bold values in the *FDR* column indicate the Top-2 highest separability. BS: best single augmentation, ST: strategy-level fusion, SC: score-level fusion, LO_{ST}: least overlap-based strategy-level fusion, LO_{SC} : least overlap-based score-level fusion

for measuring the reliability and generalizability of the PAD algorithms.

Fusion-based Results Table 11 presents the performance results of the **B**est **S**ingle augmentation (BS) for each dataset and network, and four fusion-based methods: (1) **ST**ategy-level fusion (ST) with all augmentations, (2) **SC**ore-level fusion (SC) with all augmentations, (3) **L**east **O**verlap-based strategy-level fusion (LO_{ST}), (4) **L**east **O**verlap-based score-level fusion (LO_{SC}). The augmentations used for LO_{ST} and LO_{SC} are selected by Algorithm 1 described in Sect. 3.1. It can be observed in Table 11 that strategy-level fusion has a greater probability to produce the best results than the score-level fusion method. For instance, the ST method obtains the lowest D-EER values in the Clarkson and the Notre Dame dataset by using the ResNet50 model, and LO_{ST} fusion achieves the best performance in the Clarkson by VGG16 and in the IIITD-WVU by the MobileNetV3 network. Moreover, for VGG16 and MobileNetV3 networks, our augmentation selection protocol achieves one of the two lowest ACERs for five of the six experimental setups. Although a pre-defined threshold can influence the ACER value, a higher FDR value always suggests a lower ACER value. Therefore, the higher the FDR value, the higher the reliability of the PAD algorithm.

Comparison with SoTAs We also compare our results with several SoTA algorithms in Table 12. The first three rows are the winners of the Iris-LivDet-2017 competition [42], followed by four of the latest SoTAs, and then the best results of our three networks, respectively. The detailed description of the competition and SoTA algorithms is presented in Sect. 2. The Meta-Fusion [27] approach combined 61 CNNs to classify multiple BSIF views of the iris images via SVM meta-fusion. D-NetPAD method [31] adopted a DenseNet model that is pre-trained on a private combined iris dataset. They also trained a DenseNet model on the competition datasets from scratch. We report these scratch D-NetPAD results for a fair comparison on the same data resource. MLF method [13] fused the information from multiple network layers to make a PAD decision. MSA [14,17] approach focuses on the artifacts differences in the image dynamics around the iris/sclera border area by extracting information from micro-stripes. Because MLF and MSA do not report the results on the Clarkson dataset, we mark '-' in Table 12.

For the Clarkson dataset, the lowest ACER value (0.84%) is produced by the MobileNet trained with the horizontal shift augmentation. For the IIITD-WVU dataset, our ResNet50 model trained with vertical shift generated data achieves the best result with the ACER value of 10.05%. However, the MLF [13] method achieves the best results on the Notre Dame dataset, while our solutions perform worse than Anon1, D-NetPAD, Meta-Fusion, and MSA methods. Due to the lack of training data in the Notre Dame dataset (1200 training data, 3600 testing data), even though data augmentations improve the results, the model still overfits. Therefore, we concluded that shift augmentation is worth attempting for the improvement of the PAD performance. Also, fusing vari-

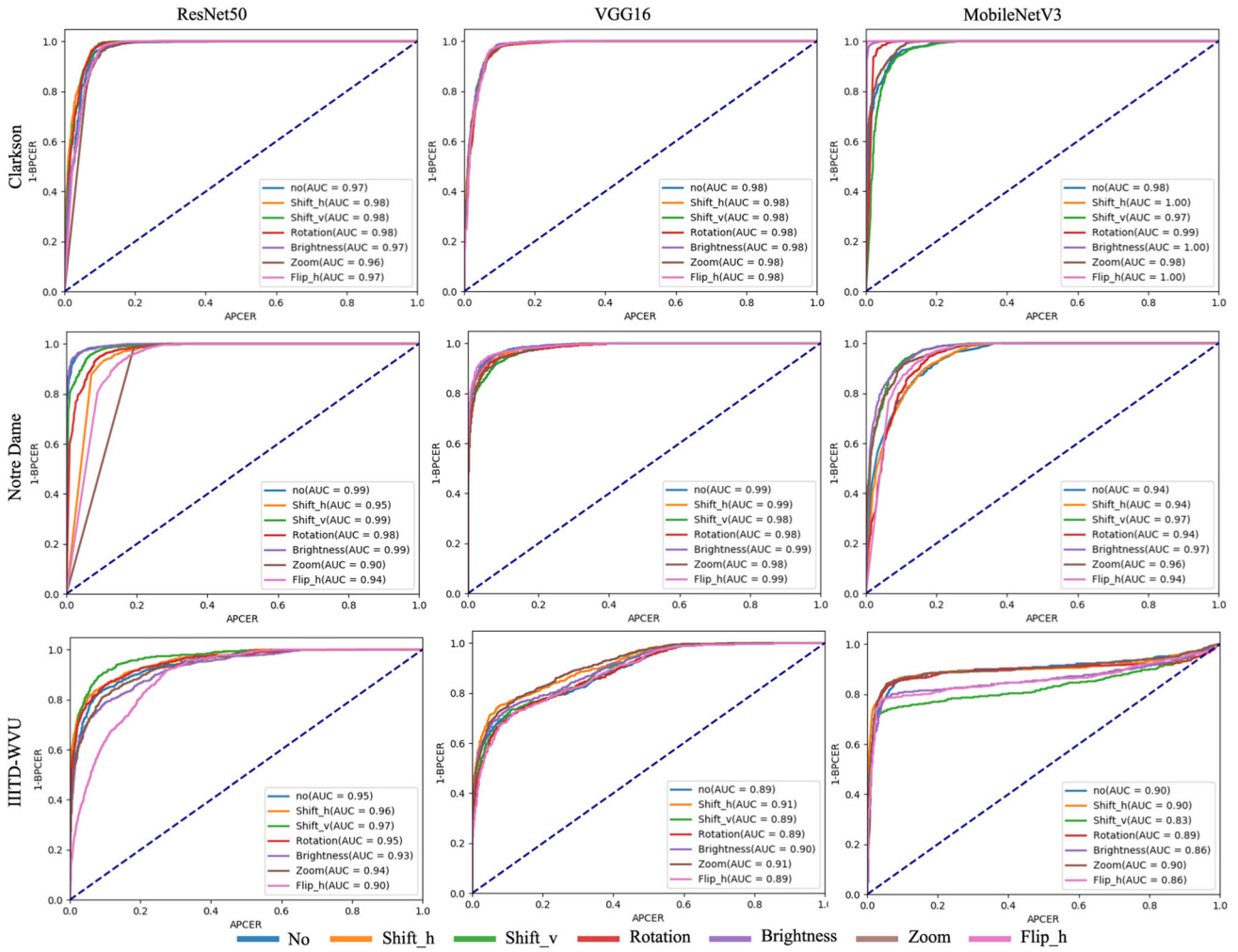

Fig. 2 ROC curves of single augmentation technique. The columns from left to right are ResNet50, VGG16, and MobileNetV3. The rows from top to bottom are for the Clarkson, Notre Dame, and IIITD-WVU datasets. The x-axis is the APCER values, and the y-axis is the $1 - BPCER$ values

ous augmentations in the strategy-level is a good start for iris PAD by considering all the previous results.

Cross-dataset evaluation In addition to inter-dataset evaluation, we also report the cross-dataset results in terms of D-EER, ACER, and FDR values in Table 13. In the cross-dataset scenario, the training data are the training subset of one dataset, while the test data are the test subset of the other two datasets. For instance, the model trained on the Clarkson dataset is used to produce the prediction scores on the test subset of Notre dame and IIITD-WVU datasets. The threshold is set to 0.5 as defined in the Iris-LivDet-2017 competition protocol. To demonstrate the generalizability of the different fusion strategies, we provide the results generated by the BS, ST, SC, LO_{ST} and LO_{SC} settings, similarly to the inter-dataset results in Table 11. In addition to fusion methods, the results of the training without augmentation technique (denoting as *No*) are also reported for comparison. The bold values in the D-EER and ACER columns are the lowest two error rates, and the FDR column's bold values indicate the Top-2 separability measured by FDR. For further comparison, we also provide a visual representation of the D-EER values achieved by the different experimental settings in Fig. 4 and the ROC curves in Fig. 5. As can be concluded from Table 13, (1) training without augmentation techniques performs worse than using augmentations in most cases. (2) BS and ST methods achieve one of the two lowest ACER values in half of the experimental setups. (3) SC augmentation method obtains one of the lowest D-EER values for nine of the eighteen experimental setups, notably eight of the nine lowest D-EER values are produced by the fine-tuned ResNet50 and VGG16. Furthermore, the reliability of the FDR value is consistent with the observation of the previous inter-dataset results that a higher FDR value hints at a lower ACER value, even though the ACER value can be affected by a pre-defined threshold. It can also be noticed in Table 13 that training MobileNetV3 from scratch with LO_{ST}

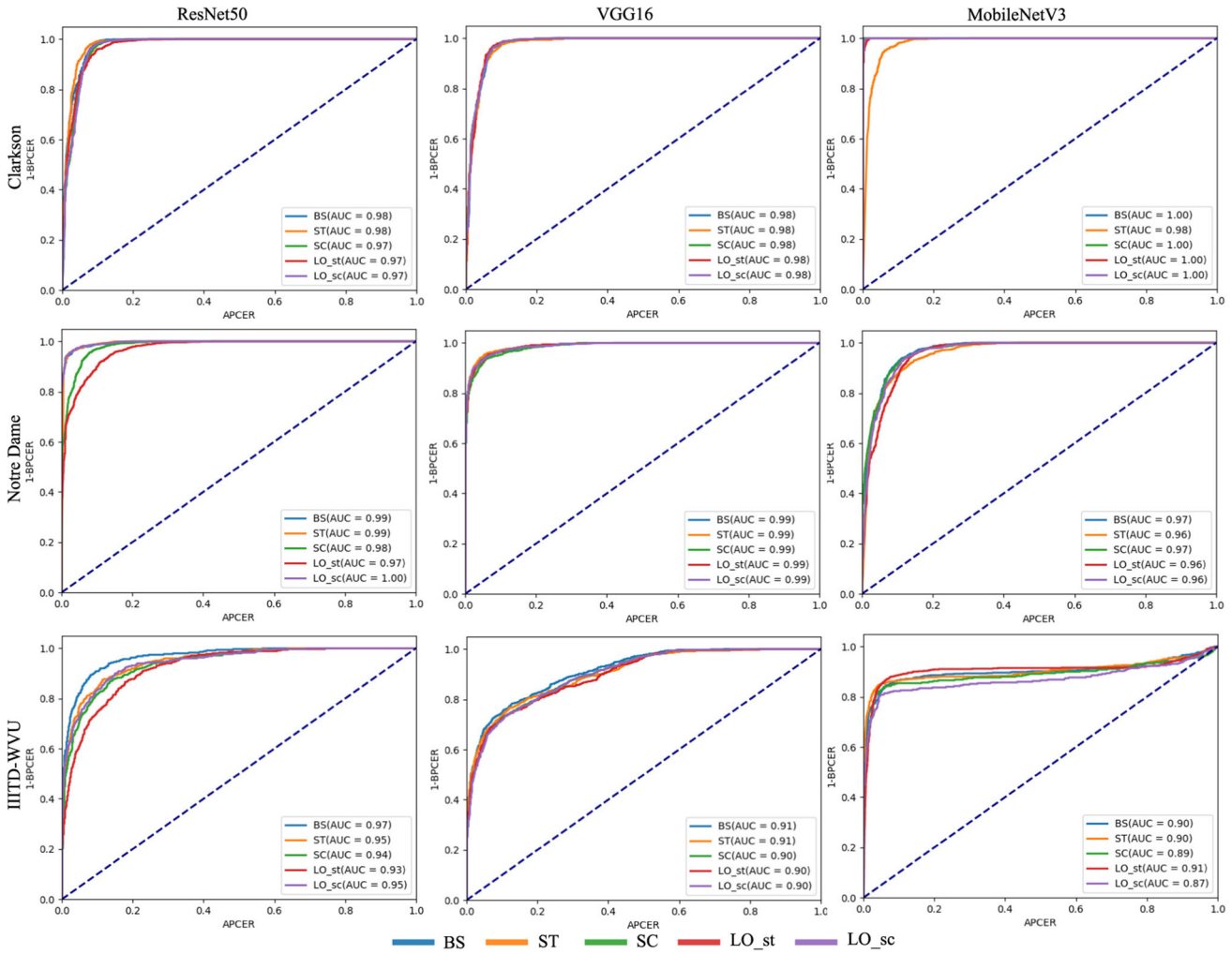

Fig. 3 ROC curves of multiple fusion methods (corresponding to Table 11. The rows from top to bottom are for the Clarkson, Notre Dame, and IIITD-WVU datasets. The x-axis is the APCER values and the y-axis is the $1 - BPCER$ values.)

performs better than with other augmentation strategies in most cases. Similar observation can be found in Fig. 4. SC (yellow) and LO_{SC} (green) methods achieve lower D-EER values than ST (grey) and LO_{ST} (navy blue) methods for ResNet50 and VGG16 networks. In contrast, SC and LO_{SC} produce higher D-EER values than ST and LO_{ST} for the MobileNetV3 network. One possible reason is the different training strategies of networks.

5.2 Analysis and discussion

This section explores if different augmentations lead to the same or different kinds of performance improvements. To do that, we analyze the overlap of misclassified samples between different augmentation protocols, including four fusion methods with the help of confusion matrices. Furthermore, the limitations and potentials of our analyses will be discussed. The confusion matrices for each dataset can be seen in Figs. 6, 7 and 8. The horizontal axis (X axis) from left to right and the vertical axis (Y axis) from top to bottom correspond to the augmentation strategies: No, Shift_h, Shift_v, Rotation, Brightness, Zoom, Flip_h, ST, SC, LO_{ST} and LO_{SC}, respectively.

The matrices from left to right are generated by the ResNet50, VGG16, and MobileNetV3 separately. The value in top matrices refers to the misclassified attacks overlap ratio $O^a_{A_{pq}}$ computed as in Eq. (1a), and the bottom matrices present the misclassified bona fides overlap ratio $O^{bf}_{A_{pq}}$ computed as in Eq. (1b) in Sect. 3.1.

As can be seen from the previous results, different augmentation strategies improve the performance on the different datasets. In this case, shift, rotation, and horizontal flip play a relatively prominent role. The most overlap values are

Table 12 Iris PAD performance (%) reported in terms of APCER, BPCER and ACER in comparison with the SoTAs

Algorithms	Clarkson			Notre Dame			IIITD-WVU		
	APCER	BPCER	ACER	APCER	BPCER	ACER	APCER	BPCER	ACER
CASIA [42]	9.61	5.65	7.63	11.33	7.56	9.45	23.16	16.10	19.63
Anon1 [42]	15.54	3.64	9.59	7.78	0.28	4.03	29.40	**3.99**	16.70
UNINA [42]	13.39	0.81	7.10	25.44	0.33	12.89	23.18	35.75	29.44
Meta-Fusion [27]	18.66	0.24	9.45	4.61	1.94	3.28	12.32	17.52	14.92
Scracth D-NetPAD [31]	5.78	0.94	3.36	10.38	3.23	6.81	36.41	10.12	23.27
MLF [13]	–	–	–	**2.71**	1.89	**2.31**	5.39	24.79	15.09
MSA [14,17]	–	–	–	12.28	**0.17**	6.23	**2.31**	19.94	11.13
ResNet50 (Best)	7.35	2.89	5.12	14.78	**0.17**	7.47	14.68	5.41	**10.05**
VGG16 (Best)	10.53	1.08	5.81	22.28	1.00	11.64	16.97	20.09	18.53
MobileNet (Best)	**1.67**	**0.00**	**0.84**	18.00	2.56	10.28	3.62	17.95	10.79

Our best results for each model are compared with the winners of Iris-LivDet-2017 competition [42]: CASIA, Anon1, UNINA and two state-of-the-art algorithms: Meta-Fusion [27] and D-NetPAD [31]. The bold values are the lowest error rates in each column

between 0.2 and 0.7 in confusion matrix plots. In general, the lower overlap rates indicate that different augmentation techniques enhance the model to adapt to different variations in iris samples. As shown in Figs. 6, 7 and 8, we can find that the overlap misclassification rate on MobileNetV3 network is lower (lighter blue) compared to the ResNet50 and VGG16 for each dataset. A general observation can be made from Figs. 6, 7 and 8, the fusion of multiple augmentation techniques (all or by our proposed augmentation selection protocol), especially on the score-level (SC and LO_{SC}), leads to higher overlap with the basic augmentation methods. This indicates our success in addressing a larger number of variations in the data simultaneously. This is not the case when we apply the strategy-fusion method, as the multiple augmentation methods used in the training phase might cause confusion.

Summing up all the results, we can see that training with augmentation techniques significantly improves PAD performance than training only with original data. Each augmentation method plays a positive role on a particular dataset or network. Shift augmentation performs better than other methods in most cases. However, the results do not exhibit an exact consistency across all networks, augmentation techniques, and datasets. One improvement can be to preserve the created images in the memory rather than randomly augment and fed them to the network during the training process The advantage is the later exact knowing numbers of original and augmented data, whereas the drawback is the higher hardware requirements. The data augmentation techniques are classed into two general categories, data warping, and oversampling. Because the images generated by oversampling methods like using the GAN network should be detected as attack images, this could easily exacerbate the imbalance in the data. For iris PAD, only data warping can be applied to augment the training data. However, there is no consensus about the best augmentation strategy, especially no best combination way. The reason is that the intrinsic bias in the capture environment, subject population, or scale of datasets is different. Consequently, the first future work is to learn an optimal augmentation strategy in an automatic way. Also, we need to find an optimal dataset size after augmentation by balancing the used strategy and the available memory for storing augmented images. Moreover, the imbalance between bona fide and attack samples can be addressed.

Table 13 Iris PAD performance reported in terms of D-EER (%), ACER (%) and FDR on cross-dataset scenarios

Train	Test	Method	ResNet50			VGG16			MobileNetV3		
			D-EER (%)	ACER (%)	FDR	D-EER (%)	ACER (%)	FDR	D-EER (%)	ACER (%)	FDR
Clarkson	Notre Dame	No	10.89	14.22	**4.11**	16.67	20.97	1.69	46.72	47.36	0.01
		BS	9.00	**14.11**	3.64	**13.72**	**16.50**	**2.67**	100.00	50.11	0.00
		ST	**6.89**	**7.89**	**7.08**	23.67	28.78	0.83	**31.77**	**40.33**	**0.23**
		SC	**8.67**	29.25	1.47	18.29	18.81	1.91	44.94	48.94	0.02
		LO_{ST}	19.27	19.80	1.54	**16.38**	**16.28**	**2.32**	**29.27**	**45.14**	**0.16**
		LO_{SC}	9.17	26.17	1.61	18.11	18.17	2.01	42.72	50.03	0.07
	IIITD-WVU	No	**15.94**	**20.18**	**2.02**	26.49	29.65	0.72	100.00	46.08	**0.08**
		BS	17.51	**19.71**	1.55	21.80	**26.56**	**0.97**	100.00	49.97	0.00
		ST	23.43	22.69	0.92	25.21	32.68	0.64	**51.51**	**44.91**	**0.08**
		SC	**14.23**	27.30	1.50	**20.65**	29.47	0.92	59.17	**48.99**	0.04
		LO_{ST}	31.40	32.48	0.46	22.36	31.35	0.71	**58.12**	49.75	0.02
		LO_{SC}	17.00	20.43	**1.98**	**20.58**	**29.04**	**0.94**	61.17	50.02	0.04
Notre Dame	Clarkson	No	16.11	18.49	2.09	**14.15**	**12.88**	**3.04**	**40.09**	41.92	0.07
		BS	14.94	13.81	2.60	15.64	16.26	2.31	42.02	**41.18**	**0.08**
		ST	13.07	12.84	2.37	14.72	**14.18**	1.62	42.37	**40.29**	**0.11**
		SC	**11.27**	**12.12**	**3.50**	15.23	15.89	2.45	48.29	48.79	0.00
		LO_{ST}	21.30	35.92	0.44	16.78	16.29	2.24	**41.23**	49.47	0.05
		LO_{SC}	**8.64**	**10.58**	**4.60**	**14.41**	15.99	**2.54**	47.27	56.37	0.05
	IIITD-WVU	No	**22.35**	31.65	0.77	12.88	34.21	0.84	11.13	10.78	3.97
		BS	30.55	26.47	0.92	14.17	35.30	0.73	15.86	15.65	2.25
		ST	25.29	**24.74**	0.92	11.81	**30.57**	**1.12**	**10.39**	**10.21**	**4.29**
		SC	**21.25**	**21.89**	**1.34**	**11.11**	33.16	1.02	**9.42**	**8.82**	**5.26**
		LO_{ST}	25.09	25.46	1.05	**10.39**	**31.65**	1.06	12.24	12.47	3.61
		LO_{SC}	23.51	25.54	**1.23**	11.29	34.44	0.95	12.52	13.21	3.73
IIITD-WVU	Clarkson	No	**15.58**	37.14	0.55	23.27	26.26	0.90	45.72	50.54	0.00
		BS	19.34	**21.98**	**1.60**	**15.26**	**18.29**	**1.87**	**37.43**	**41.03**	**0.10**
		ST	31.79	30.63	0.58	22.10	22.87	1.15	**36.35**	47.67	**0.08**
		SC	**15.07**	**22.68**	**1.96**	**15.80**	**18.59**	**1.85**	51.42	50.39	0.00
		LO_{ST}	16.84	35.56	0.60	19.31	19.51	1.56	41.29	**44.41**	0.06
		LO_{SC}	17.32	23.79	1.57	16.18	18.96	1.75	51.29	48.18	0.01
	Notre Dame	No	12.44	12.28	3.47	**17.17**	**20.22**	**1.53**	32.28	37.36	0.23
		BS	**7.38**	**7.36**	**7.30**	17.56	24.67	1.14	33.11	34.86	0.30
		ST	13.06	15.19	2.56	**13.77**	**18.53**	**2.12**	**23.06**	31.63	**0.56**
		SC	8.06	11.47	4.74	17.22	23.11	1.33	31.56	**31.61**	0.55
		LO_{ST}	21.55	22.03	1.32	21.39	26.50	0.92	30.33	**29.25**	**0.61**
		LO_{SC}	**7.83**	**10.14**	**4.97**	17.22	23.52	1.31	**27.83**	35.03	0.55

The bold values in *D-EER* and *ACER* column are the Top-2 lowest error rate, and the bold values in the *FDR* column indicate the Top-2 highest separability. The ACER value is determined by a threshold of 0.5

6 Conclusion

This paper addresses a clear research gap by providing an in-depth analysis of the data augmentation role in iris PAD. Data augmentation technique is one of the crucial steps to address the limitation of iris attack data. We explore the impact of widely used data augmentation strategies and two combination methods, strategy-level, and score-level,

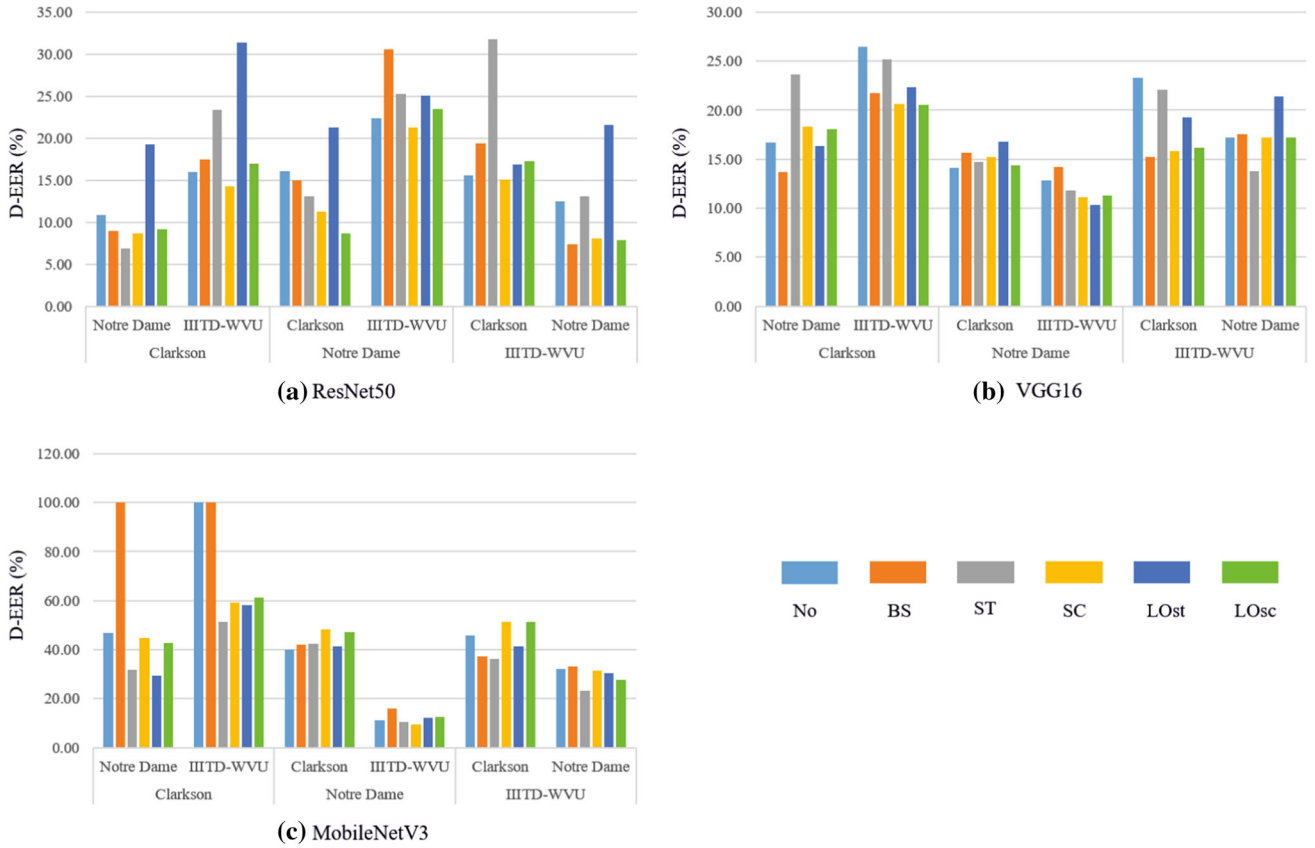

Fig. 4 The histogram of performance on cross-dataset evaluations. The x-axis the different experimental settings, and the y-axis is the D-EER (%) value

on the generalization of iris PAD. We also propose a least overlap-based augmentation selection protocol to bring different types of wrongly classified samples into the correct classification. This is based on a detailed analysis of the overlap between the effect of different augmentation techniques. The experiments are performed on three datasets in the Iris-LivDet-2017 competition [42] and with three neural networks for comparison and analysis. The experimental results linked certain data augmentation methods to significant enhancement of generalizability and indicated the relatively low-overlapping effect of these augmentations.

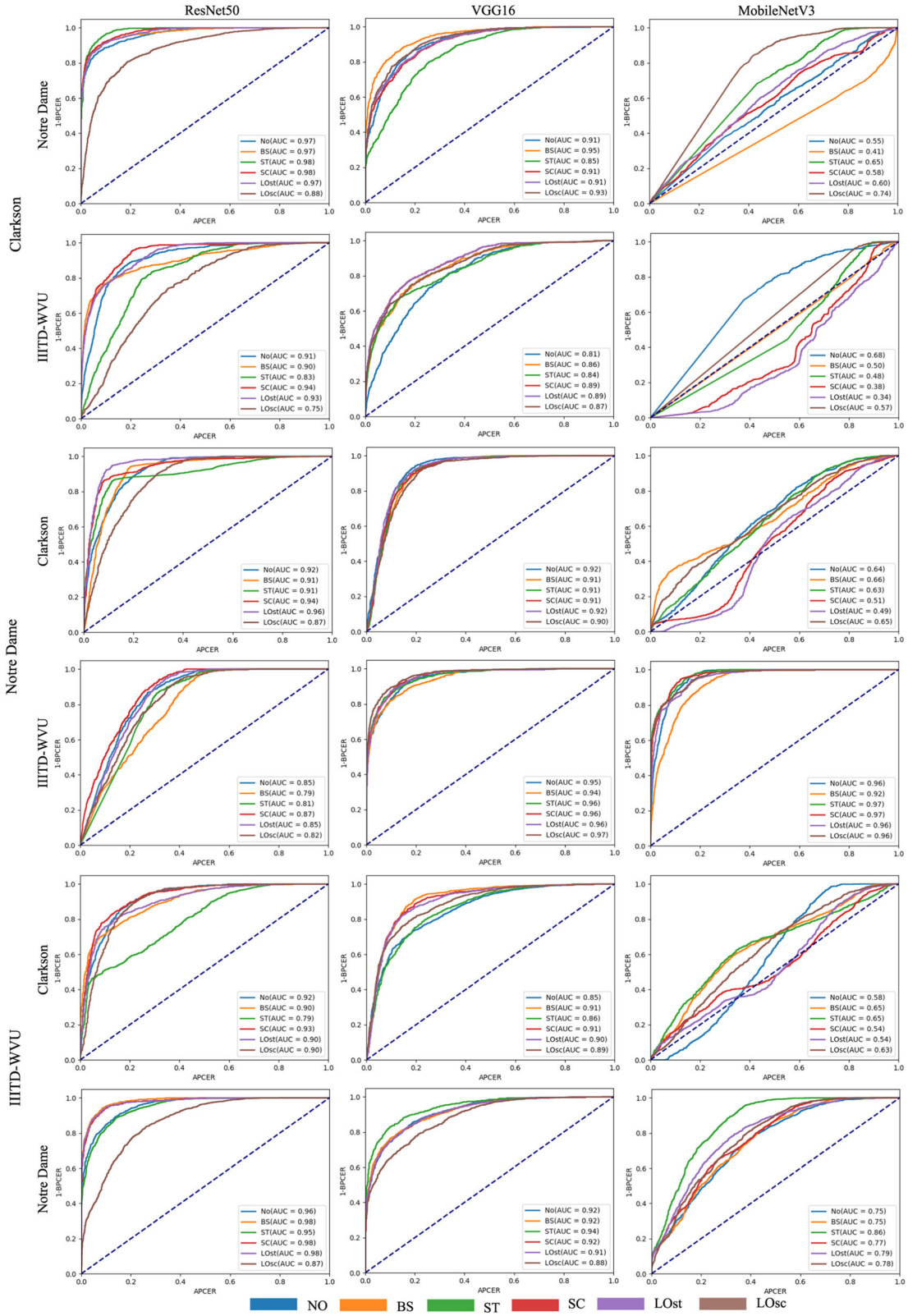

Fig. 5 ROC curves of cross-dataset evaluations (corresponding to Table 13. The rows from top to bottom are for training on Clarkson and testing on Notre Dame and IIITW-WVU case, then training on Notre Dame and testing on Clarkson and IIITD-WVU, and training on IIITD-WVU and testing on Clarkson and Notre Dame cases. The columns from left to right are ResNet50, VGG16, and MobileNetV3 networks. The row and column order are the same as in Table 13). The x-axis in each ROC plot is the APCER values, and the y-axis is the $1 - BPCER$ values

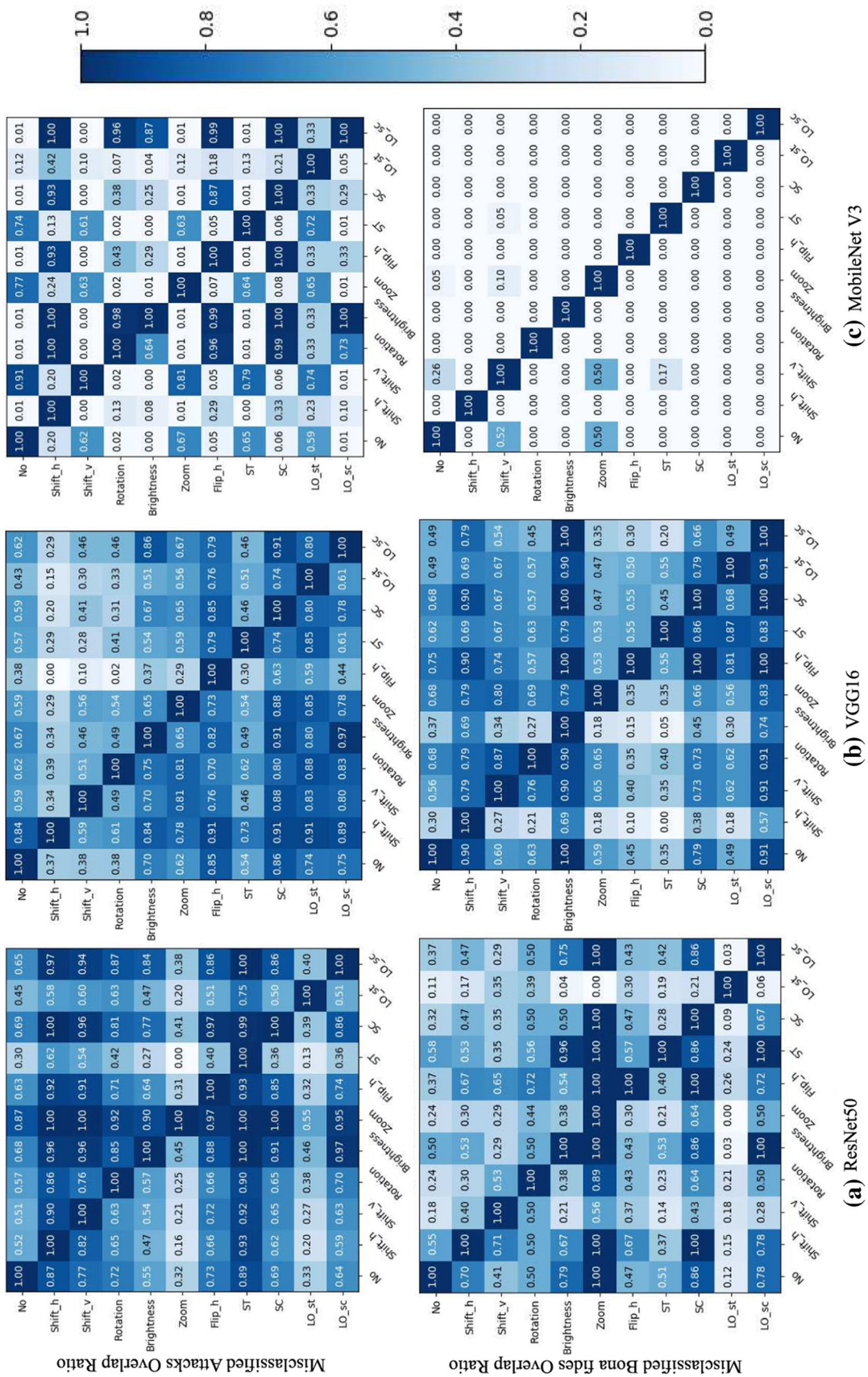

Fig. 6 Overlap confusion matrix for the Clarkson dataset. The horizontal axis (X axis) from left to right and the vertical axis (Y axis) from top to bottom correspond to the augmentation strategies: No, Shift_h, Shift_v, Rotation, Brightness, Zoom, Flip_h, ST, SC, LO_{ST} and LO_{SC}, respectively. The values in top matrices are the misclassified attacks overlap ratio $O^a_{A_{pq}}$ computed as in Eq. (1a), and the values in the bottom matrices are the misclassified bona fides overlap ratio $O^{bf}_{A_{pq}}$ computed from Eq. (1b)

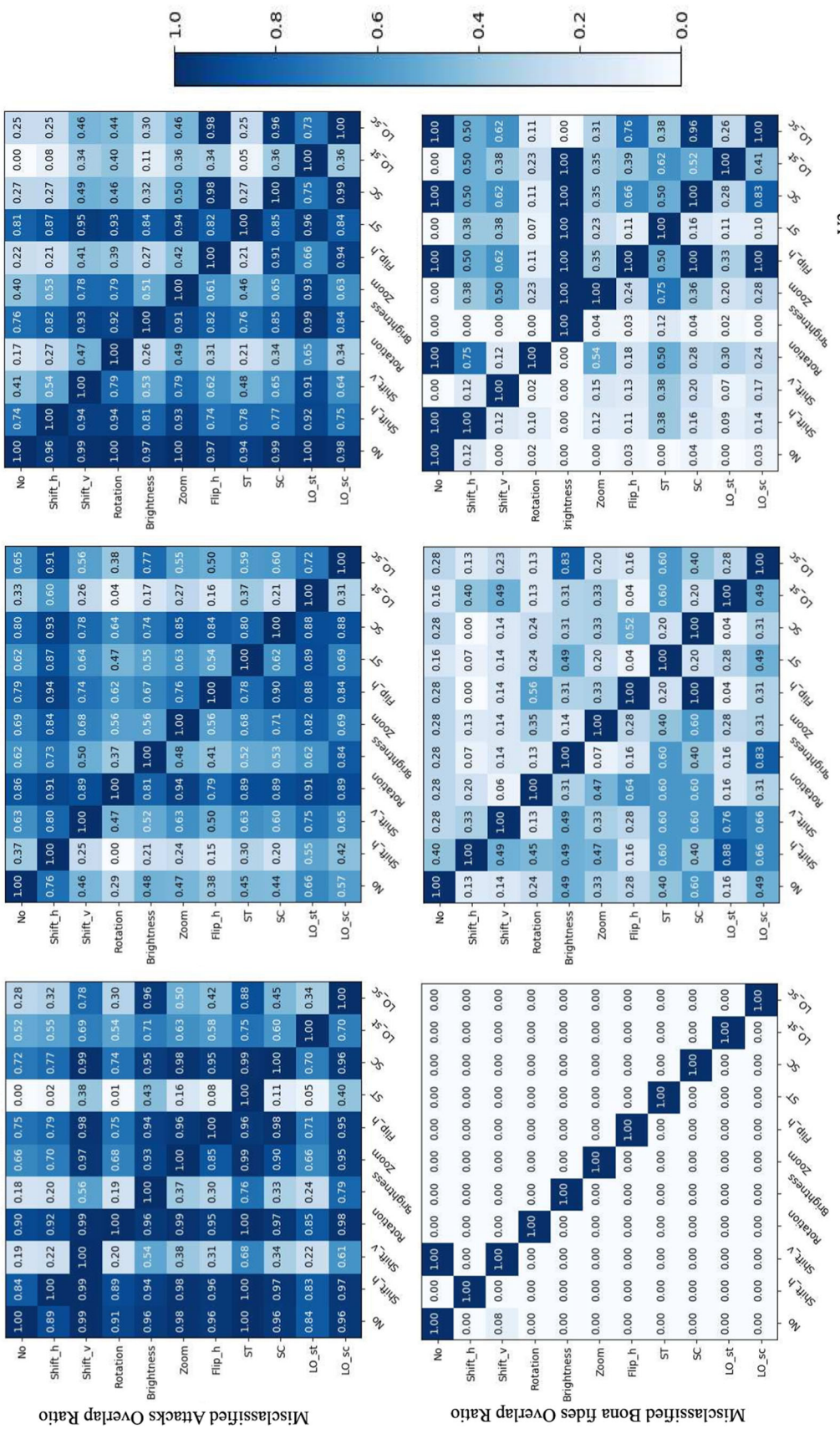

Fig. 7 Overlap confusion matrix for the Notre Dame dataset. The horizontal axis (X axis) from left to right and the vertical axis (Y axis) from top to bottom correspond to the augmentation strategies: No, Shift_h, Shift_v, Rotation, Brightness, Zoom, Flip_h, ST, SC, LO_{ST} and LO_{SC}, respectively. The values in top matrices are the misclassified attacks overlap ratio $O^a_{A_{pq}}$ computed as in Eq. (1a), and the values in the bottom matrices are the misclassified bona fides overlap ratio $O^{bf}_{A_{pq}}$ computed from Eq. (1b)

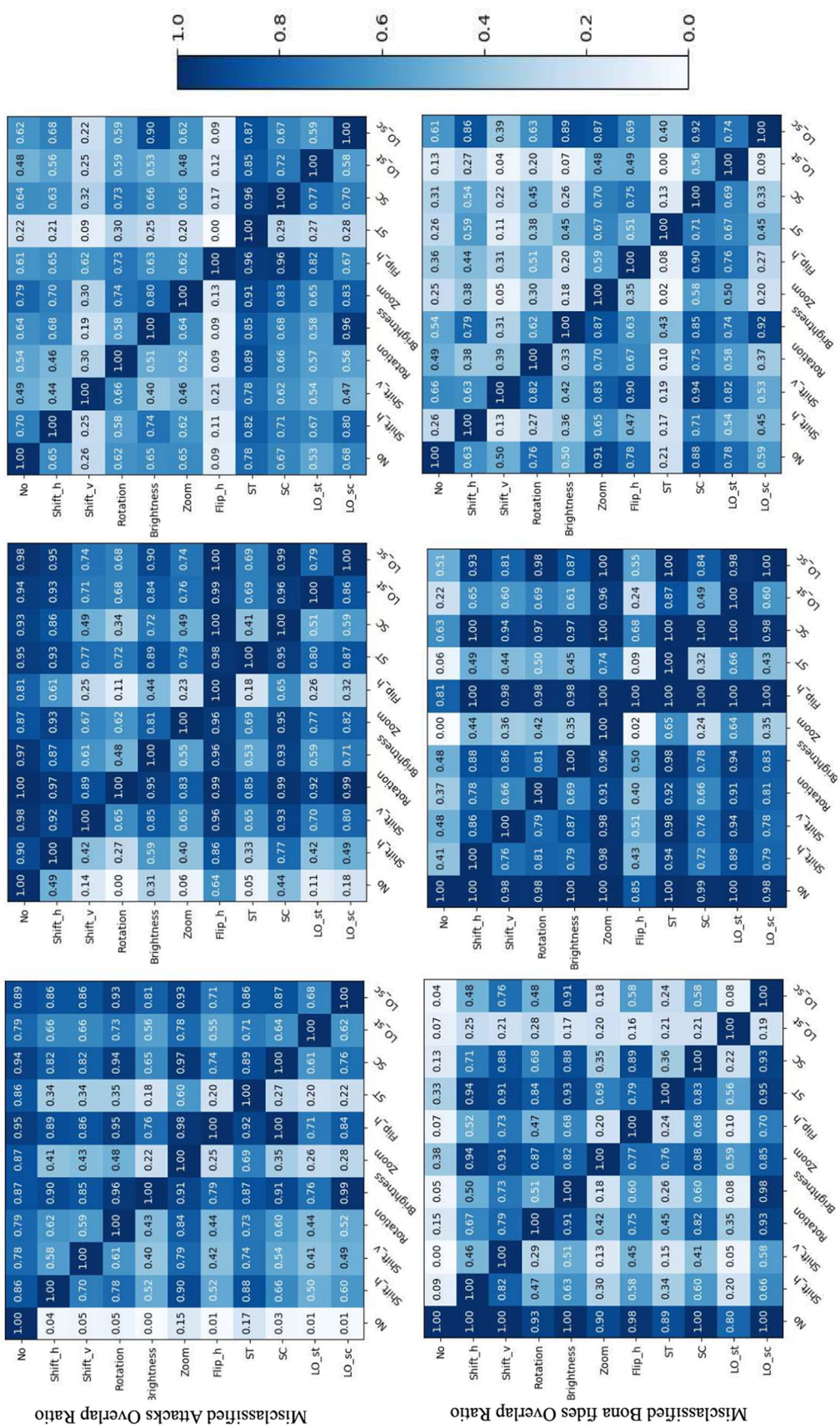

Fig. 8 Overlap confusion matrix for the IIITD-WVU dataset. The horizontal axis (X axis) from left to right and the vertical axis (Y axis) from top to bottom correspond to the augmentation strategies: No, Shift_h, Shift_v, Rotation, Brightness, Zoom, Flip_h, ST, SC, LO_{ST} and LO_{SC}, respectively. The values in top matrices are the misclassified attacks overlap ratio $O^a_{A_{pq}}$ computed as in Eq. (1a) and the values in the bottom matrices are the misclassified bona fides overlap ratio $O^{bf}_{A_{pq}}$ computed from Eq. (1b)

References

1. Baird, H.S.: Document image defect models and their uses. In: In: 2nd International Conference Document Analysis and Recognition, pp. 62–67. IEEE Computer Society, Tsukuba City, Japan (1993)
2. Bakshi, S., Mehrotra, H., Majhi, B.: Postmatch pruning of SIFT pairs for iris recognition. Int. J. Biom. **5**(2), 160–180 (2013). https://doi.org/10.1504/IJBM.2013.052965
3. Barpanda, S.S., Sa, P.K., Marques, O., Majhi, B., Bakshi, S.: Iris recognition with tunable filter bank based feature. Multim. Tools Appl. **77**(6), 7637–7674 (2018). https://doi.org/10.1007/s11042-017-4668-z
4. Boutros, F., Damer, N., Raja, K.B., Ramachandra, R., Kirchbuchner, F., Kuijper, A.: Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. Image Vis. Comput. **104**, 104007 (2020)
5. Boutros, F., Damer, N., Raja, K.B., Ramachandra, R., Kirchbuchner, F., Kuijper, A.: In: IJCB, (ed.) On benchmarking iris recognition within a head-mounted display for AR/VR applications, pp. 1–10. IEEE (2020)
6. Chen, C., Ross, A.: A multi-task convolutional neural network for joint iris detection and presentation attack detection. In: 2018 IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2018, Lake Tahoe, NV, USA, March 15, 2018, pp. 44–51. IEEE Computer Society (2018). https://doi.org/10.1109/WACVW.2018.00011
7. Choudhary, M., Tiwari, V., Uduthalapally, V.: Iris presentation attack detection based on best-k feature selection from yolo inspired roi. In: Neural Comput and Applic (2020)
8. Czajka, A., Bowyer, K.W.: Presentation attack detection for iris recognition: An assessment of the state-of-the-art. ACM Comput. Surv. **51**(4), 86:1-86:35 (2018). https://doi.org/10.1145/3232849
9. Damer, N., Opel, A., Nouak, A.: Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In: 22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014, pp. 1382–1386. IEEE (2014)
10. Dao, T., Gu, A., Ratner, A., Smith, V., Sa, C.D., Ré, C.: A kernel theory of modern data augmentation. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, *Proceedings of Machine Learning Research*, vol. 97, pp. 1528–1537. PMLR (2019)
11. Das, P., McGrath, J., Fang, Z., Boyd, A., Jang, G., Mohammadi, A., Purnapatra, S., Yambay, D., Marcel, S., Trokielewicz, M., Maciejewicz, P., Bowyer, K.W., Czajka, A., Schuckers, S., Tapia, J.E., Gonzalez, S., Fang, M., Damer, N., Boutros, F., Kuijper, A., Sharma, R., Chen, C., Ross, A.: Iris liveness detection competition (livdet-iris) - the 2020 edition. In: 2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020, pp. 1–9. IEEE (2020). https://doi.org/10.1109/IJCB48548.2020.9304941
12. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255. IEEE Computer Society (2009)
13. Fang, M., Damer, N., Boutros, F., Kirchbuchner, F., Kuijper, A.: Deep learning multi-layer fusion for an accurate iris presentation attack detection. In: IEEE 23rd International Conference on Information Fusion, FUSION 2020, Rustenburg, South Africa, July 6-9, 2020, pp. 1–8. IEEE (2020). https://doi.org/10.23919/FUSION45008.2020.9190424
14. Fang, M., Damer, N., Boutros, F., Kirchbuchner, F., Kuijper, A.: Cross-database and cross-attack iris presentation attack detection using micro stripes analyses. Image Vis. Comput. **105**, 104057 (2021). https://doi.org/10.1016/j.imavis.2020.104057
15. Fang, M., Damer, N., Boutros, F., Kirchbuchner, F., Kuijper, A.: Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In: 2021 IEEE International Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, Aug.4 - 7, 2021. IEEE (2021)
16. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Demographic bias in presentation attack detection of iris recognition systems. In: 28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021, pp. 835–839. IEEE (2020). https://doi.org/10.23919/Eusipco47968.2020.9287321
17. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Micro stripes analyses for iris presentation attack detection. In: 2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020, pp. 1–10. IEEE (2020). https://doi.org/10.1109/IJCB48548.2020.9304886
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014). http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
19. Gragnaniello, D., Sansone, C., Poggi, G., Verdoliva, L.: Biometric spoofing detection by a domain-aware convolutional neural network. In: 2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS), pp. 193–198 (2016). https://doi.org/10.1109/SITIS.2016.38
20. Gupta, P., Behera, S., Vatsa, M., Singh, R.: On iris spoofing using print attack. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014, pp. 1681–1686. IEEE Computer Society (2014). https://doi.org/10.1109/ICPR.2014.296
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE CVPR, Las Vegas, NV, USA, June 27-30, 2016, pp. 770–778. IEEE Computer Society (2016)
22. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 1314–1324. IEEE (2019)
23. Hu, B., Lei, C., Wang, D., Zhang, S., Chen, Z.: A preliminary study on data augmentation of deep learning for image classification. CoRR arXiv:1906.11887 (2019)
24. International Organization for Standardization: ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting (2017)
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: F. Bach,

D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 448–456. PMLR, Lille, France (2015)
26. Kohli, N., Yadav, D., Vatsa, M., Singh, R., Noore, A.: Synthetic iris presentation attack using idcgan. In: 2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017, pp. 674–680. IEEE (2017). https://doi.org/10.1109/BTAS.2017.8272756
27. Kuehlkamp, A., Pinto, A., Rocha, A., Bowyer, K.W., Czajka, A.: 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)
28. Lorena, A.C., de Leon Ferreira de Carvalho, A.C.P.: Building binary-tree-based multiclass classifiers using separability measures. Neurocomputing **73**(16–18), 2837–2845 (2010). https://doi.org/10.1016/j.neucom.2010.03.027
29. Nguyen, D.T., Pham, T.D., Lee, Y., Park, K.R.: Deep learning-based enhanced presentation attack detection for iris recognition by combining features from local and global regions based on NIR camera sensor. Sensors **18**(8), 2601 (2018)
30. Raghavendra, R., Raja, K.B., Busch, C.: Contlensnet: Robust iris contact lens detection using deep convolutional neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017, pp. 1160–1167. IEEE Computer Society (2017). https://doi.org/10.1109/WACV.2017.134
31. Sharma, R., Ross, A.: D-netpad: An explainable and interpretable iris presentation attack detector. 2020 IJCB, Sep. 28 - Oct. 1, 2020, online conference arXiv: 2007.01381 (2020)
32. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Y. Bengio, Y. LeCun (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
35. Thanh, K.N., Fookes, C., Ross, A., Sridharan, S.: Iris recognition with off-the-shelf CNN features: A deep learning perspective. IEEE Access **6**, 18848–18855 (2018). https://doi.org/10.1109/ACCESS.2017.2784352
36. Tolosana, R., Gomez-Barrero, M., Busch, C., Ortega-Garcia, J.: Biometric presentation attack detection: Beyond the visible spectrum. IEEE Trans. Information Forensics and Security **15**, 1261–1275 (2020)
37. Wei, Z., Tan, T., Sun, Z.: Synthesis of large realistic iris databases using patch-based sampling. In: 19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA, pp. 1–4. IEEE Computer Society (2008). https://doi.org/10.1109/ICPR.2008.4761674
38. Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding data augmentation for classification: When to warp? In: 2016 International Conference on DICTA, 2016, Gold Coast, Australia, November 30 - December 2, 2016, pp. 1–6. IEEE (2016)
39. Yadav, D., Kohli, N., Agarwal, A., Vatsa, M., Singh, R., Noore, A.: Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 572–579. IEEE Computer Society (2018). https://doi.org/10.1109/CVPRW.2018.00099
40. Yadav, D., Kohli, N., Jr., J.S.D., Singh, R., Vatsa, M., Bowyer, K.W.: Unraveling the effect of textured contact lenses on iris recognition. IEEE Trans. Inf. Forensics Secur. **9**(5), 851–862 (2014). https://doi.org/10.1109/TIFS.2014.2313025
41. Yadav, S., Chen, C., Ross, A.: Synthesizing iris images using rasgan with application in presentation attack detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 2422–2430. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPRW.2019.00297
42. Yambay, D., Becker, B., Kohli, N., Yadav, D., Czajka, A., Bowyer, K.W., Schuckers, S., Singh, R., Vatsa, M., Noore, A., Gragnaniello, D., Sansone, C., Verdoliva, L., He, L., Ru, Y., Li, H., Liu, N., Sun, Z., Tan, T.: Livdet iris 2017 - iris liveness detection competition 2017. In: 2017 IEEE IJCB, Denver, CO, USA, October 1-4, 2017, pp. 733–741. IEEE (2017)

Meiling Fang received her masters degree from Karlsruhe Institute of Technology (2019). Since August 2019, she is a researcher at Fraunhofer IGD. Her research interests are in the fields of machine learning, computer vision, and biometrics.

Naser Damer is a senior researcher at Fraunhofer IGD. He received his PhD from TU Darmstadt (2018). He is a principal investigator at the ATHENE Center, serves as an AE of TVCJ, lectures at TU Darmstadt, and represents the German Institute for Standardization (DIN) in ISO/IEC SC37 standardization committee.

Fadi Boutros received his masters degree from TU Darmstadt (2019). He was a student research assistant and later worked on his masters thesis at Fraunhofer IGD. Since March 2019, he is a researcher at Fraunhofer IGD. His research interests are in the fields of machine learning, computer vision, and biometrics.

Florian Kirchbuchner received his masters degree from TU Darmstadt in 2014. Since 2018, he has been Head of the Department for Smart Living & Biometric Technologies at the IGD. Mr. Kirchbuchner has been the spokesperson for the Fraunhofer Alliance Ambient Assisted Living AAL since 2019.

Arjan Kuijper is a member of the management of Fraunhofer IGD and is responsible for scientific dissemination. He holds the Chair in mathematical and applied visual computing with TU Darmstadt. He is the author of over 300 peer-reviewed publications, the AE of CVIU, PR, and TVCJ, and the Secretary of the IAPR.