RESEARCH ARTICLE

One-Shot Learning for Periocular Recognition: Exploring the Effect of Domain Adaptation and Data Bias on Deep Representations

KEVIN HERNANDEZ-DIAZ, (Student Member, IEEE),
FERNANDO ALONSO-FERNANDEZ, (Member, IEEE),
AND JOSEF BIGUN, (Fellow, IEEE)
Department of Intelligent Systems and Digital Design (ISDD), Halmstad University, 302 61 Halmstad, Sweden

Corresponding author: Kevin Hernandez-Diaz (kevin.hernandez-diaz@hh.se)

ABSTRACT One weakness of machine-learning algorithms is the need to train the models for a new task. This presents a specific challenge for biometric recognition due to the dynamic nature of databases and, in some instances, the reliance on subject collaboration for data collection. This paper investigates the behavior of deep representations in widely used Convolutional Neural Network (CNN) models under extreme data scarcity for One-Shot periocular recognition, a biometric recognition task. We analyze the outputs of CNN layers as identity-representing feature vectors. We examine the impact of Domain Adaptation on the network layers' output for unseen data and evaluate the method's robustness concerning data normalization and generalization of the best-performing layer. We improved state-of-the-art results that made use of networks trained with biometric datasets with millions of images and fine-tuned for the target periocular dataset by utilizing out-of-the-box CNNs trained for the ImageNet Recognition Challenge and standard computer vision algorithms. For example, for the Cross-Eyed dataset, we reduced the EER by 67% and 79% (from 1.70% and 3.41% to 0.56% and 0.71%) in the Close-World and Open-World protocols, respectively, for the periocular case. We also demonstrate that traditional algorithms like SIFT can outperform CNNs in situations with limited data or scenarios where the network has not been trained with the test classes like the Open-World mode. SIFT alone was able to reduce the EER by 64% and 71.6% (from 1.7% and 3.41% to 0.6% and 0.97%) for Cross-Eyed in the Close-World and Open-World protocols, respectively, and a reduction of 4.6% (from 3.94% to 3.76%) in the PolyU database for the Open-World and single biometric case.

INDEX TERMS Biometrics, deep representation, periocular, transfer learning, one-shot learning.

I. INTRODUCTION

Convolutional Neural Networks (CNNs) have become the standard increasingly in applications of Computer Vision and Pattern Recognition. From object detection [1], [2] to object recognition [3], data generation [4], image manipulation [5], CNNs dominate the state-of-the-art. The popularity and success of CNNs largely stem from their ability to learn

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

and extract highly discriminative features, as well as to easily adapt to different applications such as medical data [6], autonomous driving [7], or, in our case, biometric recognition [8].

Nonetheless, to achieve good results, CNNs usually require a substantial amount of varied data to allow the network to learn the abstraction of objects [9]. Since acquiring such data is often expensive and infeasible, many researchers are working to make CNNs more efficient [10]. Transfer Learning is one of the most common approaches to tackling

data scarcity. It aims to adapt a network trained for a usually more complex task for which much more training data exists to a new target domain. The idea is to take advantage of the feature extraction power of the pre-trained CNN and fine-tune it for the specific task under consideration. One-shot learning is an extreme case of Transfer Learning, where no data is available to train the network for the new target. Instead, a vector of embedding, or deep representation, is extracted from a class-sample image using a pre-trained network for comparison. Then, distance or similarity-based metrics between the deep representations are used to determine if a new image belongs to the same class. Typically, the last layer before the classification stage is used to extract such deep representations. However, as this paper and previous preliminary studies on periocular recognition show [11], [12], selecting the final layer of the network may not always be the best option. Moreover, as we also study here, the best layer selection depends heavily on the input data normalization as well as the amount and variety of data available when training the model.

With the need for One-Shot Learning appeared newer approaches like the use of Contrastive-Loss [13] and Triplet-Loss [14]. In these types of losses, the network is optimized to extract a vector of embedding that maximizes the inter-class distance and, in the case of Triplet-Loss, also minimizes the intra-class distance up to some margin. The approaches for One-Shot Learning explained in this section can, once the network has been trained on a large dataset, be used directly on other datasets for recognition. For example, to use a VGG-Face [15] network directly on a target Face dataset.

The eye region is one of the most discriminative areas of the face [16], [17]. However, it was not until 2009 when [18] first introduced the concept of periocular recognition. They described this new biometric as using the facial area in the immediate vicinity of the eye to recognize a person's identity. Besides the iris, a well-established biometric trait [19], the eye's shape, texture, and subcomponents, like eyebrows, eyelids, commissures, or skin, provide much information that one can exploit to recognize a person. This periocular area has proven to achieve high recognition performance [20], not only for identities but also for soft-biometric traits like gender, ethnicity, and age [21], [22] [23], [24], [25], while having fewer acquisition constraints than other ocular modalities like the iris. However, despite its potential, large periocular datasets are scarce [26], leading to limited research in this area. Nonetheless, due to the recent COVID-19 pandemic and the widespread use of face masks, this region has gained significant attention within the biometric community [27].

In a previous contribution [11], we evaluated a selection of CNNs for One-shot periocular recognition on the UBIPr database. Later in [28], we also analyzed the utility of a pre-trained CNN for few-shot cross-spectral periocular recognition on the IMP database. Here, we evaluate a wider selection of networks and databases. We also analyze the effect of other factors, such as image pre-processing, domain

adaptation, and data partition. Our contributions are shown below:

- State-of-the-art (SOA) periocular recognition comparison using One-Shot Learning. We report the performance per layer of six widely used CNN architectures (ResNet101v2, DenseNet121, VGG19, Inceptionv3, MobileNetv2, and Xception), as well as the most widely used hand-crafted features in periocular recognition (LBPH, HOG, SIFT) for three different datasets (IMP, PolyU, Cross-Eyed).
- We investigate the effect that Domain Adaptation had on the selected networks' deep representation when we used CNNs trained for the periocular modality for Cross-Dataset recognition. We consider the following scenarios: CNNs pretrained with the ImageNet dataset, ImageNet CNNs fine-tuned for periocular recognition with auxiliary datasets, randomly initialized CNNs, and finally, randomly initialized CNNs trained for periocular recognition with auxiliary datasets.
- We examined how the acquisition method and input image preprocessing can affect the performance of Deep Representations and best-performing layers.
- Finally, we also report the generalizability of the best layer found by showcasing how performance varies when using the best-layer information between datasets and same-dataset partitions on the Open-World (OW) and Close-World (CW) cases.

The rest of the paper is organized as follows. Section II frames our work within the related research area of periocular biometrics. Section III presents the databases, metrics, and matching strategies used. Section IV explains the paper's experimental framework. Section V shows the results, presents relevant findings, and compares our methods with the state-of-the-art for the databases used. Finally, Section VI draws the final conclusions obtained from our research.

II. RELATED WORK

This section surveys Deep Learning biometric recognition, focusing particularly on periocular biometrics and One-Shot Learning.

This paper extends two previous works [11], [28] that dealt with deep representations for periocular recognition. In [11], we compared the performance per layer of deep representations from a selection of four well-known architectures pre-trained on ImageNet or face recognition databases (AlexNet, GoogLeNet/Inception v1, ResNet, and VGG) and the results with traditional computer vision hand-crafted features. The data employed consisted of periocular images in the visible range from the UBIPr database. This paper discovered that intermediate CNN representations of such networks could outperform traditional CV methods used in biometric recognition with no additional training needed. Furthermore, we saw that biometric-trained models like VGG-Face did not perform better than their general-purpose counterparts trained for the ImageNet challenge.
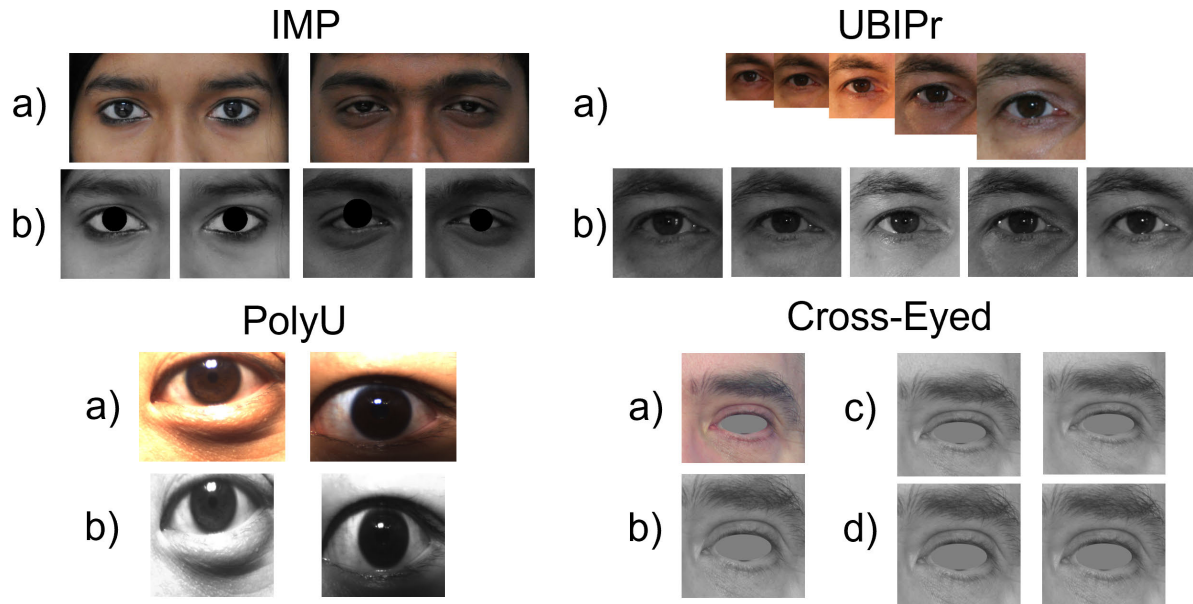
FIGURE 1. Image samples from each database. a) Original database samples, UBIPr shows the relative difference size between image samples b) normalized images after pre-processing as explained in III-A. Cross-Eyed c) and d) show the normalization difference used in Section V-B. In c), the only modifications to the images were the conversion to grayscale and cropped to be squared, while d) shows the full normalization effect as explained in III-A.

In [28], we extended the work for cross-spectral periocular recognition using the IMP database, which contains images with three different types of illumination: visible, near-infrared, and night vision. We first analyzed the changes in the performance per layer, observing that the optimal layer is different for each spectrum. We later investigated how intermediate representations could be used for cross-spectral purposes. We found that cross-spectral performance could be improved by training a fully connected network at the end of the best-performing layer of each spectrum, thus demanding a small fine-tuning step only. The cross-spectral performance was observed to improve largely, up to 65% (EER) and 87% (accuracy at 1% FAR) w.r.t. previous papers, constituting the best-published results to date on the IMP database. The work [28] focused on one CNN only (ResNet), while the present paper extends the study of the cross-spectral issue to a selection of six different architectures.

The mentioned periocular research of [11] and [28] is inspired by [29], where the authors studied the per-layer performance in iris recognition of five different ImageNet pre-trained CNNs (AlexNet, VGG, Inception, ResNet, and DenseNet). The segmented and normalized iris image is given to the CNN. The intermediate layers' output at different depths is then extracted and used as feature representations to feed an SVM for identity classification. The paper achieved state-of-the-art recognition performance on two large iris databases, LG2200 (ND-CrossSensor-Iris-2013) and CASIA-Iris-Thousand. The authors concluded that the employed Off-The-Shelf CNNs can extract rich features from iris images that could be used for recognition, thus reducing the complexity of using CNNs for the task by not having to

train them, opening the door to new iris representations. In our previous papers [11], [28], and in the present contribution, we also follow this direction for the periocular modality.

CNNs pretrained on large image datasets such as ImageNet, MS1M, and VGG-Face have been widely used as the backbone of many architectures in the literature. In [21], the authors use a frozen VGG16 trained on ImageNet with its Fully Connected layers discarded as the backbone architecture to extract periocular features later used for person recognition, for soft biometric classification, and both together in a Joint Periocular Recognition Block. They improved the SOA for periocular recognition on both UBIRISV2 and FRGC datasets, as well as the soft-biometric classification on FRGC.

In [30], the authors used a VGG16 trained for face recognition using the VGG-Face dataset, a dataset with 2,6M images and 2,622 identities, and fine-tuned the network for periocular recognition while controlling the size of the final feature vector. Once the network was adapted to the new domain, the last layers were removed, and the recognition was made by comparing the deep representations of the test images using the Euclidean distance, Spearman distance, or Cosine similarity. They demonstrated the feasibility of using the periocular area in unconstrained scenarios by achieving SOA on NICE.II and MobBIO, two datasets with images captured in uncontrolled environments in the visible spectrum. In another study [31], the authors used a similar approach to analyze the effect that iris normalization and segmentation have on Deep Representations for biometric recognition. They used two networks (VGG and ResNet50) trained for face recognition and fine-tuned them for iris

recognition by removing the last layer and incorporating two new fully connected layers. Once the training was complete, they removed the last classification layer and used the Cosine similarity between the deep representations for biometric verification, reaching a new SOA for the NICE.II dataset.

The authors of [32] used a One-Shot learning approach to extract a vector of embeddings for joint biometric and sensor recognition. They extracted the images' deep representations using an embedding network that was trained using one of three different types of losses: Cross-Entropy, Contrastive (single and double margin), and Triplet-Loss (with off-line and online triplet mining, as well as multi-class negative-pairs). They then extracted the vector of embedding from the final layer of the network (removing the classification layer from the Cross-entropy approach) and used it for recognition. They compared their results in three different biometric modalities: face, periocular, and iris, as well as for two different types of sensors: Near-Infrared iris sensors and smartphone cameras. They found that the representations were robust across the three biometric modalities and different sensors, outperforming SOA commercial approaches.

In the paper [33], authors proposed a method that consists of a periocular ROI detection model for image alignment, custom data augmentation, and illumination normalization to extract robust and generalizable periocular features using a MobileNetv2 network. They followed an Open-Set protocol in which they trained their models using the VISOB database on visible images and then evaluated the generalizability of their model on UBIRIS-V2, UBIPR, FERET, Cross-Eyed, CASIA-IRIS-TWINS, which includes adverse imaging environment and cross-spectral comparisons, by matching the features vectors extracted using Cosine similarity. They reduced the error rate up to 7 times when compared with existing models in the literature.

In [34], authors proposed to use an unsupervised convolutional auto-encoder to create subject-invariant feature representations for ocular recognition. For each image input, two augmented views were created and fed to the network. They used an L1 norm between the Deep Representations of both image views created by the encoder and between the original and reconstructed images by the decoder. They coupled the loss also with a KL-divergence term as a sparsity regularizer and two coefficients to weigh the contribution of the regularizer and Deep Representations to the loss. They followed an Open-Set cross-dataset evaluation protocol where they used the Cosine similarity or Hamming distance for matching. They achieved a 2.2% lower EER for cross-illumination conditions when compared to a supervised ResNet50.

III. DATABASES, METRICS, AND PROTOCOL

This section describes the databases, the matching protocol, and the metrics used to compare the results from our experiments and the baselines.

A. DATABASES

We employed images in the visible (VIS) range from four commonly used periocular datasets in the experimentation: IIITD Multispectral Periocular (IMP) [36], UBIPr [37], Cross-Eyed [38], [39] and PolyU [40].

UBIPr is a periocular database captured with a CANON EOS 5D digital camera with different degrees of subject-camera distance (4-8m), resolutions, illumination, poses, and occlusions in two separate sessions. We only kept the frontal images to match the same type of images as the other databases. In addition, we retained only the users that had two recorded sessions. Since both eyes are available per user per session, our final database has 86 individuals $\times 2$ sessions $\times 2$ eyes $\times 5$ distances $= 1,720$ images. Each eye is considered a different identity, thus having 172 identities. Furthermore, we resized the images using bicubic interpolations. We normalized them (with the annotated ground-truth used in [11]) to have the same average sclera radius in their distance group and aligned them by extracting a square region of $7.6R_s$ x $7.6R_s$ around the sclera center.

IMP is a cross-spectral periocular database. It offers images captured in three spectra: Near-Infrared (NIR), Visible (VIS), and Night Vision. The VIS images were captured using a Nikon SLR camera from a distance of 1.3m in a controlled environment and illumination. The database has 62 users with 5 images per user and spectrum containing both eye regions. We manually annotated the sclera center of each eye and the sclera radius. Then, we separated each eye and normalized the images to have the same sclera radius, and aligned them by cropping a squared region around its sclera center. The database thus has 62 users $\times 2$ eyes $\times 5$ images per eye $= 620$ VIS images.

Cross-Eyed is a cross-spectral periocular database captured for the 1st Cross-Spectral Iris/Periocular Competition [38]. The database was collected using a custom dual-spectrum image sensor that simultaneously captured images in both NIR and VIS at a distance of 1.5m in an uncontrolled indoor environment. It comprises images of periocular and iris regions of 120 subjects from different nationalities, ethnicities, and eye colors. There are 120 subjects $\times 8$ images $\times 2$ eyes $= 1,920$ images per spectra and modality. In this paper, we make use of VIS periocular images. Periocular images in Cross-Eyed have their iris masked to ensure pure periocular recognition. We used these masks to normalize them to have the same sclera radius, center, and orientation. They were also zero-padded and cropped, so all have the same size.

PolyU is an iris image database captured using simultaneous bi-spectral imaging. It offers iris images in NIR and VIS, where each eye has pixel correspondence between both spectrum versions. It has 209 subjects $\times 15$ images $\times 2$ eyes $= 6,270$ images per spectrum. As with the previous datasets, we only used the VIS images for this paper. Since the periocular region in this dataset is rather limited, images are just resized to be squared.
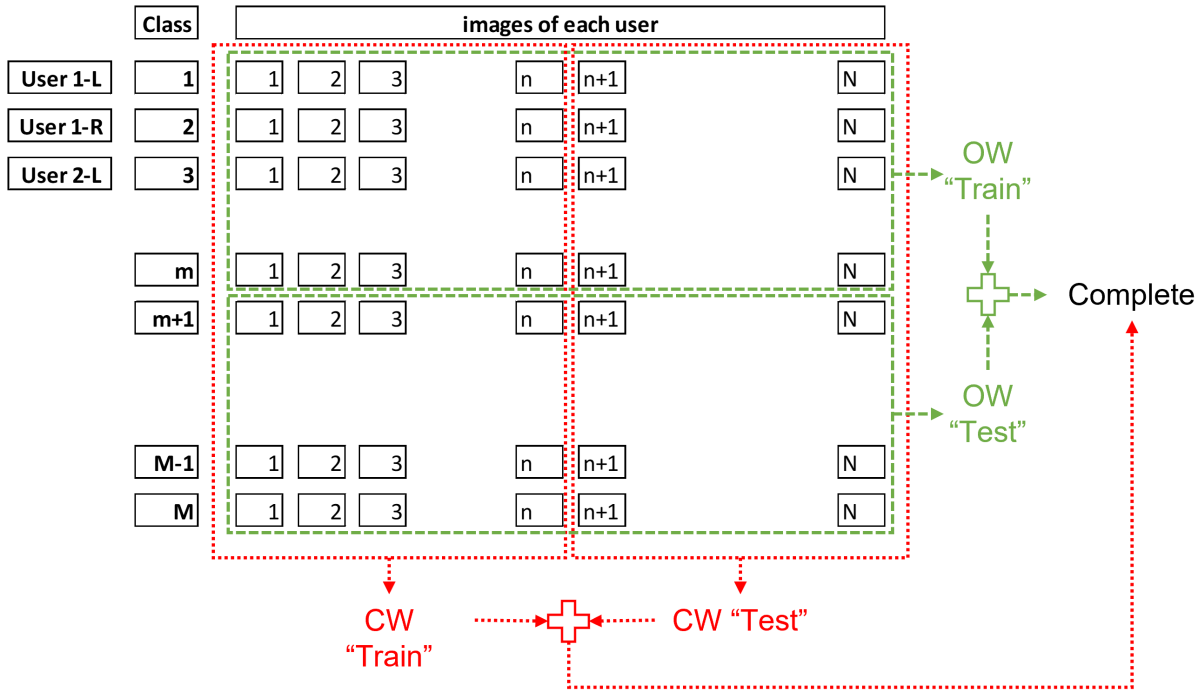
FIGURE 2. Different database partitions for the Close World (CW), Open World (OW), and Complete protocol.

TABLE 1. Summary of train/test partitions per database. The Close World (CW) and Open World (OW) protocols with PolyU and cross-eyed are defined following [35].

Database	Protocol	Train Part. Images (Classes)	Gen./Imp. Pairs	Test Part. Images (Classes)	Gen./Imp. Pairs
PolyU	CW	4,180 (418)	18,810/8,715,300	2,090 (418)	4,180/2,178,825
PolyU	OW	3,135 (209)	21,945/4,890,600	3,135 (209)	2,1945/4,890,600
Cross-Eyed	Complete	-	-	1,920 (240)	6,720/1,835,520
Cross-Eyed	CW	1,200 (240)	2,400/717,000	720 (240)	720/258,120
Cross-Eyed	OW	960 (120)	3,360/456,960	960 (120)	3,360/456,960
IMP	Complete	- (124)	-	620 (124)	1240/190,650

All images were converted to grayscale to normalize skin color across databases, padded with zeroes when images were not squared, resized using bicubic interpolation, and copied across the RGB channels to fit Imagenet networks' input size. Figure 1 shows examples of images from the different databases after this procedure.

B. METRICS

The Equal Error Rate (EER) is the most common evaluation metric for biometric verification systems. EER refers to the error at the intersection point between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) curves.

To compare two feature vectors $v1$ and $v2$, we used the cosine similarity illustrated in Equation 1 for its fast calculation, even for very high dimensionality vectors, to calculate the FAR and FRR from the CNN embeddings, as well as with the LBPH and HOG descriptors. With SIFT features, we used Equation 2, defined as the ratio of matches (M) between images over the minimum number of keypoints (K) detected in either image a and b, with epsilon being

a control parameter for any case when no keypoints where found in an image.

$$\cos(\theta) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \tag{1}$$

$$\text{ratio}_{\text{sift}} = \frac{M}{\min(K_a, K_b, \epsilon)} \tag{2}$$

C. PROTOCOLS

Although biometric identification was used for training the networks for periocular recognition, a verification setting was the choice for analyzing the performance of the proposed method. In biometric verification, one compares an input image against an image of the identity the user claims to be. If the similarity between images is above a predefined threshold, the user is considered genuine; otherwise, the user is regarded as an impostor.

All tests employed a cross-dataset One-Shot Learning approach. If a network is to be trained, we used a dataset formed from the combination of all databases introduced in III-A except the one used for testing. For instance, when
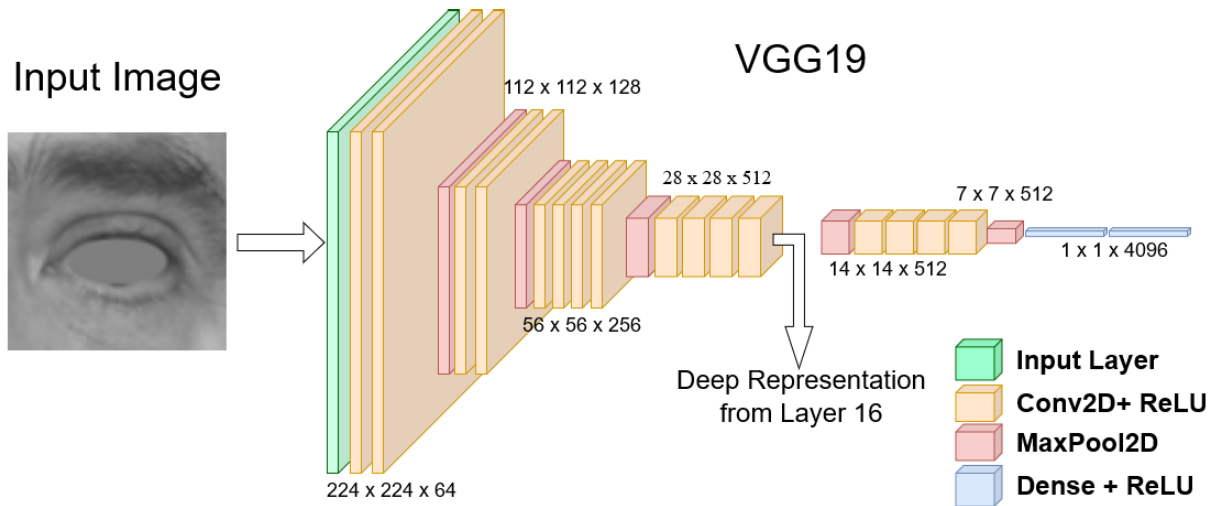
FIGURE 3. Example of a middle layer's deep representation extraction for VGG19.

calculating the EER for the Cross-Eyed, a combined dataset with all data from UBIPr, IMP, and PolyU was used to form the training set. Subsequently, we followed an all-against-all strategy to calculate the test performance, computing all pairs of genuine and impostor scores. All images in the test dataset were used to analyze the performance per layer of the networks. However, we also follow the same protocols as in previous papers employing the same databases to enable comparison. In particular, for comparison for the PolyU and Cross-Eyed, we use the same approaches carried on [35]: the closed-world (CW) protocol, where the images from each user are split into training and testing, and the open-world (OW) protocol, where the users are divided into training and test, along with all their available images in such a way that there are no images from the same user in training and test simultaneously. Figure 2 reflects the difference between the partitions in CW and OW.

In the PolyU dataset and the Close-World setup, the "Test" partition contains the last five images of each user, while the remaining ten images are included in the "Train" partition. In the Open-World approach, the subjects are divided into two halves of 209 users each, the first half used for the "Train" partition and the latter half of the subjects for the "Test" partition. Regarding the Cross-Eyed database, the CW "Test" partition includes the last three images of each user, and the remaining five images go to the "Train" partition. For the OW protocol, the users are divided into two halves, with the first 120 users for "Train" and the last 120 for "Test".

To summarize the number of classes, images, and comparisons for each partition, refer to Table 1. In the case of the IMP database, due to limited data, we only make all-against-all comparisons to calculate the performance on the complete dataset. Finally, the UBIPr dataset is only used in training due to the experiments with different networks already done in [11].

IV. METHODOLOGY

This section presents the experimentation setup used for this study. In particular, the networks, libraries, training strategies, and other algorithms used and how the data was handled and compared.

This paper investigates the performance of deep representations in the middle layers of convolutional neural networks (CNNs) for periocular recognition. We also focus on the impact of training and Transfer Learning on performance. We utilized six widely used and readily available CNNs: ResNet101v2 [41], DenseNet121 [42], VGG19 [43], Xception [44], Inceptionv3 [45], and MobileNetv2 [46]. We conducted periocular verification on the VIS images of the IMP, Cross-Eyed, and PolyU datasets presented in the previous section. We assessed how the performance per layer of each network varied as a One-Shot verification algorithm on a target dataset. To do so, four cases were considered: i) networks trained with the ImageNet dataset; ii) ImageNet networks fine-tuned for periocular recognition; iii) random initialized networks, and iv) networks trained for periocular recognition from scratch.

In cases ii) and iv), where the network requires training for periocular recognition, we do so by training it for biometric identification. The training set comprises all the available periocular datasets except the one used for testing, as indicated in the previous section. When the target (test) dataset was IMP, we combined UBIPr, Cross-Eyed, and PolyU to form a dataset with 830 classes and 9,909 images. We then split it into training and validation sets. The validation set included the last image-distance of each session and user from the UBIPr dataset, the last two images of the Cross-Eyed dataset for each user, and the last five images of each user from PolyU, resulting in training and validation partitions of 6,996 and 2,913 images, respectively. When the target dataset was Cross-Eyed, we trained the networks on a dataset combining UBIPr, IMP, and PolyU, which comprised

8,609 images from 714 classes. The training and validation split followed the same strategy as IMP for UBPIr and PolyU; for the IMP dataset, we used only the last image of each eye and user for the validation split, resulting in training and validation sets of 6,052 and 2,557 images, respectively.

We used Tensorflow-Keras to download, initialize, train, and test the networks. We retained the network's main body, altering only the final Dense layer to fit the number of training classes. We trained them using the Adam optimizer with a learning rate of 0.003, except for VGG, for which we employed Stochastic Gradient Descent with a learning rate of 0.001 and a ClipValue of 0.5, as it provided better stability during training. We trained the networks with Early Stopping, monitoring the validation loss with a patience of 20 epochs and a maximum limit of 500, saving and restoring the weights of the best-performing epoch. Due to GPU memory constraints, the batch sizes were either 16 or 32, depending on the network. We performed data augmentation by randomly rotating the images up to 30 degrees, shifting the height and width by up to 20 percent, and zooming by up to 20 percent. All training was conducted on a Windows 10 machine with 64GB of RAM and an Nvidia RTX2070 GPU with 8GB of VRAM.

After preparing all the networks, we extracted the output of the network layers as illustrated in Figure 3. We sliced the network from the input to the desired layer, inputted all the images from the target dataset, extracted the layer's output 4D matrix, and flattened the matrix while maintaining the batch dimension. Subsequently, we compared the entire dataset using an all-against-all matching strategy, employing Cosine similarity for its rapid and straightforward computation before proceeding to the next layer, as described in Section III-B.

We also use in our experiments three methods based on the most widely used features in periocular research, employed as baseline in many studies [20]: Histogram of Oriented Gradients (HOG) [47], Local Binary Patterns (LBPH) [48], and Scale-Invariant Feature Transform (SIFT) keypoints [49]. HOG and LBPH features are extracted from non-overlapped regions of the image, forming per-block histograms of 8 bins, which are then concatenated to form a feature vector of the entire image. Comparison between two images is done via Cosine similarity between their histograms. On the other hand, SIFT operates by extracting keypoints (with dimension 128 per keypoint) from the entire image. The comparison metric between two images is as explained in III-B. For LBPH and HOG extraction, we used the native Matlab implementation, while for SIFT, we employed the Matlab version available here.[1]

V. RESULTS AND DISCUSSION
This section presents the results obtained from the experiments for the different networks, databases, and modalities. We started by analyzing the performance of middle-layer

[1] https://www.vlfeat.org/overview/sift.html

representations of well-known networks for periocular verification trained for the ImageNet dataset. These pre-trained networks have become the standard starting point for most image classification tasks [50]. Once we obtained the reference results, we explored how they compare when the networks are trained for the same type of data as the target domain, both in EER and depth of the best layer. The results of this study are reported in Section V-A. Since we are comparing very high dimensional data using simple similarity scores, we also investigate the impact that alignment and preprocessing on input images can have for this type of Transfer Learning strategy. This is done in Section V-B. In the mentioned two sections, we have utilized entire datasets to compare the performance of the methods. To assess how the employed strategies generalize, we examine in Section V-C the consistency of the method in terms of layer depth and performance when changing between training and test partitions on the same dataset as well as at the best layer found on other datasets. Finally, in Section V-D, we compare our results with previous CNN-based works employing the same datasets, as well as with traditional handcrafted features.

A. TRAINING EFFECT
Tables 2 and 3, along with Figures 4, 5 and 6 show the effect that training has on the deep representation of the networks for periocular verification. using IMP and Cross-Eyed as test databases. The figures show the performance of the different CNN layers per network and per training strategy, while the tables summarize the best performance and in which layer it is obtained. Results in this subsection make use of the "Complete" partition of the databases (Table 1).

As the tables show, the best results are not necessarily obtained with fine-tuned networks (cases ii, iv). For some networks, it is better to use ImageNet weights directly (case i), or even random weights (case iii), as with MobiletNet. This is consistent with previous findings in [11], in which the VGG-Face network, a VGG network trained for face recognition on a dataset with 1 million images, achieved worse periocular recognition results than its ImageNet counterpart. In absolute numbers, ResNet for the IMP dataset and InceptionV3 for Cross-Eyed are the networks that obtained the best results (EER of 2.05 and 0.7, respectively). ResNet performs the best in the IMP dataset for both the ImageNet (case i in Table 2), and the periocular-trained network (case iv). It also ranks second best for the IMP dataset for the fine-tuned ImageNet network (case ii, Table 2) and the Cross-Eyed dataset ImageNet and fine-tuned ImageNet (cases i and ii in Table 3), albeit sharing the position in this last case with DenseNet. Conversely, InceptionV3 performs the best for these three categories in Cross-Eyed. Overall, it is thus unclear what training strategy is optimal since the best EER obtained for each network and dataset varies. For ResNet and InceptionV3, it seems better to use the ImageNet version than any other training. On the other hand, it seems better to

TABLE 2. Best EER performance per network and the layer at which is obtained for the IMP database. The results for all layers are given in Figures 4 and 6. The training strategies i)-iv) are detailed in Section IV. The numbers in bold indicate the best results per training strategy (column-wise), while the gray cells indicate the best results per network (row-wise). Results are shown for the Complete protocol defined in Table 1.

	IMP DATABASE							
	TRAINING STRATEGY							
	i) ImageNet		ii) TL ImageNet		iii) Random		iv) Trained	
	EER	Layer	EER	Layer	EER	Layer	EER	Layer
ResNet	**2.05**	218	3.31	215	5.99	371	**3.95**	108
DenseNet	3.39	101	**3.26**	175	5.38	438	5.37	129
VGG	7.82	16	5.84	16	7.33	21	4.35	21
Xception	5.07	73	4.04	73	8.26	45	4.44	53
InceptionV3	3.79	96	4.52	110	5.85	163	4.23	122
MobileNetV2	6.28	96	5.75	106	2.55	157	7.91	5

TABLE 3. Best EER performance per network and the layer at which is obtained for the cross-eyed database. The results for all layers are given in Figures 5 and 6. The training strategies i)-iv) are detailed in Section IV. The numbers in bold indicate the best results per training strategy (column-wise), while the gray cells indicate the best results per network (row-wise). Results are shown for the complete protocol defined in Table 1.

	CROSS-EYED DATABASE							
	TRAINING STRATEGY							
	i) ImageNet		ii) TL ImageNet		iii) Random		iv) Trained	
	EER	Layer	EER	Layer	EER	Layer	EER	Layer
ResNet	1.25	195	1.37	240	8.2	292	5.65	51
DenseNet	1.67	143	1.37	462	4.12	529	3.97	252
VGG	4.8	16	2.65	19	6.16	21	7.76	10
Xception	1.67	40	1.49	75	7.01	45	4.47	120
InceptionV3	0.7	132	**1.35**	283	4.04	155	**3.31**	130
MobileNetV2	2.11	64	1.53	126	0.89	157	3.51	132

fine-tune for DenseNet, VGG, Xception, and MobileNetV2. Interestingly, when the networks are fine-tuned, starting from ImageNet weights ("TL ImageNet" column in the tables or blue curve in the figures) gives better results than starting from scratch consistently ("Trained" column or green curve). This effect is much more prominent in the Cross-Eyed dataset. This corroborates previous works that suggest employing a general purpose training such as ImageNet as a starting point, especially if data available for training is limited [50].

Upon examining Figures 4 and 5, we can see that at deeper layers, the fine-tuned networks initialized with ImageNet weights (blue curves) start to perform better than the ImageNet counterpart (red curve). This may be due to the similarity in the domain and higher abstraction at deeper layers achieved by fine-tuning, which helps to close the gap between datasets. However, fine-tuned networks started from scratch (green curves) are, in some cases, even worse than ImageNet networks, especially with the Cross-Eyed database. Again, This confirms that fine-tuning ImageNet networks is a better starting point than scratch initialization, especially with limited training data. On the other hand, at early layers, ImageNet and fine-tuned networks perform very similarly in many cases. This confirms the general assumption that early layers of CNNs usually extract low-level features that are domain-agnostic in many

cases, while deeper layers become more specialized for the particular task at hand. Another very relevant result is that the very last layers of the networks always suffer a jump in error, even fine-tuned versions. Moreover, with IMP, many cases show performance degradation even earlier. Indeed, optimal performance with any network or database is attained already at the middle layers or just after the first third of the network.

We examined the performance per layer of randomly initialized networks as a control (Figure 6). However, the results are surprisingly good in some cases. As Tables 2, 3 show, the performance of networks with random weights is not as bad as it could be expected, equating to or even beating other cases involving training. The performance per layer, as seen in Figure 6, shows a very stable behavior after some initial variability. This is somewhat expected since the weights are random, so the extracted features are also. Most networks have no clear positive or negative tendency as we increase depth, but they show a relative plateau in performance, especially after a relative depth of 0.2. DenseNet, however, does exhibit a slight performance improvement the deeper the layer is, but this is in the form of steps. Finally, we can see that MobileNet has a peak performance at the very last layer. Indeed, MobileNetv2 with random weights achieves the second-best performance for both datasets. This represents an outlier, but it shows that
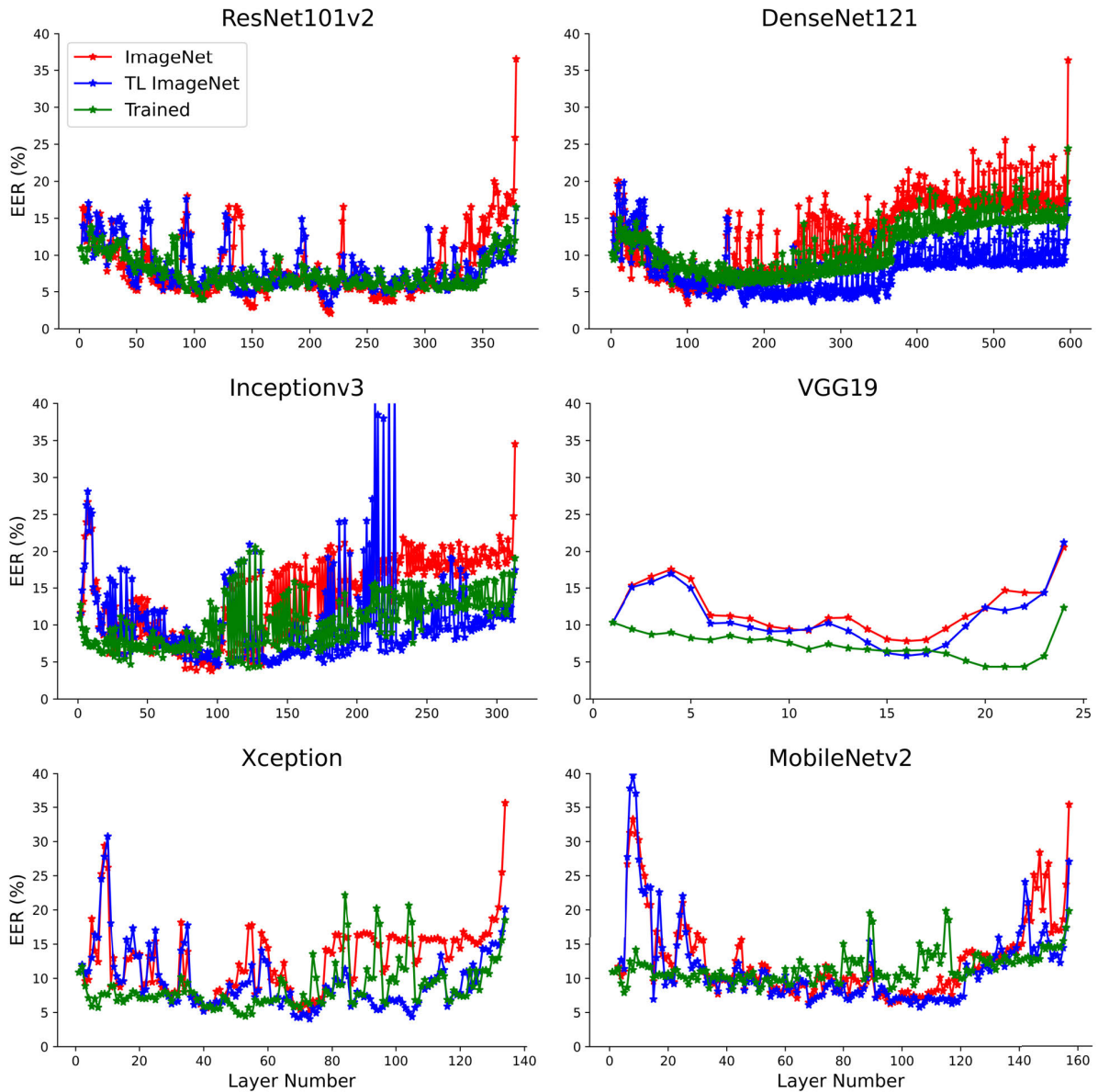
IMP DATABASE

FIGURE 4. EER per layer for the IMP dataset for the cases when: i) the network is trained with the ImageNet dataset (red curve), ii) the network is fine-tuned from ImageNet model for periocular recognition (blue), and iv) the network is trained for periocular recognition from scratch (green curve). These training strategies i), ii), and iv) are detailed in Section IV. The best cases per network and per training strategy are given in Table 2. Results are shown for the Complete protocol defined in Table 1.

the exponential behavior of the last layer can also work to one's advantage. Nonetheless, as mentioned above, some randomly initialized networks perform relatively similarly to their trained counterparts. DenseNet exemplifies this for the IMP dataset and InceptionV3 and DenseNet for Cross-Eyed, where the difference in EER is less than 1%. A notable example is VGG, which performs better in the random version than its trained one for both datasets. Moreover, as we will analyze later when comparing to other methods (Section V-C, Table 7), all randomly initialized networks

yield results comparable to baseline CV algorithms like LBPH and HOG.

B. NORMALIZATION EFFECT

We then examined the method's robustness for perturbations in the input data and how it affects the deep representation behavior. To do so, we employed normalized and unnormalized images of the Cross-Eyed database. The normalized version consisted of the images processed to have the same sclera radius, center, and orientation, as described in
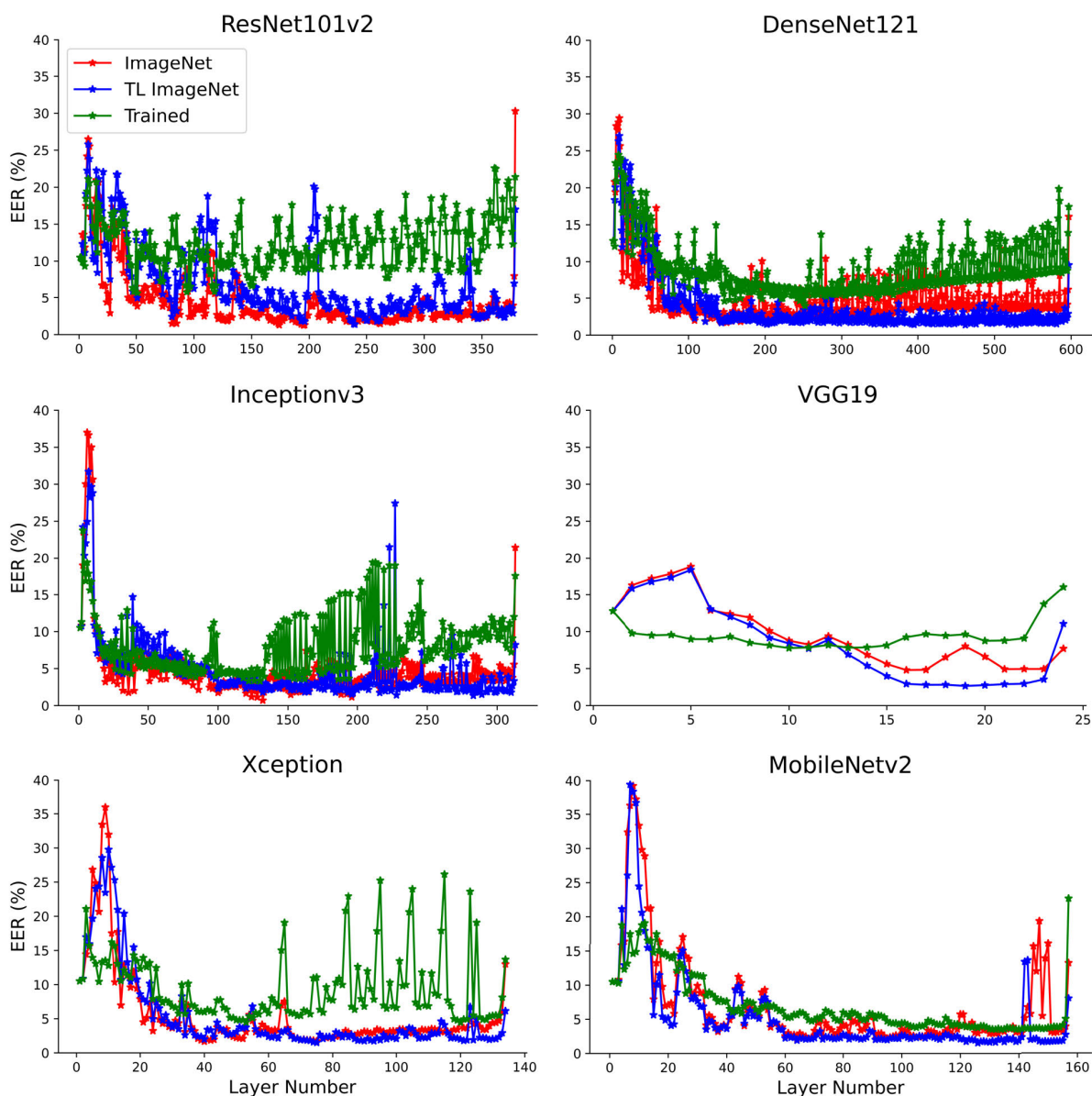
CROSS-EYED DATABASE

FIGURE 5. EER per layer for the Cross-Eyed dataset for the cases when trained with the ImageNet dataset, when: i) the network is trained with the ImageNet dataset (red curve), ii) the network is fine-tuned from ImageNet model for periocular recognition (blue), and iv) the network is trained for periocular recognition from scratch (green curve). These training strategies i), ii), and iv) are detailed in Section IV. The best cases per network and per training strategy are given in Table 3. Results are shown for the Complete protocol defined in Table 1.

Section III-A, whereas the unnormalized images are only converted to grayscale and cropped to be squared using the smallest dimension as reference size. Figure 1 shows the effect of the normalization process on the data for the Cross-Eyed database. Although the Cross-Eyed database was captured at a constant distance, small differences in scale and position between different images can appear if the image is not normalized. Figure 7 shows the effect per layer of input normalization on the performance of the deep representations. For space-saving purposes, we report results

using networks trained for ImageNet only. We can see that the EER becomes significantly worse if images are unnormalized (blue curves), especially at the early and middle layers. Only in the final layers, does the performance with unnormalized images become closer to the normalized counterparts. This is understandable since networks are susceptible to scale, orientation, and, to a minor degree, translation. Only the deeper layers can achieve a higher-level representation of the input data, contributing to closing the gap between the two cases. However, normalized data achieves the best absolute
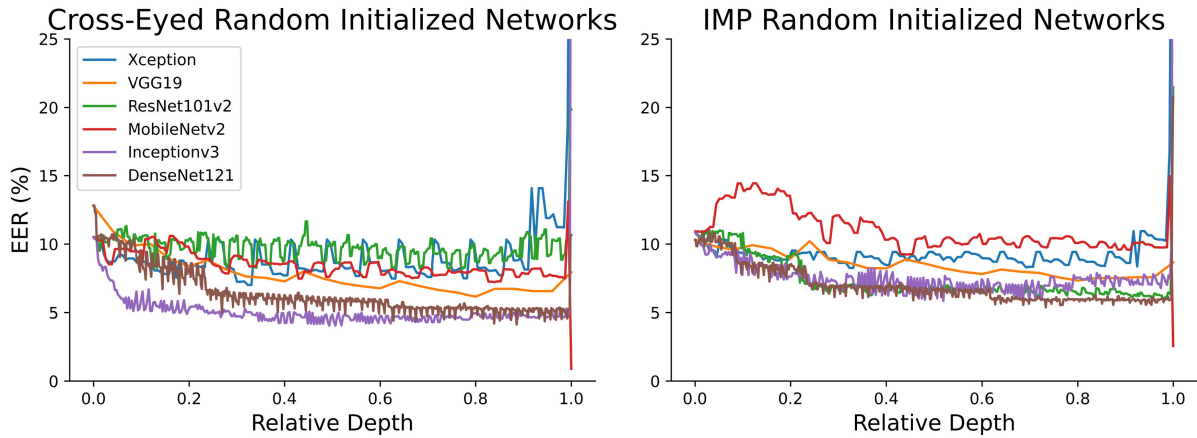
FIGURE 6. EER per layer when the networks were not trained and were just randomly initialized. The best cases per network are given in Table 2 (IMP database) and Table 3 (Cross-Eyed database). Results are shown for the Complete protocol defined in Table 1.

TABLE 4. Best EER comparison per network between normalized and not normalized input images and the layer at which it is obtained (Cross-Eyed database). The numbers in bold indicate the best results per normalization method (column-wise), while the gray cells indicate the best results per network (row-wise). The results for all layers are given in Figure 7.

| | CROSS-EYED DATABASE | | | | |
| | Normalized | | Not Normalized | | Total |
	EER	Layer	EER	Layer	Layers
ResNet	1.25	195	5.27	372	379
DenseNet	1.67	143	4.33	594	597
VGG	4.8	16	**3.68**	25	26
Xception	1.67	40	4.1	133	134
InceptionV3	**0.7**	132	3.77	294	313
MobileNetV2	2.11	64	3.73	156	**157**

performance with most networks, as shown in Table 4. Except for VGG, all networks exhibit an increase in EER between 77% (MobileNetv2) and 439% (InceptionV3) with unnormalized data. We can also see that the best-performing layer with unnormalized data becomes close to the last layer, compared to the normalized version, which usually achieves the best performance in the first half of the network. Thus, robust data normalization is key to achieving better performance.

C. PARTITION EFFECT

Although one can utilize the network without training them for a specific domain, we have seen in previous subsections that it is still essential to determine which layer yields the best performance. As we demonstrate in Tables 2 and 3, the best-performing layer can vary significantly from database to database and from network to network, even in the same periocular domain. In the present subsection, we go one step beyond and consider how the performance and the best layer can change considering different partitions of the database. In other words, we select the best layer in one data partition and test the performance in another partition.

Furthermore, we check generalizability even further by looking at how the performance changes using the best-performing layers from other datasets. For this subsection, we only retain ResNet, DenseNet, and Inception, the best-performing networks of previous subsections (Tables 2, 3, and 4). We report results using networks trained for ImageNet only for space-saving purposes. In addition, since only PolyU and Cross-Eyed have different partitions and CW/OW protocols (Table 1) [35], this section focuses on these two databases.

Table 5 shows the EER obtained on Cross-Eyed on the different partitions of the Closed-World (CW), Open-World (OW), and Complete protocols. The "Train", "Test", and "Complete" partitions of the different protocols refer to those detailed in Table 1. Recall that the main difference between CW and OW protocols (Figure 2) is that the CW protocol contains the same users in the "Train" and "Test" splits (the images of each user are split into two sets), whereas the OW protocol contains different users in the "Train" and "Test" splits (the users are split into two sets). For better viewing, Figure 8 (left) depicts the relative accuracy values (black = 0%, white = maximum EER per CNN/database considering the CW and OW experiments together).

From table 5 and Figure 8, Inception offers very stable results with Cross-Eyed, at least when the best layer is selected on a partition of the same database. The best-performing layer for each partition and, when considering the whole dataset, are quite close together, as do the EER (see the relatively similar gray colors in Figure 8, left bottom). Inception even yields the same exact layer regardless of the partition and mode used to select it. Interestingly, this is true both in the CW case (the Train/Test partitions contain images from the same users) and the OW case (the Train/Test partitions have different users), meaning that Inception generalizes very well over the Cross-Eyed database, even to unseen users. On the other hand, when the best layer is selected externally (using the IMP database), the
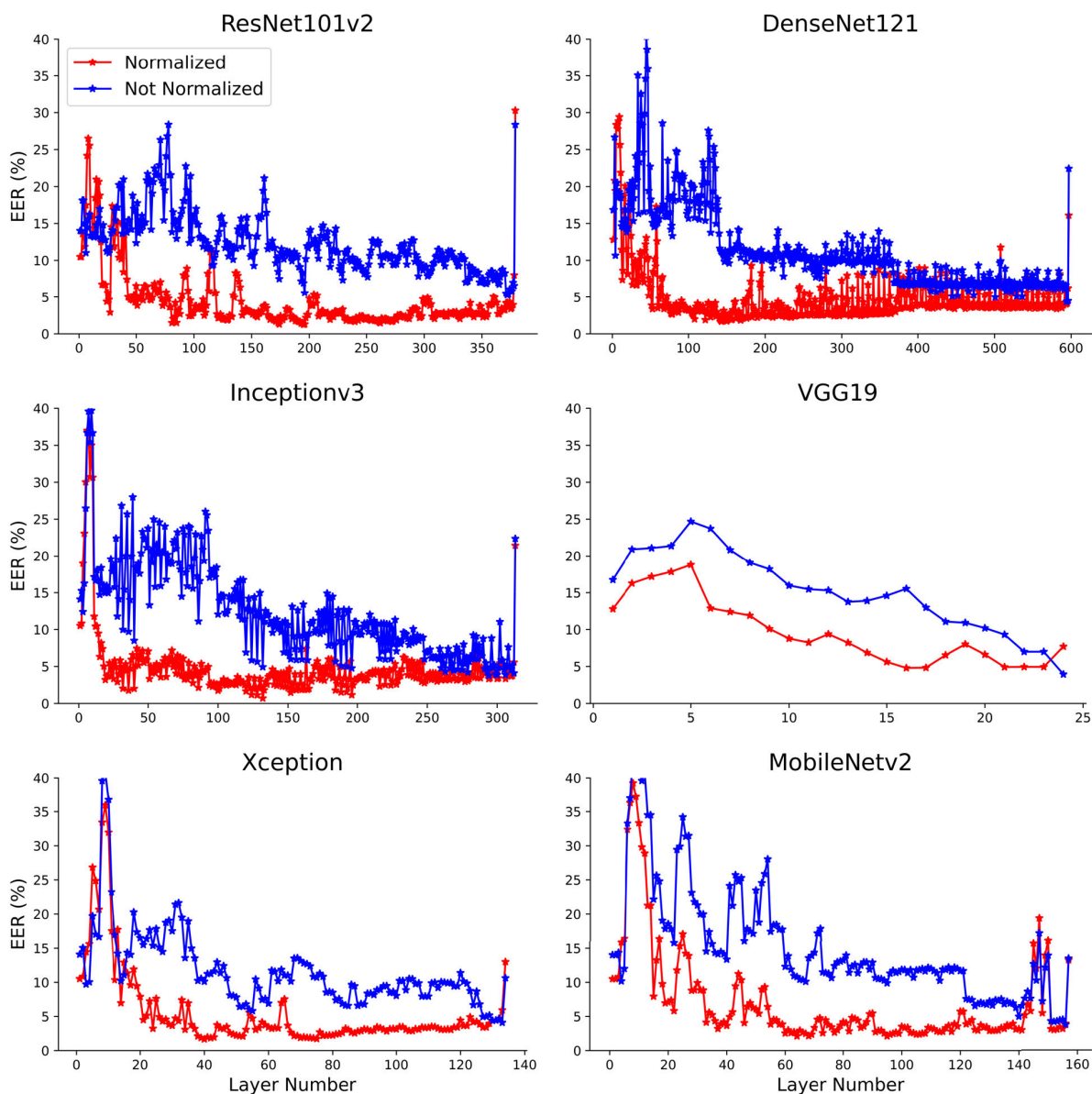
CROSS-EYED DATABASE

FIGURE 7. EER per layer for the Cross-Eyed dataset for the cases when the images were preprocessed and not. The networks employed are trained with ImageNet. The best cases per network are given in Table 4.

results degrade substantially, giving the worst EER. Indeed, this is true for any CNN or partition with the Cross-Eyed database (see the brightest boxes in Figure 8, left, which most of the time corresponds to the case when the best layer is selected using IMP). In addition, the best layer with IMP is usually very different than the best layer calculated with other partitions. This suggests that, despite using databases in the VIS spectrum, their differences (in sensor, illumination, etc.) play an important role in selecting the optimum layer.

ResNet and Densenet, on the other hand, do not appear to generalize as well as Inception over Cross-Eyed. The best layer is different depending on the partition used to select it, at least in the CW case, where it can also be seen that the performance across partitions varies. This can be appreciated in the gray variations of Figure 8, left, for the CW case with these two networks. This result is interesting because the CW case contains images from the same users in all partitions, so one would expect similar optimum layers and performance. On the other hand, the OW case

TABLE 5. Table with the best layers and associated EER for each partition (defined in Table 1) of the cross-eyed database.

	Network	Best Layer at	Layer Number	Cross-Eyed Test Partition		
				Train Part EER (%)	Test Part EER (%)	Complete EER (%)
CW	ResNet101v2	Train Part	262	0.83	1.76	1.45
		Test Part	173	1.33	1.25	1.5
		Complete	195	0.92	1.53	1.25
		IMP	218	1.17	2.19	1.94
	DenseNet121	Train Part	143	1.25	1.99	1.67
		Test Part	168	1.41	1.57	1.82
		Complete	143	1.25	1.99	1.67
		IMP	101	2.82	3.34	3.21
	Inceptionv3	Train Part	132	0.54	0.56	0.7
		Test Part	132	0.54	0.56	0.7
		Complete	132	0.54	0.56	0.7
		IMP	96	1.89	2.59	2.29
OW	ResNet101v2	Train Part	196	1.07	1.45	1.3
		Test Part	195	1.19	1.2	1.25
		Complete	195	1.19	1.2	1.25
		IMP	218	1.84	2.26	1.94
	DenseNet121	Train Part	143	1.58	1.73	1.67
		Test Part	150	1.72	1.7	1.7
		Complete	143	2.83	3.9	1.67
		IMP	101	3.33	3.2	3.21
	Inceptionv3	Train Part	132	0.69	0.71	0.7
		Test Part	132	0.69	0.71	0.7
		Complete	132	0.69	0.71	0.7
		IMP	96	2.26	2.29	2.29

contains images from different users on each partition, but the preferred layers are closer, and the performance is more constant across partitions. One observation in this regard would be that the CW protocol entails more users but with fewer images per user in each partition, whereas the OW protocol includes fewer users per partition but with more images per user. This connects with previous research [15], [51] that shows that it is better to make training decisions based on a larger number of images per user, even if it implies fewer users. This is because a larger number of samples per user allows to model the intra-class diversity better. In our case, it translates to better generalizability when the best layer is selected under the OW protocol. As seen (Figure 8, left, OW case), the gray variations between partitions, in this case, are not so high compared to CW.

Table 6 and Figure 8 show the results for the PolyU dataset. In this case, we can see that the best layers for each network and partition are close to the network's end when they are selected on a partition of the same PolyU database. This is, as explained in Section V-B, most likely due to the normalization of the PolyU dataset and its more inconsistent periocular region. Even though several papers use the PolyU dataset as a periocular one, the database was collected to be an iris database, so the surrounding ocular area

is not consistent. The periocular area, orientation, location, and scale between images change much more than in Cross-Eyed. Also, the available periocular area is reduced, making periocular recognition with this database challenging. As a result, the layers with the best performance are towards the network's end when the networks have achieved a sufficient level of abstraction. If we use an external database instead, like IMP, which contains periocular images of better quality, the best layers appear earlier for all networks. While the lower abstraction of such layers may be sufficient for IMP, their performance on PolyU is substantially worse.

Regarding the best layer of each network, it can be observed that it is approximately the same when it is selected using PolyU, no matter the partition or protocol used. This is relevant from a generalizability point of view. However, when it comes to the EER, the behavior of the Close-World and Open-World changes drastically, even if the layers at which they are calculated are the same. This can be attributed to the number of comparisons made in each mode combined with the worse quality of the PolyU images. In the Open-World case, since both partitions have the same number of users and the same number of images per user, the EER obtained for each one is very close. However, the number of images per

TABLE 6. Table with the best layers and associated EER for each partition (defined in Table 1) of the PolyU database.

	Network	Best Layer at	Layer Number	PolyU Test Partition	
				Train Part EER (%)	Test Part EER (%)
CW	ResNet101v2	Train Part	373	8.25	3.4
		Test Part	373	8.25	3.4
		Cross-Eyed Complete	195	14.14	7.49
		IMP	218	11.23	5.31
	DenseNet121	Train Part	594	5.79	2.49
		Test Part	594	5.79	2.49
		Cross-Eyed Complete	143	12.72	6.39
		IMP	101	21.08	11.96
	Inceptionv3	Train Part	293	6.8	2.68
		Test Part	299	6.96	2.67
		Cross-Eyed Complete	132	8.74	3.91
		IMP	96	15.55	8.35
OW	ResNet101v2	Train Part	373	7.74	7.76
		Test Part	373	7.74	7.76
		Cross-Eyed Complete	195	13.46	13.54
		IMP	218	10.68	10.83
	DenseNet121	Train Part	594	5.54	5.82
		Test Part	594	5.54	5.82
		Cross-Eyed Complete	143	12.09	12.34
		IMP	101	19.18	21.4
	Inceptionv3	Train Part	294	6.56	6.63
		Test Part	279	6.92	6.41
		Cross-Eyed Complete	132	8.71	8.15
		IMP	96	14.41	15.35

user and partition varies in the Close-World mode, resulting in different genuine and impostor comparisons, as shown in Table 1. As the number of comparisons is smaller in the test partition, especially for the genuine case, the EER on this partition is systematically lower than the train partition. This is because a smaller amount of genuine scores does not allow to account sufficiently for intra-class variability effects, providing a more optimistic performance when the database is of lower quality.

Lastly, it can also be seen with PolyU the negative effect of selecting the optimal layer with an external database.

The worst EER (brightest boxes in Figure 8, right) happens when Cross-Eyed or IMP are used. In addition, there is no consistency per CNN. With ResNet, the worst result is given by the optimal layer in Cross-Eyed. However, with DenseNet and Inception, the worst EER comes from the optimum IMP layers.

D. COMPARISON WITH SOA AND OTHERS
Table 7 summarizes and compares the results with previous works using the same databases [28], [35], [52] and the LBP, HOG, and SIFT hand-crafted features. We surpass the
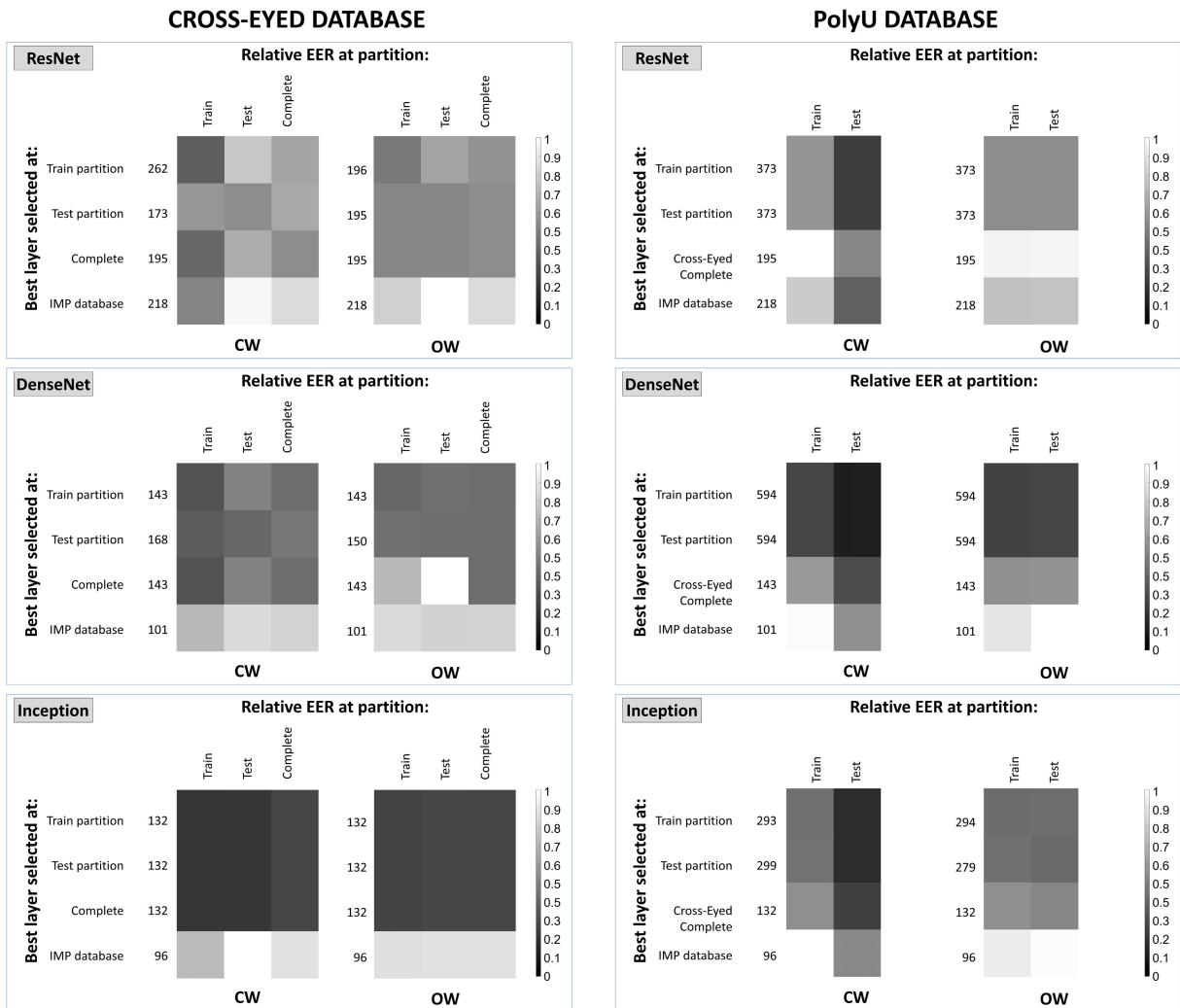
FIGURE 8. Verification accuracy for each partition (defined in Table 1) of Cross-Eyed (left) and PolyU (right). The lower the values, the better. The values are re-scaled, so the maximum EER per CNN and database of the CW/OW experiments are set to 1 (white). Black indicates 0% EER. The exact values are given in Tables 6, 5.

state-of-the-art results for the Cross-Eyed dataset achieved by [35]. In their study, they train a ResNet50 and a VGG network for cross-spectral periocular recognition but also calculate the performance for same-spectrum verification. InceptionV3 reduced the EER by 58% and 79% for Cross-Eyed in Close-World and Open-World protocols, respectively, and it even achieves superior results than their fusion of iris and periocular. ResNet and DenseNet also reduced the EER w.r.t [35] despite lacking training on the target dataset. However, for PolyU, the situation is different. None of the networks achieved comparable or superior results than [35]. The increment in the EER for our method is probably due to the higher degree of variability in the PolyU data. At the same time, the bigger gap in the results for the PolyU and Cross-Eyed dataset achieved in [35] is partially due to the amount of data available to fine-tune the network, allowing training the CNN better for the task. We also

outperformed our previous results on the IMP dataset [28]. Even if we also used a ResNet101, the version used in this study was a newer ResNet101V2 on Tensorflow-Keras, which can explain the difference.

Regarding traditional computer vision algorithms, LBPH and HOG performed worst on each occasion. Nonetheless, SIFT managed to achieve similar results for all Cross-Eyed partitions. It also achieved comparative results for PolyU in the Close-World protocol but outperformed [35] in the Open-World case when the authors only utilized the periocular region. This can be due to the weakness of machine-learning methods when confronted with users that were not present in the training data, which can benefit non-trainable algorithms like SIFT. It must also be highlighted that PolyU has the worst image quality and highest variability among the databases employed. In this case, it becomes less evident the gap between

TABLE 7. Summary of results and comparison with SOA. We reference the best results found in terms of EER for same-spectrum periocular recognition on the visible spectrum for the selected databases. [1]: authors reported the EER for the Left/Right eyes separately.

Method	IMP	Cross-Eyed Complete	Cross-Eyed CW	Cross-Eyed OW	PolyU CW	PolyU OW
ResNet	**2.05**	1.25	1.25	1.2	3.4	7.76
DenseNet	3.26	1.37	1.57	1.7	2.49	5.82
VGG	4.35	2.65	-	-	-	-
Xception	4.04	1.49	-	-	-	-
InceptionV3	3.79	**0.7**	**0.56**	**0.71**	2.67	6.41
MobileNetV2	2.55	0.89	-	-	-	-
LBPH	6.31	3.91	2.64	5.67	7.41	13.87
HOG	10.16	5.23	5.7	5.68	10.73	18.19
SIFT	3.07	0.74	0.6	0.97	**1.64**	**3.76**
[35] Single	-	-	1.70	3.41	**0.78**	3.94
[35] Fusion	-	-	1.17	2.57	**0.35**	**2.61**
[52]	-	-	2.41/1.91[1]	-	-	-
[28]	2.95	-	-	-	-	-

hand-crafted and data-driven approaches when there is limited data.

VI. CONCLUSION

This study examined the effect that training and fine-tuning have on the behavior of CNN's deep representations for One-Shot learning. We utilized well-known pretrained networks as out-of-the-box feature extraction methods for periocular recognition. We investigated the behavior per layer of the networks for different datasets and under different training modes. Additionally, we examined the approach's robustness to some natural acquisition noise and how the best layer changes in relation to an auxiliary database or sampled data from the same distribution.

There is no clear best option regarding training strategy. In our experiments, we have observed that it depends on the network used. ResNet, InceptionV3, and MobileNetV2 do better using the ImageNet weights, while the rest can benefit from fine-tuning to the target periocular task. As in previous works, ResNet typically yields one of the best performances among the networks, making it a good default option for this approach. It is worth mentioning that we outperformed CNNs specifically trained not only for the task of biometric recognition but also for the same dataset without having to fine-tune our models. Furthermore, non-CNN-based algorithms like SIFT can still outperform trained CNNs for the same dataset.

Regarding robustness, a crucial factor seems to be the normalization of the input data. Since our method relies on simple similarity scores between high dimensional matrices, misalignment will heavily penalize the performance. Normalization also affects the depth at which the best-performing layer is situated, which tends to be close to the end for not normalized data, which is when the network has achieved a

sufficient level of abstraction. When it comes to the sample set used to select the best layer, using an external database with different acquisition conditions has shown to have a very negative effect, giving much worse EER and a very different optimal layer w.r.t. using a partition of the same database. Also, it is essential to have a sufficient number of images per user to properly model intra-user variability. This is especially critical if the target dataset is of very low quality.

A limitation of this approach is that it relies on finding the best layer for the task. To accomplish this, it is necessary to have a certain amount of data to calculate the network's performance for that specific domain properly. Nonetheless, the amount of data needed could be, in principle, smaller than the amount needed to properly train a network, as we can see in Table 7, where we achieve better performance for Cross-Eyed than a network trained for biometric recognition with millions of images and then fine-tuned for the dataset. However, the greater the available data, the better results the trained network will have. Another limitation of this approach is that the deep representation matrices of middle layers can be quite big, posing challenges for large datasets or embedded systems due to memory constraints. For these reasons, the normalization process required for this method makes it a suitable option only for small-scale, easy-to-normalize scenarios. Nevertheless, using facial landmarks and iris and sclera segmentation methods, the periocular region is relatively easy to normalize for frontal images.

Regarding its advantages, our approach can use CNNs as out-of-the-box feature extractors with relatively good results. It also enables us to save resources on training, data collection, and processing power through network pruning, removing all other network parts that are not required to

get to the best-performing layer. Our future work on this approach includes investigating methods to reduce the deep representations' dimensionality or memory used, as well as exploring its potential for network pruning in transfer learning.

REFERENCES

[1] E. Arulprakash and M. Aruldoss, "A study on generic object detection with emphasis on future research directions," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7347–7365, Oct. 2022.

[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[3] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10684–10695.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[5] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[6] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[7] V. R. Kumar, S. Yogamani, H. Rashed, G. Sitsu, C. Witt, I. Leang, S. Milz, and P. Mäder, "OmniDet: Surround view cameras based multi-task visual perception network for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2830–2837, Apr. 2021.

[8] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2018, doi: 10.1145/3190618.

[9] C. Luo, X. Li, L. Wang, J. He, D. Li, and J. Zhou, "How does the data set affect CNN-based image classification performance?" in *Proc. 5th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2018, pp. 361–366.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[11] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun, "Periocular recognition using CNN features off-the-shelf," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2018, pp. 1–5.

[12] F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, J. Bigun, R. Vera-Rodriguez, and J. Fierrez, "Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach," *Inf. Fusion*, vols. 83–84, pp. 110–130, Jul. 2022.

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[14] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. 3rd Int. Workshop Similarity-Based Pattern Recognit. (SIMBAD)*. Copenhagen, Denmark: Springer, Oct. 2015, pp. 84–92.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[16] J. Royer, C. Blais, I. Charbonneau, K. Déry, J. Tardif, B. Duchaine, F. Gosselin, and D. Fiset, "Greater reliance on the eye region predicts better face recognition ability," *Cognition*, vol. 181, pp. 12–20, Dec. 2018.

[17] F. Alonso-Fernandez, J. Bigun, J. Fierrez, N. Damer, H. Proença, and A. Ross, "Periocular biometrics: A modality for unconstrained scenarios," 2022, *arXiv:2212.13792*.

[18] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–6.

[19] K. Nguyen, H. Proença, and F. Alonso-Fernandez, "Deep learning for iris recognition: A survey," *arXiv preprint arXiv:2210.05866*, 2022.

[20] F. Alonso-Fernandez and J. Bigun, "A survey on periocular biometrics research," *Pattern Recognit. Lett.*, vol. 82, pp. 92–105, Oct. 2016.

[21] V. Talreja, N. M. Nasrabadi, and M. C. Valenti, "Attribute-based deep periocular recognition: Leveraging soft biometrics to improve periocular recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1141–1150.

[22] F. Alonso-Fernandez, K. Hernandez-Diaz, S. Ramis, F. J. Perales, and J. Bigun, "Facial masks and soft-biometrics: Leveraging face recognition CNNs for age and gender prediction on mobile ocular images," *IET Biometrics*, vol. 10, no. 5, pp. 562–580, Sep. 2021.

[23] K. K. Kamarajugadda and T. R. Polipalli, "Extract features from periocular region to identify the age using machine learning algorithms," *J. Med. Syst.*, vol. 43, no. 7, pp. 1–15, Jul. 2019.

[24] A. Rattani, N. Reddy, and R. Derakhshani, "Convolutional neural networks for gender prediction from smartphone-based ocular images," *IET Biometrics*, vol. 7, no. 5, pp. 423–430, Sep. 2018.

[25] S. Khellat-Kihel, J. Muhammad, Z. Sun, and M. Tistarelli, "Gender and ethnicity recognition based on visual attention-driven deep architectures," *J. Vis. Commun. Image Represent.*, vol. 88, Oct. 2022, Art. no. 103627.

[26] L. A. Zanlorensi, R. Laroca, E. Luz, A. S. Britto, L. S. Oliveira, and D. Menotti, "Ocular recognition databases and competitions: A survey," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 129–180, Jan. 2022.

[27] R. Sharma and A. Ross, "Periocular biometrics and its relevance to partially masked faces: A survey," *Comput. Vis. Image Understand.*, vol. 226, Jan. 2023, Art. no. 103583.

[28] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun, "Cross spectral periocular matching using ResNet features," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–7.

[29] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with off-the-shelf CNN features: A deep learning perspective," *IEEE Access*, vol. 6, pp. 18848–18855, 2018.

[30] E. Luz, G. Moreira, L. A. Zanlorensi Jr., and D. Menotti, "Deep periocular representation aiming video surveillance," *Pattern Recognit. Lett.*, vol. 114, pp. 2–12, Oct. 2018.

[31] L. A. Zanlorensi, E. Luz, R. Laroca, A. S. Britto, L. S. Oliveira, and D. Menotti, "The impact of preprocessing on deep representations for iris recognition on unconstrained environments," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 289–296.

[32] S. Banerjee and A. Ross, "One-shot representational learning for joint biometric and device authentication," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5988–5995.

[33] N. Reddy, A. Rattani, and R. Derakhshani, "Generalizable deep features for ocular biometrics," *Image Vis. Comput.*, vol. 103, Nov. 2020, Art. no. 103996.

[34] N. Reddy, A. Rattani, and R. Derakhshani, "Robust subject-invariant feature learning for ocular biometrics in visible spectrum," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–9.

[35] L. A. Zanlorensi, D. R. Lucio, Alceu de Souza Britto Jr., H. Proença, and D. Menotti, "Deep representations for cross-spectral ocular biometrics," *IET Biometrics*, vol. 9, no. 2, pp. 68–77, Mar. 2020.

[36] A. Sharma, S. Verma, M. Vatsa, and R. Singh, "On cross spectral periocular recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5007–5011.

[37] C. N. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 439–445.

[38] A. Sequeira, L. Chen, P. Wild, J. Ferryman, F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, and J. Bigun, "Cross-eyed-cross-spectral iris/periocular recognition database and competition," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.

[39] A. F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira, S. Marcel, S. S. Behera, M. Gour, and V. Kanhangad, "Cross-eyed 2017: Cross-spectral iris/periocular recognition competition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 725–732.

[40] P. R. Nalla and A. Kumar, "Toward more accurate iris recognition using cross-spectral matching," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 208–221, Jan. 2017.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[44] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[47] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[48] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[50] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.

[51] F. Alonso-Fernandez, J. Barrachina, K. Hernandez-Diaz, and J. Bigun, "SqueezeFacePoseNet: Lightweight face verification across different poses for mobile platforms," in *Proc. Int. Workshops Pattern Recognit. Challenges (ICPR)*. Cham, Switzerland: Springer, Jan. 2021, pp. 139–153.

[52] R. Vyas, "Enhanced near-infrared periocular recognition through collaborative rendering of hand crafted and deep features," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 9351–9365, Mar. 2022.

KEVIN HERNANDEZ-DIAZ (Student Member, IEEE) received the B.S. degree in telecommunication engineering from Universidad de Sevilla, Spain, in 2016, and the M.S. degree in data science and computer engineering from Universidad de Granada, Spain, in 2017. He is currently pursuing the Ph.D. degree in signals and systems engineering with Halmstad University, Sweden. His thesis focuses on biometric recognition from the ocular region in unconstraint sensing environments. His research interests include AI for biometrics, particularly face and ocular recognition, and signal and image analysis, processing, generation, and feature extraction using deep learning.

FERNANDO ALONSO-FERNANDEZ (Member, IEEE) received the M.S. and Ph.D. degrees in telecommunications from Universidad Politecnica de Madrid, Spain, in 2003 and 2008, respectively.

Since 2010, he has been with Halmstad University, Sweden, where he has been an Associate Professor, since 2017, and is currently a Docent and an Associate Professor. He has been involved in multiple EU/national projects focused on biometrics and human–machine interaction. He has over 100 international contributions at refereed conferences and journals and several book chapters. His research interests include AI for biometrics and security, signal and image processing, feature extraction, pattern recognition, and computer vision.

Dr. Alonso-Fernandez is an elected member of the IEEE IFS-TC. He is involved in several roles at the European Association for Biometrics (EAB), such as co-chairing the EAB-RPC or Jury Member of the Annual EAB Biometrics Awards. He was a recipient of a Marie Curie IEF and a Postdoctoral Fellowship of the Swedish Research Council and later with a Project Research Grant for Junior Researchers of the Swedish Research Council. He co-chaired ICB2016 and the Ninth IAPR International Conference on Biometrics. He is an Associate Editor of IEEE Transactions on Information Forensics and Security and IEEE Biometrics Council Newsletter.

JOSEF BIGUN (Fellow, IEEE) received the M.S. and Ph.D. degrees from Linköping University, Sweden, in 1983 and 1988, respectively.

From 1988 to 1998, he was a Faculty Member with EPFL, Switzerland, as an "Adjoint Scientifique." He was an Elected Professor with the Signal Analysis Chair (current position), Halmstad University, Sweden, where he is currently a Full Professor. His scientific research interests include computer vision, texture and motion analysis, biometrics, and the understanding of biological recognition mechanisms. He has been a keynote speaker at several international conferences on pattern recognition, computer vision, and biometrics, including ICPR. He has served on the executive committees of several associations, including IAPR, and as an expert for research evaluations, including Sweden, Norway, and EU countries. He is a fellow of IAPR. He has co-chaired several international conferences and contributed to initiating the ongoing International Conference on Biometrics, formerly known as Audio and Video-Based Biometric Person Authentication, in 1997. He has contributed as Editorial Board Member of journals, including *Pattern Recognition Letters*, IEEE Transactions on Image Processing, and *Image and Vision Computing*.

• • •