*sensors*

MDPI

*Article*

# Adaptive Spatial Transformation Networks for Periocular Recognition

**Diana Laura Borza [1,*], Ehsan Yaghoubi [2], Simone Frintrop [2], Hugo Proença [3]**

[1] Babes Bolyai University, 1st Mihail Kogalniceanu Street, 400084 Cluj-Napoca, Romania
[2] Hamburg University, 177 Mittelweg, 20148 Hamburg, Germany; ehsan.yaghoubi@uni-hamburg.de (E.Y.); simone.frintrop@uni-hamburg.de (S.F.)
[3] IT: Instituto de Telecomunicações, University of Beira Interior, Marquês de Ávila e Bolama, 6201-001 Covilhã, Portugal; hugomcp@di.ubi.pt
* Correspondence: diana.borza@ubbcluj.ro

**Abstract:** Periocular recognition has emerged as a particularly valuable biometric identification method in challenging scenarios, where the acquired data is of poor quality. In this context, partially occluded faces (e.g., due to COVID-19 masks or in VR-applications where the face is not visible due to a head-mounted display) are particularly concerning, where facial recognition cannot be applicable. This work presents a periocular recognition framework based on deep learning architectures, that autonomously localize and analyze the most important areas in the periocular region, for recognition purposes. The main idea is to derive several parallel local branches from a neural network architecture, which in a semi-supervised manner learns the most discriminative areas in the feature map and solves the identification problem solely upon the corresponding cues. Here, each local branch learns a transformation matrix that allows for basic geometrical transformations (cropping and scaling), which is used to select a region of interest in the feature map, further analysed by a set of shared convolutional layers. Finally, the information extracted by the local branches and the main global branch is fused together for recognition. The experiments carried out on the challenging UBIRIS-v2 benchmark show that by integrating the proposed framework with various ResNet architectures, we consistently obtain an improvement in mAP of more than 4% over the "vanilla" architecture. In addition, extensive ablation studies were performed to better understand the behavior of the network and how the spatial transformation and the local branches influence the overall performance of the model. The proposed method can be easily adapted to other computer vision problems, which is also regarded as one of its strengths.

**Keywords:** periocular recognition; spatial transform; attention; biometrics



**Citation:** Borza, D.L.; Yaghoubi, E.; Proenca, H., Frintrop, S. Adaptive Spatial Transformation Networks for Periocular Recognition. *Sensors* **2023**, *1*, 0. https://doi.org/

Academic Editor: Loris Nanni

Received: 31 January 2023
Revised: 10 February 2023
Accepted: 21 February 2023
Published:



## 1. Introduction

Biometric identifiers refer to unique physical (iris, face, fingerprints, etc.), behavioural (gait, typing patterns), or physiological (EEG) traits that can be used to identify and describe individuals. The COVID-19 pandemic has caused a major decline in the performance of existing face identification systems [1], as a result of large data drift. This is because many people wear protective masks that conceal most of the face area, leaving only the periocular region and the forehead visible.

Although there is no standardised definition from organisations such as NIST or ISO/IEC, the term periocular refers to the area around the eyes (i.e., the eyebrows, eyelashes, eye-folds, skin texture, tear ducts, etc.). Figure 1 shows an overview of a periocular re-identification system, in which the area surrounding the eye is used as a cue to determine the correct match between the query and the gallery individuals.

In Ref. [2], the periocular cues are categorised into level one features, which are more dominant in nature and deal with geometry or shape (eyelids, eye corners, eyebrows),

and level two features, which are more related to colour and texture (skin appearance, skin pores, wrinkles, colour). Studies [3] have shown that the periocular region contains relevant cues for person identification both in visible (VIS) and near-infrared images (NIR), and that, in general, level two features are more efficient in VIS images, while level one features are useful in NIR images. Moreover, Ref. [4] showed that humans and computers rely on the same periocular cues for person identification, in both NIR and VIS scenarios.

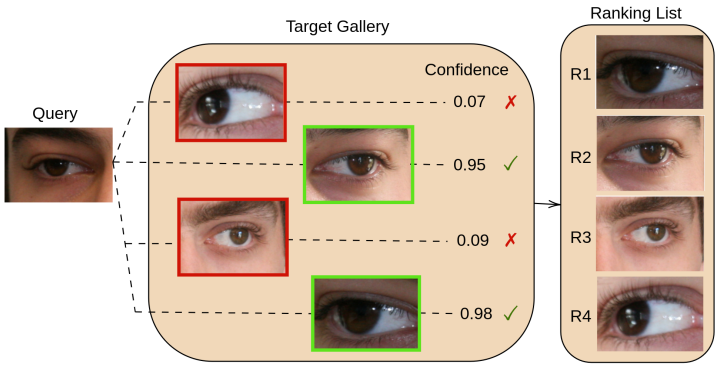


**Figure 1.** Periocular re-identification system. The goal is to find the most similar correct matches of a query person in the target gallery. R stands for rank, and confidence is the prediction probability of the biometric system.

In recent years, periocular recognition has become a prominent research area in biometric systems, because it has proved to be a valuable biometric approach and offers several advantages. First of all, it can be captured with the same imaging devices used for facial or iris identification [3], it is non-intrusive and can be performed without the need for physical contact or cooperation from the subjects. In addition, periocular recognition can be performed in different spectra (both in visible-light and infrared spectrum), making it a versatile method of biometric identification.

All the cues present in the periocular area are prone to occlusions or other factors that influence their applicability:

- The iris and the sclera are sensitive to corneal reflexions (the Purkinje images);
- Accessories—such as eyeglasses or bangs—can occlude the eyebrows;
- Head/eye movements can lead to capturing blurry data;
- Makeup can influence the overall appearance perception of the eye.

With these challenges in mind, the key idea of this paper is to design a convolutional neural network that can adapt to these challenging environmental changes and automatically focus on the most relevant cues visible in the input image. To this end, we propose a multi-branch architecture that can focus on several bio-metrical traits at different granularities, by adding several local branches to focus on different regions of the image. Each local branch learns a geometrical transformation matrix (scaling and translation), which is then used to sample a region of interest (ROI) from an intermediate feature map. These ROIs are further analysed by a set of shared local layers. and then merged (via feature map summation) with the global branch. The workflow of the proposed solution is illustrated in Figure 2. By adding additional local branches to the model, we effectively exploit different bio-metrical traits and boost their performance in challenging environmental scenarios. Although the model design involves setting several hyperparameters, the features and the spatial transformation are automatically learned, and the proposed model can be trained in an end-to-end manner. Extensive ablation studies were performed to determine the impact of each hyperparameter.

The main contributions of this article are threefold: (1) we employ a spatial transformation module, which learns, in a semi-supervised manner, to identify the most prominent regions of the periocular area; (2) the periocular area is analysed both holistically, but also at the local level within the selected regions; and (3) the global and local information are fused

to solve the recognition problem, but the loss function is applied to both the global and local branches to ensure that relevant features are extracted. Last but not least, the proposed method is generic, and can be easily adapted to other computer vision tasks.

The remainder of this manuscript is organized as follows: Section 2 reviews the existing approaches for periocular recognition, and the methodology used for conducting this study and the proposed method are introduced in Section 3. Section 4 presents the experimental results and a comparison with state-of-the-art works. The conclusions of the paper and several future research directions are summarized in Section 6.

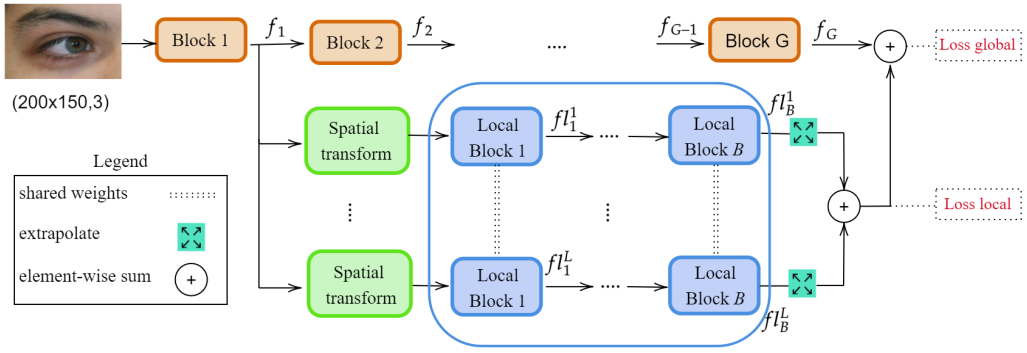


**Figure 2.** The overview of the proposed deep learning model based on convolutional neural networks. The input image is processed using several global and local convolutional blocks, and uses global and local loss functions to optimize the model performance. *Block* and *Local Block* refer to the convolutional blocks/layers in the global branch (depicted in orange) and local branch (depicted in blue), respectively. The feature map $f_i$ from some level $i$ in the network architecture is fed to $L \geq 1$ local branches. Each branch applies a *Spatial transformation* module to select an ROI within the input feature map, and then further analyses this region by applying several *shared* convolutional blocks. In the end, the feature maps of the local branches are extrapolated to the same size and fused with the global branch via element-wise summation. To ensure that the local branches extract useful features, the loss function is applied to both the global branch and the summation of the local branches.

## 2. Literature Survey

The pioneering work of Ref. [5] analysed the feasibility of periocular biometrics and proposed a system that exploits a set of global (LBP, HoG) and local (SIFT) descriptors, fused together at the score level based on a weighted sum with min-max normalisation. Early works focused on designing effective feature descriptors to capture textual information around the eyes, either at a global level—(variants of) LBP [6,7], Gabor filters [8]—or at a local level—SIFT [9], SURF [10], or SAFE [11].

Global-based methods operate holistically on the entire periocular region and extract a feature vector based on texture or shape information. Several methods rely on LBP descriptors, which determine binary patterns by comparing each pixel with its neighbours. The global descriptor is then obtained by concatenating histograms of binary patterns computed across image cells. Ref. [7] investigated several feature extraction methods and determined that LBP substantially improved the performance of both verification and identification methods. In addition, they proposed the Local Walsh-Transform Binary Pattern feature representation, an effective variant of LBP. Other works use Gabor filters—with different orientations and frequencies—to analyse the texture in the periocular region. These filters [8] proposed a matching algorithm based on Gabor filters and a feature encoding scheme that relies on three operators to extract robust features in different spectral bands. Two operators—Weber Local Descriptor (WLD) and uniform LBP—work on the magnitude of the filtered images, while another one—uniform generalised LBP operator (GLBP)—operates on the phase.

PPDM (Periocular Probabilistic Deformation Model) [12] applied a probabilistic inference model to compute 1:1 matching scores (between query and gallery images) based on correlation filters extracted from periocular image patches. Subsequently, the match

performance was improved [13] with an unsupervised method used to select discriminate regions from the periocular image. Local-based methods employ a multi-stage process: first, prominent keypoints are located within the periocular area, and then features are extracted from their vicinity. In Ref. [11], the authors adapted the Symmetry Assessment by Finite Expansion (SAFE) algorithm, previously used in fingerprint analysis, to the problem of periocular recognition. The key idea is to sample several keypoints based on a rectangular grid positioned in the eye centre, and then project ring-shaped areas of different sizes onto a space of harmonic functions used to determine symmetric curve families. A multi-modal authentication system that analyses and fuses features from the face, periocular and, if visible, iris area, is introduced in Ref. [10]. The system extracts three feature descriptors—SIFT, SURF, and Binarised Statistical Image Features (BSIF)—and explores various fusion strategies to effectively combine information from all three modalities. Ref. [14] identifies four prominent regions in the periocular area (eyebrows, upper eye fold, lower eye fold, and eye corners) and then computes a feature representation vector based on HOG, KAZE, and SING descriptors, as well as shape information. Finally, a Naïve Bayes classifier is used to perform the periocular recognition based on the extracted feature vector. The main drawback of this method is that it also requires the accurate segmentation of the features in the periocular area.

With the impressive advances of deep learning in the fields of computer vision and image recognition, recent developments in periocular biometrics focus exclusively on deep convolutional neural networks. Before CNNs, traditional pattern recognition methods used handcrafted features, such as LBP, HoG, Zernike moments [15], fast block processing feature extraction [16] etc., which were manually designed, taking into consideration the specific task and problem. On the other hand, in CNNs the features are automatically learnt from the input data by convolutional layers. In general, deep learning methods tend to outperform traditional hand-crafted feature extraction techniques in computer vision tasks due to their ability to automatically learn relevant features from the data. The interested reader can refer to Refs. [3,17,18] for an in-depth presentation on early periocular research.

In Ref. [19], using transfer learning, seven CNN architectures were trained and compared in the context of periocular recognition. Ref. [20] proposed an original augmentation strategy based on multi-class region swapping, such that the network learns to consider the iris and the sclera regions as not reliable for biometric recognition, and only focus on the information surrounding the eyes. Although the method does not involve additional parameters or an increase in inference time, it completely disregards some regions in the periocular area that contain powerful biometric traits.

Other works [21,22] employed multi-task models to boost the performance of periocular recognition systems. Ref. [22] proposed semantics-assisted convolutional neural networks (SCNN) to incorporate explicit semantic information (gender and eye side): Two separate CNNs are trained on these two tasks (identification and semantic task), and in the end are joined to obtain more powerful feature representations or to perform score fusion. Similarly, Ref. [21] introduced an end-to-end biometric system, based on a multi-task architecture. The framework features a shared convolutional backbone and two separate, dedicated branches, one for biometric identification and one for soft biometric recognition. These branches are fused together for the final periocular recognition, while also predicting soft biometrics. However, the main disadvantage of these approaches is the need for annotated datasets with identity and soft biometric attributes, which are not always available. In addition, they tend to involve more parameters for the supporting tasks.

In Ref. [23], the authors proposed a two-branch deep learning model to analyse iris and periocular cues, and then fuse the corresponding predictions through a multilayer perceptron (MLP). The training procedure is rather complicated, as it requires several stages. In addition, the inputs to each branch require different pre-processing techniques and, as two CNNs are used, the system has more learnable parameters, which leads to longer inference times.

Ref. [24] designed a multimodal biometric system to exploit facial and periocular cues. This model features a shared convolutional backbone, as well as two predictor branches to accommodate the two modalities. During training, additional loss functions are defined to decrease the distance between feature embeddings or periocular-face intra-subjects, while simultaneously maximising feature embeddings of the periocular-face inter-subjects. In Ref. [25] the authors proposed AttenMidNet, a lightweight CNN based on attention mechanisms. The building blocks of the architecture are the MCRS blocks that comprise a convolutional layer, a squeeze-and-excitation block [26], and a residual connection. Ref. [27] proposed a Siamese-like dual stream network, which analyses in parallel the left and right periocular regions of a subject, and then investigates the feature aggregation techniques of the two streams.

Table 1 provides a summary (features, highlight, brief methodology) of the related periocular recognition methods discussed in this section.

**Table 1.** Representative research papers on periocular recognition. As in Refs. [3,18], the methods are grouped based on the feature extraction strategy.

| Strategy | Reference | Highlight |
|---|---|---|
| Global features | [5] | Global features: LBP + HoG, fused together at score level |
| | [12] | probabilistic inference model based on correlation filters |
| | [8] | Gabore filters + feature encoding scheme to extract features in different spectral bands |
| Local features | ISURE [9] | Multinomial Naïve Bayes learning + Dense SIFT for nearest neighbour matching |
| | [11] | Symmetry Assessment by Finite Expansion for periocular recognition |
| | [10] | multimodal system (face, periocular, iris) based on SIFT, SURF, and BSIF fusion |
| | [14] | HOG, KAZE, SING features from four regions in the periocular area + Naïve Bayes classifier |
| Deep learning | SCNN [22] | several CNNs to incorporate semantic information (gender and eye type) in the training process |
| | DEEPRWIS [20] | augmentation strategy to focus solely on the outside eye areas |
| | [19] | transfer learning on different CNN architectures |
| | ADPR [21] | multitask CNN for periocular recognition and soft attribute prediction |
| | AttenMidNet [25] | attention mechanism (squeeze-and-excitation blocks) and mid-level features |
| | Dual-Input CNN [28] | dual-stream Siamese-like CNN + fusion scheme |

Despite their impressive performance, a major caveat of deep learning methods is their lack of explainability. As a result, some works [29,30] tackled the problem of visual explanations and interpretable artificial intelligence in the context of periocular recognition.

## 3. Materials and Methods

### *3.1. Problem Setting*

We formulate the biometric recognition problem as a re-identification problem. Consider a training image-based set of $N$ different identities, each containing $n_{id}$ samples. The purpose of a re-identification system is to learn a function that will find the best match between a query image and a gallery set of images. The query set contains images of the periocular area of a subject we want to identify in another image or set of images (the gallery set). The gallery set contains all potential matches (periocular images) for the target person in the query set.

### *3.2. Solution Outline*

The periocular area comprises various anatomical cues suitable for recognizing individuals and numerous studies have analysed their significance for this process (refer to Ref. [3] for a detailed survey on this matter). However, the applicability of the periocular biometric traits is influenced by environmental factors, and depending on the overall appearance of the eye area and the image capture modality (VIS or NIR), one cue might be more relevant than the other. As an example, the skin texture and the overall eye shape are suitable cues for VIS images [4], but make-up can influence their appearance and therefore degrade the performance of a machine learning model that has been trained to focus on such features. With this in mind, we devised a deep learning model which uses several local branches trained to locate (in a semi-supervised manner) and analyse several discriminative regions in the periocular area (Figure 2).

The proposed method can be easily integrated into any network architecture, and it can effectively boost the performance of (re-)identification systems with a small increase in inference time. Additionally, this strategy can be easily adapted to other image recognition tasks.

### *3.3. Model Architecture*

The key idea of our method is to employ $L$ branches, branched from the $i$th level of a neural network architecture, which will learn to extract prominent ROIs within the periocular area and analyse them for biometric identification. Each local branch starts with a *Spatial transformation* module, responsible for the selection of an ROI in the input feature map $f_i$. Then, a set of shared (between all the $L$ branches) convolutional layers (of size $B$) process the selected areas, and finally, their corresponding feature maps are fused through element-wise summation. This summation result is also added to the global branch. To ensure that the local branches actually learn relevant information, we apply the loss function to both the global branch's output and the summation of the local branches.

#### 3.3.1. The Local Branch

Local branches learn to spot the most relevant regions on the input feature map and attempt to solve the identification problem based solely on the features from these areas.

Inspired by Refs. [31,32], we employ a visual attention mechanism to locate the discriminative parts of a feature map $f_i \in \mathbb{R}^{h \times w \times x}$. To achieve this, each local branch starts with a *Spatial transformation* module, which learns to select an ROI from the input feature map. Inspired by Ref. [33], we generate a spatial map using grouping operations to compute two bi-dimensional maps: $f_{avg} \in \mathbb{R}^{h \times w \times 1}$ and $f_{max} \in \mathbb{R}^{h \times w \times 1}$ and aggregate them into a single map $s_i$ using element-wise summation. $s_i$ is then passed to a feed-forward multi-layer perceptron that regresses an affine spatial transformation matrix $\theta$:

$$\theta = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix}. \tag{1}$$

This matrix allows for cropping and translation (2D spatial parameters $t_x$ and $t_y$) and image scaling (scaling parameters $s_x$ and $s_y$). The transformation is not learnt explicitly from the dataset labels; instead, the model automatically optimises its parameters such that it boosts the recognition accuracy. At the beginning of the training process, the weight and biases of the linear layer are initialised with the identity transformation (i.e., all weights initialised to 0, biases for $s_x$ and $s_y$ initialised to 1, biases for $t_x$ and $t_y$ set to 0). After the transformation matrix $\theta$ (Equation (1)) is determined, the *grid generator* module computes a 2D flow-field grid based on $\theta$ to generate the coordinates from the input image corresponding to each position in the output. The *grid sampling* module applies the transformation parameters to the input and returns an ROI $r_i$ from the input feature map. The structure of this spatial transformation module is depicted in Figure 3, and, for clarity, is also detailed in Algorithm 1. As illustrated in Figure 2, the selected ROI $r_i$ is further passed through a set of shared convolutional layers. In the end, all the outputs of the local

branches are brought to agree with the shapes via an extrapolation layer, and they are added to the output of the global branch.

To guarantee that the features learnt by the local branches are in fact useful in the periocular identification task, the loss function is also applied to their summation. More precisely, during training, the model has multiple outputs, one corresponding to the global branch and one corresponding to the summation of the local branches' output. However, at test time, only the global branch output is used and evaluated. This is inspired by Ref. [34], where the authors used several small classifiers (discarded at test time) on top of come convolutional blocks to ensure that the layers in the middle of the network are also very discriminative.

**Algorithm 1:** Spatial transformation module.

**Input** : $f_i$ - feature map.
**Output**: $f_i'$ - ROI in $f_i$
`/* compute` $\mu$ `channel-wise mean */`
$\mu \leftarrow mean_{ch}(f_i)$;
`/* compute` $m$ `channel-wise maximum */`
$m \leftarrow max_{ch}(f_i)$;
`/* fuse(addition)` $\mu$ `and` $m$ `and flatten result */`
$p \leftarrow \text{flatten}(\mu + m)$;
`/* apply multi-layer perceptron to compute transformation matrix` $\theta$ `*/`
$\theta \leftarrow MLP(p)$;
`/* generate affine grid based on` $\theta$ `and sample` $f'_i$ `*/`
$f'_i \leftarrow \text{sample}(f_i, \theta)$;
**return** $f'_i$

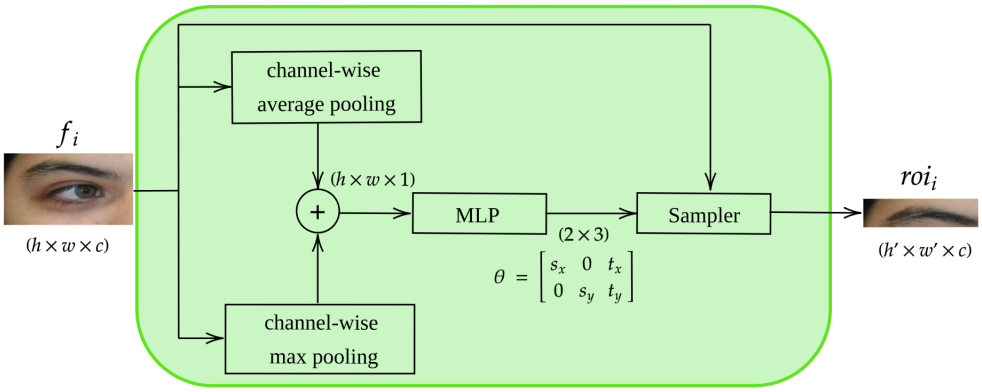


**Figure 3.** Detailed architecture of the spatial transformation module. The input convolutional feature map $f_i$ first undergoes two channel-wise pooling operations (maximum and average) to highlight the most discriminative spatial areas. Their outputs are then added together and passed to a multi-layer perceptron to learn a transformation matrix $\theta$. $\theta$ is used to sample a region of interest from $f_i$, which will be processed by the local branches. Although the spatial transformation module operates on intermediate feature maps, for illustration purposes we exemplify its function with an image corresponding to the input feature map.

### 3.3.2. Closed vs. Open-World Operating Modes

In the context of periocular identification, AI models can operate in closed or open-world modes, depending on whether the identities of the subjects to be recognised are known or not.

In the closed-world setting (i.e., when all the subjects are known in advance), the identification problem can be formulated as a classification task, and a softmax layer can be used to predict the identities. In this case, the final layer of CNN is a dense layer with

softmax activation, and its number of neurones is equal to the number of identities in the dataset. This problem can be seen as a watch-list identification problem [20], in which the model aims to spot some subjects from a predefined list.

On the other hand, in the open-world setting, when the set of identities is unknown, the model needs to be trained to distinguish between unseen subjects. In this case, the identification problems are formulated as a distance learning task or a retrieval ranking problem. In this case, the learning process aims to encode an input image into an embedding space, such that the distance between the images of the same identity is small, while the distance between images from different identities is large. Therefore, the final layer of the model is used as a feature descriptor (and not as a classification layer as in the closed-world setting. For this setup we employed the triple loss (3) to optimise the model. Once the model is trained, the recognition and verification task becomes straightforward in the embedding space, as it simply involves the computation between the computed embeddings.

We evaluated the proposed method for both closed-world and open-world settings. The forward pass of the proposed network architecture is illustrated in Algorithm 2.

### *3.4. Training Process*

The proposed method is trained in an end-to-end manner. As mentioned above, biometric identification systems can operate in either *closed-world* or *open-world* settings.

In closed-world settings, test subjects are known at train time, and the identification task becomes essentially a classification problem. For this test setup, the identification loss is the standard categorical cross entropy loss: $L_{ID}(\hat{x}, y) = -\sum y \cdot log(\zeta(\hat{x}))$, where $y$ is the ground truth identity for the $x$ sample, $\hat{x}$ is the model prediction (logits), and $\zeta$ is the softmax function: $\zeta_\tau(z_i) = \frac{e^{z_i/\tau}}{\sum_j e^{z_j/\tau}}$.

During training for closed-world scenarios, the model has two outputs: $\hat{x_g}$, the response of the classification layer in the global branch, and $\hat{x_l}$, the response of the classification layer applied on the summation of the feature maps from the local branches. The final loss function $L_{CW}$ for the closed-world setting is given in Equation (2):

$$L_{CW} = L_{ID}(\hat{x_g}, y) + \sum_{l=1}^{L} L_{ID}(\hat{x_l}, y). \tag{2}$$

On the other hand, in open-world scenarios, the output of the network does not consist of class probabilities, but in a feature vector in a lower dimensional embedding space in which the L2 distances correspond to subject similarities. In this case, a variant [35] of the triplet loss function [36] is used:

$$L_t = \sum_{\substack{a,p,n \\ y_a = y_p ! = y_n}} \max(||f(x_a) - f(x_p)||^2 - ||f(x_a) f(x_n)||^2 + \lambda, 0)). \tag{3}$$

The triplet loss ensures that given an anchor point $x_a$, the projection of a positive sample $x_p$ (belonging to the identity $y_a$) is closer to the anchor's projection than that of a negative sample of a different identity $y_n$, by at least a margin of $\lambda$. In addition, during training, we also load a final classification layer with the number of neurons equal to the number of identities in the training set, on top of the global branch $cl_g$, and on top of each local branch $cl_l, l \in \{1, ..., L\}$.

To sum up, the loss function for the open-world setting is specified in Equation (4):

$$L_{OW} = \lambda_1 (L_t(x_g) + \sum_{l=1}^{L} (L_t(x_l)) + \lambda_2 (L_{id}(cl_g) + \sum_{l=1}^{L} (L_t(cl_l)). \tag{4}$$

All models were trained using transfer learning (using pre-trained weights from ImageNet [37]) for 70 epochs, using the Adam optimiser, with an initial learning rate of 0.015, updated with a step decay scheduler at epochs 25 and 50.

**Algorithm 2:** Forward function of the proposed network architecture.

**Input** : $I$ - input image

**Data:**

BG = $\{gblock_1, .., gblock_G\}$—list of convolutional blocks in the global branch

BL = $\{bblock_1, .., bblock_B\}$—list of shared convolutional blocks in the local branches

S = $\{st_1, .., st_L\}$—list of spatial transformation modules for each local branch

$cls_g$—classification layer for global branch

$cls_l$—shared classification layer for local branches

*isTrain*—a Boolean value indicating whether the model is in train mode or not

$l$—block index from which the local branches are derived

*mode*—a string indicating if the model is designed for closed-world or open-world setting

;

$f \leftarrow gblock_1(I)$;

/* apply the global branch */

**for** $i \leftarrow 2$ **to** $G$ **do**

  $f \leftarrow gblock_i(f)$;

  **if** $i = l$ **then**

    /* store the input (feature map) for the local branches */

    $f_l \leftarrow f$;

  **end**

**end**

/* apply the local branches */

$f_{local} \leftarrow []$;

**for** $i \leftarrow 1$ **to** $L$ **do**

  /* apply the spatial transformation layer to get an ROI in $f_l$ */

  $r_i \leftarrow st_i(f_l)$;

  /* apply the shared local branches on $r_i$ */

  $l_i \leftarrow bblock_1(r_i)$;

  **for** $bi \leftarrow 2$ **to** $B$ **do**

    $l_i \leftarrow bblock_{bi}(l_i)$;

  **end**

  /* resize the feature map to the size of the last feature map in the global branch */

  $l_i \leftarrow extrapolate(l_i, size(f))$;

  $append(f_{local}, l_i)$;

**end**

/* aggregate the local branches' responses via element-wise summation */

$v_{local} \leftarrow sum(f_{local})$;

/* v - global branch information; GAP - global average pooling */

$v \leftarrow GAP(f + v_{local})$;

$v_{local} \leftarrow GAP(v_{local})$;

/* during inference solely the global branch output is used */

**if** $\neg isTrain$ **then**

  **return** $v$

**end**

**if** $mode =$ *"closed_world"* **then**

  **return** $\left(cls_g(v), cls_l(v_{local})\right)$

**end**

**if** $mode =$ *"open_world"* **then**

  **return** $(v, v_{local}), (cls_g(v), cls_l(v_{local}))$

**end**

## 4. Results

This section presents the experiments performed for evaluating the performance of the proposed method for periocular biometrics, following the methodology introduced in Section 3.

### *4.1. Experimental Setup*

The proposed method can be incorporated into any network architecture, but we chose the ResNet architecture [38] due to its widespread use in computer vision tasks. The key feature of this architecture is the use of residual connections, which allow the layers of the model to learn a residual mapping $\mathcal{H}(x) = x + \mathcal{F}(x)$, instead of directly fitting a function $\mathcal{F}(x)$. This is achieved by defining convolutional blocks (Figure 4) with shortcut connections (skipping one or more layers), which simply perform an identity mapping by adding the input of one convolutional block to its output. With this strategy, it is ensured that the network can learn an identity function when the residual connections are not required, and therefore much deeper architectures can be created.

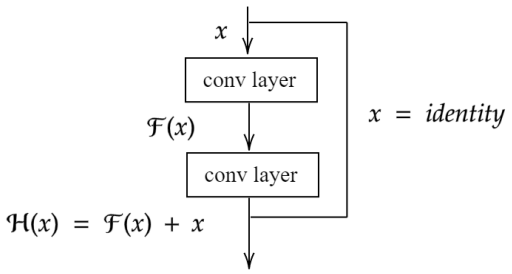


**Figure 4.** The structure of a ResNet block, adapted from Ref. [38]

The ResNet architecture comprises a chain of such blocks with residual connections and its generation follows some simple rules: the convolutional layers in the ResNet blocks have a size of $3 \times 3$; the depth of a layer (number of convolutional filters) is the same if the feature maps have the same spatial size; and when the spatial size of the feature maps is halved, the number of filters of the layer is doubled. Several architectures [38] with different depths have been defined, depending on the number of blocks chained together. In this paper, we experimented with ResNet-18 and ResNet-34 architectures (Table 2).

**Table 2.** Resnet architectures.

| Network | Architecture | | | | |
|---|---|---|---|---|---|
| ResNet-18 | $(7 \times 7, 64)/2$ | $2 \times \begin{pmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{pmatrix}$ | $2 \times \begin{pmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{pmatrix}$ | $2 \times \begin{pmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{pmatrix}$ | $2 \times \begin{pmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{pmatrix}$ |
| ResNet-34 | $(7 \times 7, 64)/2$ | $3 \times \begin{pmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{pmatrix}$ | $4 \times \begin{pmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{pmatrix}$ | $6 \times \begin{pmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{pmatrix}$ | $3 \times \begin{pmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{pmatrix}$ |

The number of local branches used for these experiments is $B = 3$, and they are added from the 13th layer of the network.

The proposed method can be easily adapted for other network architecture, as it involves only adding the local branches on top of a feature map extracted from some layer in the network.

### *4.2. UBIRIS-v2 Dataset*

UBIRIS-v2 [39] is a periocular image dataset designed to evaluate iris recognition models in challenging environments. The dataset contains 11,102 periocular images corresponding to the regions of the left and right eye of 261 subjects. In addition, each image in the dataset is also annotated with the distance to the camera, and the subject's gaze.

We have used this dataset to evaluate the performance of our model in open-world and closed-world settings.

### *4.3. Results*

**Closed-World Setting**

For the closed-world setting—i.e., when the test identities are known in advance—we employed the following data split protocol: the dataset is divided into two disjoint subsets, such that 80% of the samples are used in the training process, and the remaining 20% are used for performance evaluation. The splits were then automatically validated to check that they contain images belonging to all of the subjects in the dataset, and thus the closed-world assumption is satisfied.

Figure 5 illustrates the evolution of the metrics in the training process for the closed-world setting.

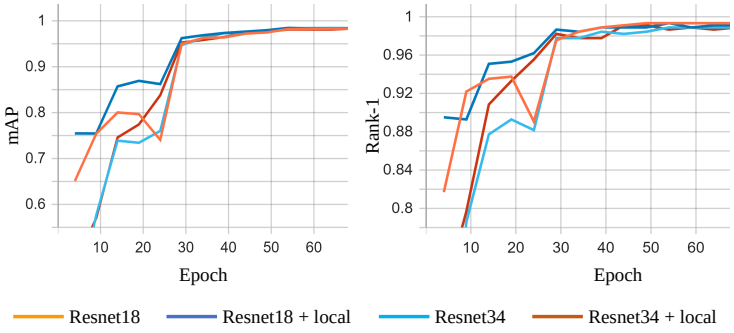


**Figure 5.** Metrics (mAP and rank-1) evolution during the training process for the closed-world setting.

In Table 3, ResNet18(baseline) refers to a ResNet18 architecture trained based on the schedule described in Section 3.4 and ResNet18 + local is a ResNet-18 architecture with 3 local branches added after the 13th convolutional layer. In terms of Rank-1, the proposed method is almost identical to the ResNet-18 baseline, but since the Rank-1 of the baseline is close to 100%, it is difficult to achieve notable improvements. On the other hand, the mAP of the proposed method surpasses the baseline by 10.15%.

**Table 3.** Recognition results on UBIRIS-v2 dataset for the closed-world setting.

| Method | mAP | Rank-1 |
|---|---|---|
| SCNN [22] | - | 79.30% |
| DEEPPRWIS [20] | - | 87.64% |
| ADPR (PR) [21] | - | 83.95% |
| ADPR (JPR )[21] | - | 92.68% |
| ResNet18 (baseline) | 87.81% | 99.33% |
| ResNet18 + local (**ours**) | **97.96%** | **99.33%** |
| ResNet34 (baseline) | 95.81% | 98.88% |
| ResNet34 + local (**ours**) | **96.47%** | **99.10%** |

The proposed method surpasses other periocular recognition methods by a considerable amount. This is due to the fact that by also training the local branches in the network to solve the periocular recognition problem based on the selected periocular regions, the discriminative power of the network is increased and therefore it improves the overall periocular recognition performance.

However, a direct numerical comparison is not completely fair, as other methods use different network architectures and training setups. Both SCNN [22] and ADPR [21]

rely on multi-task learning models and these methods require labelled data not only for identification but also for auxiliary tasks. The additional classification streams added in these networks seem to have a high impact on the classification performance; in Ref. [21] if solely the periocular recognition stream is branch, without the attribute classification branch, the Rank-1 drops from 92.68% to 83.95% (with 8.7%). In addition, the training procedure of Ref. [21] is not performed in an end-to-end manner, and it involves several training steps. This is a cumbersome process, as at each step the appropriate optimisation hyperparameters need to be found, and the overall training takes longer. DEEPPRWIS [20] relies on an original augmentation strategy used solely during the training process, in which the model is automatically trained to focus on the area outside the eye (i.e., ignoring the iris and the sclera). One reason for the lower performance of DEEPPRWIS is that in many images in the UBIRISv2 dataset the area outside the eye is barely visible and the images are misaligned. As the key idea of DEEPPRWIS is to rely on the outside eye area, it is expected to have a lower performance in such cases.

**Open-World Setting**

For the open-world setting, we selected the first 19 subjects as test samples, and the rest of the images (belonging to 142 subjects) were used to train the model. To avoid a friendly split of the dataset, and inspired by the problem of person re-identification [40], we considered the gaze of each subject when splitting the images into sets and gallery sets.

Figure 6 illustrates the evolution of the metrics over the training process for the open world setting.

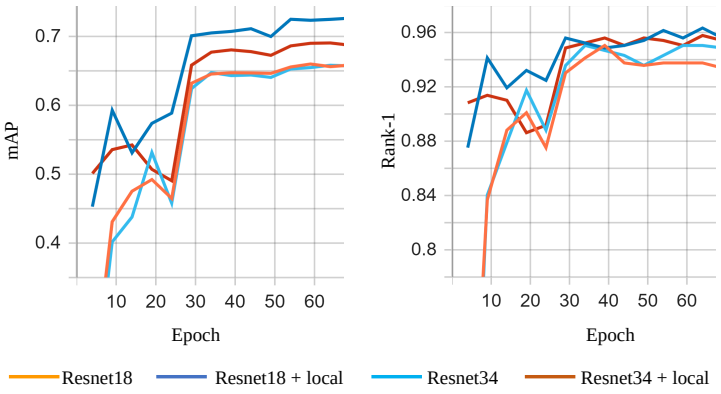


**Figure 6.** Metrics (mAP and rank-1) evolution during the training process for the open world setting.

Table 4 reports the results in the open-world setting. ResNet18 (baseline) and Resnet34 (baseline) refer to the "vanilla" Resnet architectures trained for the problem of iris re-identification.

**Table 4.** Recognition results in the UBIRIS-v2 dataset for the open-world setting.

| Method | mAP | Rank-1 |
|---|---|---|
| ResNet18(baseline) | 66.08% | 93.02% |
| ResNet18 + local (**ours**) | **72.65**% | **95.59**% |
| ResNet34(baseline) | 65.81% | 95.04% |
| ResNet34 + local (**ours**) | **69.03**% | **96.33**% |

The proposed method exceeds the baselines by a large amount in terms of both mAP and Rank-1. For the ResNet18 architecture, the mAP improved from 66.08% to 72.65% (with 6.57%) and the gain in Rank-1 is 2.57%. For the ResNet-34 architecture, the improvements are slightly lower: 3.22% in mAP and 1.29 in Rank-1.

The performance boost is more visible in the open-world scenario. In this case, the baseline ResNet model has lower accuracy allowing for greater improvement, both in terms of mAP and Rank-1, as opposed to the closed-world setting where the Rank-1 is saturated at around 99%. In the open-world scenario, the model learns to project the images into a lower-dimensional embedding space in which samples belonging to the same identity are close together, while the ones belonging to different identities have larger distances. Another possible explanation resides in the fact that in such scenarios, the fine-grained details extracted by the local branches are more useful in the recognition process.

## 5. Ablation Studies

The idea of this paper is to add additional local branches to a neural network architecture, which will learn to identify and analyse the most relevant regions in the input. There are several hyper-parameters that can influence the effectiveness of the proposed method, and, in this section, we will analyse their effect on the model's performance. All experiments for this ablation study were performed using the ResNet-18 architecture.

The first hyper-parameter is the depth (layer number) at which the local branches are added to the model. In Table 5, the column *depth* indicates the value of this parameter. As discussed in Section 4.1, ResNet-18 architecture comprises a convolutional layer, followed by four types of ResNet block, each repeated twice (Table 2). In other words, the local branches were added before the second, third, and fourth ResNet block pairs of the ResNet-18 architecture. The results indicate that better results are obtained when the local branches are added deeper into the network architecture. This is somewhat expected, as deeper layers operate with more semantically meaningful features. In addition, this setup allows the model to use the shallower layers as a shared local feature extractor, and therefore to reduce the inference time and learnable parameters.

Another experiment we performed is related to the way the feature map from the global branch is being preprocessed by the *Spatial transformation* module (Figure 3). To this end, we experimented with three strategies: *Bottleneck*, *Channel*, and *Spatial*.

For the *Bottleneck* setup (in Table 5, *Bottleneck* in the column *Preprocessing*), the feature map is first passed through a $1 \times 1$ convolutional layer to reduce its depth to $ch = 32$, thus making the computations more feasible (Equation (5)). Next, the processed feature map $f_i^{'}$ is flattened and then processed by the MLP to compute the transformation matrix $\theta$.

$$f_i^{'} = conv_{1\times1}^{32} f_i, \tag{5}$$

where $conv_{1\times1}^{32}$ denotes a $1 \times 1$ convolutional layer with 32 filters.

The results show that this strategy achieves the lowest performance—even lower than the baseline ResNet18 architecture. The MLP layer needs to process a high dimensional ($h_{f_i} \times w_{f_i} \times 32$, where $h_{f_i} \times w_{f_i}$ represents the spatial size of the feature map) flattened feature vector. This can lead to optimisation issues (overfitting) as the input layer in the MLP has a large number of neurones. Moreover, local branches might fail to extract semantically meaningful information from the flattened input feature.

Another strategy is to use global average pooling operations as a preprocessing step, as given in Equation (5):

$$f_i^{'} = gap(f_i) + gmp(f_i), \tag{6}$$

where $gap$ and $gmp$ represent the global average pooling and global max pooling operators, respectively.

This is similar to the Channel Attention from Ref. [33]. In this case, the input size of the MLP layer is $1 \times 1 \times c_{f_i}$, where $c_{f_i}$ is the number of channels in the input feature map and this vector models the information about the channels with the most prominent features responses. With this setup, when local branches are added to deeper layers of the network, the baseline ResNet-18 architecture is surpassed in both mAP and Rank-1.

However, the best result is obtained with the *Spatial* preprocessing (Equation (7)). For this setup, the MLP takes as input the sum between the channel-wise average and maximum pooling operations, similar to the spatial attention component from Ref. [33]:

$$f_i^{'} = cap(f_i) + cmp(f_i), \tag{7}$$

where $cap$ and $cmp$ represent the channel-wise average pooling and max pooling, respectively. With this configuration, the proposed method surpasses the baseline ResNet-18 architecture with more than 6% in mAP and 2.56% in rank-1. Intuitively, this seems to be the most promising choice, since the goal of the *Spatial transformation* module is to select the most discriminative periocular areas. Equation (7) returns a map of dimensions $h_{f_i} \times w_{f_i}$, in which higher values indicate areas with prominent features.

**Table 5.** Ablation study of the depth at which local branches are added to the network and the strategy for preprocessing the global branch feature map.

| Method | Preprocessing | Depth | mAP | Rank-1 |
|---|---|---|---|---|
| ResNet18(baseline) | – | – | 66.08% | 93.03% |
| ResNet18 + local | Bottleneck | 5 | 61.66% | 93.76% |
| ResNet18 + local | Bottleneck | 9 | 60.38% | 93.76% |
| ResNet18 + local | Bottleneck | 13 | 61.50% | 93.76% |
| ResNet18 + local | Channel | 5 | 62.64% | 95.04% |
| ResNet18 + local | Channel | 9 | 68.78% | 96.14% |
| ResNet18 + local | Channel | 13 | 71.19% | 95.22% |
| ResNet18 + local | Spatial | 5 | 61.26% | 93.57% |
| ResNet18 + local | Spatial | 9 | 66.09% | 94.86% |
| ResNet18 + local | Spatial | 13 | **72.64**% | **95.59**% |

## 6. Conclusions

This paper presented a periocular recognition system suitable for VIS images based on convolutional neural networks. The main contribution is to equip (a generic) model architecture with additional local branches that will learn, in a semi-supervised manner, to focus on the most discriminative features around the eyes. Unlike existing models, the proposed model is able to extract and exploit multiple regions in the periocular area in a lightweight end-to-end network architecture. Extensive experiments and ablation studies, both for closed-world and open-world setups, prove the effectiveness of the proposed solution. The limitations of the current work are related to the additional hyperparameters that need to be fine-tuned, and also to an increase in the number of learnable parameters.

In future work, we plan to experiment with other strategies for combining the response of the global and the local branches, such as more complex ensembling methods or attention mechanisms.

**Author Contributions:** Conceptualization, D.L.B. and E.Y.; methodology, D.L.B., E.Y., H.P. and S.F.; software, D.L.B., E.Y; validation, D.L.B., E.Y., H.P. and S.F.; investigation, X.X.; resources, X.X.; writing—original draft preparation, D.L.B., E.Y; writing—review and editing, D.L.B., H.P., E.Y. and S.F.; supervision, H.P., S.F.; funding acquisition, H.P., E.Y. and S.F. All authors have read and agreed to the published version of th manuscript.

**Funding:** The contributions due to Ehsan Yaghoubi and Simone Frintrop were funded by the German Science Foundation (DFG) in the project Crossmodal Learning, TRR 169, and the contributions due to Hugo Proença in this work were funded by FCT/MCTES through national funds and co-funded EU funds under the project UIDB/EEA/50008/2020.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** In this article, the publicly available UBIRIS.v2 dataset was used: http://iris.di.ubi.pt/ubiris2.html (accessed on 29 January 2023). The starting code for this work is from Ref. [41].

**Conflicts of Interest:** The authors declare no conflict of interest.

**Sample Availability:** Samples of the compounds are available from the authors.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| BSIF | Binarised Statistical Image Features |
| CNN | Convolutional Neural Network |
| LBP | Local Binary Patterns |
| mAP | mean Average Precision |
| MLP | Multi Layer Perceptron |
| NIR | Near-infrared (spectrum) |
| ROI | Region Of Interest |
| SAFE | Symmetry Assessment by Feature Expansion |
| SIFT | Scale-Invariant Feature Transform |
| SING | Sub-Image based Neighbor Gradient |
| SURF | Speeded Up Robust Features |
| VIS | Visible (spectrum) |

## Notations

The following notations are used in this manuscript:

| | |
|---|---|
| $a_i$ | indexing. element $i$ of vector a, with indexing starting at 1 |
| $B$ | number with convolutional blocks in the shared local branches |
| $BG$ | list with convolutional blocks in the global branch |
| $BL$ | list with shared convolutional blocks in the local branches |
| $cap$ | channel-wise average pooling |
| $cmp$ | channel-wise max pooling |
| $conv$ | convolution layer |
| $f$ | feature maps |
| $f'$ | processed feature maps |
| $G$ | number of convolutional blocks in the global branch |
| $h, w, c$ | height, width, channel |
| $I$ | input image |
| $L$ | number of local branches |
| $L_{CW}$ | closed-world loss function |
| $L_{OW}$ | open-world loss function |
| $L_t$ | total loss function |
| $L_{ID}$ | identity loss function |
| $\lambda$ | margin hyperparameter |
| $n_{id}$ | number of samples per individual |
| $N$ | number of identities |
| $S$ | list with spatial transformation modules for each local branch |
| $\theta$ | spatial transformation matrix |
| $\tau$ | temperature of the softmax function |
| $\hat{x}$ | model prediction |
| $x$ | model input sample |
| $x_a$ | anchor input sample |
| $x_p$ | positive input sample |
| $x_n$ | negative input sample |

| | |
|---|---|
| $x_p$ | positive input sample |
| $x_n$ | negative input sample |
| $y$ | ground truth label |
| $\zeta(.)$ | softmax function |
| $z$ | input variable of the softmax function |
| $\|\|.\|\|^2$ | L2 distance |

## References


1. Damer, N.; Grebe, J.H.; Chen, C.; Boutros, F.; Kirchbuchner, F.; Kuijper, A. The effect of wearing a mask on face recognition performance: An exploratory study. In Proceedings of the 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 16–18 September 2020; pp. 1–6.
2. Woodard, D.L.; Pundlik, S.J.; Lyle, J.R.; Miller, P.E. Periocular region appearance cues for biometric identification. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 162–169.
3. Sharma, R.; Ross, A. Periocular biometrics and its relevance to partially masked faces: A survey. *Comput. Vis. Image Underst.* **2023**, *226*, 103583.
4. Hollingsworth, K.P.; Darnell, S.S.; Miller, P.E.; Woodard, D.L.; Bowyer, K.W.; Flynn, P.J. Human and machine performance on periocular biometrics under near-infrared light and visible light. *IEEE Trans. Inf. Forensics Secur.* **2011**, *7*, 588–601.
5. Park, U.; Ross, A.; Jain, A.K. Periocular biometrics in the visible spectrum: A feasibility study. In Proceedings of the 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Washington, DC, USA, 28–30 September 2009; pp. 1–6.
6. Adams, J.; Woodard, D.L.; Dozier, G.; Miller, P.; Bryant, K.; Glenn, G. Genetic-based type II feature extraction for periocular biometric recognition: Less is more. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 205–208.
7. Xu, J.; Cha, M.; Heyman, J.L.; Venugopalan, S.; Abiantun, R.; Savvides, M. Robust local binary pattern feature sets for periocular biometric identification. In Proceedings of the 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Tampa, FL, USA, 10–13 September 2010; pp. 1–8.
8. Cao, Z.X.; Schmid, N.A. Matching heterogeneous periocular regions: Short and long standoff distances. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4967–4971.
9. Ahuja, K.; Bose, A.; Nagar, S.; Dey, K.; Barbhuiya, F. ISURE: User authentication in mobile devices using ocular biometrics in visible spectrum. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 335–339.
10. Raja, K.B.; Raghavendra, R.; Stokkenes, M.; Busch, C. Multi-modal authentication system for smartphones using face, iris and periocular. In Proceedings of the 2015 International Conference on Biometrics (ICB), Phuket, Thailand, 19–22 May 2015; pp. 143–150.
11. Mikaelyan, A.; Alonso-Fernandez, F.; Bigun, J. Periocular recognition by detection of local symmetry patterns. In Proceedings of the 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems, Marrakech, Morocco, 23–27 November 2014; pp. 584–591.
12. Smereka, J.M.; Boddeti, V.N.; Kumar, B.V. Probabilistic deformation models for challenging periocular image verification. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 1875–1890.
13. Smereka, J.M.; Kumar, B.V.; Rodriguez, A. Selecting discriminative regions for periocular verification. In Proceedings of the 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), Sendai, Japan, 29 February–2 March 2016; pp. 1–8.
14. Ramachandra, S.; Ramachandran, S. Region specific and subimage based neighbour gradient feature extraction for robust periocular recognition. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 7961–7973.
15. Deng, A.W.; Wei, C.H.; Gwo, C.Y. Stable, fast computation of high-order Zernike moments using a recursive method. *Pattern Recognit.* **2016**, *56*, 16–25.
16. Abdulhussain, S.H.; Mahmmod, B.M.; Flusser, J.; AL-Utaibi, K.A.; Sait, S.M. Fast overlapping block processing algorithm for feature extraction. *Symmetry* **2022**, *14*, 715.
17. Nigam, I.; Vatsa, M.; Singh, R. Ocular biometrics: A survey of modalities and fusion approaches. *Inf. Fusion* **2015**, *26*, 1–35.
18. Alonso-Fernandez, F.; Bigun, J. A survey on periocular biometrics research. *Pattern Recognit. Lett.* **2016**, *82*, 92–105.
19. Kumari, P.; Seeja, K. Periocular biometrics for non-ideal images: With off-the-shelf deep cnn & transfer learning approach. *Procedia Comput. Sci.* **2020**, *167*, 344–352.
20. Proença, H.; Neves, J.C. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Trans. Inf. Forensics Secur.* **2017**, *13*, 888–896.
21. Talreja, V.; Nasrabadi, N.M.; Valenti, M.C. Attribute-Based Deep Periocular Recognition: Leveraging Soft Biometrics to Improve Periocular Recognition. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 4041–4050.

22. Zhao, Z.; Kumar, A. Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 1017–1030.
23. Wang, K.; Kumar, A. Periocular-assisted multi-feature collaboration for dynamic iris recognition. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 866–879.
24. Ng, T.S.; Low, C.Y.; Chai, J.C.L.; Teoh, A.B.J. Conditional Multimodal Biometrics Embedding Learning For Periocular and Face in the Wild. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 812–818.
25. Zou, Q.; Wang, C.; Yang, S.; Chen, B. A compact periocular recognition system based on deep learning framework AttenMidNet with the attention mechanism. *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
27. Abate, A.; Cimmino, L.; Nappi, M.; Narducci, F. Fusion of Periocular Deep Features in a Dual-Input CNN for Biometric Recognition. In Proceedings of the Image Analysis and Processing—ICIAP 2022: 21st International Conference, Lecce, Italy, 23–27 May 2022; pp. 368–378.
28. Yang, K.; Xu, Z.; Fei, J. Dualsanet: Dual spatial attention network for iris recognition. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 19–25 June 2021; pp. 889–897.
29. Brito, J.; Proença, H. A Deep Adversarial Framework for Visually Explainable Periocular Recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1453–1461.
30. Brito, J.; Proença, H. A Short Survey on Machine Learning Explainability: An Application to Periocular Recognition. *Electronics* **2021**, *10*, 1861.
31. Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
32. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
34. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
35. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
36. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
37. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Proença, H.; Alexandre, L.A. UBIRIS: A noisy iris image database. In Proceedings of the International Conference on Image Analysis and Processing, Cagliari, Italy, 6–8 September 2005; pp. 970–977.
40. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893.
41. Zhou, K.; Xiang, T. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv* **2019**, arXiv:1910.10093.