Improving Periocular Recognition by Explicit Attention to Critical Regions in Deep Neural Network

Zijing Zhao, Ajay Kumar

***Abstract*—Periocular recognition has been emerging as an effective biometric identification approach especially under less constrained environments where face and/or iris recognition is not applicable. This paper proposes a new deep learning based architecture for robust and more accurate periocular recognition which incorporates attention model to emphasize important regions in the periocular images. The new architecture adopts multi-glance mechanism, in which part of the intermediate components are configured to incorporate emphasis on important semantical regions, *i.e.*, eyebrow and eye, within a periocular image. By focusing on these regions, the deep convolutional neural network (CNN) is able to learn additional discriminative features which in turn improves the recognition capability of the whole model. The superior performance of our method strongly suggests that eyebrow and eye regions are important for periocular recognition, and deserve special attention during the deep feature learning process. This paper also presents a customized verification-oriented loss function, which is shown to provide higher discriminating power than conventional contrastive/triplet loss functions. Extensive experiments on six publicly available databases are performed to evaluate the proposed approach. The *reproducible* experimental results indicate that our approach significantly outperforms several state-of-the-art methods for the periocular recognition.**

***Index Terms*—Periocular recognition, deep learning, attention model, region of interest.**

I. Introduction

AUTOMATED human identification under less constrained environment has become one of the key research areas in biometric recognition in recent years. Periocular recognition has been receiving increasing attention for its promising performance especially under less constrained conditions [1-9]. Although there is no strict definition or standard, the periocular region usually refers to the region around the eye, which preferably includes the eyebrow [8]. Periocular region has been validated to be highly discriminative for different persons, and is considered as an effective alternative or supplement to face and/or iris recognition especially when the complete face or clear iris images are not available [11] [15]. Some researchers also point out that the periocular region suffers less impact from expression variations [8] and aging [10], as compared with the entire face.

In spite of usefulness of periocular recognition, matching periocular images accurately under less constrained environments remains a challenging problem in the community. This is largely due to the fact that this region reveals less information than the whole face, and may suffer from severe interference from artifacts like glasses and hair. By reviewing the recent development of periocular recognition algorithms [1-9], we can conclude that there is still considerable space for the matching accuracy improvement in order to meet the need for large scale real applications, and therefore further research efforts are necessary to advance state-of-the-art performance for periocular recognition.

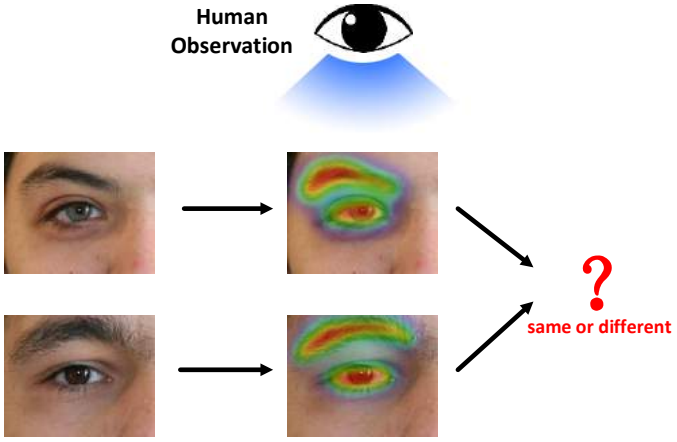


Figure 1: Illustration of implicit human visual attention while performing recognition tasks such as periocular verification. Critical regions that can provide more discriminative information attract more attention, especially for the find-grained recognition.

A. Our Work and Contributions

In this paper we propose the attention based deep learning architecture, referred to as *AttNet*, for more accurate and robust periocular recognition under less constrained environments. The key assumption of our approach is that, the eyebrow and eye region are critical for periocular recognition and should attract additional attention for feature learning. This is inspired by human perception as well as the recent trend in the deep learning community, which suggests that incorporating visual attention to potentially more important regions can significantly benefit the performance for a number of image understanding tasks [23-26]. As illustrated in Fig. 1, when human performs recognition tasks, salient regions such as eye and eyebrow within periocular may provide more discriminative information, and naturally attract more attention than the surrounding regions.

With such assumption, we develop the explicit attention based deep neural network, which incorporates a region of interest detection network and attention implication module. The proposed framework is shown to extract more comprehensive periocular features with higher discriminative

capability. The main contributions of our work can be summarized as follows: 1) the proposed approach achieves superior accuracy for periocular recognition under less constrained environments with visible and near-infrared (NIR) imaging. Extensive experimental results on four publicly available databases suggest that our attention based model outperforms several state-of-the-art methods significantly. Such results provide strong support to our assumption on the importance of critical regions, *i.e.*, eye and eyebrow, for more accurate periocular recognition. 2) We also present a customized loss function, referred to as *Distance-driven Sigmoid Cross-entropy* (*DSC*) loss. The *DSC* loss is shown to offer a marginal effect for both positive and negative training samples during the verification oriented learning, which results in more effective supervision compared with other loss functions such as contrastive loss and triplet loss.

The trained models and source codes of our approach are provided in [38] for reproducing our experimental results, so that other researchers can easily follow our work for further research progress on periocular recognition.

B. Related Work

Continuous research efforts have been devoted into investigating periocular recognition algorithms under different environments [39] [40]. The early feasibility study on using periocular region for human identification was performed by Park *et al*. [1] in 2009, and promising results have been reported, which provides support to subsequent research. Bharadwaj *et al*. [3] further ascertained the usefulness of periocular recognition, especially when iris recognition fails. Some of the later research focuses on cross-spectrum periocular matching [7] using techniques of neural network. Above explorative works have motivated further research efforts to continuously improve the accuracy of periocular recognition. One of the state-of-the-art approaches is proposed by [2] in 2013, which exploited DSIFT features of periocular images, followed by K-means clustering for dictionary learning and representation. This work also explored score level fusion of iris and periocular recognition and reported encouraging results. However, this approach did not investigate periocular-specific feature representation, and the employed DSIFT feature is computationally expensive. Smereka *et al*. [8] has proposed the Periocular Probabilistic Deformation Model (PPDM) in 2015, which provided a sound modelling for potential deformation existing between periocular images. Inference of the captured deformation using correlation filter is utilized for matching periocular pairs. Later in 2016, the same group of researchers improved their basic model by selecting discriminative patch regions for more accountable matching [41]. These two methods achieved promising performance on multiple datasets. Nevertheless, both of them rely on patch-based matching scheme, and therefore are less resistant to scale variation or misalignment that often violate the patch correspondence but is more likely to happen during the real deployments.

Deep learning techniques, especially convolutional neural networks (CNN), have gained immense popularity for computer vision and pattern analysis tasks in recent years. CNN-based methods have been impressively successful for handwritten character recognition [12], object detection [16] [17], image classification [18-20], face recognition [21] [22], palmprint matching [42] and many others. However, surveys on periocular recognition [39] [40] suggest that few studies have exploited deep learning techniques for boosting periocular matching accuracy. Reference [9] proposed semantics-assisted CNN (SCNN) in 2017 for utilizing latent semantical information from periocular images to improve the feature representation. By leveraging additional supervision from semantical information (gender and side) of the training samples, the SCNN has shown to offer better discriminating power with limited training data, and achieved promising performance under cross-database training/testing scenarios. More recently, Proença and Neves [45] claimed that iris and sclera regions may be less reliable for periocular recognition and proposed Deep-PRWIS. In their work, periocular images are augmented with inconsistent iris and sclera regions for training a deep CNN, so that the network implicitly degrades the iris and sclera features during learning. Good results were reported from the Deep-PRWIS on two public databases.

Despite the significant and encouraging research progress gained by aforementioned studies, the performance of periocular recognition still needs to be further improved in order to meet the expectation for real applications. Besides, existing periocular feature extraction methods seldom consider the underlying regional significance that may exist in periocular images. In summary, the following aspects require further research in order to facilitate the performance of periocular recognition:

- Hand-crafted features and shallow learning models are still in the majority of focus for periocular recognition algorithms. Advanced deep learning architectures and technologies, whose effectiveness has already been largely ascertained, have immense potential but not yet been fully exploited in this area, possibly due to the need for large amount of training data;
- Several studies already revealed the importance of eye and eyebrow regions for periocular recognition, but most of existing approaches only consider including these regions for the input/acquired images, and little effort has focused on emphasizing these regions during feature extraction process.

Based on the above facts as well as earlier studies on the human visual attention, this paper proposes an attention based CNN architecture for more accurate and robust periocular feature learning, under the assumption that eyebrow and eye regions preserve higher importance and deserve more attention than the surrounding skin areas. As discussed earlier, employing visual attention mechanism may address the regional significance for the deep feature extraction and benefit the recognition accuracy [23-26]. Besides, several mechanisms including customized network structure, pair-wise training and dynamic data augmentation are adopted to relax the need for training data.

The rest of this paper is organized in the following way: Section II details the methodology of the proposed approach,

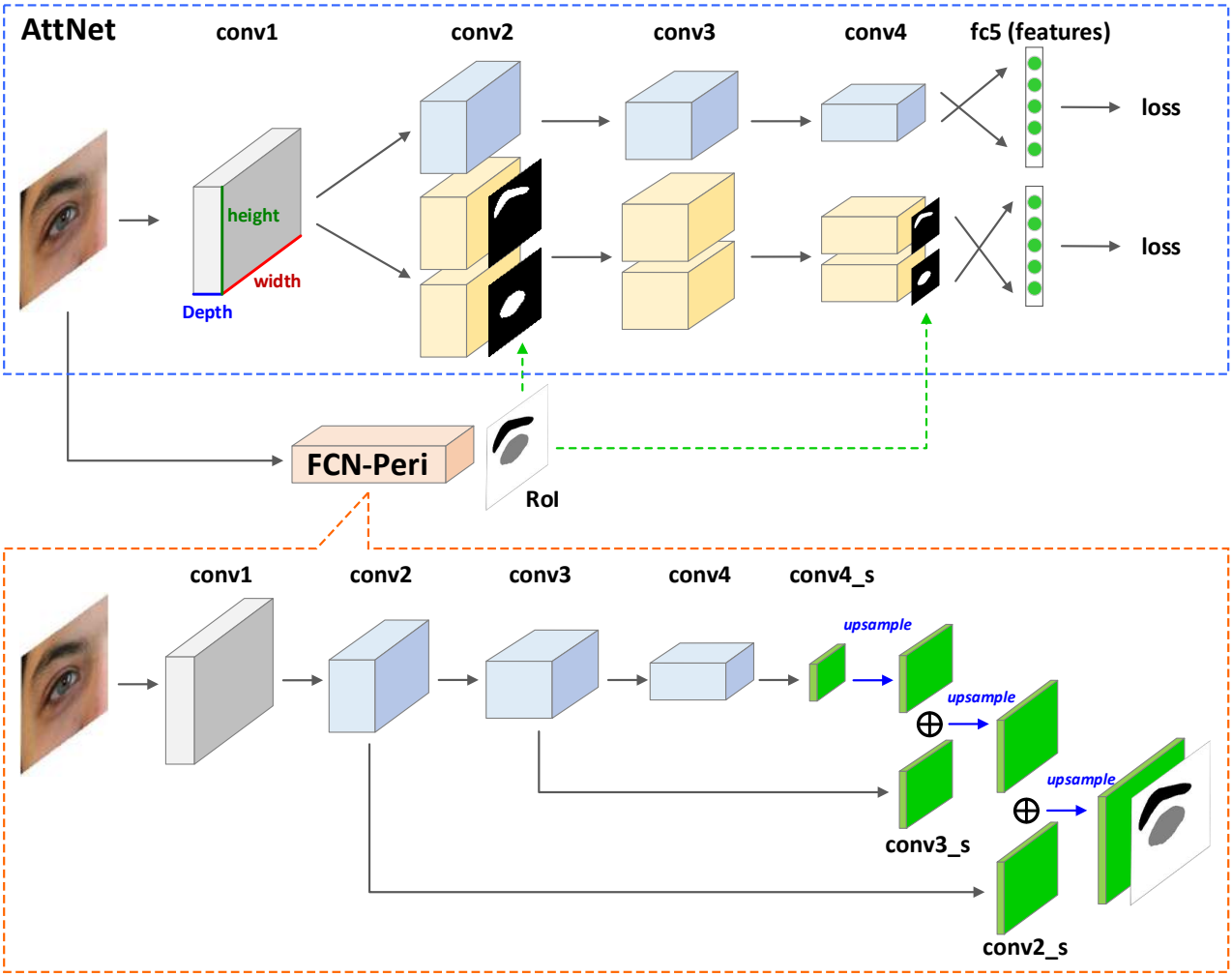


Figure 2: Architecture of the proposed attention based convolutional neuron network, referred to as *AttNet* (top), and the utilized fully convolutional network for specific region detection, called *FCN-Peri* (bottom).

including the visual attention based model and the customized *DSC* loss function; Section V provides the experimental configurations and the analysis on the results; Section VI draws conclusions of this work and introduces our future research goals.

II. Methodology

As discussed earlier, the key innovation of our method is the incorporation of attention model which draws the network attention to specific region of interest (RoI) during feature learning and matching for the periocular recognition. The overall framework is illustrated in Fig. 2. The proposed network structure, referred to as *AttNet* in this paper, firstly exploits a convolutional unit (*i.e.*, conv1) for extracting low-level features from the input image. The network is then split into two branches, where the first branch process the bottom inputs as usual CNNs, while the second branch incorporates RoI information in its intermediate layers (*i.e.*, conv2 and conv4) so that higher attention is imparted to the specific areas of the input periocular image. The first branch without utilizing attention mechanism is designed to recover global features that a typical CNN can perform, which is able to maintain the robustness of the network when RoI information is incorrect, and improve overall performance by feature conjunction. The RoI information is provided by a fully convolutional network (FCN) [30], *i.e.*, *FCN-Peri* in Fig. 2. The detailed layer configuration of these two networks are provided in Table 1. It is worth noting that both networks employed in this work are relatively simple compared with popular and very deep architectures such as VGG [27] and ResNet [19], considering the availability of training data. Besides, we adopt the Siamese infrastructure for training the network in end-to-end verification protocol, and develop a new compositional loss function which is referred to as Distance-driven Sigmoid Cross-entropy (DSC) loss. This new DSC loss has shown to offer superior performance than traditional verification oriented loss functions like contrastive loss and triplet loss.

In this section, the detailed mechanisms for RoI detection and attention implication are explained in Section A and Section B respectively; Section C presents the newly developed *DSC* loss function, followed by the details on the training and test configurations in Section D.

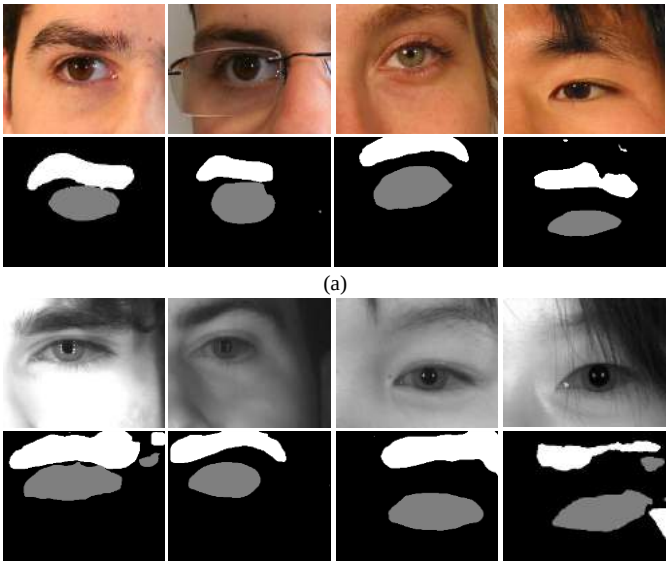
Table 1: Detailed layer configurations for *AttNet* and *FCN-Peri*.

Unit	Layer	Type	#Output channels	Kernel size	Stride
AttNet					
conv1	conv1_1	convolution	32	5×5	1
	relu1_1	ReLU	/	/	/
	conv1_2	convolution	32	5×5	1
	relu1_2	ReLU	/	/	/
	pool1	max pooling	/	2×2	2
conv2	conv2_1	convolution	32	3×3	1
	relu2_1	ReLU	/	/	/
	conv2_2	convolution	32	3×3	1
	relu2_2	ReLU	/	/	/
	pool2	max pooling	/	2×2	2
	att2*	attention	/	/	/
conv3	conv3_1	convolution	64	3×3	1
	relu3_1	ReLU	/	/	/
	conv3_2	convolution	64	3×3	1
	relu3_2	ReLU	/	/	/
	pool3	max pooling	/	2×2	2
conv4	conv4_1	convolution	64	3×3	1
	relu4_1	ReLU	/	/	/
	conv4_2	convolution	64	3×3	1
	relu4_2	ReLU	/	/	/
	pool4	max pooling	/	2×2	2
	att4*	attention	/	/	/
fc5	fc5	fully connected	64	/	/
FCN-Peri					
conv1	conv1	convolution	16	5×5	1
	relu1	ReLU	/	/	/
	pool1	max pooling	/	2×2	2
conv2	conv2	convolution	32	3×3	1
	relu2	ReLU	/	/	/
	conv2_s	convolution	3	1×1	1
	pool2	max pooling	/	2×2	2
conv3	conv3	convolution	64	3×3	1
	relu3	ReLU	/	/	/
	conv3_s	convolution	3	1×1	1
	pool3	max pooling	/	4×4	2
conv4	conv4	convolution	128	3×3	1
	relu4	ReLU	/	/	/
	conv4_s	convolution	3	1×1	1

***** Two branches of *AttNet* as shown in Fig. 2 have the same layer configuration, but attention layers are only placed in the second branch.

A. FCN-Peri – Semantical Region Detection

The key issue for incorporating visual attention model is to identify potentially important regions that deserve more attention than other regions during learning. In general image classification/understanding, the inference of important regions is often jointly learned with the specific tasks [23] [26], as the input data generally involves significantly different background information and those regions could not be predefined. Such strategies, however, require huge amount of training data with sufficient variation to limit the freedom of learning. For fine-grained tasks such as periocular recognition, predefined region detection is preferred [25] as prior knowledge about the input images is usually available, so that the learning process can be better regularized with limited training data. In our approach, based on human perception model, we assume that the regions containing eyebrow and eye are relatively important for periocular recognition. Under such assumption, we firstly exploit a fully convolutional network (FCN) to detect the eyebrow and eye regions.



(a)

(b)

Figure 3: Samples outputs of *FCN-Peri* for test images with visible (a) and near infrared (b) imaging. The black pixels represent predicted background, and the white and gray pixels identify predicted eyebrow and eye respectively.

The FCN employed in our work was firstly proposed for the semantic segmentation in [30]. Different from common CNNs, FCN does not contain fully connected layers, and the upsampling layers are utilized to integrate intermediate convolutional feature maps at different scales. The spatial correspondence between the input image and the output features is therefore maintained to achieve pixel-to-pixel prediction. The FCN is supervised by a pixel-wise *softmax* loss function using groundtruth labels. In our approach, we employed a simplified version of the FCN proposed in [30] for segmenting eyebrow and eye from background in the input periocular image, which we refer to as *FCN-Peri*. The detailed architecture of *FCN-Peri* is illustrated in Fig. 2 (bottom), which contains about 0.1M parameters.

The original FCN in [30] was developed to classify each pixel into one of 21 classes. In our work, eyebrow and eye are regarded as two different classes, and pixels in the original input image are to be segmented into three classes, *i.e.*, eye, eyebrow and background. We manually labeled the eyebrow and eye regions for about 100 images from the training sets of visible and near infrared (NIR) data (details of datasets are in Section III) respectively as the ground truths to train *FCN-Peri* from scratch. It should be noted that by "eye region", we refer to the region including the iris, sclera, eyelid and eyelash, *etc*., rather than just the iris region. Fig. 3 shows several region segmentation results from trained *FCN-Peri* on the test sets. It

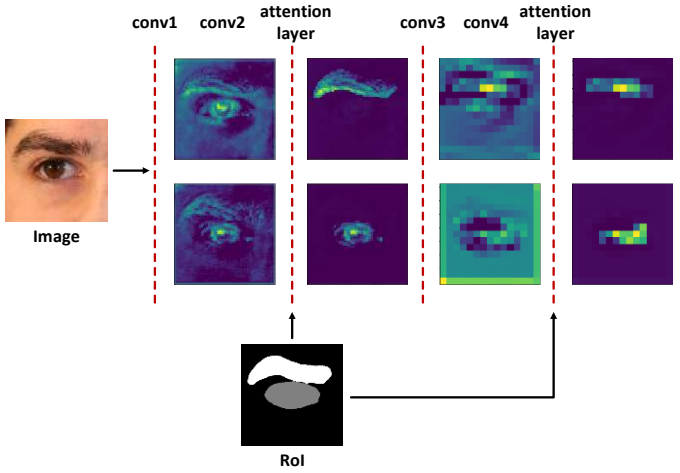


Figure 4: Visualization of convolutional features from intermediate layers before and after attention layers. *Attention layers* increase the feature values within the RoI, and meanwhile decrease those in background. Feature maps of different scales are upsampled to the same size for better illustration.

can be observed that the region predictions are quite robust despite that it makes some mistakes for some challenging samples. The proposed attention based deep neural network, *i.e.*, *AttNet*, is however expected to be tolerant to such level of errors in a few samples. It should also be noted that the networks for visible and NIR spectrums are separately trained.

B. AttNet – Incorporating Visual Attention for Periocular Feature Learning

With the detected regions containing eyebrow and eye for an input image from *FCN-Peri*, we then incorporate the resulting RoI in *AttNet* for attention model implementation. As shown in Fig. 2, after convolutional units conv2 and conv4, the output map from *FCN-Peri* indicating eyebrow and eye positions is utilized to adjust the convolutional features. There is no standard procedures for accomplishing attention in deep neural networks. Some methods use the RoI for affine transformation and alignment [24], while others consider bluring/masking the background for the input images or intermediate features [26], or feed cropped areas into multiple deep networks [25]. In our approach, we apply a straightforward yet effective mechanism for emphasizing important areas inferred by *FCN-Peri*, *i.e.*, increasing the magnitudes of the convolutional features within the RoI and decreasing those outside the RoI. More specifically, an *attention layer* is placed after a convolutional unit and performs the follow operation:

$$f'_{x,y} = \begin{cases} \alpha f_{x,y} & , \text{if } (x,y) \in R \\ \frac{1}{\alpha} f_{x,y} & , \text{otherwise} \end{cases} \quad (1)$$

where R is the set of x-y coordinates where the current position is considered as RoI, $\boldsymbol{f}$ is the convolutional feature map from the previous layer, $\boldsymbol{f}'$ is the processed feature map before entering the next layer, and α is a positive parameter controlling the intensity of adjustment. It was empirically fixed to 5 for all our experiments. Such operation attempts to simulate human visual attention by weighting the features within the RoI more than those in the background for the subsequent layers of the network. The feature adjustments for eyebrow and eye are separately performed, each on half of the channels of the feature maps respectively, as these two regions present quite different characteristics. We selectively incorporate such attention mechanism for conv2 and conv4 to account for both low-level and high-level convolutional features. Since conv1 is shared by the RoI-aware and common branches, conv2 is therefore more appropriate to incorporate for the low-level attention. On the other hand, conv4 is right before the fully connected layer fc5 (*i.e.*, the layer generating feature vectors) and is also judicious to be selected to impart high-level attention. Fig. 4 visualizes the effect of the employed attention model for the features from the two convolutional units. It can be observed that the background features which do not belong to the RoI "fade" after the operation by attention layers. In this way, the foreground features make more impacts on the feature extraction process by subsequent layers. Although simply increasing the feature magnitudes inside the RoI may not be an *optimal* approach to incorporate visual attention, it is quite scientific and easy-to-implement scheme to achieve key objective of our research, *i.e.*, to investigate and evaluate the importance of eye and eyebrow regions to advance periocular recognition through the deep periocular feature extraction.

C. Distance-driven Sigmoid Cross-entropy (DSC) Loss for Verification Oriented Supervision

We adopt Siamese-like pair-wise network infrastructure for training our *AttNet*, *i.e.*, instead of classifying a single image into a standalone class, a pair of images are jointly evaluated to predict whether they belong to the same class or not. Such configuration is illustrated by Fig. 5. Contrastive loss [28] or triplet loss [29] are often used for the pair-wise training. Compared with the classification training protocol which usually uses a *softmax* loss function for supervision, the pair-wise protocol is closer to the verification problem (one-to-one matching) which is a fundamental application scenario for most biometric systems. A classification based model, in contrast, may require additional transfer learning to make itself more effective and scalable, such as in [29]. Besides, the pair combination from training samples introduce more data variation, which can is likely to reduce the overfitting of trained model. In the following, we present a brief introduction to conventionally used loss functions for the pair-wise training, followed by our newly designed *DSC* loss function.

1) Conventional Verification Oriented Loss Functions

The conventional contrastive loss function for training Siamese network is formulated as follows:

$$L_{con} = td^2 + (1-t)\max(0, m-d)^2 \quad (2)$$

where t is the label of the current pair, *i.e.*, t=1 if the two samples come from a same class and t=0 otherwise, and d is simply the Euclidean distance between the two input feature vectors $\boldsymbol{f}_X$ and $\boldsymbol{f}_Y$:

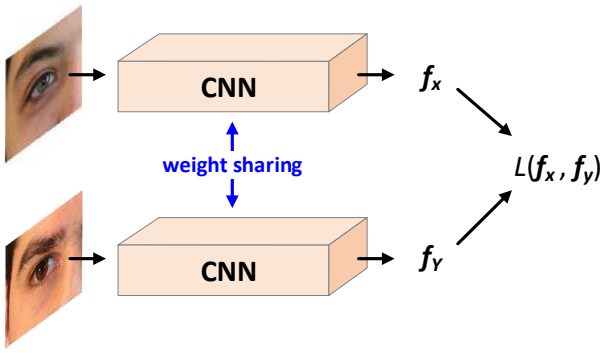


Figure 5: Illustration of Siamese architecture for training CNN in verification protocol. Two identical CNNs are placed in parallel to process a pair of samples. Specific pair-wise loss function (*e.g.*, contrastive loss) is employed to supervise the training, and the weights (parameters) of the two networks are kept the same (weight sharing) during the entire training process.

$$d = \left\| \boldsymbol{f}_X - \boldsymbol{f}_Y \right\|_2 \tag{3}$$

m is a preset margin for regularizing the distance from a negative pair (*i.e.*, a pair for samples from different classes). The contrastive loss is designed to reduce the distance between a positive pair as a quadratic energy term, while for negative pairs, the distance between a negative pair would be increased until it exceeds the hard margin m. The effect of m is to force the network to concentrate on relatively challenging negative pairs only. However, there is no regularization on the positive pair samples. As the training progresses, more and more negative pairs do not produce any losses due to the hard margin, while all the positive pairs still have continues impact on the backpropagation. This causes unbalanced training for positive and negative pair samples.

The above side effect is to some extent alleviated by triplet loss, which can be considered as a variant of contrastive loss. Instead of evaluating a simple pair, the triplet loss composes positive and negative pair into a triple structure, and measures the loss by:

$$L_{tri} = \max(\left\| \boldsymbol{f}_{X_1} - \boldsymbol{f}_{X_2} \right\|_2^2 - \left\| \boldsymbol{f}_{X_1} - \boldsymbol{f}_Y \right\|_2^2 + m', 0) \tag{4}$$

where $\boldsymbol{f}_{X_1}$ and $\boldsymbol{f}_{X_2}$ are features from a same class while $\boldsymbol{f}_Y$ is extracted from another class. Different from contrastive loss, which uses an absolute margin to regularize negative pairs, the triple loss relies on a relative margin m' to enlarge the difference between the positive pair distance and negative pair distance. In this way, the balance of positive and negative pair samples is always retained during the training process. Verification oriented applications, however, mostly use an absolute value as threshold instead of relative margin for decision making, and therefore slight inconsistency exists between the training process supervised by triplet loss and the actual test (matching) process.

2) Distance-driven Sigmoid Cross-entropy (DSC) Loss

In order to address the above limitations, in this paper we introduce a customized compositional loss function called

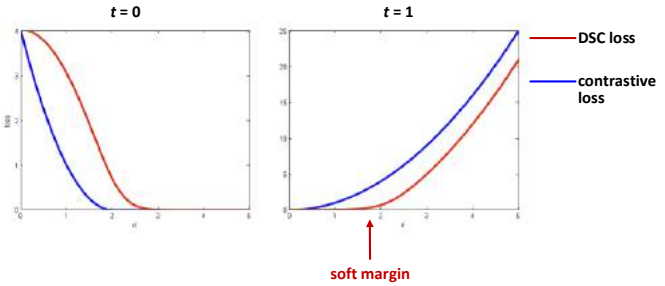


Figure 6: Comparison of *DSC* loss (a = 1, b = 4) and conventional contrastive loss (m = 2) with respect to d. The *DSC* loss provides a (soft) margin for positive cases (t = 1) which achieves better regularization for genuine pairs, such that the learning process mainly focuses on challenging samples.

Distance-driven Sigmoid Cross-entropy (*DSC*) loss. Given the distance d between a pair of features to be evaluated, we firstly perform following mapping on it:

$$s = b - ad^2 \tag{5}$$

$$p = \frac{1}{1+e^{-s}} \tag{6}$$

where a and b are positive constants which are used for linear transformation on the square of the Euclidean distance, p is obtained by a sigmoid function on the transformed s and can be regarded as the probability that the two samples come from a same class. The motivation of using sigmoid function is that it maps any real value into (0, 1), and varies significantly near zero but much slower at two ends. Such property essentially enables a kind of soft margins for the low and high values of s. In this way, the learning process for both positive and negative pairs can be regularized, so that it mainly focus on challenging samples with s values near zero. The loss for the obtained probability p is then measured by the cross-entropy function:

$$L_{DSC} = -[t \log p + (1-t)\log(1-p)] \tag{7}$$

The sigmoid cross-entropy loss is widely used when the task is to predict probabilities of certain events. In this case, we regard our task as predicting the probability of a binary event – same class or different classes. Different from common approaches which feeds a single neuron output spanning over $(-\infty, +\infty)$ into the sigmoid function, we originally map the Euclidean distance d to a term s that spans over $(-\infty, b]$, then transfer to approximated probability p. The constant b should be selected such that its sigmoid value $1/(1+e^b)$ is very close to one. Such transfer is the key to the new *DSC* loss function which utilizes the soft margins of sigmoid function in a straightforward way.

Fig. 6 demonstrates the comparison of the newly developed *DSC* loss function and conventional contrastive loss function w.r.t d, for both positive (t = 1) and negative (t = 0) cases. It can be clearly observed that for negative cases the two losses have similar distribution that, when d is greater than certain values, the losses approach to zero. Such marginal effects make sure that the learning process does not waste energy on unchallenging negative pairs that already have large distance. For positive cases, however, notably different characteristics

are presented by the two losses. The contrastive loss simply evaluates the distance with a quadratic term, which results in the fact that unchallenging positive samples would have continuous impact on the learning process. In contrast, a number of negative samples would be ignored due to the hard margin *m*. Such imbalance may mislead the training process to focus too much on positive samples, even for unchallenging ones. On the other hand, our *DSC* loss provides a (soft) marginal effect for positive cases as well, *i.e.*, when *d* is in certain small range, it produces a loss close to zero. Such minor loss values indicate that the current samples are typically unchallenging, and they do not generate noticeable gradients for the backpropagation of the training process. In this way, the learning keeps focusing on challenging samples, for both positive and negative cases, to maximally increase the discriminating capability of the network.

As would be shown from the experiments in Section V, the proposed *DSC* loss contributes to better discriminating power than conventional contrastive loss and triplet loss, especially for lower false acceptance rates.

D. Training and Test Configuration

In order to improve the network generalizability and feature effectiveness, we have adopted several commonly used data augmentation techniques for the training process, as well as feature composition during the matching phase. These measures are explained in the following.

1) Training Data Augmentation

All the training images are resized to 300×240 in advance. Besides, we have performed several *on-the-fly* image augmentation approaches. These approaches are randomly applied before each image is fed into the network, and are described in the following:

- Scaling – There is 80% probability for each image to be enlarged, with a factor randomly drawn from a uniform distribution over (1, 1.3).
- Cropping – Each image is cropped with a window of 240×240 that is randomly placed across the entire image region.
- Color/intensity jittering – For an RGB image (visible imaging), a color augmentation method called Fancy PCA as described in [13] is applied. For a grayscale image (NIR imaging), a random value drawn from N (0, 0.02) is added to its pixel intensities to simulate illumination variation.

Above random parameters are drawn once for each image in the mini-batch during the training process. When a same image appears again in a later iteration, the parameters will be randomly drawn again to create a different variant of that image. In this way, one source image can produce a good amount of different versions without consuming much of the storage space. Such augmentation measures can effectively reduce the risk of over-fitting when training deep neural networks, especially when the number of training samples is not very large.

2) Test Feature Composition

As mentioned earlier, our network model accepts 240×240 square image as the input. On the other hand, the source periocular images used in our experiments have rectangular aspect ratios close to 5:4. During the test phase, we adopt feature composition similar to [21] and [27], to make our model adaptive to (slightly) different resolutions / aspect ratios, and also to obtain multi-scale feature representation. The composition process is described sequentially in the following:

a) The input image is resized to *w*×240, where *w* is larger than 240 and subject to the image's original aspect ratio.
b) The resized image is cropped with three 240×240 windows that are placed on the left end, center and right end of it respectively.
c) The resized image is enlarged with a factor of 1.2, then another 240×240 window is placed in the center of it, to create the fourth cropped version.
d) Four cropped versions are fed into the network separately, each generating a 128-D feature vector. These four vectors are then concatenated into a 512-D vector for the matching.

The Euclidean distance between two vectors is regarded as the dissimilarity score. Above feature composition process can cover the entire image region and account for the multi-scale feature representation to certain extent.

III. Analysis on Region Selection

In this section, we provide justification on the selection of pre-defined regions for the visual attention enhancement. As outlined earlier, we select eyebrow and eye as the RoI mainly due the following two reasons:

a) Inspired by human perception, eyebrow and eye regions will attract most of attention when humans observe periocular images. It is useful to note that many machine learning / deep learning algorithms are inspired by human perception/ behaviors, including neural networks, reinforcement learning, long-short term memory (LSTM) and also the referenced attention models in this paper.
b) The importance of eyebrow and eye characteristics for periocular recognition has been ascertained by a number of earlier research works [1],[3],[8],[15],[47], where excluding or masking eyebrow or eye regions leads to performance degradation in most cases.

In order to statistically ascertain the effect of selecting these areas for attention enhancement, we also trained different versions of AttNet by adjusting the feature weights α in Equation (1), detailed as follows:

- Eye + Eyebrow: $\alpha_{\text{eye}} = \alpha_{\text{eyebrow}} = 5$
- Eye only: $\alpha_{\text{eye}} = 5,\ \alpha_{\text{eyebrow}} = 1$
- Eyebrow only: $\alpha_{\text{eye}} = 1,\ \alpha_{\text{eyebrow}} = 5$
- No attention: $\alpha_{\text{eye}} = \alpha_{\text{eyebrow}} = 1$

The above settings enable a preliminary investigation for the effect of selected regions, for attention enhancement, on the recognition results. The results from such comparative analysis using UBIPr database are shown in Fig. 7. It can be observed

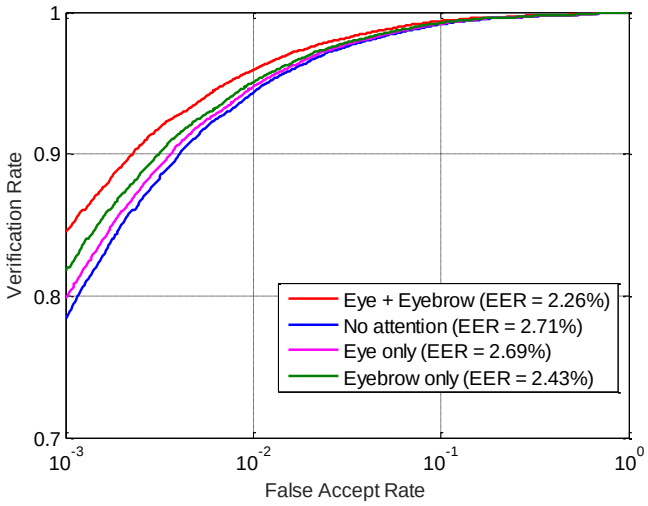


Figure 7: Comparison of different weights on the selected regions of interest for attention incorporation.

that with explicitly enhanced attention on eye and eyebrow regions simultaneously, we can largely benefit the recognition accuracy. Emphasizing the eyebrow regions separately yields higher improvement than focusing only on the eye regions. This is likely due to the fact that the eyebrow characteristics are more stable and resistant to the illumination variations, eyeball movements, *etc*. The above observations have validated the positive effect of incorporating visual attention within the detected eyebrow and eye regions during deep feature extraction for more accurate periocular recognition.

IV. Analysis on Training

The effectiveness of training scheme is a key aspect for the success of deep learning based approaches, which is related to a number of factors such as the classification task, network complexity, volume of training data and learning algorithm. As compared with typical deep learning solutions for ImageNet classification [18-20], semantic segmentation [30], *etc.*, one of the most critical challenges when researchers explore deep learning's potential for biometrics problems are on the availability of *large* amount of labeled training data. Insufficient training data can cause severe over-fitting, *i.e.*, the model fits too well on the small scale of training data but is not able to properly classify test data which was unseen during training phase. In this section, we present analysis on the training processes for *AttNet* and *FCN-Peri* to validate that our models are adequately trained and the level of over-fitting is within acceptable range.

A. Training of AttNet

There is no definite conclusions so far on the minimum required numbers to properly train a CNN for the classification problem. Generally, it is accepted that when there are more parameters to learn and the problem is more complicated, the required amount of training data will be larger in order to avoid over-fitting. A practical way is to refer to some typical architectures and the training configuration which have been widely adopted by researchers/developers in the literature. Table 2 presents the summary of scale of our networks as well as some existing architectures for different classification tasks.

Table 2: Comparison of network configurations for our work and other typical architectures.

Architecture	Problem	#Classes	#Param.	# Train Images
AlexNet [13]	Image class.	1,000	60M	~1M
VGG-16 [27]	Image class.	1,000	138M	~1M
ResNet-152 [19]	Image class.	1,000	60M	~1M
DeepIrisNet [48]	Iris recog.	356	138M	~30K
PRWIS [46]	Periocular recog.	518	248M	~8K
AttNet	Periocular recog.	224	**7.7M**	~3K
FCN [30]	Semantic segm.	21	134M	~8K
FCN-Peri	Semantic segm.	3	**0.1M**	100

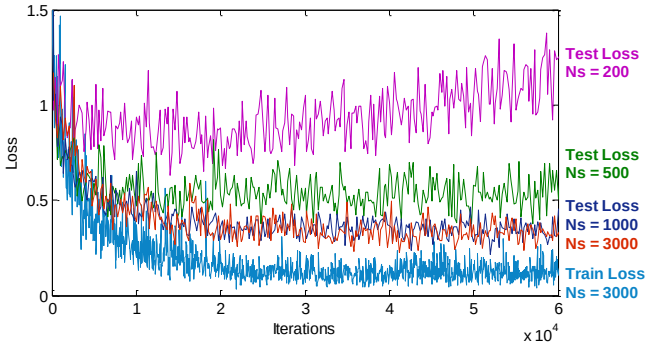


Figure 8: Learning status of *AttNet* with different number of training samples (N_S). With N_S no less than 1,000, test loss converges to a stable level. Train losses with different N_S are similar and therefore only one is plotted for clarity. *Best viewed in color*.

It can be inferred from Table 2 that (1) our network is much smaller than other typical network architectures in terms of parameter scale, and it is therefore reasonable to assume that the required number of training samples should be less than other examples in this table; (2) For the general image classification tasks such as in [13] or [27], intra- and inter-class variation are dramatically high and therefore a large volume of training data should be devoted for sufficient learning. On the other hand, for typical biometric problems such as for iris or periocular recognition, relatively small amount of training has been employed to achieve promising results. This is probably because smaller inter-image variation for biometric recognition may not require that many training samples to supply over-complex information. The periocular recognition problem discussed in this paper belongs to such category of problems. Considering the above two factors, our configuration for training the small *AttNet* with about 3,000 (on UBIPr dataset which will be detailed in the next section) images is justifiable.

In order to statistically evaluate the convergence condition of our configuration, we vary the number of training samples to train *AttNet* on UBIPr database, for several times, and observe the convergence status. These results are shown in Fig. 8. It can be observed that employing several hundreds of training images may easily cause over-fitting as there is a large gap between the train loss and the test loss. However, when this number

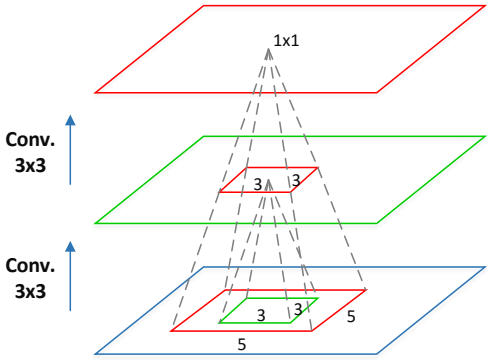


Figure 9: Illustration of receptive fields. Through one or more convolutional or pooling layers, each output neuron in the top layer is determined by a patch in the bottom/input layer.

increases to 1,000 or above, test loss converge to similar level and the gap becomes smaller. Note that it is difficult to totally eliminate the gap for most deep learning approaches. The above results indicate that the actual configuration we incorporated in this paper, in which approximately 3,000 images were used for training *AttNet*, is practically appropriate for sufficient training.

B. Training of FCN-Peri

The problem of training an FCN for semantic segmentation is quite different from training a CNN for image classification. Semantic segmentation (*e.g.*, detecting eyebrow and eye regions in this paper) is a pixel-wise classification task, rather than entire image classification task. In other words, with semantic segmentation, each pixel in the input image is classified into one of several pre-defined classes. Therefore, analysis on the number of training samples, or data points, should be casted at pixel level instead of image level. However, not all the pixels should be considered as independent data points, as adjacent pixels will have highly redundant information. The concept of *receptive field* can help to more scientifically estimate meaningful data points in an image when training FCN.

In a single or multiple regular convolution/pooling operations, one output element or pixel is computed from a certain region in the input image/map, and this region is referred to as the *receptive field*. For example, with one convolutional layer in CNN/FCN having a 3×3 kernel, the receptive field is 3×3. With two such convolutional layers, the receptive field from input to output is 5×5. Fig. 9 can help to illustrate the concept. Since FCN mainly comprises convolutional layers and pooling layers, the output of each element/pixel is determined by a patch from the input rather than the entire image. We can therefore compute the receptive field of *FCN-Peri* first to estimate the approximate number of non-redundant data points available in the training process.

The receptive field can be computed in a top-down manner to identify the region at bottom layer determining one pixel at the topmost layer. Following the longest path from input to output in *FCN-Peri*, this process is illustrated in the following:

Layer	Kernel, Stride	Receptive Field
output	-	1×1
upsample×3	-	2×2
conv4	3×3, 1	4×4
pool3	4×4, 4	16×16
conv3	3×3, 1	18×18
pool2	2×2, 2	36×36
conv2	3×3, 1	38×38
pool1	2×2, 2	76×76
conv1	5×5, 1	80×80

These observations indicate that each output pixel of *FCN-Peri* is determined by a patch of 80×80 from the input image. We can roughly assume that two patches can be considered as independent data points when the overlap between them is no less than 25% (otherwise the information will be highly redundant). As a result, a 300×240 image we used as input can provide approximately 108 (9×12) non-redundant data points. As discussed earlier, we have labelled about 100 images for training *FCN-Peri*, generating approximately 10,000 data points for learning classification of three classes (*i.e.*, eyebrow, eye and background). On average, about 3,000 training samples per class are available for training. Note that network is more than 1,000 times smaller than the original FCN as revealed from Table 2, which suggest that the number of available training samples should be sufficient. In fact, the segmentation results on test data shown in Fig. 3, which were visually appropriate, can also validate that our *FCN-Peri* has been properly trained.

V. Experiments and Results

Thorough experiments have been performed to evaluate the proposed approach from various perspectives, and comparisons are made with several state-of-the-art methods. Our experimental results are reproducible via [38]. We have conducted two sessions of experiments, which focuses on *Open-World problem* and *Closed-World problem* respectively. In this section we detail the problem definition, experimental configurations as well as observation and analysis on the results.

A. Open-World vs. Closed-World Verification

The open-world problem refers to the configuration that the subjects to be enrolled into the gallery in the deployment process may be *unseen* during the training phase. On the other hand, the closed-world problem has a constraint that all the subjects to be recognized in the deployment process are already *known* during the training phase.

The open-world problem is apparently more challenging but closer to the real deployment environments for most applications, such as citizen authentication, general access control and searching for missing people, as it is not feasible for these systems to collect data from all possible subjects in advance during training/development phase. The closed-world setting may result in higher recognition accuracy as more precise data adaptation can be achieved during training, but the system may be less scalable for the deployment, which is also clarified by [45].

Table 3: Summary of the employed databases for training and testing. The training sets of FRGC and CASIA.v4-distance are used for training [2] and [8]. Our method and [9] only adopt UBIPr and FOCS for training.

Database	**UBIPr**		**FRGC**		**FOCS**		**CASIA.v4-dist.**		**UBIRIS.v2**		**VISOB**	
Spectrum	visible		visible		NIR		NIR		visible		visible	
Imaging distance	4 – 8m		N/A		N/A		≥3m		3 - 8m		8 - 12 in.	
World scenario	Open		Open		Open		Open		Open/closed		Closed	
Division	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
#Subjects	224	120	13	150	80	56	10	131	518	518	484	475
#Images	3,359	1767	40	500	3,262	1,530	79	998	8,886	2,215	5,270	5,103
#Genuine scores (Test)	12,351		826		39,614		3,371		2,215		4,914	
#Imposter scores (Test)	1,547,910		123,425		1,130,071		494,132		1,145,155		2,464,938	

It should be clarified that the approach presented in this paper, especially the newly developed DSC loss function, are proposed for the open-world problem. However, we noticed that some recent method and contest [43] [45] in the literature focus on closed-world problem only, and therefore we investigate the performance under both settings.

B. Baseline Methods

Several state-of-the-art methods, *i.e.*, [2], [8], [9] and [45], are selected as baselines to evaluate the performance of proposed approach. These methods are used as baselines because they focus on the same problem with us, *i.e.*, less constrained periocular recognition, and report state-of-the-art performance on multiple datasets in the recent years and with judicious theoretical significance. It should be noted that the methods in [2], [8], [9] and also ours are adaptive to the open-world setting, while [45] is only developed for closed-world setting as also clarified in their paper.

C. Datasets and Protocols

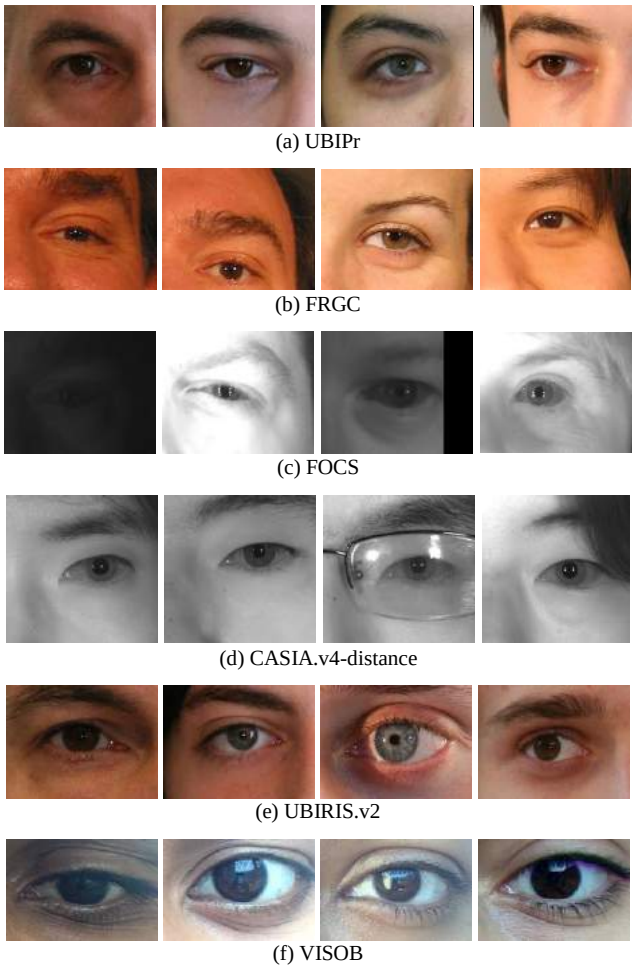
We employ six publicly available databases for the experiments. Four of them are acquired under visible spectrum while the other two are with NIR imaging. The brief information of the employed datasets are described in the following.

● UBIPr [36]

This database contains 5,126 left and 5,126 right periocular images from 344 subjects, and simulates less constrained periocular acquisition environment under visible spectrum. Noticeable amount of images from this dataset present occlusion, off-angle or illumination variation. For the experiments, only left periocular images are used. We employed the same training set of 3,359 images as used in [9] for model learning. The remaining 1,767 left images are used for test phase for performance evaluation. This database is used for *open-world* experiments and therefore no subjects are overlapping between the training and test sets.

● Face Recognition Grand Challenge (FRGC) [31]

This dataset is released by the National Institute of Standards and Technology (NIST), and was primarily provided for evaluating advanced algorithms for the face recognition. Similar to as in [2] and [9], the periocular regions are automatically extracted from the source face images within a subset of 540 samples, using publicly available face and eye detection algorithms [32]-[33]. The 540 right eye images from 163 subjects are employed in our experiments, from which the first 40 images form the training set and the rest 500 form the test set. Experiments on this dataset also adopt the *open-world* configuration.



(a) UBIPr (b) FRGC (c) FOCS (d) CASIA.v4-distance (e) UBIRIS.v2 (f) VISOB

Figure 10: Sample images from the employed databases, which present noticeable pose, illumination variation and occlusions due to the less constrained imaging environments.

● Face and Ocular Challenge Series (FOCS) [34]

The FOCS dataset is also released by NIST, and comprises face, ocular images and videos acquired under NIR imaging spectrum. We employ 4,792 left periocular images from 136 subjects of the "OcularStillChallenge1" part for the experiments. The imaging condition for this dataset is highly challenging that many of the images suffer from significant

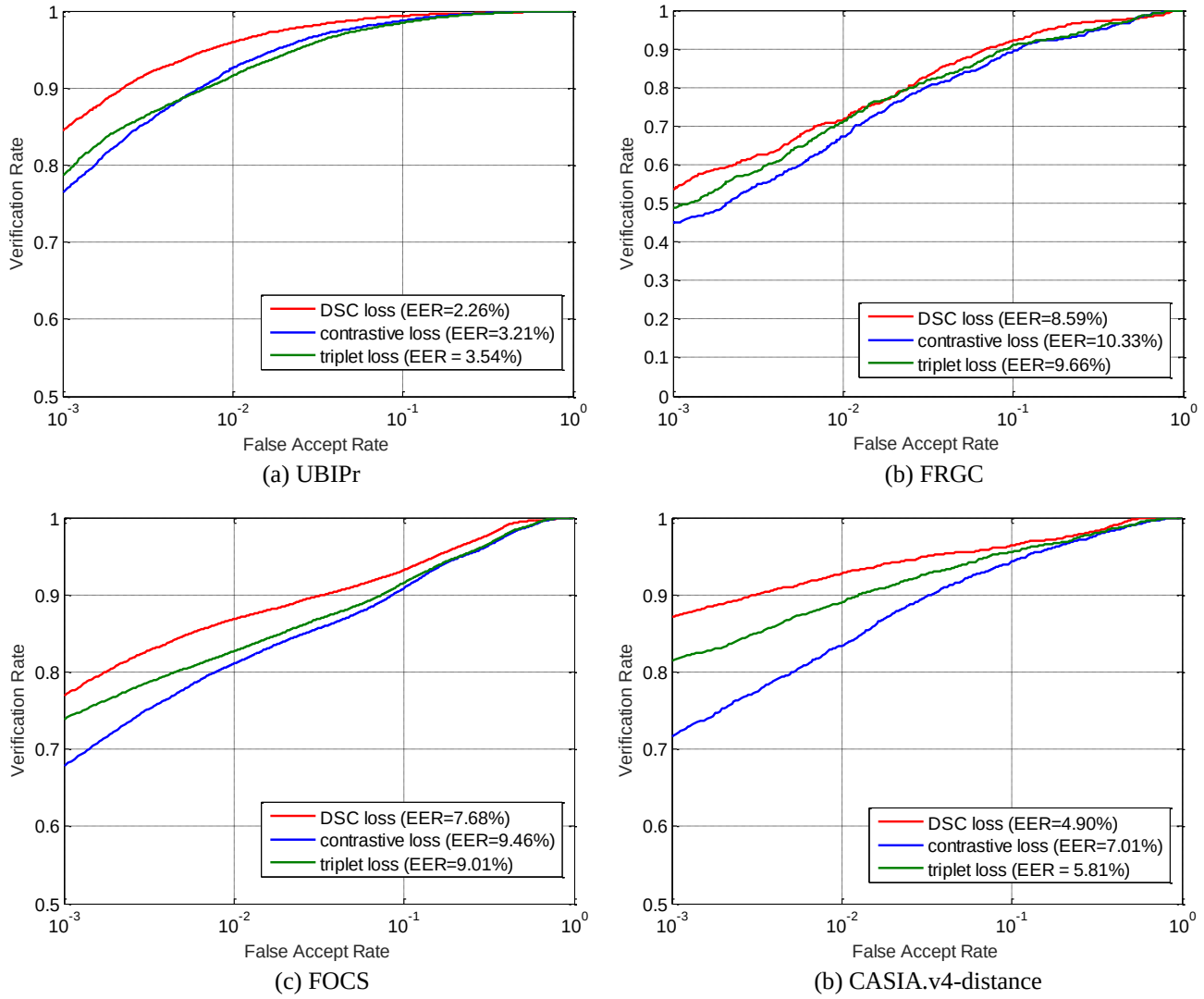


Figure 11: ROCs of training *AttNet* with *DSC* loss and conventional losses on four employed databases. The parameters of DSC loss are empirically set to a=10 and b=5; margins for contrastive loss and triplet loss are tuned among {1, 2, 3, 4} and the best performing ones are used here for comparison, which are m=3 and m'=4.

illumination variation and misalignments. It can be observed that some extremely low-quality samples are included in this dataset, which brings great challenges to accurate recognition. We use 3,262 left periocular images from the first 80 subjects for training, and the remaining 1,530 images from 56 subjects for testing. *Open-world* configuration is applied for this dataset.

● CASIA.v4-distance [35]

This database contains 2,567 upper face images from 142 subjects in single session, which are acquired under NIR spectrum at a distance (> 3m). Similar as for FRGC, the publicly available eye detector [32]-[33] is applied to automatically segment the left periocular images which are used in our experiments. The first eight samples of each subject, excluding a few very poorly segmented images, form the subset of 1,077 images used in our experiments. The first 79 samples are used for training while the remaining 998 samples are used during the test phase. Experiments on this dataset also follow the *open-world* protocol.

● UBIRIS.v2 [46]

This dataset is released for noisy iris recognition under visible spectrum. The full set contains 11,101 eye images from 518 subjects, which are acquired from 3-8 meters away. Experiments on this dataset is mainly set for *closed-world* verification and comparison with method [45], but will also attach *open-world* results for comparative study. In the closed-world setting (as in [45]), 80% of images from all 518 subjects are used for training and the remaining 20% are selected for testing. In the open-world setting, images from the first 400 subjects are used for training while the remaining are used for testing.

● VISOB [43]

This competition dataset comprises ocular images captured with three different smartphones under three illumination conditions. The Visit-1 involves 550 subjects and was released for algorithm development. The Visit-2 has images from 290 subjects and was used for performance evaluation in the competition. It is important to note that the competition organizers have only provided Visit-1 part of this database in public domain and was also downloaded by us. Therefore our

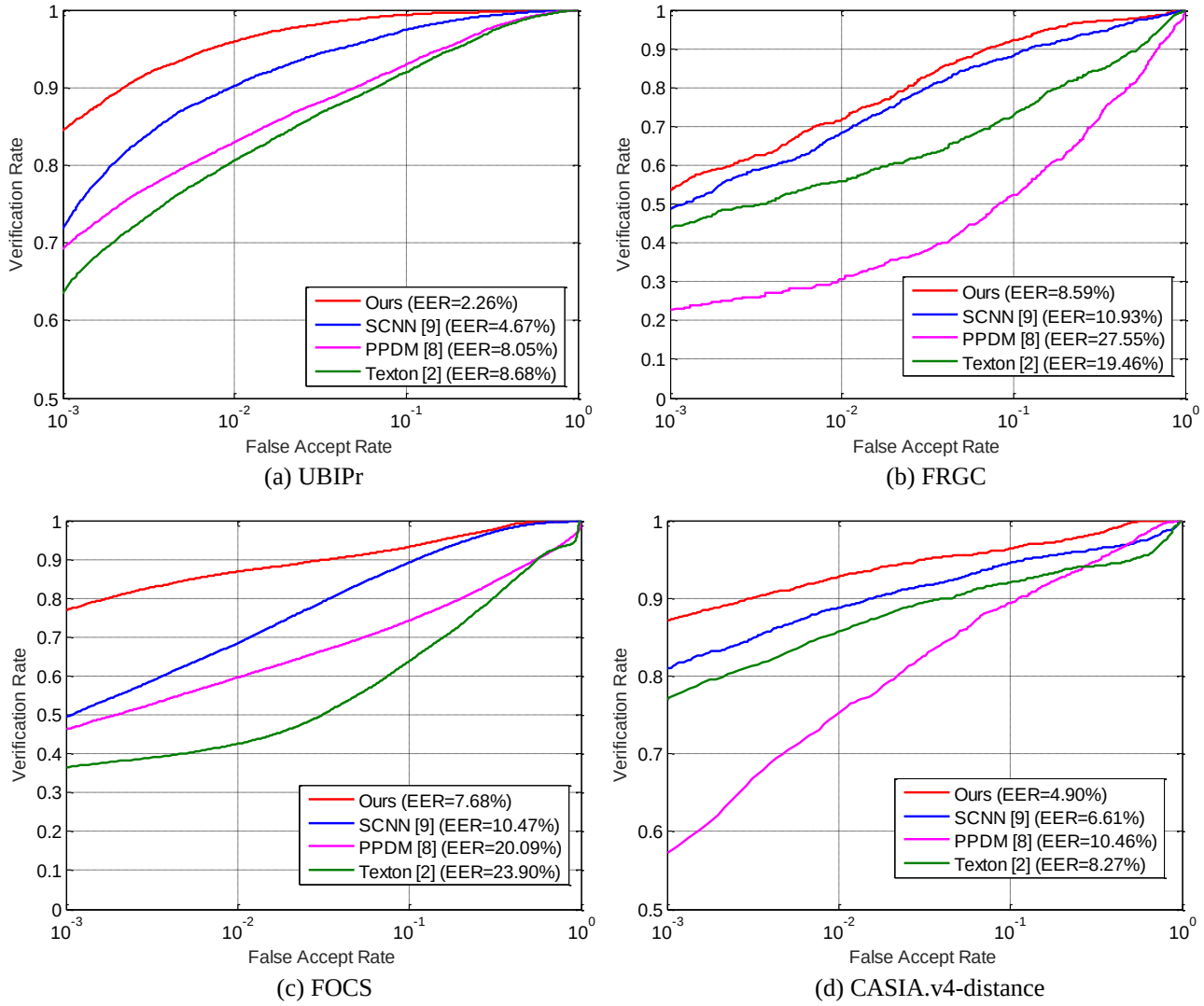


Figure 12: ROC curves of the periocular verification using our method and comparison with other state-of-the-art methods on different databases.

experimental results were obtained on Visit-1 part only and should not be directly compared with the published ranked methods in [43]. *Closed-world* setting was applied on the experiments on this dataset.

Above datasets cover both visible and NIR spectrums, and were acquried under varying and less constrained imaging environments that are close to real world application scenarios. A few sample images from them are provided in Fig. 10. More detailed information about the employed databases and training/test set division is provided in Table 3.

For experiments carried out under open-world configuration, it is important to clarify the reasonable difference of training mechanisms for the four methods: a) For our method and [9], the visible models are trained on UBIPr database and tested on UBIPr and FRGC databases; the NIR models are trained on FOCS and tested on FOCS and CASIA.v4-distance datasets. In other words, experiments on FRGC and CASIA.v4 are under *cross-database* scenarios. Such a training/test configuration is identical to the original one in [9], which therefore provides a fair comparison. b) For methods [2] and [8], the required training efforts are less, and it is observed that the *within-database* training and testing manner offers better results for these two methods. Therefore the training and testing are performed on the same dataset for them. Aforementioned experimental configuration is also the same as used in [9], and justification has been provided to incorporate the best possible performance from these two baseline methods and ensure fairness in the performance comparisons.

Table 4: Results of significance test for comparison of our method and [9]. *p*-value indicates the probability of the null hypothesis, *i.e.*, two sets of data do not differ significantly.

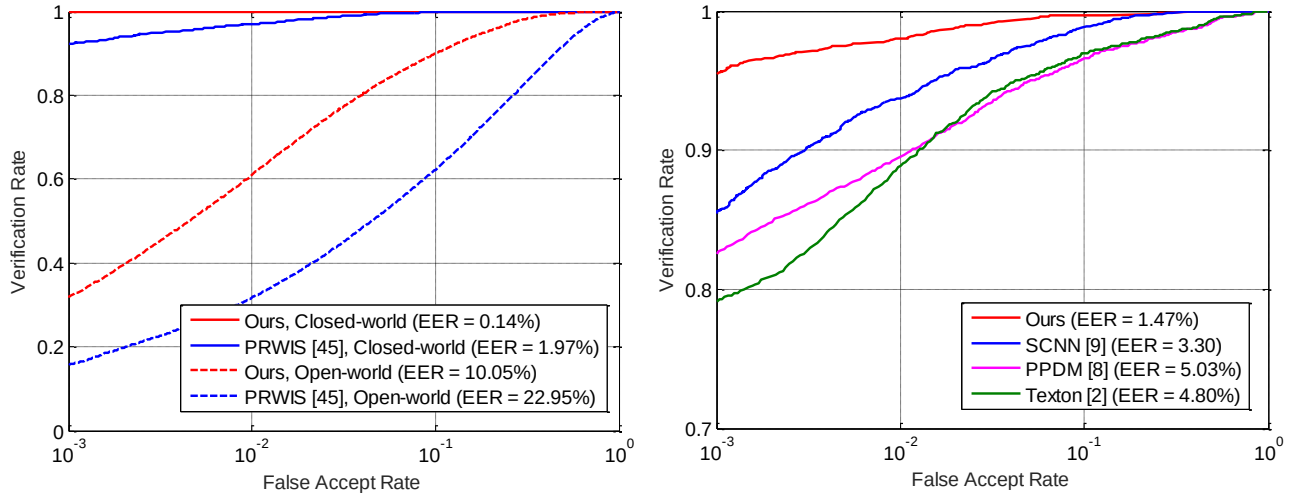
Comparison with [9]	UBIPr	FRGC	FOCS	CASIA.v4 -distance
z-statistic	14.323	3.859	25.259	8.829
p-value*	$<10^{-4}$	1.14×10^{-4}	$<10^{-4}$	$<10^{-4}$

* *p* will be denoted as $<10^{-4}$ if the computed *z* is too large such that the corresponding *p* is too small for the computer to return exact value.

D. Open-World Performance

1) Effectiveness of DSC Loss Function

The performance of the proposed *DSC* loss function which is designed for *open-world* verification is firstly examined. We compare it with conventional contrastive loss and triplet loss,



(a) UBIRIS.v2 (open-/closed-world) (b) VISOB (closed-world)

Figure 13: ROC curves on UBIRIS.v2 database and VISOB database (iphone-day-light-short subset). Note that the *AttNet* result under closed-world setting on UBIRIS.v2 is close to line $y = 1$. *Best viewed in color*.

which are also designed for 1:1 verification purpose. The experiment is performed on UBIPr, FRGC, FOCS and CASIA.v4-distance. Three *AttNet* models with identical structures are trained with *DSC* loss, contrastive loss and triplet loss respectively. When training with contrastive loss and triplet loss, the margins are discretely tuned from {1, 2, 3, 4}, and the ones providing best performance are used for comparison. The receiver operating characteristic (ROC) curves are shown in Fig. 11.

It can be observed that *DSC* loss delivers noticeable and consistent improvements over the other two loss functions, especially for lower false acceptance rates (FAR). The performance at low FAR is regarded as more important for biometric verification systems, and the key factor to this metric is the ability to verify challenging cases, *i.e.*, highly dissimilar genuine pairs and similar imposter pairs. The superiority of *DSC* loss is mainly attributed to the marginal effects for both positive and negative pair samples during the feature learning process, such that more training efforts can be put into challenging cases.

2) *Comparison with State-of-the-art Works*

As discussed earlier, the performance of the proposed approach has been comparatively evaluated with state-of-the-art methods [2], [8] and [9] in the literature. The resulting ROC curves are provided in Fig. 12. We can observe from these results that our method consistently outperforms the other three baseline methods on all of the four employed databases. It is important to note that the advancements from our method are particularly significant at lower FAR, which indicates the outstanding capability of our method for verifying challenging periocular samples. Even under the challenging cross-database training and test protocol, the proposed method has exhibited high level of robustness. The promising results from the proposed attention-based model have further validated the importance of eyebrow and eye regions for the periocular recognition.

We have also performed significance tests to ascertain the statistical significance of the improvements from our method. The method for the significance test is described in [14], which

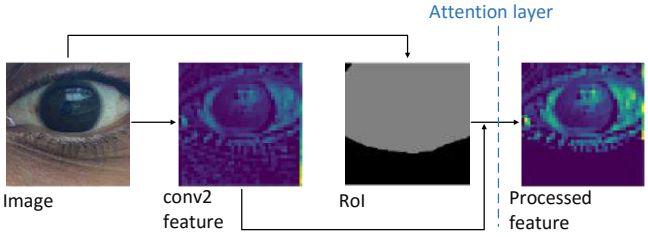


Figure 14: Visualization of convolutional features on a VISOB image which does not contain eyebrow. In this case the attention mechanism does not much impact on the feature distribution, and *AttNet* will basically act like a common CNN to guarantee fundamental performance.

is based on the area under the curve (AUC) of the ROC statistics. Comparison has been made with [9] only, as this method delivers the best performance among the three baselines. The results from the tests are provided in Table 4. It can be inferred that, with widely used confidence level of 95%, the improvements from our method are statistically significant over its competitors.

E. Closed-World Performance

As discussed earlier, the proposed approach is mainly designed for open-world verification problem. However, some recent methods/competitions also adopt or focus on closed-world setting, in which all the subjects to be recognized are known during training/development phase, and it is usually allowed to use the gallery set for the training process. Typical examples include [43][44][45]. Despite the fact that closed-world setting is less challenging, it may be feasible for some applications to know all the interested subjects in advance during training phase, such as watchlist system. Hence, we also supplement experiments under the closed-world configuration, which were conducted on UBIRIS.v2 and VISOB databases.

Under the closed-world setting, we maintained the architecture of *AttNet* but trained it in a different way. Similar to [45], we added a softmax layer after the feature layer (fc5 in Fig. 2) with N_C output neurons, where N_C is the number of classes (subjects) to be recognized. As closed-world setting is applied, N_C is consistent during training and test phases. Each output neuron at the softmax layer is regarded as the probability

that input sample belongs to a specific subject, and therefore is used as the verification score. Fig. 13 provides ROCs for the verification results on UBIRIS.v2 with comparison to [45], and on VISOB with comparison to [2], [8] and [9]. Note that for experiments on UBIRIS.v2, we also attached open-world results for comparative study. To obtain the comparative open-world results from [45], we used the l2-norm distance between the feature vectors from fc7 layer as suggested in their paper.

From the results on UBIRIS.v2, we can observe that our approach consistently outperforms the recently published state-of-the-art method [45]. Under the closed-world settings, our results have scored significantly high accuracy (0.14% EER), due to reason that class-specific recognition has been learned with softmax loss function for given and fixed set of subjects (and same for the baseline method). In contrast, when switched to open-world setting, both [45] and our method suffer from obvious performance degradation, which reflects that open-world problem inherently brings more challenges compared with the closed-world problem. However, our appraoch can still achieve superior results over that from [45].

The results on VISOB dataset reveal that our method still consistently outperforms other methods investigated in this paper. It should be noted that the eye images in this database do not include the eyebrow region, and the eye region occupies most the image area (Fig .10f). This implies that the proposed visual attention mechanism may not benefit much the recognition performance. Fig. 14 visulizes the intermediate features learned by *AttNet* on such data, from which we can observe that enhancing attention within the eye region does not affect much the feature contents. In this case, *AttNet* can serve as a common CNN for backing up the perfomance even if desired regions are absent or can not be correctly segmented. Another aspect worth noticing is that, as discussed earlier, *only* Visit-1 subset (550 subjects) is provided in public domain but not the Visit-2 (290 subjects) part that was used for benchmarking in [43]. Therefore it will be unfair to directly compare the results provided in this paper with those in [43].

VI. Conclusions

This paper has developed an attention based CNN architecture for more accurate and robust periocular recognition. The proposed framework includes *FCN-Peri*, which can accurately detect eyebrow and eye regions as key regions of interest, and *AttNet*, which makes use of the RoI information for more discriminative feature learning. A newly developed verification oriented loss function, referred to as *DSC* loss, has also been introduced in this paper. The new loss function has shown to provide marginal effects for both positive and negative training samples during learning, which contributes to more robust feature representation for matching challenging periocular image pairs. Extensive experiments on four publicly available databases presented in Section V of this paper indicate that, the proposed attention-based framework achieves significantly better results than several state-of-the-art methods for the periocular recognition. The effectiveness of the newly designed *DSC* loss function was also separately validated through comparison with conventional contrastive loss and triplet loss. The experimental results provide strong support to our assumption that, information within eyebrow and eye regions are critical to periocular recognition, and deserve more attention during feature learning and matching. The trained models and source for reproducing our experimental results are made publicly available via [38].

Despite success in simulating human visual attention model for the automated periocular recognition, as illustrated from promising results on multiple databases in this paper, a lot more work needs to be done, *e.g*., to develop on-the-fly and more intelligent RoI learning through the feedback from the feature learning process, on the basis of pre-trained *FCN-Peri*. More robust and adequate visual attention mechanisms, in addition to the currently used feature adjustment strategy, is also expected to further improve the performance and therefore pursued in the future extension of this work. Last but not least, the separate impact from each of eyebrow and eye regions is another interesting and important aspect to investigate.

References

[1] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study", in *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2009, pp. 1-6, DOI: 10.1109/BTAS.2009.5339068.

[2] C. W. Tan and A. Kumar, "Towards online iris and periocular recognition under relaxed imaging constraints", *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3751-3765, 2013.

[3] S. Bharadwaj, H. S. Bhatt, M. Vatsa and R. Singh, "Periocular biometrics: When iris recognition fails", in *2010 IEEE 4th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2010, pp. 1-6, DOI: 10.1109/BTAS.2010.5634498.

[4] C. N. Padole and H. Proenca. "Periocular recognition: Analysis of performance degradation factors." in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 439-445, DOI: 10.1109/ICB.2012.6199790.

[5] G. Santos and H. Proenca, "Periocular biometrics: An emerging technology for unconstrained scenarios", in *2013 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, 2013, pp. 14-21, DOI: 10.1109/CIBIM.2013.6607908.

[6] L. Nie, A. Kumar and S. Zhan, "Periocular recognition using unsupervised convolutional RBM feature learning", in *2014 22nd International Conference on Pattern Recognition (ICPR)* , pp. 399-404, 2014, DOI: 10.1109/ICPR.2014.77.

[7] A. Sharma, S. Verma, M. Vatsa and R. Singh, "On cross spectral periocular recognition", in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5007-5011, DOI: 10.1109/ICIP.2014.7026014

[8] J. Smereka, V. Boddeti and B. Vijaya Kumar, "Probabilistic deformation models for challenging periocular image verification", *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 9, pp. 1875-1890, 2015, DOI: 10.1109/TIFS.2015.243421.

[9] Z. Zhao and A. Kumar, "Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network", *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1017-1030, 2017, DOI: 10.1109/TIFS.2016.2636093.

[10] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui and C. Y. Suen, "Investigating age invariant face recognition based on periocular biometrics", in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1-7, DOI: 10.1109/IJCB.2011.6117600.

[11] R. Jillela and A. Ross. "Mitigating effects of plastic surgery: Fusing face and ocular biometrics", in *2012 IEEE 5th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012, pp. 402-411, DOI: 10.1109/BTAS.2012.6374607.

[12] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998, DOI: 10.1109/5.726791.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in Neural Information Processing Systems (NIPS) 2012*, 2012, pp. 1097-1105.

[14] E. DeLong, D. DeLong and D. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", *Biometrics*, vol. 44, no. 3, p. 837, 1988, DOI: 10.2307/2531595.

[15] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, and A. Ross, "On the fusion of periocular and iris biometrics in non-ideal imagery", in *2010 20th IEEE International Conference on Pattern Recognition (ICPR)*, 2010, pp. 201-204, DOI: 10.1109/ICPR.2010.58.

[16] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580-587, DOI: 10.1109/CVPR.2014.81.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", in *Advances in neural information processing systems (NIPS) 2015*, 2015, pp. 91-99.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9, DOI: 10.1109/CVPR.2015.7298594.

[19] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2016, pp. 770-778, DOI: 10.1109/CVPR.2016.90.

[20] G. Huang, Z. Liu, M. Laurens, W. Kilian Q, "Densely connected convolutional networks", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700-4708, DOI: 10.1109/CVPR.2017.243.

[21] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes", in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1891-1898, DOI: 10.1109/CVPR.2014.244.

[22] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification", in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701-1708, DOI: 10.1109/CVPR.2014.220.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention", in *Proceedings of the 32nd International Conference on Machine Learning* (*ICML*), 2015, pp. 2048-2057.

[24] V. Mnih, N. Heess and A. Graves, "Recurrent models of visual attention", in *Advances in Neural Information Processing Systems (NIPS) 2014*, 2014, pp. 2204-2212.

[25] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 842-850,

[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, "Residual attention network for image classification", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450-6458, DOI: 10.1109/CVPR.2017.683.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[28] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification", in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1 pp. 539-546.

[29] F. Schroff, K. Dmitry and P. James, "Facenet: A unified embedding for face recognition and clustering", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815-823.

[30] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440, DOI: 10.1109/CVPR.2015.7298965.

[31] "FRGC dataset", 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/frgc.cfm. [Accessed: 29- Mar- 2016].

[32] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 1, pp. I-511-I-518, DOI: 10.1109/CVPR.2001.990517.

[34] "FOCS dataset", 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/focs.cfm. [Accessed: 29- Mar- 2016].

[35] "CASIA.v4 dataset", 2016. [Online]. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=4. [Accessed: 29- Mar- 2016].

[36] "UBIPr dataset", 2016. [Online]. Available: http://socia-lab.di.ubi.pt/~ubipr/. [Accessed: 29- Mar- 2016].

[37] "Biometric Evaluations Homepage", *Nist.gov*, 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/biometric_evaluations.cfm. [Accessed: 29- May- 2016].

[38] Web link to download the implementation for Attention Network detailed in this paper, http://www.comp.polyu.edu.hk/~csajaykr/attnet.htm.

[39] F. Alonso-Fernandez and J. Bigun, "A survey on periocular biometrics research", *Pattern Recognition Letters*, vol. 82, pp. 92-105, 2016, DOI: j.patrec.2015.08.026.

[40] A. Rattani and R. Derakhshani, "Ocular biometrics in the visible spectrum: A survey", Image and Vision Computing, vol. 59, pp. 1-16, 2017. DOI: 10.1016/j.imavis.2016.11.019.

[41] J. Smereka, B. Vijaya Kumar and A. Rodriguez, "Selecting discriminative regions for periocular verification", in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016, pp. 1-8, DOI: 10.1109/ISBA.2016.7477247.

[42] A. Kumar and K. Wang, "Identifying humans by matching their left palmprint with right palmprint images using convolutional neural network," *Proc. DLPR 2016, Intl. Workshop on Deep Learning in Biometrics*, Cancun, Mexico, Dec. 2016..

[43] A. Rattani, R Derakhshani, S. K. Saripalle and V. Gottemukkula, "ICIP 2016 competition on mobile ocular biometric recognition", in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 320-324, DOI: DOI: 10.1109/ICIP.2016.7532371.

[44] R. Raghavendra and C. Busch, "Learning deeply coupled autoencoders for smartphone based robust periocular verification", in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 325-329, DOI: 10.1109/ICIP.2016.7532372.

[45] H. Proença and J. C. Neves, "Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks", *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 888-896, 2018, DOI: 10.1109/TIFS.2017.2771230.

[46] H. Proenca, S. Filipe, R. Santos, J. Oliveira and L. Alexandre, "The UBIRIS.v2: A Database of Visible Wavelength Iris Images Captured On-the-Move and At-a-Distance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1529-1535, 2010, DOI: 10.1109/tpami.2009.66.

[47] J. M. Smereka, and B. V. Kumar, "What Is a 'Good' Periocular Region for Recognition?" In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 117-124, DOI: 10.1109/CVPRW.2013.25.

[48] A. Gangwar and A. Joshi, "DeepIrisNet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition", In *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2301-2305.