Joint Iris Segmentation and Localization Using Deep Multi-task Learning Framework

Caiyong Wang, Yuhao Zhu, Yunfan Liu, Ran He, *Senior Member, IEEE*, and Zhenan Sun, *Member, IEEE*

***Abstract*—Iris segmentation and localization in non-cooperative environment is challenging due to illumination variations, long distances, moving subjects and limited user cooperation, etc. Traditional methods often suffer from poor performance when confronted with iris images captured in these conditions. Recent studies have shown that deep learning methods could achieve impressive performance on iris segmentation task [1]–[5]. In addition, as iris is defined as an annular region between pupil and sclera, geometric constraints could be imposed to help locating the iris more accurately and improve the segmentation results. In this paper, we propose a deep multi-task learning framework, named as IrisParseNet, to exploit the inherent correlations between pupil, iris and sclera to boost up the performance of iris segmentation and localization in a unified model. In particular, IrisParseNet firstly applies a Fully Convolutional Encoder-Decoder Attention Network to simultaneously estimate pupil center, iris segmentation mask and iris inner/outer boundary. Then, an effective post-processing method is adopted for iris inner/outer circle localization. To train and evaluate the proposed method, we manually label three challenging iris datasets, namely CASIA-Iris-Distance, UBIRIS.v2, and MICHE-I, which cover various types of noises. Extensive experiments are conducted on these newly annotated datasets, and results show that our method outperforms state-of-the-art methods on various benchmarks. All the ground-truth annotations, annotation codes and evaluation protocols are publicly available at https://github.com/xiamenwcy/IrisParseNet.**

***Index Terms*—Iris segmentation, iris localization, attention mechanism, multi-task learning, iris recognition**

I. Introduction

Iris recognition has been considered as one of the most stable, accurate and reliable biometric identification technologies [6], hence it is widely applied in various biometric applications including intelligent unlocking, border crossing control, security and crime screening, etc. A complete iris recognition system often consists of four sub-processes: iris image acquisition, iris preprocessing, feature extraction and matching. Both Iris segmentation and iris inner/outer circle localization (iris localization) are part of the iris preprocessing step [7], [8].

The supplemental material is made available via https://drive.google.com/open?id=1Fo3Nmv_ha5-d5jC2vcbAtbjMyJ_aa_fL

Caiyong Wang is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: wangcaiyong2017@ia.ac.cn.

Yuhao Zhu, Yunfan Liu, Ran He and Zhenan Sun(**Corresponding author**) are with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: {yuhao.zhu, yunfan.liu}@cripac.ia.ac.cn, {rhe, znsun}@nlpr.ia.ac.cn.

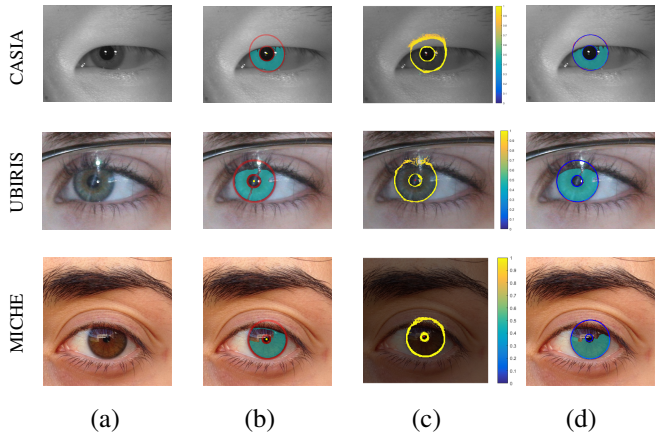


Fig. 1. The first column (a) shows iris images from three datasets (as described in Sec. IV-A) collected in different environments. The second column (b) illustrates the ground truths of pupil center, iris inner/outer boundary and iris segmentation mask, highlighted in yellow, red and aqua, respectively. The third column (c) shows the predicted pupil center (marked as red) and iris inner/outer boundary (highlighted in a color bar where the hotter color indicates the higher probability of a pixel belonging to the actual iris boundary). By utilizing the inherent correlation of pupil center, iris mask (highlighted in aqua in the column (d)) and iris inner/outer boundary, we further eliminate the noise of detected iris boundaries. As shown in the fourth column (d), with the help of refined iris boundaries and pupil center, we could extract coarse iris contours (highlighted in red) as the fitting points, then locate iris inner/outer circle (highlighted in blue) with the least-squares circle fitting algorithm [9]. Best viewed in color.

As shown in Fig. 1 (a) and (b), iris refers to an annular region between pupil and sclera. Iris boundaries are approximately defined by two circles, i.e. an inner circle that divides pupil and iris (also called pupillary boundary), and an outer circle that separates iris and sclera (also called limbic boundary). Iris segmentation aims to isolate valid iris texture region from other components, such as pupil, sclera, eyelashes, eyelids, reflections, and occlusions in an eye image to obtain a binary mask, where valid iris pixels are classified as foreground and other pixels are regarded as background. Iris localization refers to estimating the parameters (center and radius) of iris inner and outer circular boundaries. After obtaining parameters of the iris region, normalization is carried out to get normalized image and mask, then followed by feature extraction and match operations to produce the final recognition result. As the beginning of iris recognition flow, accurate segmentation and localization has a great impact on subsequent processes [10], [11]. Therefore, a segmentation and localization algorithm with high performance is the key to the success of the entire iris recognition system.

Earlier iris recognition systems require user cooperation and highly controlled imaging conditions, which restricts

the applications of iris recognition technology. Hence, it is necessary to develop less constrained iris recognition systems. However, images captured in less constrained scenarios (*e.g.* long distances, moving subjects, using mobile devices, and limited user cooperation) are often of poor quality and introduce various kinds of noise, such as partial occlusions due to eyelids or glasses and blur caused by motion and defocus, as shown in Fig. 2.

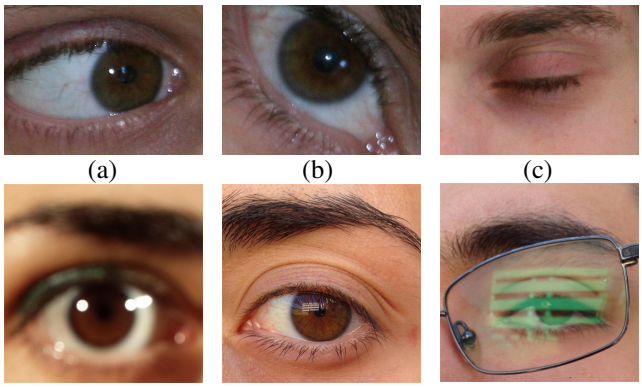


Fig. 2. Examples of degraded iris images with different types of noises. (a) gaze deviation; (b) rotation images; (c) absence of iris; (d) defocus blur; (e) specular reflections; (f) iris occlusions due to glasses.

Over the past decades, a number of methods have been proposed for iris segmentation and localization, such as Hough transform [7], [12], Active Contours [13], and GrowCut [14]. However, these methods could not work well when dealing with degraded images. Compared with these traditional approaches, deep learning models, especially Convolutional Neural Networks (CNNs), have shown incomparable advantages in tasks such as image classification [15] and object detection [16]. To be specific, hierachical semantic representations of the input image could be automatically learned in an end-to-end manner without requiring extra human efforts. Since the rapid development of deep learning, a large amount of studies using CNNs have been proposed for iris segmentation [1]–[5], iris bounding box detection [4], and pupil center detection [17]–[19]. However, to the best of our knowledge, little research attention has been devoted to locating iris inner and outer boundaries based on deep learning technology. In addition, the geometric structure of iris, i.e. the pupil center is inside the inner boundary of the iris and the iris mask is located in between the inner and outer boundaries of the iris, could serve as priori constraints in designing iris segmentation and localization algorithms.

Based on these observations, we propose a deep multi-task learning framework for simultaneous pupil center detection, iris segmentation and iris inner/outer boundary detection, followed by an effective post-processing operation for iris localization, as shown in Fig. 1 (c) and (d). Compared with single objective learning, joint learning of multi-modal eye structures makes the network learn more discriminative and essential features.

To train and evaluate the proposed model, we collect three challenging public iris datasets: CASIA-Iris-Distance [20], UBIRIS.v2 [21] and MICHE-I [22]. All these datasets contain segmentation annotations provided by other literatures. We also manually label pupil center and iris inner/outer boundary as additional ground truths for each iris image. These datasets contain various categories of noises such as blur, off-axis, occlusions and specular reflections, which could evaluate the robustness of the proposed method. To promote the research on iris preprocessing, we have made our manually annotated labels freely available to the community.

Main contributions of this paper are summarized as follows:

1) This paper introduces a novel multi-task framework which consists of two parts: the first part is a Fully Convolutional Encoder-Decoder Network equipped with attention modules which could learn more discriminative features for producing multiple probability maps. By optimizing focal loss [23] and balanced sigmoid cross-entropy loss [24], the model could alleviate the class-imbalanced problem and converge quickly. The second part is an effective post-processing method including edge denoising, Viterbi-based coarse contours detection [25] and least-squares circle fitting [9] for iris localization.
2) We select three representative iris datasets and label the pupil center as well as inner/outer boundary for each iris image. Furthermore, we build comprehensive evaluation protocols for evaluating the performance of iris segmentation and localization algorithms.
3) The proposed method achieves state-of-the-art results on various iris benchmarks. Moreover, it has strong robustness and generalization ability, providing a good foundation for subsequent iris recognition processes.

The paper is organized as follows. In Section II, we briefly review related work on iris segmentation and iris localization. Technical details of the proposed method are elaborated in Section III. Section IV introduces three databases and the annotation method that we adopt. Section V describes the evaluation protocols and analyzes experimental results. Finally, we conclude the paper and discuss future work in Section VI.

II. Related Work

This section provides an overview of literatures on iris segmentation, semantic edge detection and iris localization.

A. Iris Segmentation

Over the past decades, a number of methods are proposed for iris segmentation. In general, these segmentation methods could be classified into two main categories: boundary-based methods and pixel-based methods [1]. Boundary-based methods mainly locate pupillary, limbic and eyelid boundaries to isolate iris texture regions. On the contrary, pixel-based methods directly distinguish iris pixels from non-iris pixels according to the pixel-level appearance information.

For boundary-based methods, Daugman's integro-differential operator [26] and Wilde's circular Hough transforms [7] are the two most well-known algorithms. The most critical and fundamental assumption these two methods made is that pupillary and limbic boundaries are circular

contours. The integro-differential operator searches for the largest difference of intensity over the parameter space which normally corresponds to pupil and iris boundaries, while Hough transforms find optimal curve parameters by a voting procedure in a binary edge image. Although these methods have achieved good segmentation performance in iris images captured in controlled environments, they are time consuming and not suitable for degraded iris images. To overcome these problems, many noise removal [12], coarse iris location [27], [28] and multiple models selection [29] methods have been proposed to improve the robustness and efficiency of bounding-based iris segmentation methods. Besides, since the pupil and iris boundaries are not strictly circular, some works attempted to use geodesic active contours [30] or elliptic contours [13] to replace the circular assumption.

On the other hand, pixel-based methods exploit low-level visual information of individual pixel, such as intensity and color, to classify the pixels of interest from the background of the image. The most promising method in this category use commonly known pixel-level techniques, such as Graph Cut [13], [14], to pre-process the image and traditional classification methods, such as SVMs [31], to classify the iris pixels from non-iris pixels.

Current boundary-based and pixel-based methods are designed mainly based on prior knowledge and require much pre- and post-processing effort. Deep learning models, especially Convolutional Neural Networks (CNNs), provide a powerful end-to-end solution to effectively solve these problems.

Semantic segmentation could be considered as a pixel-wise image classification task, i.e. each pixel in the image is assigned an object class. In 2005, Long [32] *et. al.* firstly proposed Fully Convolutional Network (FCN) for semantic segmentation. Afterwards, a number of semantic segmentation methods based on FCN have been proposed, such as DeepLab series [33]–[35], U-Net [36], and PSPNet [37] to improve the performance of semantic segmentation. FCN-based methods take the whole image as input and produce a probability density map through a series of convolutional layers without involving fully connected layers. The whole model is end-to-end, which does not require any manual processing, and could achieve state-of-the-art performances of the time. Iris segmentation could be regarded as a special binary semantic segmentation problem. Hence, many FCN-based segmentation methods could be directly applied on iris images, such as [1]–[3], [5]. Inspired by the success of U-Net on binary semantic segmentation task [38]–[40], in this paper, we propose a Fully Convolutional Encoder-Decoder Attention Network for iris segmentation.

B. Semantic Edge Detection & Iris Localization

Edge detection is a classical challenge in computer vision. Previous to the rapid development of deep learning, well-known Sobel detctor and Canny detector [41] etc are widely adopted. However, traditional methods are difficult to deal with semantic edges, i.e. edges which we are interseted in. Therefore, a lot of deep learning based methods [24], [42] are proposed to solve the semantic edge detection problem. Most of these methods adopt Fully Convolutional Networks (FCNs) and directly concatenate the features of different stages to extract semantic edges. In this paper, we mainly concentrate on iris inner/outer boundary detection using deep learning models.

Classical iris localization methods usually involve Daugman's integral differential operator [26], Wildes's circular Hough Transform [7] and their variants, as described in Sec. II-A. The main idea of these methods is directly searching for the optimal parameters of inner and outer circular boundaries of iris in the parameter space. These methods are efficient but only suitable for iris images without severe distortions and noises. Different from these methods, edge detection based iris localization methods have demonstrated their superiorities on non-ideal iris images. In [28], the author adopted coarse-to-fine strategy to localize inner and outer boundaries of iris. Inner boundary is coarsely detected using an iterative search method by exploiting dynamic thresholds and multiple local cues, and outer boundary is first approximated in polar space using adaptive filters, then refined in the cartesian space. As a result, these two boundaries are robust against noises and distortions in iris images, which facilitates the subsequent circle fitting process. In [25], the Viterbi algorithm is applied on gradient maps of iris images to find coarse low-resolution contours which means selecting the least number of noisy gradients points as possible, then followed by least-squares circle fitting [9] for iris localization. Experiment results indicate that the method is accurate and robust, and does not require refined parameter adaptation to various degradations encountered. In this paper, we adopt the method proposed in [25] as the main body of our post-processing step, and use real iris boundaries extracted by deep learning models in replace of gradient maps.

III. Technical Details

In this section, we firstly introduce the whole pipeline of our method. After that, we elaborate on the proposed multi-task network framework based on Fully Convolutional Network and attention mechanism, followed by an effective post-processing approach. Finally we describe our training objectives of the proposed model.

A. Pipeline

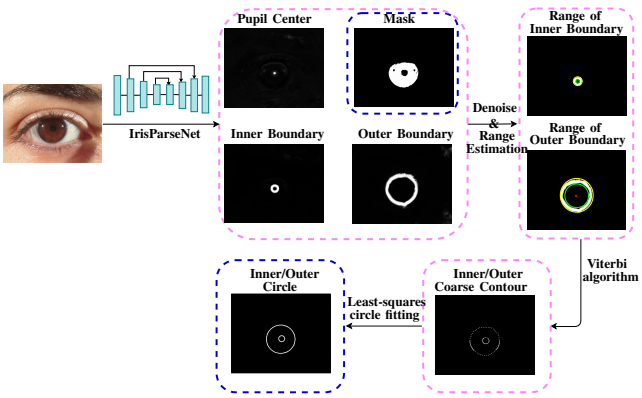


Fig. 3. The pipeline of proposed method: network output and post-processing.

The pipeline of the proposed method is illustrated in Fig. 3. IrisParseNet predicts probability maps of pupil center, iris

segmentation mask and iris inner/outer boundary. Then, we further utilize the prior geometry relations of these elements to exclude mispredicted results, remove outliers and get the range of iris inner/outer circular boundary(i.e. circle center, minimum/maximum radius). Subsequently, Viterbi algorithm [25] is used to extract coarse iris inner/outer contour. Finally iris inner/outer circle is localized by fitting on these coarse iris contours.

B. Multi-task Network Framework

Recently, Fully Convolutional Networks (FCNs) have been widely applied in many tasks such as semantic segmentation [32]–[36], edge detection [42] and salient object detection [43]. FCNs are built only with locally connected layers, such as convolution, pooling and upsampling layers, and no dense layers such as fully connected layer are used. Hence, FCNs could take images of arbitrary size as input and produce corresponding-sized output, which is desired in spatially dense prediction tasks.

Accordingly, we propose a multi-task Fully Convolutional Encoder-Decoder Attention Network framework, shown in Fig. 4, which contains an Encoder path and a Decoder path. The Encoder path encodes feature maps of CNN models by convolution, ReLu, etc., to capture semantic information. The Decoder path decodes the feature maps to recover spatial information lost in the pooling layers by concatenation with feature maps of the Encoder path.

The Encoder path adopts VGG-16 [44] as the encoding network. We remove the fully convolutional layers and the remaining network is used to learn hierarchical features. The whole encoding network could be divided into 5 stages and every stage is composed of a serial of convolutional layers, batch normalization layers, ReLU layers, and max-pooling layers which gradually reduce the size of feature maps. In lower stages, the feature maps contain more low-level spatial information such as edges but lack semantic information due to small receptive fields. In higher stages, bigger receptive fields extract more semantic information and embed it in the feature maps. In fact, many similar networks, such as ResNet [45] and DenseNet [46], could also be used as the encoding network.

As described in [37], the size of receptive fields could roughly indicate how much the context information is taken into consideration. For dense prediction task, we need to consider both the local spatial features and global, non-local semantic features. Encouraged by the high performance of DeepLab [35] and PSPNet [37] on semantic segmentation task, we directly adopt atrous spatial pyramid pooling (ASPP) and Pyramid Pooling Module (PSP) for effectively extracting multi-scale receptive fields to reflect multi-scale context information, respectively.

In order to further focus on the most important information and suppress distracting noise, we apply attention mechanism to ASPP and PSP. Attention mechanism allows us to adjust the weights of different channels in feature maps and also re-estimates the spatial distribution of feature map according to the context [47]–[51]. Hence, more discriminative features could be learned. Different from [50], we do not apply channel and spatial attention module sequentially, instead, 3D attention maps that integrating cross-channel and spatial information are directly computed.

After the attention module, we gradually up-sample the feature maps to recover the spatial information. Before up-sampling, we need to fuse feature maps from two different layers: the Encoder layer at the same stage and the Decoder layer in the previous stage. The Decoder layer encodes rich context semantic information while the Encoder layer contains the detailed spatial information. The Decoder layer in the previous stage firstly applies two sequential convolutional layers with kernel size of 3×3, batch normalization layers and ReLU layers to further refine features and reduce the number of output channels to half of the number of channels of the Encoder layer at the same stage. Then we fuse the two features by element-wise concatenation.

After fusing the feature maps of the final stage, we apply a sequence of 3×3 convolutional layer, each followed by a batch normalization layer and a ReLu layer to summarize the final semantic feature. Then, a 1×1 convolutional layer with 4 filters and a per-pixel sigmoid function are adopted to generate probability maps of pupil center, iris segmentation mask, iris inner boundary and iris outer boundary.

1) ASPP Attention Module: Atrous Spatial Pyramid Pooling (ASPP) is first proposed in DeepLab V2 [34] which is inspired by the success of spatial pyramid pooling in image classification. In ASPP, dilated convolution (or atrous convolution) with different dilation rates is adopted to extract multi-scale contextual information while keeping the spatial resolution of feature maps unchanged. The original ASPP in DeepLab V2 contains four parallel dilated convolutions with increasing dilation rate, such as 6,12,18,24, on top of the last feature map of the model. In DeepLab V3 [35], ASPP is improved in three aspects: (1) batch normalization layer is included for scale adjustment; (2) 1×1 convolution is adopted to replace the degenerated dilated convolution with a higher dilation rate, such as 24; and (3) global average pooling is connected to the last feature maps of the model to capture the global contextual information. We will incorporate the improved ASPP with attention module to effectively extract important and discriminative features. The detailed structure of ASPP Attention module is illustrated in Fig. 5.

Given an intermediate feature map F as input, a pooling layer with kernel size 3×3 and stride 1 is used to get the same sized feature map P as the new input map. Then, five parallel modules are used, including one 1×1 convolution with 256 filters (as in Eq. (1)), three dilated convolution with 256 filters and dilation rate set to 6,12,18, respectively (as in Eq. (2)-Eq. (4)), and one global average pooling layer followed by one 1×1 convolution with 256 filters and a upsampling layer, mapping the feature map back to the desired dimension (as in Eq. (5)). It is worth noting that all the convolutional layers are followed by a batch normalization layer and a ReLu layer sequentially. These five modules could be mathematically

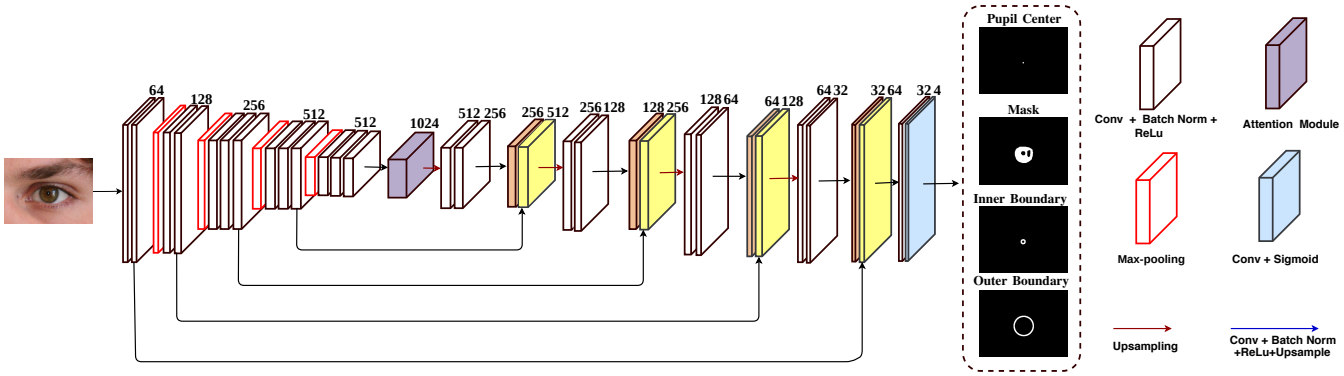


Fig. 4. Overview of Multi-task Attention Network Architecture. Best viewed in color.

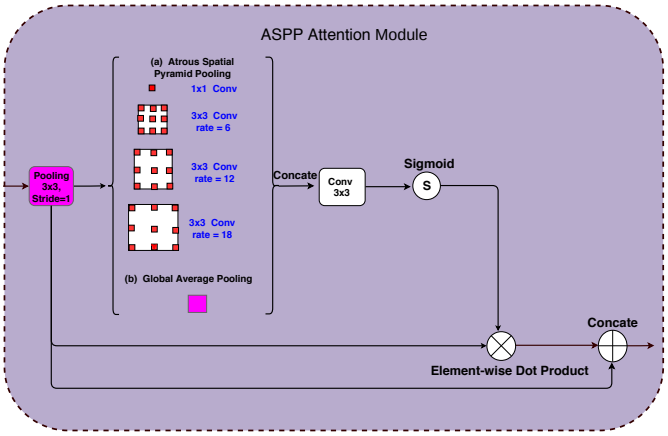


Fig. 5. An illustration of ASPP Attention Module. We extract multi-scale context features using multiple parallel filters with different dilation rates along with global average pooling. Afterwards, visual attention map is computed through one single convolution followed by a sigmoid function. Subsequently, the critical regions of input feature map are highlighted by element-wise dot production with obtained attention map. Finally, we concatenate the pooled input feature map before and after attention to get refined features.

described as follows:

$$D_1(P) = ReLu(BN(Conv_{1\times1}(P))) \quad (1)$$
$$D_2(P) = ReLu(BN(Conv^{6}_{3\times3}(P))) \quad (2)$$
$$D_3(P) = ReLu(BN(Conv^{12}_{3\times3}(P))) \quad (3)$$
$$D_4(P) = ReLu(BN(Conv^{18}_{3\times3}(P))) \quad (4)$$
$$G(P) = Up(ReLu(BN(Conv_{1\times1}(AvgPool(P))))) \quad (5)$$

The above feature maps are fused as:

$$H = D_1(P) \oplus D_2(P) \oplus D_3(P) \oplus D_4(P) \oplus G(P) \quad (6)$$

where $\oplus$ represents channel-wise concatenation. Then, we apply one single 3×3 convolution to refine the fused feature maps and reduce the number of output channel to 512 to match with the input feature map F. The final 3D attention map $M(F)$ is produced by applying a per-pixel sigmoid operation to refined feature maps. As a result, values of attention map $M(F)$ are bounded in [0,1], where the bigger value indicates the higher importance.

To focus on the more discriminative features of input feature map, the final fusion operation is defined as:

$$F' = P \oplus (P \otimes (M(F))) \quad (7)$$

where $\otimes$ represents element-wise dot product operation. The above design makes fused feature maps focus only on the most important parts of an input signal. At the same time, the original input is also concatenated to the fused ones to keep other valuable information in the original input signal.

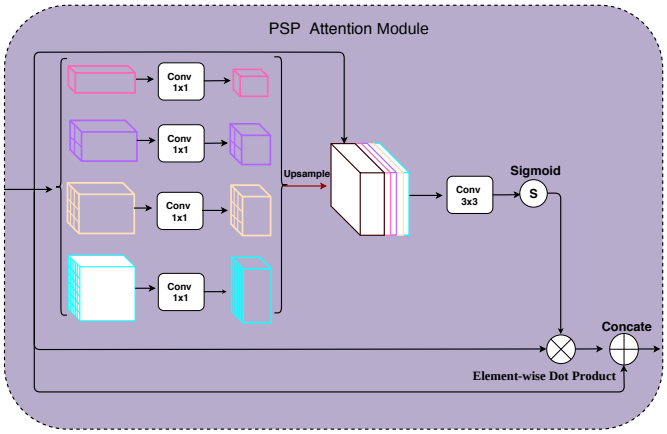


Fig. 6. An illustration of PSP Attention Module. We extract both local and global context information by concatenating the input feature map with several sub-region representations of different scales. Then, an attention processing similar to the ASPP Attention module is applied to fused feature maps to get refined features.

2) PSP Attention Module: The Pyramid Pooling Module is proposed in PSPNet [37] for semantic segmentation. The module fuses multiple features under different pyramid scales which could be controlled by varying bin sizes of pooling. By setting bin sizes to 1×1, 2×2, 3×3 and 6×6, an input feature map could be pooled to four different scales. To be concrete, the first pooling operation is actually global average pooling which captures the global contextual information, whereas the other three pooling operations divide the feature maps into different sub-regions and form multi-scale pooled representation for different localizations. Then, a 1×1 convolution (and batch normalization, ReLu) is applied to the global and local context representations to reduce the number

of output channels to a quarter of the input feature map F. To further fuse with original input feature map, we must ensure that the pooled feature maps should have the same resolution as the input feature map. Hence, we upsample the pooled maps to be of the same size as the input feature map via bilinear interpolation. Finally, upsampled feature maps are concatenated with the original input feature map as the final pyramid pooling features H. After that, an attention processing similar to ASPP Attention module is applied. The detailed structure of PSP Attention module is illustrated in Fig. 6.

C. Post-Processing

Probability maps of pupil center, iris segmentation mask and iris inner/outer boundary could be obtained by forwarding the iris image though the network. Then, we get coarse iris inner/outer contour by using Viterbi algorithm [25] and further fit iris inner/outer circle by using least-squares circle fitting algorithm [9]. Before searching the coarse contours, we remove the noise from predicted probability maps and get the range of iris inner/outer circular boundary by a serial of robust image processing operations.

1) Edge Denoising & Boundary Range Estimation: Different from thin contours produced by traditional edge detection methods such as Canny detector [41], etc., deep learning based edge detector always produces thick, noisy and blurred edges which are not well aligned to actual image boundaries [24]. To eliminate noisy edges, we utilize the prior geometric constraint of pupil center, iris segmentation mask and iris inner/outer boundary and adopt threshold segmentation, connected-component analysis and nearest neighbor search to do the job.

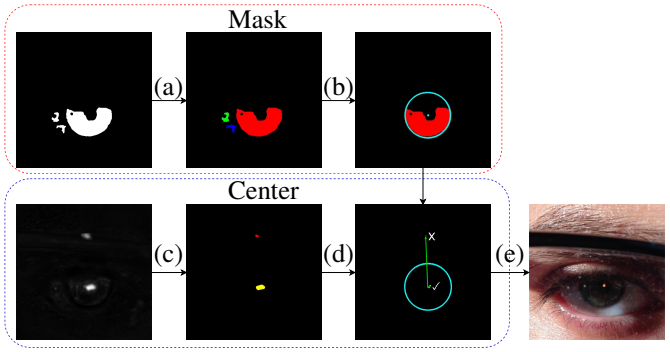


Fig. 7. Overview of pupil center localization. (a) and (c): threshold segmentation and connected-component analysis; (b): get the circumcircle of max-area mask subregion; (d): nearest neighbor search; (e): get actual pupil center.

To be specific, we locate the pupil center in the first place, as shown in Fig. 7. Among the four outputs of the network, iris segmentation mask is the most accurate and max-area iris mask connected subregion has the highest confidence. For pupil center localization, the point with the highest score in the probability map of pupil center could be considered as a good initialization. However, there may be more than one candidate center point with high confidence score for some noisy iris images and the highest score could even be achieved by a noisy pixel. Therefore, we present a more robust alternative for pupil center localization. Considering the real pupil center point is adjacent to iris mask, the pupil center is located by searching the nearest pupil center subregion from the circumcircle center of max-area iris mask subregion. Before searching, the probability map of iris mask is segmented using global threshold (200-255) to get iris mask regions with higher confidence. In addition, the probability map of pupil center is segmented by using lower threshold (150-255) to get more candidate regions. After that, we compute connected components of pupil center and iris mask, and then perform nearest neighbor search. Once the nearest connected component of pupil center is found, we consider its geometric center as the estimated pupil center. Since iris center is approximately close to pupil center in most of the cases except for serious deformation, we simply initialize iris center using the coordinates of the pupil center.

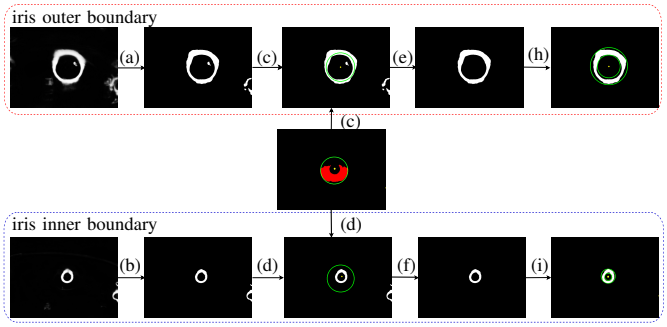


Fig. 8. Edge denoising & boundary range estimation. (a) and (b): threshold segmentation; (c) and (d): generate target region; (e) and (f): edge denoising; (h) and (i): boundary range estimation.

Afterwards, the range of iris inner/outer circular boundary is estimated. Although the majority of noisy edges are removed via applying threshold segmentation, some edges with high-intensity still exist. According to the geometric relationship between iris mask and boundaries, regions where iris boundary is impossible to be located in are further eliminated. More specifically, an enclosing circle close to actual iris outer boundary is generated by taking estimated pupil center as its origin and the maximum distance between the origin and max-area iris mask as radius. Then, for the iris outer boundary, noisy edges completely falling into the inside and outside of the enclosing circle are excluded. For iris inner boundary, those noisy edges completely falling into the outside of the enclosing circle are also excluded. Finally, we compute the minimum and maximum distances between the pupil/iris center and the refined iris inner/outer boundary. The detailed process is illustrated in Fig. 8.

2) Iris Inner/Outer Circle Localization: We modify the original Viterbi algorithm [25] by replacing radial gradient maps with refined probability maps of iris inner and outer boundaries, as well as adopting the estimated range of iris inner/outer circular boundary to output coarse iris inner and outer contours. Then, least-squares circle fitting algorithm [9] is applied on coarse contours to estimate the parameters of iris inner/outer circular boundary.

D. Training Objectives

We optimize all the outputs of IrisParseNet in an end-to-end manner simultaneously. More formally, given an input image

$X = \{x_j, j = 1, ..., |X|\}$ of arbitrary size, we are interested in obtaining probability maps of pupil center, iris segmentation mask, iris inner boundary and iris outer boundary, each of the same size as X.

1) Pupil Center Detection: We denote $P = \{p_j, j = 1, ..., |X|\}$ as the predicted probability map of pupil center, in which $p_j \in [0, 1]$ indicates the probability of pixel x_j being the pupil center, and index j samples every possible spatial location in the input image X.

The ground truth of pupil center, denoted by $\bar{P} = \{\bar{p}_j, j = 1, ..., |X|\}$, is a binary image, where pixel value $\bar{p}_j$ being 1 suggests that the pixel p_j belongs to the pupil region, otherwise is part of the background. Due to shortcomings of deep learning models for dense prediction task, the labeled ground truth of pupil center is not a single pixel but a set of pixels located in the neighborhood of the actual pupil center, see Sec. IV-B.

Due to the extreme imbalance of the number of positive and negative samples in the result of pupil center detection (most of the pixels are background), we use focal loss [23] as the objective function to alleviate this problem. Focal loss introduces two hyper parameters, i.e. α and γ, to be tuned for better performance:

$$\begin{aligned}\mathcal{L}_{\text{pupil}} &= l(P, \bar{P}) \\ &= \sum_j \Big[-\alpha(1-\tilde{p}_j)^\gamma \log(\tilde{p}_j) \Big],\end{aligned} \tag{8}$$

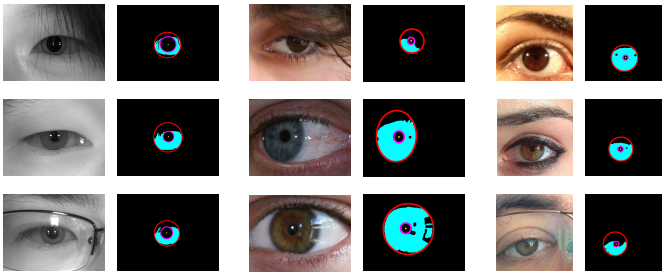
where

$$\tilde{p}_j = \begin{cases} p_j & \text{if } \bar{p}_j = 1 \\ 1 - p_j & \text{otherwise.} \end{cases} \tag{9}$$

2) Iris Segmentation: Since iris segmentation can be seen as a binary semantic segmentation task, we simply adopt a standard binary cross-entropy loss to supervise the training process. Let $S = \{s_j, j = 1, ..., |X|\}$ denote the predicted probability map of iris segmentation mask, where s_j represents the probability of pixel x_j locating in the iris area. The corresponding binary ground truth of iris segmentation mask is denoted as $\bar{S} = \{\bar{s}_j, j = 1, ..., |X|\}$, where $\bar{s}_j$ is set to 1 if pixel s_j is part of the iris region, otherwise $\bar{s}_j$ equals to 0. The cross-entropy loss for iris segmentation can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{seg}} &= l(S, \bar{S}) \\ &= \sum_j \Big[-\bar{s}_j \log(s_j) - (1-\bar{s}_j) \log(1-s_j) \Big],\end{aligned} \tag{10}$$

3) Iris Inner/Outer Boundary Detection: Inspired by CASENet [42], we define iris inner/outer boundary detection as a two-class edge detection problem. To address the problem of positive/negative imbalancing in edge detection, we use the class-balanced cross-entropy loss function which is firstly introduced in HED [24]. Suppose the probability maps of iris inner/outer boundary are denoted as $\{E^1, E^2\}$, in which $E^k = \{e_j^k, j = 1, ..., |X|, k = 1, 2\}$ and e_j^k represents the probability of pixel x_j belonging to iris inner boundary ($k = 1$) or iris outer boundary ($k = 2$). We also manually label the inner and outer boundaries for each iris image, and the ground-truth boundaries are denoted as $\{\bar{E}^1, \bar{E}^2\}$, where $\bar{E}^k = \{\bar{e}_j^k, j = 1, ..., |X|, k = 1, 2\}$ is a binary image indicating the distribution of iris boundaries. The class-balanced cross-entropy loss is formulated as:



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 9. Example images and corresponding ground truths (including iris center(chartreuse), iris inner boundary(magenta), iris outer boundary(red), iris segmentation mask(aqua)) of three iris datesets. Best viewed in color.

$$\begin{aligned}\mathcal{L}_{\text{edge}} &= l(E^1, E^2; \bar{E}^1, \bar{E}^2) \\ &= \sum_k \sum_j \Big[-\beta \bar{e}_j^k \log(e_j^k) \\ &\quad - (1-\beta)(1-\bar{e}_j^k) \log(1-e_j^k) \Big],\end{aligned} \tag{11}$$

where β is the percentage of non-edge pixels in the iris image.

The overall loss function can be expressed as follow:

$$\begin{aligned}\mathcal{L}\big(h(X|W), G\big) &= \lambda_1 \mathcal{L}_{\text{pupil}} + \lambda_2 \mathcal{L}_{\text{seg}} + \lambda_3 \mathcal{L}_{\text{edge}} \\ &= \lambda_1 l(P, \bar{P}) + \lambda_2 l(S, \bar{S}) + \lambda_3 l(E^1, E^2; \bar{E}^1, \bar{E}^2)\end{aligned} \tag{12}$$

where $\{P, S, E^1, E^2\} = h(X|W)$ is the prediction from IrisParseNet, $G = \{\bar{P}, \bar{S}, \bar{E}^1, \bar{E}^2\}$ is the corresponding ground truth. $h(X|W)$ is the model hypothesis taking image X as input, parameterized by W. We can obtain the optimal parameters by minimizing the overall loss function as follow:

$$(W)^* = \text{argmin}\, \mathcal{L}. \tag{13}$$

The hyper-parameters α, γ, λ_1, λ_2 and λ_3 are set to 0.95, 2, 10, 1, 1 in our experiments, respectively .

IV. Datasets and Annotation Methods

In this section, we present detailed descriptions of three challenging and popular datasets: CASIA-Iris-Distance [20], UBIRIS.v2 [21] and MICHE-I [22] and our annotation methods.

A. Datasets

1) ***CASIA-Iris-Distance (CASIA)*** contains 2576 images from 142 subjects with resolution of 2352×1728 pixels. The images are captured by self-developed cameras and sample iris images are shown in the first column of Fig. 9 (a). In this dataset, iris images are captured from a distance of more than 3 metres under near infrared illumination (NIR) and meanwhile the subject is moving. A subset which was manually labeled by the author [1] is selected. This subset includes 400 iris images from the first 40 subjects and to speed up processing, images are resized to 640×480 pixels. We follow the same settings

as in [1] to select first 300 images from the first 30 subjects for training, and the last 100 images from the last 10 subjects are left for testing in the experiments.

2) ***UBIRIS.v2 (UBIRIS)*** consists of 11102 images from 261 subjects which are acquired under visible light illumination (VIS). Images in this dataset are captured on-the-move and at-a-distance with Canon EOS 5D camera and involve realistic noises, such as illumination variance, motion/defocus blur and occlusion of glasses and eyelids. In NICE. I competition, a subset of 1000 UBIRIS.v2 images was used. All images were resized to 400×300 pixels and their segmentation ground truths were manually annotated. According to the protocol of NICE.I competition, 500 images are selected for training and another disjoint testing set of 500 images are used for testing. However, the testing set provided by the organizers of the NICE.I competition has only 445 images. The first column of Fig. 9 (b) shows some examples of images in UBIRIS.v2.
3) ***MICHE-I (MICHE)*** dataset was created to evaluate and develop algorithms for colour iris images captured by mobile devices. Images in MICHE-I were captured by three mobile devices including iPhone5 (abbreviated IP5, 1262 images), Samsung Galaxy S4 (abbreviated GS4, 1297 images), and Samsung Galaxy Tab2 (abbreviated GT2, 632 images) in uncontrolled conditions with visible light illumination (VIS) and without the assistance of any operator [52]. Following by [53], 140 images are selected for training and another 429 images are used for testing. Besides, we also use the manually labeled segmentation ground truths provided by [53]. To speed up processing and preserve the aspect ratio, the width of all iris images is resized to 400 and height is resized to maintain the same proportions as the original image. Finally, the size of resized image is approximately 400×400. The first column of Fig. 9 (c) shows some examples of iris images in MICHE-I.

The images from these adopted datasets were acquired under different types of less-constrained environments, thus various kinds of noises are taken into consideration. In addition, the imaging light source contains near infrared light and visible light. In summary, these datasets are representative in a variety of iris recognition applications, so it is convicing and reasonable to evaluate the performance of the proposed method using these datasets.

B. Annotation Methods

Training the proposed model requires ground truths of iris segmentation, iris inner/outer boundary and pupil center. Since the ground truth of iris segmentation has already been provided by other literatures, we only need to obtain annotations of the other three objects. In the whole labeling process, we use the interactive development environment (HDevelop) provided by the machine vision software, i.e. MVTec Halcon [54], which significantly facilitates our annotation work.

We firstly load iris images in a sequence, and then locate the iris inner and outer boundaries by positioning two ellipses close to explicit iris inner and outer boundaries as the ground truth. After that, the center of iris inner elliptical boundary is regarded as the pupil center.

The initially labeled ground-truth boundaries are too thin, with the width equals to one pixel, but the predicted boundaries from deep models are rather thick. The same problem also occurs in pupil center detection. To tackle this inconsistency, inspired by [55], ground-truth images of training set are dilated using morphologic dilation operator with a circular structuring element of radius 3.

Some examples of manually labeled ground truths can be seen in the second column of Fig. 9 (a), (b), (c), which are sampled from CASIA-Iris-Distance, UBIRIS.v2 and MICHE-I iris datasets, respectively. We sought to accurately locate the iris inner and outer boundaries as well as eliminate all noise presentd to separate the actual iris pixels.

V. Experiments and Analysis

In this section, extensive experiments are conducted on three manually annotated datasets mentioned as in Sec. IV to evaluate the proposed model. The implementation details and data augmentation methods are firstly demonstrated, and then the evaluation protocols are described. Subsequently, the comparisons of our approach with state-of-the-art iris segmentation and localization methods are presented. Finally, we analyze the contribution of each individual module of the proposed model by ablation study.

A. Implementation Details

We implement the proposed architecture based on the publicly available *caffe* [56] framework and the whole network is initialized using the VGG-16 model [57] pretrained on ImageNet. We train the network using mini-batch stochastic gradient descent (SGD) [15] with batch size of 4, momentum of 0.9 and weight decay of 0.0005. Inspired by [33], we use the "poly" learning rate policy where the learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with $power$ set to 0.9, initial learning rate set to $1e^{-3}$ and maximal iteration of 30000. All experiments are conducted on a NVIDIA TITAN Xp GPU with 12GB memory and an Intel(R) Core(TM) i7-6700 CPU.

Data augmentation is a simple yet effective way to enrich training data. During training, we augment training data with random combination of different geometric transformations (scaling, translation, flip, rotation, cropping) and image variations (blur) on-the-fly. Detailed augmentation operations are: (1) shuffle images (and gt maps) when reaching the end of an epoch; (2) randomly resize images (and gt maps) to 7 scales (0.5, 0.75, 1, 1.25, 1.5, 1.75, 2.0); (3) randomly blur images (mean filter, gaussian blur, median blur, bilateral filter, box blur); (4) randomly translate images (and gt maps) in x and y axis by a uniform factor between -30 and 30; (5) randomly left or right flip images (and gt maps); (6) randomly rotate images (and gt maps) by a uniform factor between -60 and 60; and (7) random crop images (and gt maps) to a fixed size (321 × 321) at last. For testing, we drop all augmentation operations and directly apply the model on the original image.

B. Evaluation Protocols

To quantitatively evaluate the proposed method, we introduce several evaluation protocols for iris segmentation, iris inner/outer circle localization and iris recognition. The details are described as follows:

1) Iris segmentation: The NICE. I competition [58] provides two metrics to evaluate the accuracy of iris segmentation. The first measurement is the average segmentation error rate, which could be formulated as follows:

$$E1 = \frac{1}{n \times c \times r} \sum_{c'} \sum_{r'} G(c', r') \otimes M(c', r') \quad (14)$$

where n is the number of test images of r rows and c columns. In addition, G and M are the ground truth mask and the predicted iris mask, respectively, and c', r' are the column and row coordinates of pixels in G and M. The operator $\otimes$ represents the XOR operation to evaluate the inconsistent pixels between G and M.

The second error measure aims to compensate the disproportion between the apriori probabilities of "iris" and "non-iris" pixels in the images. To be specific, it averages the false positives (fp) and false negatives (fn) rates as follows:

$$E2 = \frac{1}{2 \times n} \sum_{i} (fp + fn) \quad (15)$$

where n is the number of testing images.

We also report the following F-Measure (F1) (the harmonic mean of precision and recall) [59] and mean Intersection over Union (mIOU) to provide a comprehensive analysis of the propose method.

The values of E1 and E2 are bounded in $[0, 1]$, where the smaller value indicates the better result. Values of F1 and mIOU also fall in the same interval, but the greater value suggests the higher performance in these cases.

2) Iris inner/outer circle localization: Inspired by [60], we compute the Hausdorff distance between detected iris inner/outer circle (denoted as D) and labeled iris inner/outer boundary (denoted as G) to measure the shape similarity, which could be defined as:

$$H(G, D) = \max\{\sup_{x \in G} \inf_{y \in D} \parallel x - y \parallel, \sup_{y \in D} \inf_{x \in G} \parallel x - y \parallel\} \quad (16)$$

Smaller Hausdorff distances correspond to higher shape similarity between detected circles and ground truths, suggesting higher detection accuracy. We report the mean Hausdorff distance (mHdis) for iris inner circle and outer circle to evaluate the performance of localization. The average value of the two mean Hausdorff distances demonstrates the overall accuracy of iris localization, thus we include it in the evaluation protocol.

Besides, inspired by [61], we also report the detection rate with respect to an error threshold given by the Hausdorff distance between detected iris inner/outer circle and ground truths.

3) Iris recognition: To verify that our iris segmentation and localization framework is able to improve the performance of iris recognition, we conduct iris recognition experiments with all components but iris segmentation and localization methods fixed. We use the equal error rate (EER) and Daugman's decidability index (DI) [26] to quantitatively evaluate the performance of iris recognition. Higher DI values correspond to better discriminative ability of iris recognition systems, meanwhile the iris recognition system with the lowest EER is considered the most accurate.

C. Method Comparison

1) Benchmarks: We select four representative iris segmentation and localization approaches, including both traditional methods and deep learning based methods, as the benchmark. In particular, T. Tan *et. al.* [62] proposed an efficient and robust segmentation method to deal with noisy iris images and it could be roughly divided into four processes: clustering based coarse iris localization, pupillary and limbic boundary localization based on a novel integrodifferential constellation, eyelid localization and eyelash/shadow detection. The method was ranked the first place in NICE.I competition [58]. Since there is no source code available, we only report the result presented in the paper.

RTV-L^1 [12] proposed a novel total-variation based segmentation framework which used l^1 norm regularization to robustly suppress noisy texture pixels to obtain clear iris images. Then, an improved circular Hough transform was used to detect iris and pupil circles on noise-free iris images. Finally, the authors developed a series of robust post-processing operations to locate iris boundaries more accurately. We apply the method on above mentioned three datasets using the source code provided by the authors*.

Haindl and Krupička [27] proposed an unsupervised segmentation method for colored eye images obtained through mobile devices. The method was ranked *first* in the Mobile Iris Challenge Evaluation (MICHE)-I [63] and also outperformed the NICE.I competition winning algorithm, namely T. Tan *et. al.* [62], with average segmentation error rate $E1$ of 1.24% on UBIRIS.v2 dataset. We directly use the executable program† provided by the authors to test on UBIRIS.v2 and MICHE-I datasets except CASIA-Iris-Distance, as images in CASIA-Iris-Distance are not captured under visible lights.

Besides, MFCNs [1] was the first method that applied fully convolutional network for iris segmentation and achieved better results than previous state-of-the-art methods on CASIA-Iris-Distance and UBIRIS.v2 datasets. We reproduce the method and apply it to our labeled three datasets.

Note that except for RTV-L^1, other baseline methods only provide the comparison of iris segmentation mask due to lack of the outputs of iris inner and outer circles.

*The implementation is made available via https://www4.comp.polyu.edu.hk/~csajaykr/tvmiris.htm

†The executable program is made available via http://biplab.unisa.it/MICHE/MICHE-II/PRL_Haindl_Krupicka.zip

2) Evaluation of Iris Segmentation and Localization: Tab. I and Tab. II, Fig. 10, Fig. 11 provide summaries of the performance comparison of the proposed method with baseline approaches on iris segmentation and iris inner/outer circle localization under the proposed evaluation protocols. We also report the storage space of the model and runtime in order to further evaluate the practicability of the proposed method.

As can be seen from Tab. I, IrisParseNet outperforms other approaches on the task of iris segmentation. Especially, IrisParseNet achieves average segmentation error rates of 0.40%, 0.84%, 0.81% on CASIA-Iris-Distance, UBIRIS.v2 and MICHE-I, respectively. Hence, our method ranks first according to the NICE. I competition protocol(**E1**). Besides, IrisParseNet (including ASPP-type and PSP-type) also achieves better results in terms of mean value (greater than 91%) and standard deviation (less than 10%) on F1 metric than other approaches, demonstrating that our approach is highly accurate and robust. The same superiority is also observed on E2 and mIOU (approximately 85%) metrics. The parameters of RTV-L^1 are optimized for each dataset, which makes RTV-L^1 consistently achieve the good segmentation results on three iris datasets. It is worth noting that the performance of Haindl and Krupička [27] is not promising, which is inconsistent with the description in their paper. Although we directly use the execute program provided by the authors when conducting the experiments, we are not able to achieve average segmentation error rates of 1.24% as described in the original paper for UBIRIS.v2, instead a much higher error rate (3.24%) is obtained.

From Tab. II, we could see that for the task of iris inner/outer circle localization, IrisParseNet consistently outperforms RTV-L^1 on all three datasets under mean Hausdorff distance. Besides, It could be seen from Fig. 10 and Fig. 11 that our method performs comparably to or better than RTV-L^1 across the majority of threshold range on all three datasets.

In terms of two types of attention module, IrisParseNet (ASPP) achieves better results on the task of iris segmentation, but IrisParseNet (PSP) shows higher performance on the task of iris inner/outer circle localization.

As for the runtime, the proposed method takes approximate 0.3s, 0.1s, 0.1s for the forward propagation of the network, and 0.4s, 0.4s, 0.4s for post-processing on CASIA-Iris-Distance, UBIRIS.v2 and MICHE-I, respectively. Compared with traditional approaches, IrisParseNet is more time-efficient (In GPU time), as the overall runtime is less than 0.7s. Closer observation would reveal that the post-processing step is the most time-comsuming operation, and the runtime of the framework is directly proportional to resolution of input images.

Although our method achieves good segmentation and localization performance, it consumes relative large storage space (approximately 100MB), that limits its application on mobile platforms. To solve this problem, methods such as parameter pruning and sharing, low-rank factorization, knowledge distillation [64], etc., could be adopted to compress the model and further accelerate the training process.

In summary, the proposed IrisParseNet framework demonstrates noticeable superiority over other methods in accuracy, robustness and usability for the task of iris preprocessing.

TABLE I
COMPARISON OF DIFFERENT APPROACHES ON THE TASK OF IRIS SEGMENTATION USING THE PROPOSED PROTOCOLS.

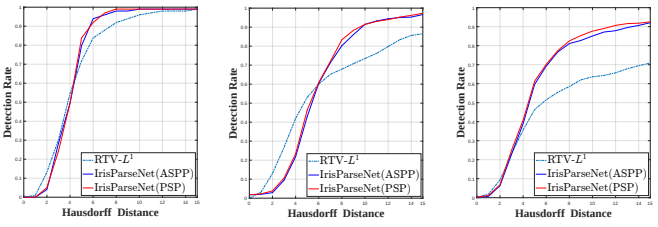
Method	Dataset	E1 (%)	E2 (%)	F1 μ(%)	F1 σ(%)	mIOU (%)	Average Runtime(s)
T. Tan *et. al.* [62]	UBIRIS	1.31	N/A	N/A	N/A	N/A	N/A
RTV-L^1 [12]	CASIA	0.68	0.44	87.55	4.58	78.11	2.46
	UBIRIS	1.21	0.83	85.97	8.72	74.01	1.07
	MICHE	2.27	1.13	77.10	14.71	64.21	1.58
Haindl and Krupička [27]	UBIRIS	3.24	1.62	77.03	20.67	65.08	14.33
	MICHE	5.08	2.54	62.19	25.28	49.79	21.94
MFCNs [1]	CASIA	0.50	0.25	93.14	2.97	87.30	0.47†
	UBIRIS	0.92	0.46	90.78	4.70	81.92	0.32†
	MICHE	0.96	0.48	88.70	8.98	80.63	0.38†
IrisParseNet (ASPP)	CASIA	**0.40**	**0.20**	**94.30**	**3.70**	**89.40**	**0.25†**
	UBIRIS	**0.84**	**0.42**	**91.82**	**4.26**	**85.39**	**0.11†**
	MICHE	0.82	**0.41**	91.33	8.04	84.79	**0.13†**
IrisParseNet (PSP)	CASIA	0.41	0.21	94.20	3.16	89.19	0.30†
	UBIRIS	0.85	**0.42**	91.63	4.06	85.07	**0.11†**
	MICHE	**0.81**	**0.41**	**91.50**	**8.01**	**85.07**	**0.13†**

† GPU time.

TABLE II
COMPARISON OF DIFFERENT APPROACHES ON THE TASK OF IRIS INNER/OUTER CIRCLE LOCALIZATION USING THE PROPOSED PROTOCOLS.

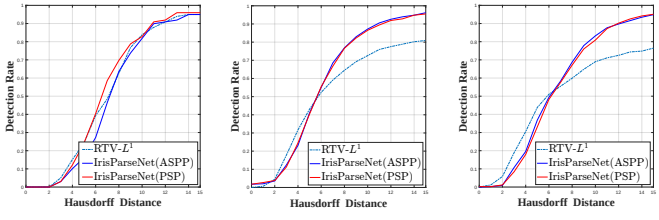
Method	Dataset	mHdis of Iris Inner Circle	mHdis of Iris Outer Circle	Overall mHdis	Average Runtime (s)	Overall Runtime (s)[1]
RTV-L^1 [12]	CASIA	4.24	7.74	6.08	N/A	2.46
	UBIRIS	8.48	11.72	10.10	N/A	1.07
	MICHE	11.96	15.49	13.73	N/A	1.58
IrisParseNet (ASPP)	CASIA	4.13	7.80	5.96	0.42†	**0.67‡**
	UBIRIS	6.06	**6.48**	**6.27**	0.37†	0.49‡
	MICHE	5.67	**7.33**	**6.50**	0.41†	0.54‡
IrisParseNet (PSP)	CASIA	**4.04**	**7.24**	**5.64**	**0.38†**	0.68‡
	UBIRIS	**5.99**	6.61	6.30	**0.32†**	**0.43‡**
	MICHE	**5.41**	7.60	**6.50**	**0.38†**	**0.51‡**

† GPU time.
‡ GPU time + CPU time.
[1] The overall runtime is the sum of the runtime of iris segmentation and iris inner/outer circle localization.



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 10. Performance comparison of iris inner circle localization against RTV-L^1 [12] on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 11. Performance comparison of iris outer circle localization against RTV-L^1 [12] on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.

3) Evaluation of Iris Recognition: To perform iris recognition (more accurately, iris verification) experiments, we use the full set iris images of CASIA-Iris-Distance, UBIRIS.v2 and MICHE-I datasets. To speed up processing, for CASIA-Iris-Distance and MICHE-I datasets, we use classical Viola-Jones eye detector [65] provided by OpenCV to extract the eye region in images, and all eye regions are resized to 400×400. Iris images in UBIRIS.v2 are already scaled to 400×300. We use single eye in iris recognition experiments and detailed settings of the experiments are provided in Tab. III.

The proposed IrisParseNet framework is firstly applied for iris segmentation and localization, then Daugman's rubber sheet normalization method [8] is used to produce normalized iris image and iris mask for feature extraction and matching. We adopt the 1-D log Gabor filter to extract iris codes and compute Hamming Distance of iris codes to verify whether two iris are from the same class [‡]. The same normalization, feature extraction and matching processes are also adopted in experiments with RTV-L^1 [12] and Haindl and Krupička [27].

Evaluation results of iris recognition are shown in Tab. IV. From Tab. IV, we could see that experiments using the proposed IrisParseNet framework achieve lower EER and higher DI than those using other methods, especially for CASIA-Iris-Distance, UBIRIS.v2, MICHE-I:iPhone5 and MICHE-I:SamsungGalaxyS4. Experiment results illustrate that our IrisParseNet method greatly improves the performance of iris recognition.

TABLE III
DETAILED SETTINGS OF IRIS RECOGNITION EXPERIMENT.

Dataset	CASIA	UBIRIS	MICHE		
			IP5	GS4	GT2
No. of subjects	119	259	75	75	75
No. of classes	238	518	150	150	150
No. of images	2280	11100	995	764	438
Resolution	400×400	400×300	400×400		

TABLE IV
COMPARISON OF DIFFERENT APPROACHES ON THE TASK OF IRIS RECOGNITION USING THE PROPOSED PROTOCOLS.

Dataset	Method	EER	DI
CASIA	RTV-L^1 [12]	0.2708	1.1116
	IrisParseNet (ASPP)	**0.0392**	**3.4474**
	IrisParseNet (PSP)	0.0412	3.4039
UBIRIS	RTV-L^1 [12]	0.3303	0.9096
	Haindl and Krupička [27]	0.4249	0.5069
	IrisParseNet (ASPP)	0.3107	0.9642
	IrisParseNet (PSP)	**0.3096**	**0.9871**
MICHE:IP5	RTV-L^1 [12]	0.2279	1.3343
	Haindl and Krupička [27]	0.3154	1.0004
	IrisParseNet (ASPP)	0.2045	**1.4994**
	IrisParseNet (PSP)	**0.1984**	1.4896
MICHE:GS4	RTV-L^1 [12]	0.2386	1.2569
	Haindl and Krupička [27]	0.3329	0.8993
	IrisParseNet (ASPP)	0.2038	1.3908
	IrisParseNet (PSP)	**0.2029**	**1.4011**
MICHE:GT2	RTV-L^1 [12]	**0.2370**	1.3509
	Haindl and Krupička [27]	0.2948	1.1415
	IrisParseNet (ASPP)	0.2553	**1.3751**
	IrisParseNet (PSP)	0.2487	1.3310

[‡]We use the open source iris recognition software package (USIT Version 2) for feature extraction and matching, which is made available via http://www.wavelab.at/sources/Rathgeb16a/

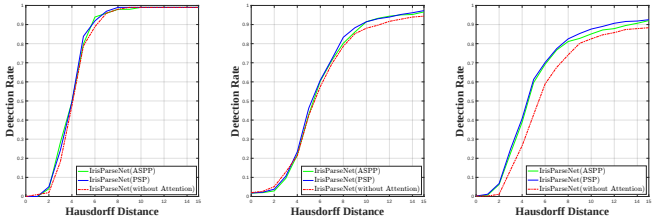
D. Ablation Study

We further explore the contribution of each individual module of the proposed model by conducting ablation study.

1) Effectiveness of Attention Mechanism: To verify the effectiveness of the attention module, we replace it with two sequential convolutional layers with 256 filters and 512 filters (along with batch normalization layer and ReLU layer). Experiment results are shown in Tab. V, Fig. 12, and Fig. 13. From Tab. V, we could see that compared with the original IrisParseNet framework, its variants without attention module suffer from significant performance drop on the task of iris segmentation for all datasets. As for the task of iris localization, removing attention module would result in a significant performance decrease on UBIRIS.v2 and MICHE-I, as shown in Fig. 12 and Fig. 13.

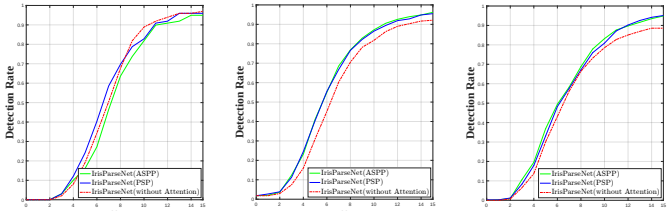
TABLE V
COMPARISON OF IRISPARSENET WITH/WITHOUT ATTENTION MODULE.

Dataset	Method	E1 (%)	E2 (%)	mean F1 (%)	mIOU (%)	Overall mHdis of Iris Localization
CASIA	ASPP-type	**0.40**	**0.20**	**94.30**	**89.40**	5.96
	PSP-type	0.41	0.21	94.20	89.19	**5.64**
	without Attention	0.43	0.21	94.10	89.00	5.76
UBIRIS	ASPP-type	**0.84**	**0.42**	**91.82**	**85.39**	**6.27**
	PSP-type	0.85	**0.42**	91.63	85.07	6.30
	without Attention	0.94	0.47	90.87	83.49	7.34
MICHE	ASPP-type	0.82	**0.41**	91.33	84.79	**6.50**
	PSP-type	**0.81**	**0.41**	**91.50**	**85.07**	**6.50**
	without Attention	0.87	0.44	90.46	83.12	8.07



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 12. Performance comparison of iris inner circle localization with/without attention module on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

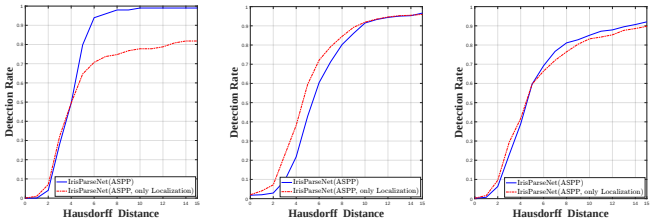
Fig. 13. Performance comparison of iris outer circle localization with/without attention module on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.

2) Effectiveness of Joint Segmentation and Localization: To evaluate the contribution of joint segmentation and localization, we compare three IrisParseNet framework variants:

original IrisParseNet (ASPP), IrisParseNet only with localization part or segmentation part, as shown in Tab. VI, Fig. 14 and Fig. 15, respectively. Experiment results show that joint learning of iris segmentation and localization helps to improve the performance on both iris segmentation and iris localization tasks.

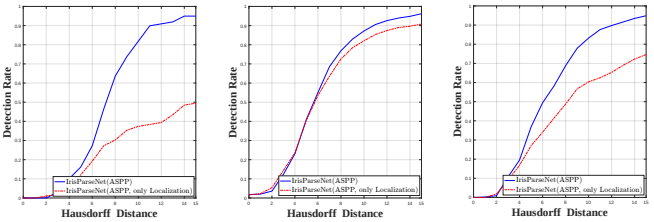
TABLE VI
COMPARISON OF IRISPARSENET WITH/WITHOUT JOINT TRAINING.

Dataset	Method	E1 (%)	E2 (%)	mean F1 (%)	mIOU (%)	Overall mHdis of Iris Localization
CASIA	ASPP-type	**0.40**	**0.20**	**94.30**	**89.40**	**5.96**
	only Localization	N/A	N/A	N/A	N/A	11.91
	only Segmentation	0.41	0.20	94.08	89.18	N/A
UBIRIS	ASPP-type	**0.84**	**0.42**	**91.82**	**85.39**	**6.27**
	only Localization	N/A	N/A	N/A	N/A	7.39
	only Segmentation	0.85	0.42	91.70	83.37	N/A
MICHE	ASPP-type	**0.82**	**0.41**	**91.33**	**84.79**	**6.50**
	only Localization	N/A	N/A	N/A	N/A	10.70
	only Segmentation	0.82	0.41	91.32	84.72	N/A



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 14. Performance comparison of iris inner circle localization with/without joint learning on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.



(a) CASIA-Iris-Distance (b) UBIRIS.v2 (c) MICHE-I

Fig. 15. Performance comparison of iris outer circle localization with/without joint learning on the labeled three iris datasets. Success rate is thresholded on the Hausdorff distance error.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel deep multi-task learning framework for joint iris segmentation and localization. In this framework, a Fully Convolutional Encoder-Decoder Attention Network and an effective post-processing operation which exploit the priori geometric constraints of pupil, iris and sclera, are proposed to improve the performance of iris segmentation and localization. Meanwhile, we have collected manual labels of three challenging iris datasets and established comprehensive evaluation protocols, which are publicly available. The proposed method is compared with state-of-the-art methods on the three annotated iris datasets, and shows a leading performance. As for future work, we would explore improving the efficiency of the post-processing step or integrate it into the iris segmentation and localization system to form an end-to-end model.

REFERENCES

[1] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan, "Accurate iris segmentation in non-cooperative environments using fully convolutional networks," in *IAPR International Conference on Biometrics*. IEEE, 2016, pp. 1–8.

[2] E. Jalilian, A. Uhl, R. Kwitt, E. Jalilian, A. Uhl, R. Kwitt, E. Jalilian, A. Uhl, and R. Kwitt, "Domain adaptation for cnn based iris segmentation," in *International Conference of the Biometrics Special Interest Group*, 2017, pp. 1–6.

[3] S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end deep neural network for iris segmentation in unconstrained scenarios," *Neural Networks*, vol. 106, pp. 79–95, 2018.

[4] E. Severo, R. Laroca, C. Bezerra, L. A. Z. Junior, D. Weingaertner, G. Moreira, and D. Menotti, "A benchmark for iris location and a deep learning detector evaluation," *CoRR*, vol. abs/1803.01250, 2018. [Online]. Available: http://arxiv.org/abs/1803.01250

[5] M. Arsalan, R. A. Naqvi, S. K. Dong, H. Nguyen, M. Owais, R. Kang, and Park, "Irisdensenet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light camera sensors," *Sensors*, vol. 18, no. 5, 2018.

[6] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of biometrics*. Springer Science & Business Media, 2007.

[7] R. P. Wildes, "Iris recognition: an emerging biometric technology," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1348–1363, 1997.

[8] J. Daugman, "How iris recognition works," in *The essential guide to image processing*. Elsevier, 2009, pp. 715–739.

[9] N. Chernov and C. Lesort, "Least squares fitting of circles," *Journal of Mathematical Imaging and Vision*, vol. 23, no. 3, pp. 239–252, 2005.

[10] H. Hofbauer, F. Alonso-Fernandez, J. Bigun, and A. Uhl, "Experimental analysis regarding the influence of iris segmentation on the recognition rate," *Iet Biometrics*, vol. 5, no. 3, pp. 200–211, 2016.

[11] H. Proença and L. A. Alexandre, "Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage," *Image & Vision Computing*, vol. 28, no. 1, pp. 202–206, 2010.

[12] Z. Zhao and K. Ajay, "An accurate iris segmentation framework under relaxed imaging constraints using total variation model," in *IEEE International Conference on Computer Vision*, 2015, pp. 3828–3836.

[13] S. Banerjee and D. Mery, "Iris segmentation using geodesic active contours and grabcut," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2015, pp. 48–60.

[14] A. Radman, N. Zainal, and S. A. Suandi, "Automated segmentation of iris images acquired in an unconstrained environment using hog-svm and growcut," *Digital Signal Processing*, vol. 64, pp. 60–70, 2017.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] 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*, 2015, pp. 91–99.

[17] W. Chinsatit and T. Saitoh, "Cnn-based pupil center detection for wearable gaze estimation system," *Applied Computational Intelligence and Soft Computing*, vol. 2017, 2017.

[18] F. Vera-Olmos and N. Malpica, "Deconvolutional neural network for pupil detection in real-world environments," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2017, pp. 223–231.

[19] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," *CoRR*, vol. abs/1805.04771, 2018. [Online]. Available: http://arxiv.org/abs/1805.04771

[20] B. I. Test, "Casia.v4 database," http://www.idealtest.org/dbDetailForUser.do?id=4, Last Accessed (Oct 2018).

[21] H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. 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.

[22] M. De Marsico, C. Galdi, M. Nappi, and D. Riccio, "Firme: Face and iris recognition for mobile engagement," *Image and Vision Computing*, vol. 32, no. 12, pp. 1161–1172, 2014.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[24] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[25] G. Sutra, S. Garcia-Salicetti, and B. Dorizzi, "The viterbi algorithm at different resolutions for enhanced iris segmentation," in *IAPR International Conference on Biometrics*. IEEE, 2012, pp. 310–316.

[26] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993.

[27] M. Haindl and M. Krupička, "Unsupervised detection of non-iris occlusions," *Pattern Recognition Letters*, vol. 57, pp. 60–65, 2015.

[28] A. Gangwar, A. Joshi, A. Singh, F. Alonso-Fernandez, and J. Bigun, "Irisseg: A fast and robust iris segmentation framework for non-ideal iris images," in *IAPR International Conference on Biometrics*. IEEE, 2016, pp. 1–8.

[29] Y. Hu, K. Sirlantzis, and G. Howells, "Improving colour iris segmentation using a model selection technique," *Pattern Recognition Letters*, vol. 57, pp. 24–32, 2015.

[30] S. Shah and A. Ross, "Iris segmentation using geodesic active contours," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 824–836, 2009.

[31] T. Rongnian and W. Shaojie, "Improving iris segmentation performance via borders recognition," in *International Conference on Intelligent Computation Technology and Automation*, vol. 2. IEEE, 2011, pp. 580–583.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.7062

[34] ——, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[35] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[38] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *CoRR*, vol. abs/1707.03718, 2017. [Online]. Available: http://arxiv.org/abs/1707.03718

[39] V. Iglovikov and A. Shvets, "Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation," *CoRR*, vol. abs/1801.05746, 2018. [Online]. Available: http://arxiv.org/abs/1801.05746

[40] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," *CoRR*, vol. abs/1803.01207, 2018. [Online]. Available: http://arxiv.org/abs/1803.01207

[41] J. F. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[42] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "Casenet: Deep category-aware semantic edge detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 21–26.

[43] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5300–5309.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[47] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[48] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[49] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," *CoRR*, vol. abs/1804.09337, 2018. [Online]. Available: http://arxiv.org/abs/1804.09337

[50] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018. [Online]. Available: http://arxiv.org/abs/1807.06521

[51] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *CoRR*, vol. abs/1807.06514, 2018. [Online]. Available: http://arxiv.org/abs/1807.06514

[52] M. De Marsico, M. Nappi, and H. Proença, "Results from miche ii–mobile iris challenge evaluation ii," *Pattern Recognition Letters*, vol. 91, pp. 3–10, 2017.

[53] Y. Hu, K. Sirlantzis, and G. Howells, "Improving colour iris segmentation using a model selection technique," *Pattern Recognition Letters*, vol. 57, no. 1, pp. 24–32, 2015.

[54] M. S. GmbH, *halcon/hdevelop reference manual*, 11st ed., 2012.

[55] Y. Liu and M. S. Lew, "Learning relaxed deep supervision for better edge detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 231–240.

[56] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[58] H. Proença and L. A. Alexandre, "The nice. i: noisy iris challenge evaluation-part i," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2007, pp. 1–4.

[59] H. Hofbauer, F. Alonso-Fernandez, P. Wild, J. Bigun, and A. Uhl, "A ground truth for iris segmentation," in *International Conference on Pattern Recognition*. IEEE, 2014, pp. 527–532.

[60] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X. Qi, P. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[61] W. Fuhl, D. Geisler, T. Santini, W. Rosenstiel, and E. Kasneci, "Evaluation of state-of-the-art pupil detection algorithms on remote eye images," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1716–1725.

[62] T. Tan, Z. He, and Z. Sun, "Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition," *Image and vision computing*, vol. 28, no. 2, pp. 223–230, 2010.

[63] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015.

[64] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks." *CoRR*, vol. abs/1710.09282, 2017. [Online]. Available: http://arxiv.org/abs/1710.09282

[65] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE conference on computer vision and pattern recognition*, vol. 1. IEEE, 2001, pp. I–I.