Eye-MMS: Miniature Multi-Scale Segmentation Network of Key Eye-Regions in Embedded Applications

Fadi Boutros [12], Naser Damer[12], Florian Kirchbuchner[12], Arjan Kuijper[12]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Germany
[2]Technische Universität Darmstadt, Germany
Email: {FirstName.LastName}@igd.fraunhofer.de

Abstract

Segmentation of the iris or sclera is an essential processing block in ocular biometric systems. However, human-computer interaction, as in VR/AR applications, requires multiple region segmentation to enable smoother interaction and eye-tracking. Such application does not only demand highly accurate and generalizable segmentation, it requires such segmentation model to be appropriate for the limited computational power of embedded systems. This puts strict limits on the size of the deployed deep learning models. This work presents a miniature multi-scale segmentation network consisting of inter-connected convolutional modules. We present a baseline multi-scale segmentation network and modify it to reduce its parameters by more than 80 times, while reducing its accuracy by less than 3%, resulting in our Eye-MMS model containing only 80k parameters. This work is developed on the OpenEDS database and is conducted in preparation for the OpenEDS Semantic Segmentation Challenge.

1. Introduction

Segmentation of different ocular regions is essential to enable accurate processing of biometric modalities, such as the iris or the sclera [16]. This segmentation also provides valuable information for eye tracking and enhances the computer-human interaction (HCI) experience, especially in virtual reality (VR) and Augmented Reality (AR) applications [19]. Driven by biometric recognition needs, the ocular segmentation focused either on the iris or sclera regions. However, multiple region segmentation is essential for HCI applications.

Achieving accurate multi-region ocular segmentation in the AR/VR (or embedded systems generally) context faces two main challenges. The first is the high variation in the region appearance because of the independent movement of different parts (eyeball, pupil dilation, eyelid, eyebrows) and the demographic-related appearance variations, which requires a highly generalized solution. The second challenge is to limit the computational needs of the segmentation solution by the minimalistic hardware specifications available in such applications. Given a solution based on neural networks, one of the main computational limitations is the segmentation model size, represented by the number of learned parameters.

This work is conducted in preparation for the OpenEDS Semantic Segmentation Challenge and therefore build its solution on the challenge database. We propose a solution based on multi-scale inter-connected convolutional modules that considers the image information at multiple scales and thus reduce the parameters needed to learn low-scale image properties. We minimize our initial model from 6574k to 80k parameters by taking advantage of the fact that segmentation takes an image from a highly detailed space to a space with small number of discrete labels. This allowed focusing on larger image changes (rather than small details) by reducing the feature maps sizes, resulting in smaller convolutional layers, and thus significantly smaller model. Despite this large reduction in the model size (more than 80 times reduction), our Eye-MMS model achieved over 90% mean intersection over union on the four label regions on evaluation data that is identity-disjoint from the training data. This is less than three percentage points lower than our initially proposed model with 6574k parameters.

2. Related work

Previous works addressing semantic image segmentation in the ocular region focused mainly on iris or sclera segmentation. Sclera can be a biometric characteristic, but its segmentation acts also as a way to detect the outer boundaries of the iris. This had been motivated mainly by the high interest in iris recognition, as one of the most accurate biometrics characteristics [10]. The localization (segmentation) accuracy of the iris significantly effects the iris recognition performance [16]. Earlier works suggested segmenting the iris region by defining its boundaries, e.g. by Hough transforms [23]. More recent works followed the

trend in generic segmentation and detected the iris region by utilizing Fully Convolutional Network (FCN) [12, 3] or U-Net [15].

Sclera segmentation has been addressed by a series of competitions in the last five years, since 2015 [7]. The latest competition [6] focused on variations in the capture angle and the use of mobile devices. The winning team utilized U-Net structure [21] modified by a channel attention module as described by Yu et al. [24].

Eye tracking can benefit greatly from multiple region semantic segmentation of the ocular area. However, only recent activities have targeted this problem and provided appropriate research databases. One of these is the iBUG Eye Segmentation Dataset [17] where relatively low resolution ocular regions are segmented into two labels, iris and pupil as one class, and sclera as the second class. The work also proposed a segmentation solution based on convolution neural network followed by refinement by conditional random field. Rot et al. [22] also addressed the multi-region (iris, sclera, pupil, periocular, eyelashes, and canthus) segmentation issue by building a convolutional encoder-decoder solution, however, with a a database of a limited size. Very recently, Garbin et al. [8] presented a research oriented database that addresses different issues related to eye tracking. One aspect of this database is the multi-region semantic segmentation of the eye region, which is the bases of this work and the OpenEDS Semantic Segmentation Challenge. This database addresses shortcomings in previous databases by providing a larger number of subjects and images, along with high resolution images to address VR applications.

Generic image segmentation solutions have achieved increasingly impressive performances since the rise of deep learning. Main advances in this regard are segmentation based on FCN [14], U-Net [21], Feature Pyramid Network [11], Mask R-CNN [9], DeepLabv3+ [4], Path Aggregation Network [13], and most recently the Context Encoding Network [25]. However, few works have addressed segmentation solutions constrained by very limited computational resources (e.g. embedded systems). The latest of such works is the Fast-SCNN [20] that result in a model with 1.1 million parameters, which is still large for some embedded devices.

3. Methodology

The goal of our solution is to create an accurate segmentation for a given eye region image despite appearance variations. The created model should be of a small size (around or below 1MB) to enable application in embedded environments, such as AR/VR applications. In this section, we present two segmentation models, the first is built to demonstrate the idea of multi-scale segmentation and the second aims at maintaining (to a large degree) the performance of the first model, while being significantly smaller (smaller number of learned parameters).

We start by proposing a **Multi-scale segmentation solutions (Eye-MS)**. This model aims at extracting more general information at lower image scales, and thus minimizing the model size required to extract such information. It also process the image at higher scales to analyze detailed image information. The presented solution is influenced by the cascaded refinement network introduced by Chen and Koltun [5] as an image synthesis tool. Our proposed architecture is a convolutional neural network that consists of inter-connected refinement modules. Each module consists of only two convolutional layers (last module contains 3 convolutional layers), each followed by layer normalization [1] and a LReLU non-linearity [18]. The first module considers the lowest resolution space (40x25 in our model). This resolution is increased in the successor modules until the last module (640x400 in our case), matching the target image resolution. The input of each module is the output of the previous module up-sampled to the proper input size of the current module, concatenated with the source image down-sampled to the proper input size of the current module. Our Eye-MS model uses 4x4 convolutions and a feature map (FM) of the size 256 for the first three modules and 128 for the last two modules. A summery of the network details is presented in Table 1. Our Eye-MS model size 6574k parameters, making it relatively smaller than conventional solutions such as the real-time ICNet (6680k) [26] and SegNet (29460k) [2]. However, such a model size might be too large for embedded applications.

As we aim at producing an accurate segmentation model, however with a much smaller size, we point out that we are moving from a higher detailed space (captured eye image) to a space with lower variation (segmentation of four classes). Thus, we can neglect minor details in the image and focus on major changes across the image space. This can help us reduce the less important (for segmentation) learned parameters. We induce this notion by reducing the feature map size of the convolutional layers of the Eye-MS model. We designed our **Miniature Multi-scale Segmentation Network (Eye-MMS)** by setting the feature map size to 32 for the first two modules and 16 for the last three modules. This reduction in the feature map size lead to a reduction in the size of the subsequent convolutional layers, therefore, a significant reduction in the number of learned parameters. This model contains 80081 learned parameters, and thus, will be noted as Eye-MMS80. The model architecture is provided in Table 2.

Both networks (Eye-MS and Eye-MMS80) are trained using an $L2$ loss on the pixel-level between the produced segmentation and the ground-truth label. The networks were trained with a patch size of one and a learning rate of 10e-4. The output layer produced 2-D array of float numbers to enable a smooth learn conversion. The predicted

	Eye-MS (6574k parameters)		
Module	Input size	layer	Output size
Module 0	40x25x1	g_25_conv1 (filter:[4x4], FM:256), LN, LReLU	40x25x256
		g_25_conv2 (filter:[4x4], FM:256), LN, LReLU	
Module 1	80x50x257	g_50_conv1 (filter:[4x4], FM:256), LN, LReLU	80x50x256
		g_50_conv2 (filter:[4x4], FM:256), LN, LReLU	
Module 2	160x100x257	g_100_conv1 (filter:[4x4], FM:256), LN, LReLU	160x100x256
		g_100_conv2 (filter:[4x4], FM:256), LN, LReLU	
Module 3	320x200x257	g_200_conv1 (filter:[4x4], FM:128), LN, LReLU	320x200x128
		g_200_conv2 (filter:[4x4], FM:128), LN, LReLU	
Module 4	640x400x129	g_400_conv1 (filter:[4x4], FM:128), LN, LReLU	640x400x1
		g_400_conv2 (filter:[4x4], FM:128), LN, LReLU	
Output		g_400_conv100 ([1x1], FM:1)	

Table 1: The detailed structure of the multi-scale segmentation network Eye-MS (6574k). The input of each of the 5 modules is the source image and the output of the previous module (not for Module 0), down-sampled and up-sampled subsequently to the input size of the current module. FM (Feature map), LN (Layer Normalization), and CON (Concatenate)

	Eye-MMS80 (80k parameters)		
Module	Input size	layer	Output size
Module 0	40x25x1	g_25_conv1 (filter:[4x4], FM:32), LN, LReLU	40x25x32
		g_25_conv2 (filter:[4x4], FM:32), LN, LReLU	
Module 1	80x50x33	g_50_conv1 (filter:[4x4], FM:32), LN, LReLU	80x50x32
		g_50_conv2 (filter:[4x4], FM:32), LN, LReLU	
Module 2	160x100x33	g_100_conv1 (filter:[4x4], FM:16), LN, LReLU	160x100x16
		g_100_conv2 (filter:[4x4], FM:16), LN, LReLU	
Module 3	320x200x17	g_200_conv1 (filter:[4x4], FM:16), LN, LReLU	320x200x16
		g_200_conv2 (filter:[4x4], FM:16), LN, LReLU	
Module 4	640x400x17	g_400_conv1 (filter:[4x4], FM:16), LN, LReLU	640x400x1
		g_400_conv2 (filter:[4x4], FM:16), LN, LReLU	
Output		g_400_conv100 ([1x1], FM:1)	

Table 2: The detailed structure of the miniature multi-scale segmentation network Eye-MMS80. The input of each of the 5 modules is the source image and the output of the previous module (not for Module 0), down-sampled and up-sampled subsequently to the input size of the current module. FM (Feature map), LN (Layer Normalization), and CON (Concatenate).

segmentations are rounded to the nearest integer to represent the discrete labels.

4. Experimental setup

This work used the OpenEDS [8] data captured using a virtual-reality HMD with two eye-facing cameras. The segmentation data included 152 subjects and 12759 images annotated of 400x640 pixels resolution. The data is split into training, validation, and test identity-disjoint splits as described in [8]. Evaluation on the test split is only possible through an online portal and with limited frequency.

The segmentation performance is evaluated here as the intersection over union (IoU) of each of the four segmented regions i={pupil, iris, sclera, background} between the predicted segmentation (P) and the ground-truth label (L) and is given by

$$IoU_i = \frac{L_i \cap P_i}{L_i \cup P_i}. \quad (1)$$

To get an overall performance measure, we also report the IoU_{mean}, the unweighted mean of the four IoU_i values.

We report the results on the validation split (2403 images) of the database to provide more detailed experiments. The identity-disjoint validation split was not used to train the reported models. The possible experiments are limited on the test split by the competition rules to one evaluation per day, limiting the possibility of multiple experiments.

The results are reported for the model Eye-MS and the miniature model Eye-MMS80 after 8 epochs (reached loss: for 0.0175 Eye-MMS80 and 0.0104 for Eye-MS) of training and after 15 epochs of training (reached loss: 0.0121 for Eye-MMS80 and 0.0085 for Eye-MS). All models were trained on the training split, containing 8916 pairs of eye images and corresponding ground-truth labels.

It should be mentioned that we evaluate the output of the network without any significant post-processing (only rounding to nearest integer) to enable a clear evaluation of the network performance. Post-processing (non-learned) steps such as morphological operations or contour finding (e.g. marching squares) and masking might enhance the segmentation results, however, with a computational load. Post-processing learned refinement is also possible, e.g. conditional random field, however not applied here to maintain low computational requirements.

5. Results

Figure 1 shows samples of the validation images along with the predicted segmentation by our Eye-MS and Eye-MMS80 (both trained for 15 epochs), and the segmentation ground-truth. One can notice flakes of sclera label in

	Eye-MMS80 (80k)		Eye-MS (6574k)	
Region	8 epochs	15 epochs	8 epochs	15 epochs
IoU(BG)	0.9831	0.9857	0.9892	0.9896
IoU(Sclera)	0.7825	0.8084	0.8476	0.8519
IoU(Iris)	0.9105	0.9223	0.9391	0.9408
IoU(Pupil)	0.8958	0.9105	0.9275	0.9276
IoU-mean	0.8931	0.9068	0.9258	0.9275

Table 3: The performance, given as IoU, on different ocular regions and a mean IoU to represent general performance of our proposed models, Eye-MS and Eye-MMS80, at two different stages of the training process. It is noticed that despite the significant reduction in the model size the performance is only slightly effected. BG refers to the background region.

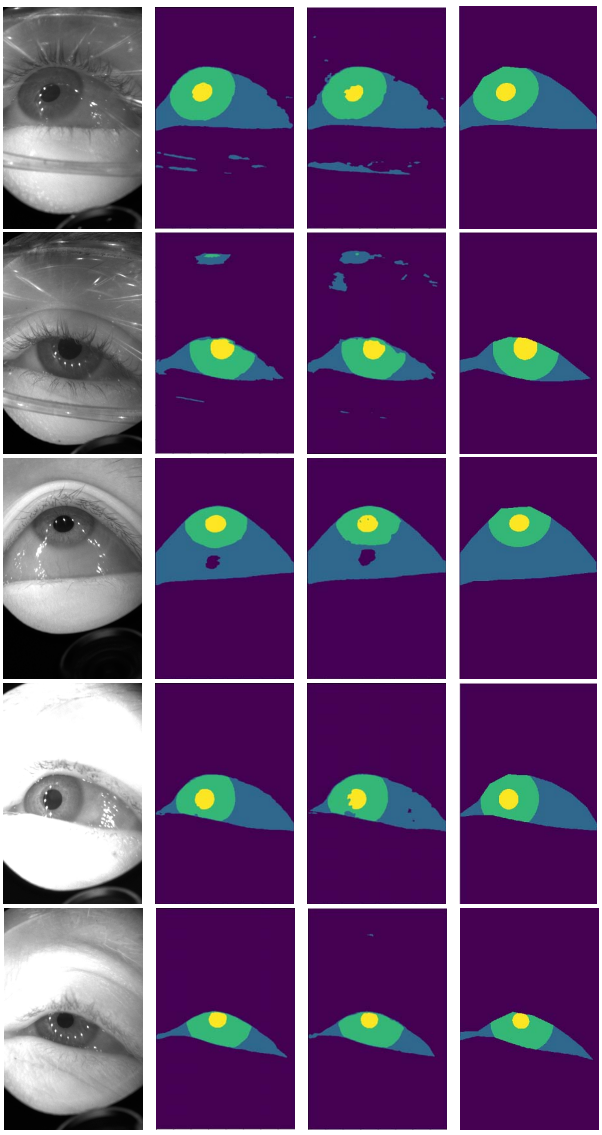
the background in the cases where the eye is covered by highly reflective eyeglasses. Such flakes might, if needed, be removed by post-processing operations (contouring and masking), which is not performed in the reported results. Extreme eye-gaze, illumination, and small opening of the eye seems to have no large effect on the performance of our proposed models. These situations lead to small flakes in the background and sclera regions. These flakes seems to be slightly more frequent or larger in the miniature Eye-MMS80 model, but without effecting the overall segmentation structural boundaries.

Table 3 lists the performances, given as IoU, for each individual region (label) and as a mean over the four regions. This performance comparison is made for both the Eye-MS and the Eye-MMS80 and in two points of the training process. It is noticeable from the table, and in all experimental settings, that the IoU(background) achieves the highest value, this might be based on the relatively large area of the background and thus the lower probable ratio of non-intersection to the union area, between the ground-truth and prediction. The IoU(iris) and IoU(pupil) achieve closer values, with the IoU(iris) slightly over performing the later. The IoU(sclera) scores significantly lower than the other eye regions. This might be due to the confusion between the sclera and background, especially with images containing highly reflective glasses. The Eye-MMS80 with 15 epoch training achieved 89.5 % mean IoU on the test data split through the online evaluation of the OpenEDS Semantic Segmentation Challenge.

Table 3 also shows that increasing the training to fifteen epochs improves the performance of both models. This points out that further training might further increase the performance without over-fitting, especially if accompanied by data augmentation (which is not implemented in the reported results). The Eye-MMS80 generally performs only slightly worse than the Eye-MS model while having less than 1/80 of its parameters.

6. Conclusion

This work proposes a multiple eye regions semantic segmentation solution containing only 80k parameters.



(a) Input image (b) Eye-MS (c) Eye-MMS80 (d) Ground-truth

Figure 1: Samples of the input images, segmentation produced by our Eye-MS and Eye-MMS80 models, and the ground-truth segmentation. These images are selected to have variations of eye glasses with reflections, extreme gaze direction, extreme illumination, and eyes with small opening.

This aims at enabling deployment in computationally restricted embedded systems, such as VR/AR applications. We initially proposed a multi-scale segmentation network based on multi-scale inter-connected convolutional modules. Then we took advantage of the nature of the segmentation task to lower the number of its learned parameters from 6574k to 80k, while only lowering the mean intersection over union (over the four regions of the eye) from 92.8% to 90.7%. We point out the validity of our approach and the possibility of enhancing the overall performance by post-processing the segmentations, augmenting the training data, and further training.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the National Research Center for Applied Cybersecurity CRISP.

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 2

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 2

[3] Shabab Bazrafkan, Shejin Thavalengal, and Peter Corcoran. An end to end deep neural network for iris segmentation in unconstrained scenarios. *Neural Networks*, 106:79–95, 2018. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018. 2

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1520–1529. IEEE Computer Society, 2017. 2

[6] Abhijit Das, Umapada Pal, Michael Blumenstein, and zhun sun. Sclera segmentation benchmarking competition in cross-resolution environment. In *International Conference on Biometrics, ICB 2019, 4-7 June, 2019, Crete, Greece*. IEEE, 2019. 2

[7] Abhijit Das, Umapada Pal, Miguel A. Ferrer, and Michael Blumenstein. SSBC 2015: Sclera segmentation benchmarking competition. In *IEEE 7th International Conference on Biometrics Theory, Applications and Systems, BTAS 2015, Arlington, VA, USA, September 8-11, 2015*, pages 1–6. IEEE, 2015. 2

[8] Stephan J. Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S. Talathi. Openeds: Open eye dataset. *CoRR*, abs/1905.03702, 2019. 2, 3

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 2

[10] Anil K. Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Techn.*, 14(1):4–20, 2004. 1

[11] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017. 2

[12] N. Liu, H. Li, M. Zhang, Jing Liu, Z. Sun, and T. Tan. Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, June 2016. 2

[13] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8759–8768. IEEE Computer Society, 2018. 2

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society, 2015. 2

[15] J. Lozej, B. Meden, V. Struc, and P. Peer. End-to-end iris segmentation using u-net. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6, July 2018. 2

[16] Jus Lozej, Dejan Stepec, Vitomir Struc, and Peter Peer. Influence of segmentation on deep iris recognition performance. In *7th International Workshop on Biometrics and Forensics, IWBF 2019, Cancun, Mexico, May 2-3, 2019*, pages 1–6. IEEE, 2019. 1

[17] Bingnan Luo, Jie Shen, Yujiang Wang, and Maja Pantic. The ibug eye segmentation dataset. In Edoardo Pirovano and Eva Graversen, editors, *2018 Imperial College Computing Student Workshop, ICCSW 2018, September 20-21, 2018, London, United Kingdom*, volume 66 of *OASICS*, pages 7:1–7:9. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2018. 2

[18] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. 2

[19] Thammathip Piumsomboon, Gun A. Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In Maud Marchal, Robert J. Teather, and Bruce H. Thomas, editors, *2017 IEEE Symposium on 3D User Interfaces, 3DUI 2017, Los Angeles, CA, USA, March 18-19, 2017*, pages 36–39. IEEE Computer Society, 2017. 1

[20] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *CoRR*, abs/1902.04502, 2019. 2

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 2

[22] P. Rot, . Emeri, V. Struc, and P. Peer. Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE*

International Work Conference on Bioinspired Intelligence (IWOBI), pages 1–8, July 2018. 2

[23] R. P. Wildes. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363, Sep. 1997. 1

[24] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1857–1866. IEEE Computer Society, 2018. 2

[25] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7151–7160. IEEE Computer Society, 2018. 2

[26] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pages 418–434. Springer, 2018. 2