

RITnet: Real-time Semantic Segmentation of the Eye for Gaze Tracking

Aayush K. Chaudhary* Rakshit Kothari* Manoj Acharya* Shusil Dangi
 Nitinraj Nair Reynold Bailey Christopher Kanan Gabriel Diaz Jeff B. Pelz
 Rochester Institute of Technology, USA

(akc5959, rsk3900, ma7583, sxd7257, nrn2741, rjbvcs, kanan, gabriel.diaz, jeff.pelz)@rit.edu

Abstract

Accurate eye segmentation can improve eye-gaze estimation and support interactive computing based on visual attention; however, existing eye segmentation methods suffer from issues such as person-dependent accuracy, lack of robustness, and an inability to be run in real-time. Here, we present the RITnet model, which is a deep neural network that combines U-Net and DenseNet. RITnet is under 1 MB and achieves 95.3% accuracy on the 2019 OpenEDS Semantic Segmentation challenge. Using a GeForce GTX 1080 Ti, RITnet tracks at > 300Hz, enabling real-time gaze tracking applications. Pre-trained models and source code are available¹.

1. Introduction

Robust, accurate, and efficient gaze estimation is required to support a number of critical applications such as foveated rendering, human-machine and human-environment interactions, as well as inter-saccadic manipulations, such as redirected walking [16]. Recent non-intrusive, video-based eye-tracking methods involve localization of eye features such as the pupil [7] and/or iris [17]. These features are then regressed onto some meaningful representation of an individual’s gaze. Convolutional neural networks (CNNs) have demonstrated high accuracy [9, 18] and robustness in unconstrained lighting conditions [1] and an ability to generalize under low resolution constraints [12, 13].

In an effort to engage the machine learning and eye-tracking communities in the field of eye-tracking for head-mounted displays (HMD), Facebook Reality Labs issued the Open Eye Dataset (OpenEDS) Semantic Segmentation challenge which addresses part of the gaze estimation pipeline: identifying different regions of interest (e.g., pupil, iris, sclera, skin) in close-up images of the eye. Such

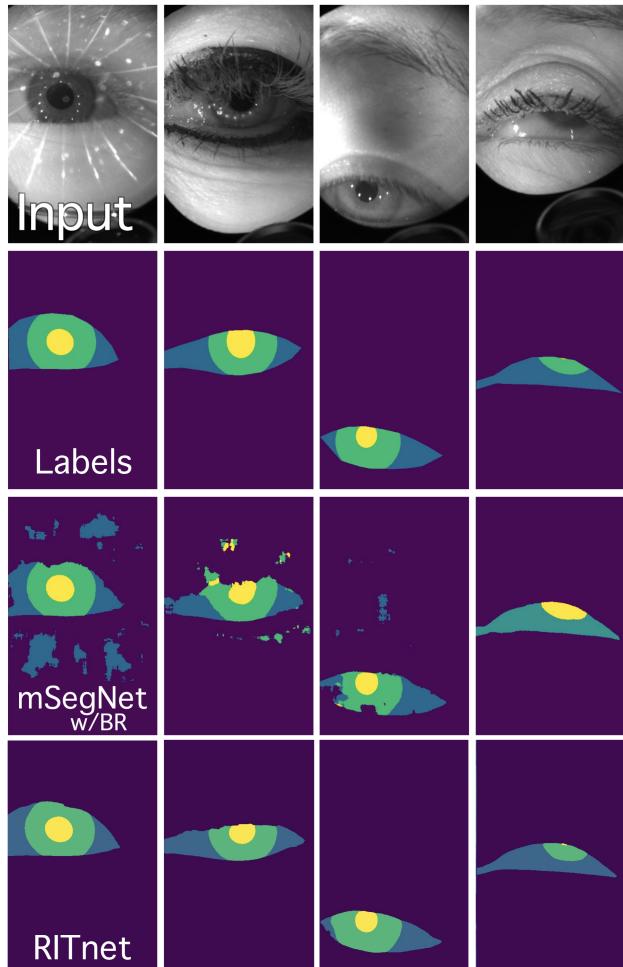


Figure 1. Comparison of model performance on difficult samples in the OpenEDS test-set. Top-row left to right shows eyes obstructed due to prescription glasses, heavy mascara, dim light and partial eyelid closure. Rows from top to bottom show input test images, ground truth labels, predictions from mSegNet w/BR [4] and predictions from RITnet, respectively.

*Equal Contribution.

¹<https://bitbucket.org/eye-ush/ritnet/>

semantic segmentation of these regions enables the extrac-

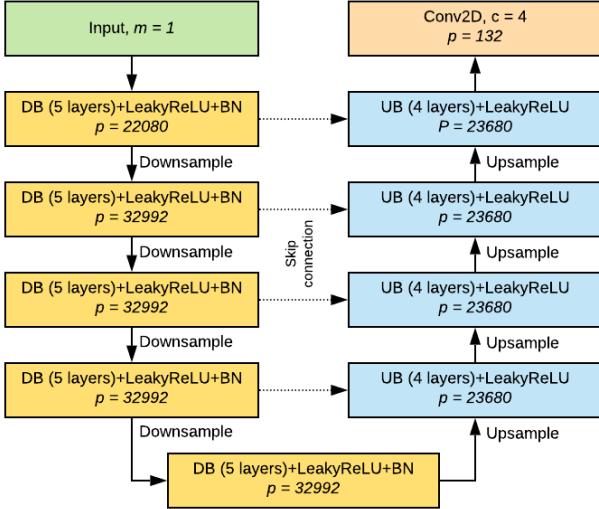


Figure 2. Architecture details of RITnet. DB refers to *Down-Block*, UB refers to *Up-Block*, and BN stands for batch normalization. Similarly, m refers to the number of input channels ($m = 1$ for gray scale image), c refers to number of output labels and p refers to number of model parameters. Dashed lines denote the skip connections from the corresponding *Down-Blocks*. All of the Blocks output tensors of channel size $m=32$.

tion of region-specific features (e.g., iridial feature tracking [2]) and mathematical models which summarize the region structures (e.g., iris ellipse [17, 1, 13], or pupil ellipse [7]) used to derive a measure of gaze orientation.

The major contributions of this paper are as follows:

1. We present RITnet, a semantic segmentation architecture that obtains state-of-the-art results on the 2019 OpenEDS Semantic Segmentation Challenge with model size of **only 0.98 MB**. Our model performs segmentation at 301 Hz for 640x400 images on an NVIDIA 1080Ti GPU.
2. We propose domain-specific augmentation schemes which help in generalization under a variety of challenging conditions.
3. We present boundary aware loss functions with a loss scheduling strategy to train Deep Semantic Segmentation models. This helps in producing coherent regions with crisp region boundaries.

2. Previous Works

Recently developed solutions for end-to-end segmentation involve using Deep CNNs to produce a labeled output irrespective of the size of the input image. Such architectures consist of convolution layers with a series of down-sampling followed by progressive upsampling layers. Downsampling operations strip away finer information that

is crucial for accurate pixel-level semantic masks. This limitation was mitigated by Ronneberger et al. by introducing skip-connections between the encoder and decoder [14]. Jergou et al. proposed TiramisuNet [6], a progression of dense blocks [5] with skip connections between the up- and down-sampling pathways. TiramisuNet demonstrated reuse of previously computed feature maps to minimize the required number of parameters. Dangi et al. proposed the DenseUNet-K architecture [3] for image-to-image translation based on simplified dense connected feature maps with skip connections. The RITnet model presented in this paper is based on the DenseUNet-K architecture².

3. Proposed Model: RITnet

Recently, segmentation models based on Fully Convolutional Networks (FCN) have performed well across many datasets [6, 14]. That success, however, often comes at the cost of computational complexity, restricting their feasibility for real-time applications where rapid computation and robustness to illumination conditions is paramount [4]. In contrast, RITnet has 248,900 trainable parameters which require less than 1MB storage with 32-bit precision (see Figure 2) and has been benchmarked at >300 Hz.

RITnet has five *Down-Blocks* and four *Up-Blocks* which downsample and upsample the input. The last *Down-Block* is also referred to as the *bottleneck* layer which reduces the overall information into a small tensor $1/16^{\text{th}}$ of the input resolution. Each *Down-Block* consists of five convolution layers with LeakyReLU activation. All convolution layers share connections from previous layers inspired by DenseNet [5]. We maintain a constant channel size as in DenseUNet-K [3] with $K=32$ channels to reduce the number of parameters. All *Down-Blocks* are followed by an average pooling layer of size 2×2 . The *Up-Block* layer upsamples its input by a factor of two using the nearest neighbor approach. Each *Up-Block* consists of four convolution layers with LeakyReLU activation. All *Up-Blocks* receive extra information from their corresponding *Down-Block* via skip connections, an effective strategy which provides the model with representations of varying spatial granularity.

3.1. Loss functions

Each pixel is classified into one of four semantic categories: *background*, *iris*, *sclera*, or *pupil*. Standard cross-entropy loss (CEL) is the default choice for applications with a balanced class distribution. However, there exists an imbalanced distribution of classes with the fewest pixels representing pupil regions. While CEL aims to maximize the output probability at a pixel location, it remains agnostic to the structure inherent to eye images. To mitigate these issues, we implemented the following loss functions:

²<https://github.com/ShusilDangi/DenseUNet-K>

Generalized Dice Loss (GDL): Dice score coefficient measures the overlap between the ground truth pixel and their predicted values. In cases of class imbalance [11], weighting the dice score by the squared inverse of class frequency [15] showed increased performance when combined with CEL.

Boundary Aware Loss (BAL): Semantic boundaries separate regions based on class labels. Weighting the loss for each pixel by its distance to the two nearest segments introduces edge awareness [14]. We generate boundary pixels using a Canny edge detector which are further dilated by two pixels to minimize confusion at the boundary. We use these edges to mask the CEL.

Surface Loss (SL): SL is based on a distance metric in the space of image contours which preserves small, infrequent structures of high semantic value [8]. BAL attempts to maximize the correct pixel probabilities near boundaries while GDL provides stable gradients for imbalanced conditions. Contrary to both, SL scales the loss at each pixel based on its distance from the ground truth boundary for each class. It is effective in recovering smaller regions which are ignored by region based losses [8].

The total loss \mathcal{L} is given by a weighted combination of these losses as $\mathcal{L} = \mathcal{L}_{\text{CEL}}(\lambda_1 + \lambda_2 \mathcal{L}_{\text{BAL}}) + \lambda_3 \mathcal{L}_{\text{GDL}} + \lambda_4 \mathcal{L}_{\text{SL}}$.

4. Experimental Details

4.1. Dataset and Evaluation

We train and evaluate our model on the OpenEDS Semantic Segmentation dataset [4] consisting of 12,759 images split into *train* (8,916), *validation* (2,403) and *test* (1,440) subsets. Each image had been hand annotated with four semantic labels; *background*, *sclera*, *pupil*, & *iris*.

Per OpenEDS challenge guidelines, our *overall score* metric uses the average of the mean Intersection over Union (mIoU) metric for all classes and model size (S) calculated as a function of number of trainable parameters in megabytes (MB). The *overall score* is given as $\frac{\text{mIoU} + \min(\frac{1}{S}, 1)}{2}$.

4.2. Training

We trained our model using Adam [10] with a learning rate of 0.001 and a batch size of 8 images for 175 epochs on a TITAN 1080 Ti GPU. We reduced the learning rate by a factor of 10 when the validation loss plateaued for more than 5 epochs. The selected model with the best validation score was found at the 151st epoch. In our experiments, we used $\lambda_1 = 1, \lambda_2 = 20, \lambda_3 = (1 - \alpha)$ and $\lambda_4 = \alpha$, where $\alpha = \text{epoch}/125$ for $\text{epoch} < 125$ otherwise 0. This loss scheduling scheme gives prominence to GDL during initial iterations until a steady state is achieved, following which SL begins penalizing stray patches.

4.3. Data Pre-processing

To accommodate variation in individual reflectance properties (e.g., iris pigmentation, eye makeup, skin tone or eyelids/eyelashes) [4] and HMD specific illumination (the position of infrared LEDs with respect to the eye), we performed two pre-processing steps. These steps were based on the difference in the train, validation and test distributions of mean image brightness (Figure 11 in Garbin et al [4]). Pre-processing reduced these differences and also increased separability of certain eye features. First, a fixed gamma correction with an exponent of 0.8 was applied to all input images. Second, we applied local Contrast Limited Adaptive Histogram Equalization (CLAHE) with a grid size of 8x8 and clip limit value of 1.5 [19]. Figure 3 shows an image before and after pre-processing.

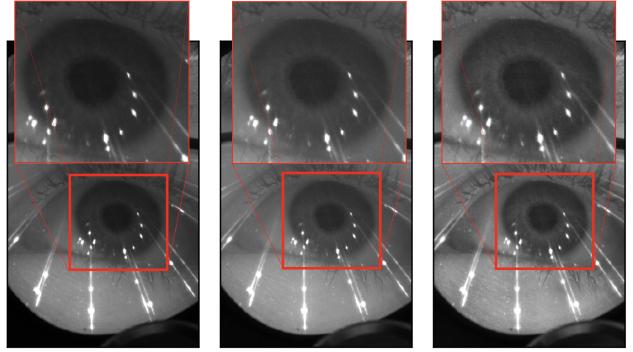


Figure 3. Left to right: Original image, image after gamma correction, image after CLAHE is applied. Note that in the rightmost image, it is comparatively easier to distinguish iris and pupil.

To increase the robustness of the model to variations in image properties, training data was augmented with the following modifications:

- Reflection about the vertical axis.
- Gaussian blur with a fixed kernel size of 7x7 and standard deviation $2 \leq \sigma \leq 7$.
- Image translation of 0-20 pixels in both axes.
- Image corruption using 2-9 thin lines drawn around a random center ($120 < x < 280, 192 < y < 448$)
- Image corruption with a structured *starburst* pattern (Figure 4) to reduce segmentation errors caused by reflections from the IR illuminators on eyeglasses. Note that the *starburst* image is translated by 0-40 pixels in both directions.

Each image received at least one of the above-mentioned augmentations with a probability of 0.2 on each iteration. The probability that an image would be flipped horizontally was 0.5.

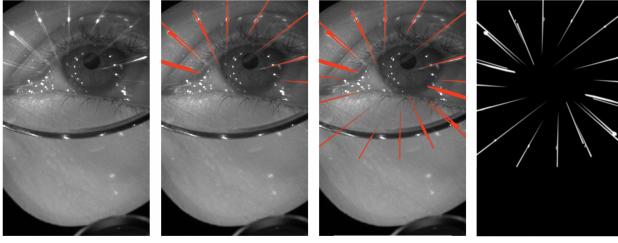


Figure 4. Generation of a *starburst* pattern from the training image 000000240768. Left to Right: Original image, selected reflections, concatenating with its 180° rotation, final pattern mask (best viewed in color).

5. Results

We compare our results against SegNet [4], another fully convolutional encoder-decoder architecture. mSegNet refers to the modified SegNet with four layers of encoder and decoder. mSegNet w/BR refers to mSegNet with Boundary Refinement as residual structure and mSegNet w/SC is a lightweight mSegNet with depthwise Separable Convolutions [4]. As shown in Table 1, our model achieves a $\sim 6\%$ improvement in mIoU score while the complexity is reduced by $\sim 38\%$ compared to the baseline model mSegNet w/SC. However, our model’s segmentation quality was impacted at higher values of motion blur and image defocus (Figure 5), Figure 1 demonstrates that our model generalizes to some challenging cases where other models fail to produce coherent results.

Model	Mean F1	mIoU	Model Size (S)	No. of parameters (million)	Overall Score
mSegNet*	97.9	90.7	13.3	3.5	0.491
mSegNet* w/BR	98.3	91.4	13.3	3.5	0.495
mSegNet* w/SC(B)	97.4	89.5	1.6	0.4	0.762
Ours	99.3	95.3	0.98	0.25	0.976

Table 1. Performance comparison on the test split of the OpenEDS dataset. The metrics and comparison models (*) are used as reported in [4].

6. Discussion

Our model achieves state-of-the-art performance with a small model footprint. The final architecture was arrived at after exploring a number of architectural variations. Reducing the channel size from 32 to 24 and increasing the number of convolution layers in the *Down-Block* did not affect the results. Surprisingly, increasing the channel size to 40 and removing one convolutional layer in the *Down-Block* degraded performance, resulting in spurious patches

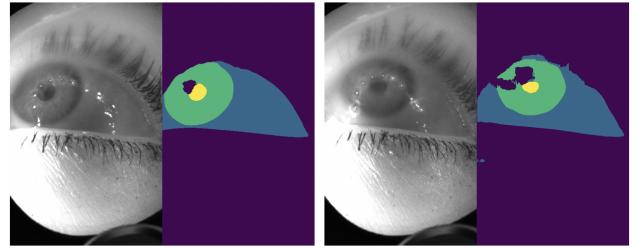


Figure 5. Our model struggles to do an accurate segmentation when eye masks are heavily blurred or defocused.

in output regions. Performance was influenced by the choice of loss functions and the adjustment of their relative weights. By setting the boundary-aware loss at a relatively higher weight, we observed sharp boundary edges and consequently improved our test mIoU from 94.8% to 95.3%.

We speculate that some aspects of our model were successful because they accounted for labeling artifacts in the openEDS dataset. For example, although pupil-to-iris boundaries were defined using ellipse fits to multiple points selected on the boundaries [4], sclera-to-eyelid boundaries were created using a linear fit between adjacent points marked on the eyelids. It is perhaps for this reason that the use of nearest-neighbor interpolation outperformed bilinear interpolation in the process of upsampling. Although the smoother curves that result from bilinear interpolation resulted in more accurate detection of the iris and pupil, it was less accurate in segmentation of the sclera.

Finally, data prepossessing had a significant impact on model performance. Introduction of CLAHE and gamma correction resulted in an overall improvement of 0.2% in the validation mIoU score. Augmentation helped in noisy cases such as reflections from eyeglasses, varying contrast, eye makeup, and other image distortions.

7. Conclusion

We designed a computationally efficient model for the segmentation of eye images. We also presented methods for implementing multiple loss functions that can tackle class imbalance and ensures crisp semantic boundaries. We showed several methods for incorporating pre-processing and augmentation techniques that can help mitigate against image distortions. RITNet attained 95.3% on the OpenEDS test set with a model size < 1 MB and benchmarks an impressive 301Hz on a NVIDIA 1080Ti.

Acknowledgements

We thank Anjali Jogeshwar, Kishan KC, Zhizhuo Yang, and Sanketh Moudgalya for providing valuable input and feedback. We would also like to thank the Research Computing group at RIT for providing access to GPU clusters.

References

- [1] W. F. B, W. Rosenstiel, and E. Kasneci. *500,000 Images Closer to Eyelid and Pupil Segmentation*, volume 11678 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2019.
- [2] A. Chaudhary and J. Pelz. Motion tracking of iris features to detect small eye movements. *Journal of Eye Movement Research*, 12(6), 2019.
- [3] S. Dangi and C. Linte. DenseUNet-K: A simplified Densely Connected Fully Convolutional Network for Image-to-Image Translation. https://github.com/ShusilDangi/DenseUNet-K/blob/master/DenseUNet_K.pdf, 9 2019.
- [4] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. OpenEDS: Open Eye Dataset. 4 2019.
- [5] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017-Janua, pages 2261–2269. IEEE, 7 2017.
- [6] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:1175–1183, 2017.
- [7] M. Kassner, W. Patera, and A. Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1151–1160, 2014.
- [8] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed. Boundary loss for highly unbalanced segmentation, 2018.
- [9] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. NVGaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. *Conference on Human Factors in Computing Systems - Proceedings*, 12:1–12, 2019.
- [10] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *Journal of neuroscience methods*, 148(2):167–76, 12 2014.
- [11] F. Milletari, N. Navab, and S. A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 2016.
- [12] S. Park, S. De Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot Adaptive Gaze Estimation. 5 2019.
- [13] S. Park, A. Spurr, and O. Hilliges. Deep Pictorial Gaze Estimation. volume 11217 LNCS, pages 741–757. 2018.
- [14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention. Springer, Cham*, 2015., 9351:234–241, 2015.
- [15] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. 10553 LNCS:240–248, 7 2017.
- [16] Q. Sun, A. Kaufman, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. McGuire, and D. Luebke. Towards virtual reality infinite walking. *ACM Transactions on Graphics*, 37(4):1–13, 7 2018.
- [17] E. Wood and A. Bulling. EyeTab: Model-based gaze estimation on unmodified tablet computers. *Eye Tracking Research and Applications Symposium (ETRA)*, pages 207–210, 2014.
- [18] Z. Wu, S. Rajendran, T. van As, J. Zimmermann, V. Badri-narayanan, and A. Rabinovich. EyeNet: A Multi-Task Network for Off-Axis Eye Gaze Estimation and User Understanding. 8 2019.
- [19] K. Zuiderveld. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*. 1994.