MinENet: A Dilated CNN for Semantic Segmentation of Eye Features

Jonathan Perry, Amanda S. Fernandez
Department of Computer Science
University of Texas at San Antonio
One UTSA Circle, San Antonio, Texas 78249
https://cs.utsa.edu/~fernandez/vail

Abstract

Fast and accurate eye tracking is a critical task for a range of research in virtual and augmented reality, attention tracking, mobile applications, and medical analysis. While deep neural network models excel at image analysis tasks, existing approaches to segmentation often consider only one class, emphasize classification over segmentation, or come with prohibitively high resource costs. In this work, we propose MinENet, a minimized efficient neural network architecture designed for fast multi-class semantic segmentation. We demonstrate performance of MinENet on the OpenEDS Semantic Segmentation Challenge dataset, against a baseline model as well as standard state-of-the-art neural network architectures - a convolutional neural network (CNN) and a dilated CNN. Our encoder-decoder architecture improves accuracy of multi-class segmentation of eye features in this large-scale high-resolution dataset, while also providing a design that is demonstrably lightweight and efficient.

1. Introduction

Semantic segmentation research has a wide applicability to medical image analysis [7, 17, 8], searching [14], mobile applications, home-automation, autonomous vehicles, and virtual and augmented reality [2, 10]. With this breadth of applications, segmentation architectures are proposed with an equally broad range of goals, rarely converging to related utility.

A key limitation to deep neural networks applied to eye-region trait-segmentation is the limited number of high resolution annotated datasets of large-scale. While early computer vision approaches and smaller machine learning models can work with a smaller number of images, deep learning models require larger datasets in order to extract features. The Open Eye Dataset (OpenEDS) is a recent large-scale dataset of eye-images with corresponding masks annotating the iris, pupil, and sclera regions [2]. This dataset provides a controlled lighting environment, synchronization of left and right eyes, optometric data, and anonymized metadata from its participants.

In this work, we propose MinENet (Minimized ENet [10]), a reduced-scale efficient neural network architecture to increase accuracy of semantic segmentation of the human eye, in response to the Facebook: OpenEDS Semantic Segmentation Challenge. The quantitative assessments chosen in this event include model accuracy and model complexity, and MinENet demonstrates significant improvements over the proven baseline model in both measures. In addition to the baseline, we compare our proposed model with standard implementations of models typically leveraged in image segmentation: a convolutional neural network (CNN) and the same CNN with dilation in its layers. The simple/vanilla CNN is shown to approximate the total score of the provided baseline, and the addition of dilation to its layers significantly improves its accuracy for segmentation. In contrast, our proposed MinENet improves accuracy while requiring fewer parameters.

A visual comparison of eye image segmentation is shown in Figure 1. The figure shows first the original and ground-truth images from the OpenEDS image dataset, followed by segmentation predictions from the vanilla CNN and MinENet.

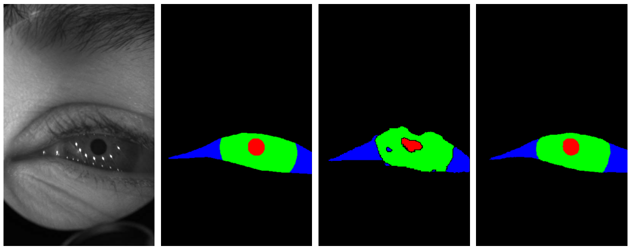

Figure 1. Comparison of segmentation results. From left: Original image, ground truth, dilated CNN prediction, our proposed model (MinENet) prediction.

2. Semantic Segmentation

Semantic segmentation has been a well-studied area of research interest for decades. As large datasets and computing resources continue to increase, machine and deep learning models continue to improve accuracy in new applications. Convolutional neural networks (CNNs) in particular have been a focus of segmentation research efforts, due to their applicability to spatially-oriented tasks. One mechanism for fine-tuning these particular segmentation networks is to increase the receptive field of the convolution by adding dilation. Model complexity is a challenge in image segmentation, and while dilation does not increase the number of parameters, it is not a reduction technique. In this section, related approaches to semantic segmentation are reviewed, with the primary goals of reducing the model size and maintaining accuracy. Finally, our proposed MinENet architecture is introduced to this end, prior to experimental evaluation.

2.1. Related Work

To the detriment of eye-tracking studies, a primary limitation of existing works in human eye image segmentation is the applicability to either only image classification tasks, or to single-class segmentation rather than multiple classes [2]. More general multi-class segmentation approaches such as entangled decision forests [8] and convolutional neural networks [13, 15, 7] demonstrate high accuracy on image segmentation tasks. Region proposal networks such as FastR-CNN [3] and FasterR-CNN [11] improve upon these image classification tasks by identifying a designated subset of regions within an image, rather than a larger and unlikely set. Recent approaches have demonstrated further accuracy gains by incorporating adversarial learning [17], encoder-decoder frameworks [9, 1], and dilated convolutions [4, 6, 16, 18]. In some cases, the accuracy improvements of these approaches are weighed against the added model complexity, reducing their effectiveness for constrained applications. The UNet [12] architecture combined fully convolutional networks with patterned downsampling, upsampling, and bottlenecks to improve efficiency. More recently, ENet [10] leveraged an encoder-decoder architecture with similarly patterned bottlenecks to improve model complexity for real-time segmentation in mobile applications.

2.2. MinENet Architecture

The main contribution of this work is a proposed Minimized ENet (MinENet) neural network designed for accurate and efficient semantic segmentation. The architecture of our proposed MinENet model is outlined in Table 1. Our encoder-decoder model consists of five bottleneck modules with dilated and asymmetric convolutions. This encoder-decoder design is illustrated in Figure 2.

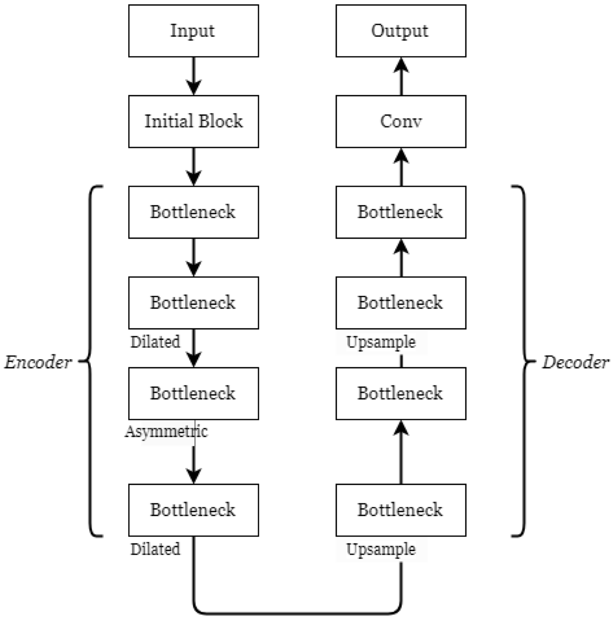

Figure 2. Visualization of the encoder-decoder architecture (MinENet). Footnotes included on the bottlenecks to differentiate between components.

The bottleneck components represented in Figure 2 are comprised of both encoder bottlenecks and decoder bottlenecks. The encoder and decoder bottlenecks have standard designs, shown in Figures 3 and 4, respectively. The footnotes on both encoder bottlenecks and decoder bottlenecks capture variations from the standard design, which are shown as alternative paths in each figure.

The model has a total of 222,440 parameters, where 217,160 are trainable parameters. The sample sizes shown demonstrate color input images of size [512 × 512], in our application to semantic segmentation of human eye regions.

MinENet leverages dilated convolutions to provide context to the segmented regions, in this application to regions in the human eye. As noted, dilation in the convolutional layers increases the receptive field, and the model learns relative locations within an image of the pupil versus sclera versus iris, distinguishing them from the background.

Given the general goal of reduction of model complexity, a primary divergence from the [10] model is the removal of redundancies within layers in the center of the model. While complex scene analysis may require a more localized evaluation for image features, applications in eye tracking more often operate on closely cropped images. For this reason, the higher dilations in the encoder framework are not recommended for images where features tend to be global rather than local.

Nonlinear operations are applied prior to convolutions in this model, and the common rectified linear units (ReLU) are replaced with Parametric Rectified Linear Units

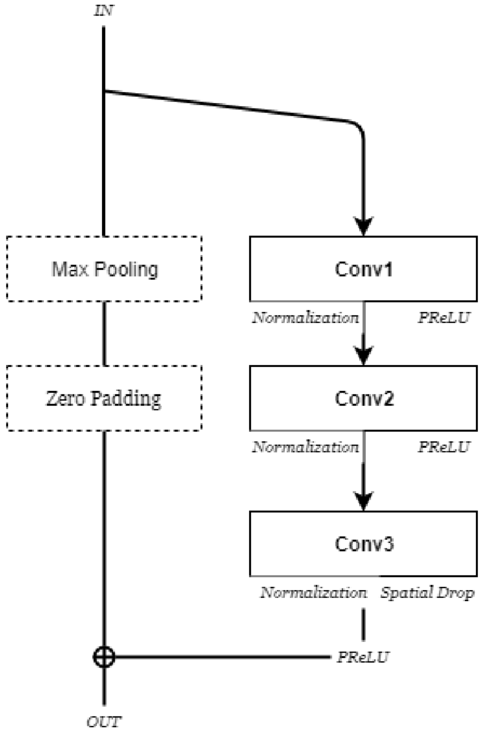

Figure 3. Realization of a single encoder bottleneck. Footnotes included on the convolution layers to highlight steps taken with each convolution.

(PReLU) [5]. In feature mapping, PReLU tracks the negative slope of nonlinearities as an additional mechanism for prediction.

Finally, in order to address common dataset imbalances, the loss function employed in MinENet is a Jaccard distance loss. Particular to the OpenEDS dataset, this loss assists the model as it encounters eye images in open, almost closed, closed, and misaligned states, where *open* would be the most capable of prediction. Imbalances in human eye datasets may also occur due to age, skin reflectivity, glasses, and makeup application.

3. Experiments & Results

In evaluation of our proposed model, we leverage metrics and data provided by the OpenEDS Challenge, and compare results against a range of comparable architectures. The baseline model is a derivation of the SegNet[1] architecture, and employing separable convolution [2]. A simple CNN model is constructed, denoted as a vanilla CNN as it consists only of a deep network of fully convolutional layers. Finally, a vanilla CNN with the same definition as the previous model is modified only by adding dilation to its convolutional layers.

The OpenEDS [2] dataset consists of 356,649 $[400 \times 640]$ images at 200Hz frame rate from 152 participants ages 19-65, with and without glasses, and under controlled lighting conditions. The images were manipulated from its original state for the purpose of matching MinENet's input. As both a pre-processing and post-processing step, the original $[400 \times 640]$ image shape is reshaped to $[512 \times 512]$, which used zero padding as a technique to preserve ratios. The images were collected via a head mounted display (HMD) with synchronized cameras under high frame rates. Annotations for the images are included in the dataset in the form of masks identifying the iris, pupil, and sclera regions of the human eyes.

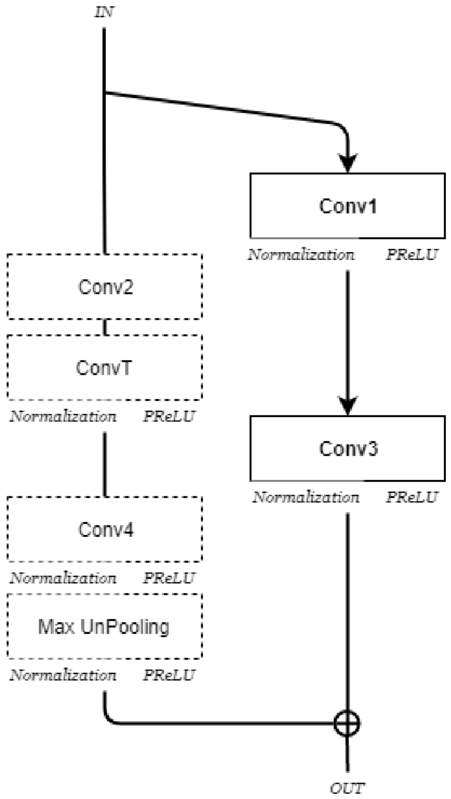

Figure 4. Realization of a single decoder bottleneck. Footnotes included on the convolution layers to highlight steps taken with each convolution.

Metrics for the OpenEDS Challenge focus on model accuracy and complexity. Model accuracy is measured as unweighted mean intersection-over union score (mIoU) for all classes in the test set, where $0 < P \leq 1$ and C is the number of learned model parameters. Model complexity $S > 0$ is measured as the number of model parameters, the unit of model size in megabytes.

$$S = C \times 4.0/(1024 * 1024)$$

A *Total* score incorporates accuracy and complexity to evaluate models in the eye segmentation challenge, where $0 <$

Name	Type	Output Size
input		$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
bottleneck1.1		$64 \times 128 \times 128$
bottleneck1.2		$64 \times 128 \times 128$
bottleneck1.3		$64 \times 128 \times 128$
bottleneck1.4		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated (2)	$128 \times 64 \times 64$
bottleneck2.3	asymmetric (5)	$128 \times 64 \times 64$
bottleneck2.4	dilated (4)	$128 \times 64 \times 64$
bottleneck2.5	dilated (8)	$128 \times 64 \times 64$
bottleneck3.0		$128 \times 64 \times 64$
bottleneck3.1	dilated (2)	$128 \times 64 \times 64$
bottleneck3.2	asymmetric (5)	$128 \times 64 \times 64$
bottleneck3.3	dilated (4)	$128 \times 64 \times 64$
bottleneck3.4	dilated (8)	$128 \times 64 \times 64$
bottleneck4.0	upsampling	$64 \times 128 \times 128$
bottleneck4.1		$64 \times 128 \times 128$
bottleneck4.2		$64 \times 128 \times 128$
bottleneck5.0	upsampling	$64 \times 256 \times 256$
bottleneck5.1		$64 \times 256 \times 256$
conv		$3 \times 512 \times 512$

Table 1. Architecture of MinENet. Output sizes are provided for input size of $512 \times 512 \times 3$.

$M \leq 100$:

$$M = 50\left(P + \min(\frac{1}{S})\right)$$

Table 2 displays these metrics, comparing our proposed model against the baseline provided in the challenge, as well as a vanilla CNN architecture and the same with dilation layers. The CNN and DilatedCNN models tested are of the same architecture and therefore have the same complexity. The difference in accuracy (mIoU) between these two models demonstrates the impact of dilation for semantic segmentation applications, versus fully convolutional layers. All tested models have a lower complexity than the provided baseline model, and both the dilated CNN and MinENet additionally improve upon accuracy.

A comparison of predictions and prediction masks on the OpenEDS image data is shown in Figure 5. The original image is immediately followed by the provided ground truth annotation mask, then the result of a dilated CNN model, and the prediction from MinENet. The eye is fully open in the first image, and demonstrates a more accessible case for all tested models. The dilated CNN does indicate an additional prediction below the eye, which is eliminated in the MinENet prediction. The second and third images depict the eye in upward and downward gazes respectively, and demonstrate more difficult cases for segmentation. The dilated CNN has difficulty identifying the pupil in the second image, whereas MinENet is able to identify and refine the boundaries against the background. The third image illustrates a challenging case for segmentation models, where heavy eyelashes mask reduce the visibility into the eye region. In this we find MinENet more capable of defining the pupil and iris regions than the CNN, however with room for improvement in clarifying the sclera.

Figure 6 further compares the results of our model on a challenging image case within the data: reflection visible in the image, caused by the participants' eye glasses.

	mIoU	Model Complexity	Total
Baseline	0.89478	416,088	0.76240
CNN	0.81711	371,688	0.76120
DilatedCNN	0.92792	371,688	0.81660
MinENet	0.92301	222,440	**0.96150**

Table 2. Comparison of semantic segmentation approaches on the OpenEDS dataset. The Baseline is the given from the OpenEDS Semantic Segmentation Challenge. CNN & DilatedCNN are simple CNNs, with/out dilations.

4. Conclusions

In this work, we propose MinENet, a dilated convolutional encoder-decoder architecture for improved semantic segmentation. Our model is demonstrated on the OpenEDS dataset and compared with state-of-the-art methods: a (provided) baseline model, a simple/vanilla CNN, and a vanilla CNN with dilation. MinENet minimizes model complexity, while improving accuracy on semantic eye segmentation tasks.

In future work, we will continue to improve the accuracy of the proposed model by training it in an adversarial environment. While the model accuracy is high, there are categories of challenging images within the dataset, for example those with reflections caused by glasses. A simple discriminator will be constructed in order to improve our model, the generator, in handling these corner-cases.

In addition, the integration of long short term memory (LSTM) units should assist in improving its overall accuracy against this and other challenging semantic segmentation datasets. This augmentation would allow the model a mechanism for storing prior knowledge in training, and in this specific task, assist in recalling the importance and general relative locations of eye regions from one image to the next.

Finally, the prevalence of adversarial examples motivates extending this work beyond the focus of model accuracy and complexity, to include the inherent security and explainability of the model.

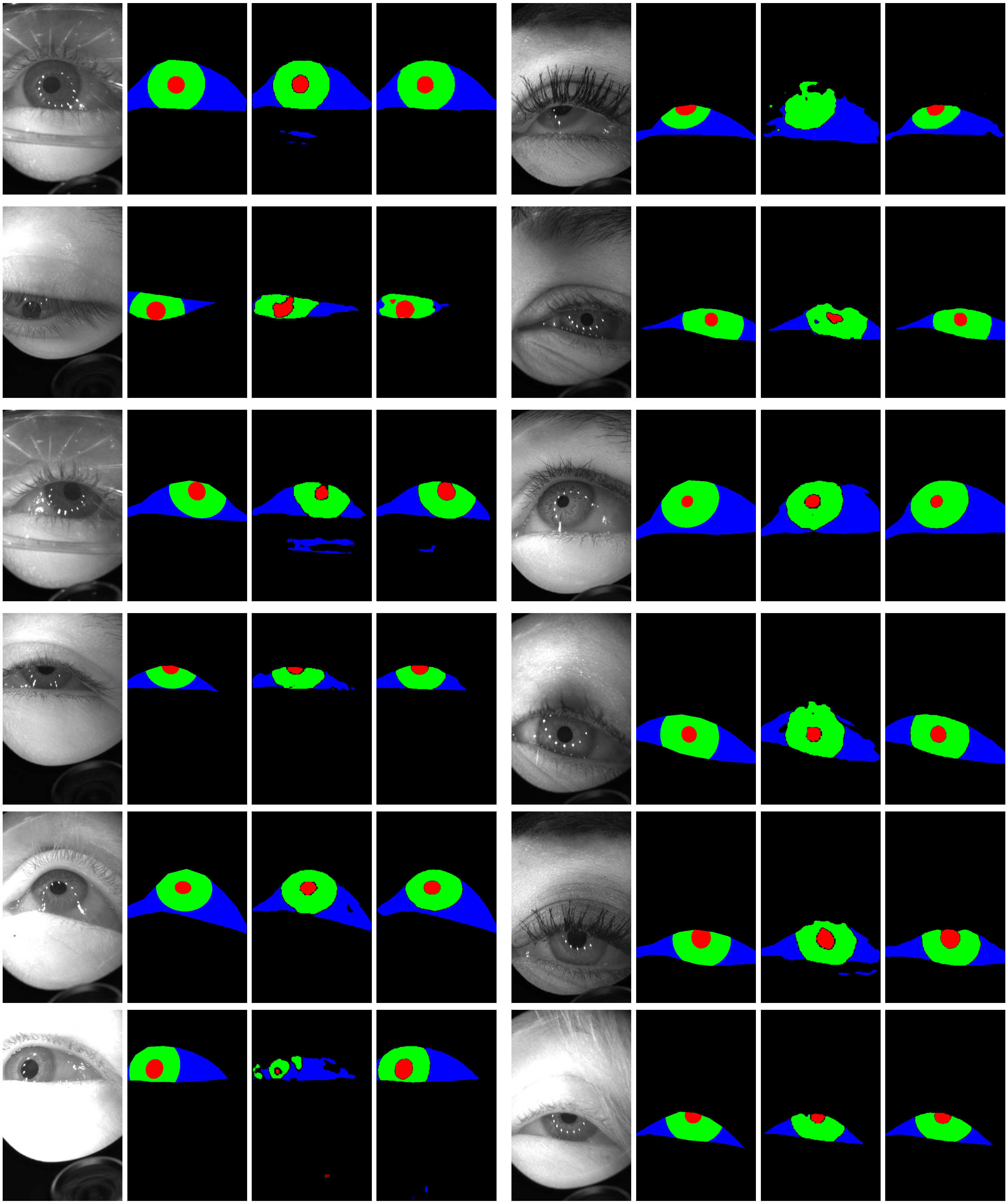

Figure 5. Comparison of segmentation results on OpenEDS data. From left: Original image, ground truth, dilated CNN result, MinENet. Best viewed in color.

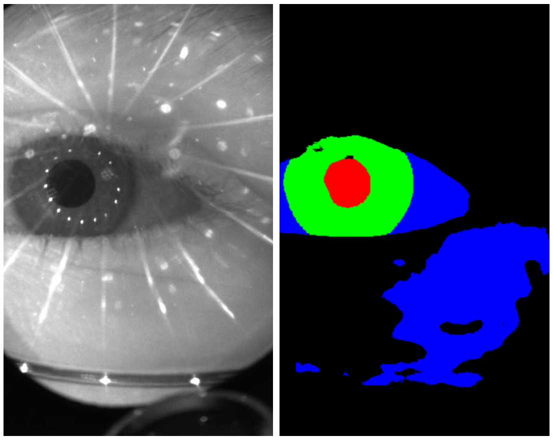

Figure 6. Segmentation predictions on an example challenging image - eye glasses on the participant cause reflections. From left: Original image, MinENet.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2, 3

[2] Stephan J Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019. 1, 2, 3

[3] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 2

[4] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1442–1450. IEEE, 2018. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, Dec 2015. 3

[6] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018. 2

[7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 1, 2

[8] Albert Montillo, Jamie Shotton, John Winn, Juan Eugenio Iglesias, Dimitri Metaxas, and Antonio Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 184–196. Springer, 2011. 1, 2

[9] Y. G. Naresh, S. Little, and N. E. O'Connor. A residual encoder-decoder network for semantic segmentation in autonomous driving scenarios. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1052–1056, Sep. 2018. 2

[10] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 1, 2

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015. 2

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[13] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):640–651, 2017. 2

[14] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *2011 International Conference on Computer Vision*, pages 1879–1886, Nov 2011. 1

[15] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018. 2

[16] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2

[17] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018. 1, 2

[18] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016. 2