EasyChair Preprint

№ 4378

Semantic Segmentation with Peripheral Vision

Mohammad Hamed Mozaffari Maaref and Won-Sook Lee

October 12, 2020

Semantic Segmentation with Peripheral Vision

M. Hamed Mozaffari[1 [0000−0002−2297−6114]] and Won-Sook Lee[1]

School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Canada
mmoza102@uottawa.ca, wslee@uottawa.ca

Abstract. Deep convolutional neural networks exhibit exceptional performance on many computer vision tasks, including image semantic segmentation. Pre-trained networks trained on a relevant and large benchmark have a notable impact on these successful achievements. However, confronting a domain shift, usage of pre-trained deep encoders cannot boost the performance of those models. In general, transfer learning is not a general solution for various computer vision applications with small accessible image databases. An alternative approach is to develop stronger deep network models applicable to any problem rather than encouraging scientists to explore available pre-trained encoders for their computer vision tasks. To deviate the direction of the research trend in image semantic segmentation toward more effective models, we proposed an innovative convolutional module simulating the peripheral ability of the human eyes. By utilizing our module in an encoder-decoder configuration, after extensive experiments, we achieved acceptable outcomes on several challenging benchmarks, including PASCAL VOC2012 and CamVid.

Keywords: Semantic segmentation · Convolutional Neural Networks · Dilated convolution · Deep learning · Peripheral vision · Encoder-Decoders · Image processing.

1 Introduction

Semantic segmentation is a fundamental step in a large group of applications, from scene understanding in self-driving vehicles to delineation of lesions in medical image analysis [19]. The aim of semantic segmentation is to assign one label for multiple objects of the same type. The main complication of semantic segmentation is closely related to scene and label variety [30] as well as the requirement of laborious works for manual labelling. However, in recent years, several groundbreaking deep learning methods based on Fully Convolutional Networks (FCNs) [21] have been exploited for the problem of semantic segmentation with astonishing advancements in several benchmarks [19] over systems relying on hand-crafted features [7].

Researchers conclude that the crucial elements for success of semantic segmentation methods are one of the two factors [4, 30, 7] of using multi-scale features, where features concatenated from intermediate layers using skip connections (e.g. spatial pyramid pooling) [30] or utilizing multi-scale input images to

a shared network [4, 16]. Moreover, embedding different Convolutional Neural Networks (CNNs) in "cascade" (deeper [11]) and "cascode" (shallower [30, 7]) configurations have boosted the performance of CNN models. Recently, combinations of encoder-decoder architectures [25] and other techniques such as spatial pyramid pooling [30] architectures with dilated convolution [7, 6, 22], and also post-processing methods [5] provide sharper object boundaries for several image segmentation benchmarks.

The major success of deep learning models in computer vision area owes to domain adaptation [12] where weights of a pre-trained model [28, 13] employed for fine-tuning of another model [29]. In designing almost all state-of-the-art image semantic segmentation models, the default routine is to adopt a publicly available classification encoder [1, 3], trained on a large database such as ImageNet [8]. Although this approach demonstrates considerable improvement in both accuracy [30, 7, 14] and speed [27], the impact of elaborating a model pre-trained on the current task as a relevant feature extractor is always ignored in many studies [27]. Moreover, using a model designed for classification tasks, pre-trained on a large dataset, cannot be a reliable approach for fine-tuning of another model which designed for image semantic segmentation task. This issue becomes even more critical when the target domain is entirely different from the source domain [18] (e.g. Pre-trained VGG16 model [28] on ImageNet [8], fine-tuned for medical image segmentation).

On the other hand, publicly available encoders are trained for specific tasks, and there are usually restrictions for using available pre-trained weights [18]. For instance, a non-modifiable network structure with a fixed-sized input image (e.g. PSPNet [30], DeepLabV3+ [7], and VGG16 [28] require squared sized images of 384×384, 513×513, and 224×224, respectively), forces researchers to manipulate (crop or interpolation) training data. An alternative technique is optimizing network architectures and improving their effectiveness [18]. For example, variants of U-net [25] model are optimized, dominated and applied in many medical image analysis tasks with outstanding results [10], even without using pre-trained encoders.

In this work, we demonstrate that the performance of recent scene parsing frameworks strongly depends on their pre-trained encoder block despite their outstanding results in many studies. At the same time, we demonstrate that for small-sized networks, pre-trained models cannot even boost the performance [24]. As a result of this dependency, there is not yet one prevailing deep learning model applicable to different types of databases. Towards designing a general deep learning model for semantic segmentation task, we proposed a new convolutional module inspired by human peripheral vision [26] (named RetinaConv), embedded into a new deep convolutional encoder-decoder architecture called IrisNet. Several novel scenes parsing framework [30, 25, 22, 3, 21] and IrisNet model evaluated on different databases, PASCAL VOC 2012 [9] and CAMVID [2] without employing any pre-trained model. Experimental results demonstrate that our proposed model can predict similar or even better instances in comparison with other techniques.

2 Methodology and Network architecture

The human brain can process different scenes in a fraction of a second, and it can detect objects and movements outside of the direct line of sight, away from the center of gaze (known as peripheral vision [26]). With the aid of this ability, we can detect and sense objects without turning our heads or eyes, resulting in fewer computations for our brain. Moreover, the human eye has a limited field of view, whereas the scene is sharper in the center and more blurry around edges [26]. Simulating the peripheral vision property of the human eye, we designed a new convolutional module called RetinaConv. RetinaConv module is presented in Figure 1, where the center of the filter (mimicking center of the human eye gaze) is stronger than neighbours. A RetinaConv kernel is created by adding two convolutional kernels, one standard and another dilated type. Similar to a Gaussian filter, RetinaConv can have different standard deviations with varying dilation and stride rates in both type of kernels.

For the implementation of RetinaConv, we benefit from the distributivity property of convolution operators $f * (g + h) = f * g + g * h$, where f is input feature, g and h are standard and dilated convolutional kernels, respectively. Different concentrations of peripheral vision can be generated by changing the hyper-parameters of RetinaConv. One advantage of RetinaConv is that it has two different effective receptive fields simultaneously. With the RetinaConv block, we propose our end-to-end IrisNet model (see Figure 2) for solving semantic segmentation tasks. IrisNet detects and emphases the core features of an input image easier than individual convolutional block due to the use of both standards and dilated convolutions in each RetinConv block. The minimum performance of the IrisNet is guaranteed at least to the extent of the U-net [25] network using unit dilation rates.

Due to the particular configuration of the RetinaConv block, IrisNet benefits from the receptive filed of both standard and dilated convolution. For instance, applying two times a standard convolutional kernel to an input image with a filter size of 3×3, padding size of 1, and stride of 2×2, the effective receptive field [20] for each feature is 3×3 and 7×7 for the first and second time, respectively. On the other hand, using a RetinaConv with a dilation rate of 2 and the same settings as the previous example, corresponding effective receptive filed is 5×5 and 13×13 with more concentration on features near to the center of the receptive fields. Stacking several layers of RetinaConv mimics the peripheral vision ability of the human brain (see Figure 1).

3 Experimental evaluations

Our proposed method is successful on scene parsing and semantic segmentation of different database types. One strength capacity of our model is its ability to train on different types of image data with acceptable results without employing any pre-trained model. We evaluate the proposed method in this section on two different databases, including PASCAL VOC 2012 or general semantic segmentation [9] and CamVid for pedestrian and vehicle segmentation [2].

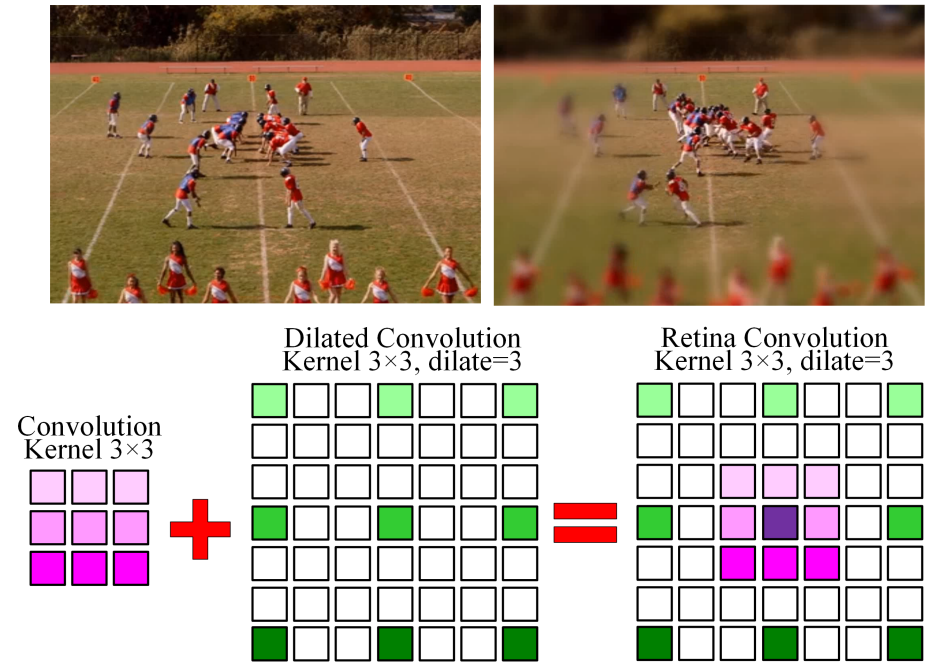

Fig. 1. Peripheral vision in the human eye. The center of gaze is sharper due to more light detectors on Retina (dense kernel) and around is blurry because of fewer detectors on Retina (sparse kernel).

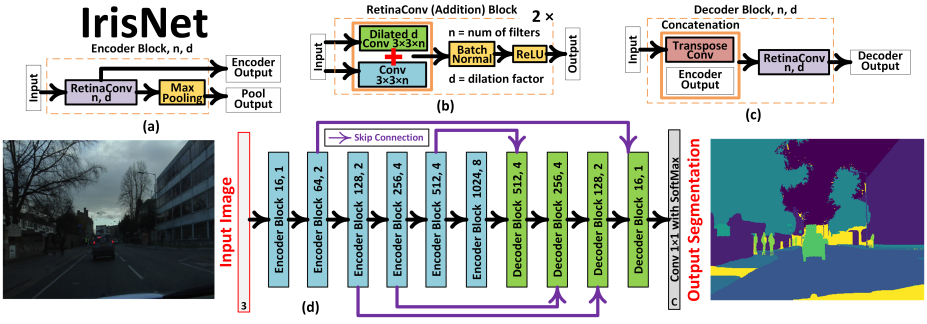

Fig. 2. Network architecture of IrisNet with different embedded blocks.

We implemented RetinaConv and IrisNet on the public platform Tensorflow. All models in this study were optimized using categorical cross-entropy loss by Adam optimization method with first (β_1) and second (β_2) momentum of 0.9 and 0.999, respectively. In the last layer of all networks, we used "Softmax"

activation functions. The learning rate value for all models was exponentially variable with iterations, initially set by 0.001 with the decay factor of 10^{-6}. The performance might be slightly improved by increasing the epoch number, which is set by 100 for CamVid and 150 for PASCAL VOC. For data augmentation, we adopt horizontal flipping, scaling between 0.5 to 1.5, and shift with 10 percent in all directions, randomly.

Furthermore, for the CamVid dataset, we added a random Gaussian blurring filter with a variance noise ranges of 0.2. We cropped images during our online data augmentation process to 320×320 for CamVid and 224×224 for PASCAL VOC. Following [15], we employed batch normalization instead of drop-out layers between each convolutional layer. For network configuration and hyperparameter-tuning of our model, we used default values from each publication or publicly available codes. For IrisNet, we followed the configuration of common encoder-decoders [25, 1, 23, 6] in the literature for a fair comparison between models. Activation function for IrisNet was ReLU, and due to the limited computational resources (GPU power), we selected the "batch-size" to 20 during training. The ratio of train, validation, and test sets are 90%, 5%, and 5%, respectively. For the comparison study, we keep the best models by saving checkpoints during the training and validation stage. In the test step, raw data are fed to each network with their original sizes.

IrisNet works satisfyingly on scene parsing challenge of PASCAL VOC 2012 benchmark where the dataset has 20 objects categories and one background. Online augmentation of the PASCAL VOC dataset results in $7,863K$, $438K$, and $438K$ images, cropped by 224×224 for training, validation, and testing. Table 1 shows the comparison results of IrisNet with several advanced methods on each benchmark.

Table 1. Performance of models in evaluation study on the PASCAL VOC 2012 test set in terms of IOU and mean IOU. The number of trainable parameters for each model is in millions.

Model	parameters	Airplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mIOU
BowNet	0.92	**62.7**	**41.5**	16.3	17.6	**53.8**	35.6	41.9	67.2	13.8	67.8	12.9	**85.4**	63.2	49.1	38.4	51.7	67.2	23.8	27.6	21.9	51.2
UNET [25]	23.7	61.8	39.6	56.8	27.3	53.5	74.6	48.6	**73.8**	18.3	**74.9**	10.2	84.3	**65.3**	53.0	79.0	**70.4**	**72.5**	**70.6**	31.4	17.8	55.7
FCN8 [21]	13.2	56.9	40.2	23.9	34.2	40.6	42.6	50.2	64.8	20.9	59.9	13.8	79.2	52.1	54.9	68.2	69.1	60.9	59.4	29.7	20.6	55.1
LinkNet [3]	20.3	55.2	32.1	34.2	**35.0**	35.2	68.5	39.1	53.6	38.9	49.6	20.3	30.2	30.6	52.6	56.8	59.7	55.8	22.7	22.0	13.9	55.8
FPN [17]	17.5	38.4	28.7	62.6	42.1	36.8	60.0	23.8	36.1	21.3	32.7	13.4	75.8	50.7	49.9	55.7	49.8	48.7	35.9	28.1	12.2	53.5
IrisNet	71.7	44.9	30.2	**66.3**	24.3	44.5	**75.6**	**52.4**	65.8	**43.5**	72.1	**42.8**	80.3	51.6	**80.3**	**82.7**	51.8	55.5	62.6	**42.6**	**27.8**	**57.2**

From table 1, IrisNet outperforms other methods in terms of mean intersection over union (mIOU). The number of trainable parameters for each model is also reported in this table. As can be seen, models such as BowNet and UNET with fewer parameters can predict acceptable results due to their efficient structures. For this reason, optimizing all sections of a network structure (even en-

coder block) is just as crucial as investigating other Influential aspects, such as decoder block. Several instances predicted by each encoder-decoder network are illustrated in Figure 3. Although the results for all models are not considerable, IrisNet predicted instances with more details. For instance, the tail of the "cat" and ears/grass for "cow" have more details than other models.

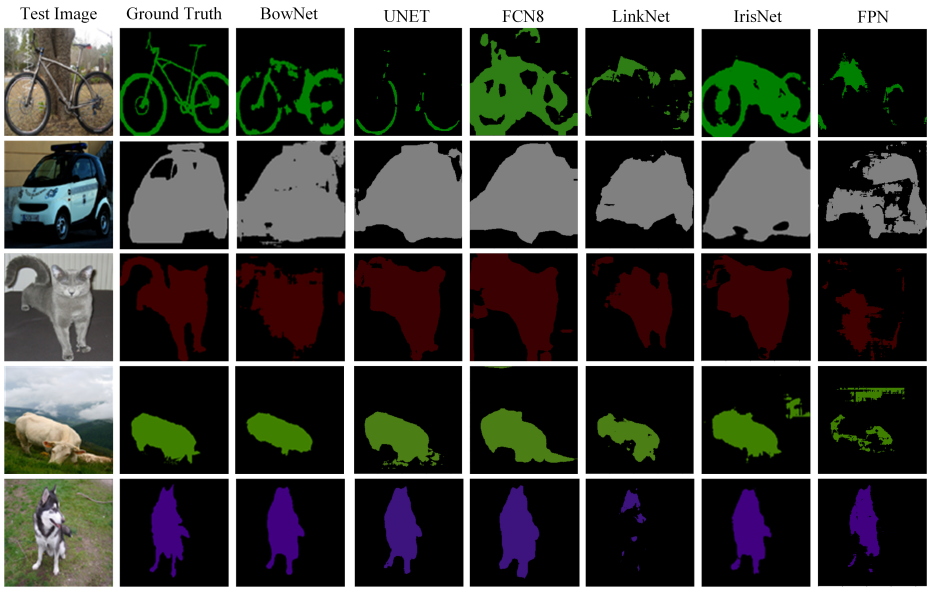

Fig. 3. Results of each model in terms of per-class results on the PASCAL VOC 2012 [9] testing set. All models are trained on the dataset with random weight initialization (no pre-trained encoder).

CamVid dataset has 32 semantic classes for urban scene understanding. To compare each model on a more straightforward dataset, we employed a subset of CamVid contains three classes of "background", "car", and "pedestrian". In our evaluation study, there were 367, 101, and 233 annotated images for training, validation, and testing sets, while after online augmentation, training and validation sets were increased to $734K$ and $202K$ images, respectively. All models except PSPNet (384×384) were trained with cropping sizes of 320×320.

Table 2 reports our assessment results of six networks on the CamVid dataset in three configurations (random initialization, initializing with pre-trained weights, and fine-tuning by freezing encoder parameters) while, except BowNet and IrisNet, the backbone network is VGG16 encoder network pre-trained on ImageNet dataset. For each configuration, we presented three evaluation criteria mIOU, F1, and Categorical Cross-Entropy. From the table, IrisNet could predict better instances than other models, while random initialization was used for each

Table 2. Quantitative results of the CamVid test set. Methods without available pre-trained models are indicated by n/a. None means the network could not train features from the dataset, without providing any predicted instance.

Method	Backbone	Random initial			Pre-trained weights			Fine-tuning		
	ImNet	mIOU	F1	Loss	mIOU	F1	Loss	mIOU	F1	Loss
BowNet [22]	n/a	50.52	0.58	0.48	n/a	n/a	n/a	n/a	n/a	n/a
UNET [25]	VGG16	35.80	0.38	0.87	37.58	0.41	0.86	0.40	0.43	0.81
PSPNet [30]	VGG16	32.18	0.33	0.93	**46.50**	**0.54**	**0.65**	54.80	0.63	0.50
LinkNet [3]	VGG16	38.83	0.44	0.80	39.19	0.44	0.80	65.06	0.73	0.36
FPN [17]	VGG16	None	None	None	None	None	None	**68.44**	**0.76**	**0.32**
IrisNet	n/a	**55.77**	**0.61**	**0.45**	n/a	n/a	n/a	n/a	n/a	n/a

model. Definitely, all models might achieve better results by optimizing all aspects of the experiment and training for more epochs. Some examples of this evaluation study are displayed in figure 4. From the figure can be seen that although IrisNet performs better in comparison with other models, it is weak in dealing with large objects in the scene.

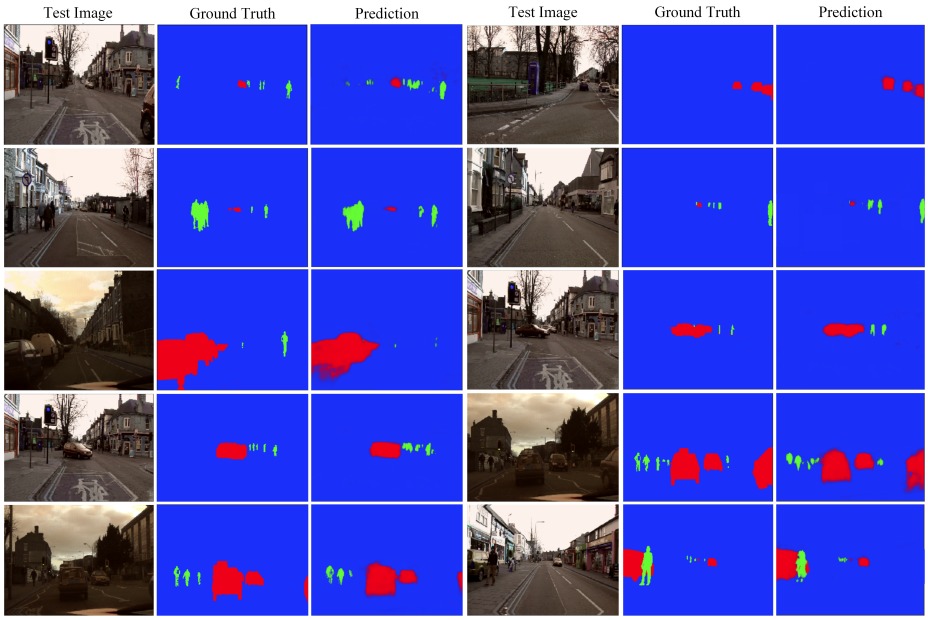

Fig. 4. Results of assessment of each model on CamVid test set.

4 Conclusion

Our proposed encoder-decoder model (IrisNet) employs a new convolutional module (RetinaConv) to mimic the nature of peripheral vision in human eyes. Specifically, benefits from the effective receptive field of RetinaConv, IrisNet encodes multi-scale information superior to other cutting-edge deep learning models. The primary motivation behind IrisNet architecture using the RetinaConv module was the need to implement an efficient deep learning model for semantic segmentation, which works independently from pre-trained models while capable of applying on several types of datasets. To address this desired model, we improved the feature extraction ability of a ubiquitous encoder-decoder model (UNET) by employing RetinaConv. Our experimental results show that the dependency of the proposed method from using pre-trained encoder blocks is significant, and it achieves comparable performance with other state-of-the-art models in similar configurations on several challenging benchmarks. Generalization capability of the IrisNet in image segmentation task on datasets with different distributions and context was evaluated with an acceptable achievements. We believe that optimized, universal, and efficient deep network architectures will stay longer in literature than models with just higher accuracy and performance.

References

1. Badrinarayanan, V., et al.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Brostow, G.J., et al.: Segmentation and recognition using structure from motion point clouds. In: ECCV (1). pp. 44–57 (2008)
3. Chaurasia, A., et al.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
4. Chen, L.C., et al.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on CVPR. pp. 3640–3649 (2016)
5. Chen, L.C., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
6. Chen, L.C., et al.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the ECCV. pp. 801–818 (2018)
8. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on CVPR. pp. 248–255. Ieee (2009)
9. Everingham, M., et al.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
10. Falk, T., et al.: U-net: deep learning for cell counting, detection, and morphometry. Nature methods **16**(1), 67 (2019)
11. Fu, J., et al.: Stacked deconvolutional network for semantic segmentation. IEEE Transactions on Image Processing (2019)

12. Hamed Mozaffari, M., Lee, W.S.: Domain adaptation for ultrasound tongue contour extraction using transfer learning: A deep learning approach. The Journal of the Acoustical Society of America **146**(5), EL431–EL437 (2019)
13. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on CVPR. pp. 770–778 (2016)
14. He, K., et al.: Mask r-cnn. In: Proceedings of the IEEE ICCV. pp. 2961–2969 (2017)
15. Ioffe, S., et al.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
16. Lin, G., et al.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on CVPR. pp. 1925–1934 (2017)
17. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on CVPR. pp. 2117–2125 (2017)
18. Liu, S., et al.: Deep learning in medical ultrasound analysis: a review. Engineering (2019)
19. Liu, X., et al.: Recent progress in semantic image segmentation. Artificial Intelligence Review **52**(2), 1089–1106 (2019)
20. Liu, Y., et al.: Understanding the effective receptive field in semantic image segmentation. Multimedia Tools and Applications **77**(17), 22159–22171 (2018)
21. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on CVPR. pp. 3431–3440 (2015)
22. Mozaffari, M.H., Lee, W.S.: Encoder-decoder cnn models for automatic tracking of tongue contours in real-time ultrasound data. Methods (2020)
23. Noh, H., et al.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE ICCV. pp. 1520–1528 (2015)
24. Poudel, R.P., et al.: Fast-scnn: fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019)
25. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on MICCAI. pp. 234–241. Springer (2015)
26. Rosenholtz, R.: Capabilities and limitations of peripheral vision. Annual Review of Vision Science **2**, 437–457 (2016)
27. Siam, M., et al.: Rtseg: Real-time semantic segmentation comparative study. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 1603–1607. IEEE (2018)
28. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Tan, C., et al.: A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. pp. 270–279. Springer (2018)
30. Zhao, H., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on CVPR. pp. 2881–2890 (2017)