Adaptive Context Selection for Polyp Segmentation

Ruifei Zhang[1], Guanbin Li[1(✉)], Zhen Li[2], Shuguang Cui[2], Dahong Qian[3], and Yizhou Yu[4]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[2] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Guangdong, China
[3] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[4] Deepwise AI Lab, Beijing, China

Abstract. Accurate polyp segmentation is of great significance for the diagnosis and treatment of colorectal cancer. However, it has always been very challenging due to the diverse shape and size of polyp. In recent years, state-of-the-art methods have achieved significant breakthroughs in this task with the help of deep convolutional neural networks. However, few algorithms explicitly consider the impact of the size and shape of the polyp and the complex spatial context on the segmentation performance, which results in the algorithms still being powerless for complex samples. In fact, segmentation of polyps of different sizes relies on different local and global contextual information for regional contrast reasoning. To tackle these issues, we propose an adaptive context selection based encoder-decoder framework which is composed of Local Context Attention (LCA) module, Global Context Module (GCM) and Adaptive Selection Module (ASM). Specifically, LCA modules deliver local context features from encoder layers to decoder layers, enhancing the attention to the hard region which is determined by the prediction map of previous layer. GCM aims to further explore the global context features and send to the decoder layers. ASM is used for adaptive selection and aggregation of context features through channel-wise attention. Our proposed approach is evaluated on the EndoScene and Kvasir-SEG Datasets, and shows outstanding performance compared with other state-of-the-art methods. The code is available at https://github.com/ReaFly/ACSNet.

1 Introduction

Colorectal cancer is a serious threat to human health, with the third highest morbidity and mortality among all cancers [15]. As one of the most critical precursors of this disease, polyp localization and segmentation play a key role in the early diagnosis and treatment of colorectal cancer. At present, colonoscopy is the most commonly used means of examination, but this process involves manual and thus expensive labor, not to mention its higher misdiagnosis rate [16].

Therefore, automatic and accurate polyp segmentation is of great practical significance. However, polyp segmentation has always been a challenging task due to the diversity of polyp in shape and size. Some examples of polyp segmentation are displayed in Fig. 1.

In recent years, with the prevalence of deep learning technology, a series of convolutional neural network variants have been applied to polyp segmentation and have made breakthrough progress. Early fully convolutional neural networks [1,2,9,12] replaced the fully connected layers of the neural network with convolutional ones. In order to enlarge the receptive field of the neurons, the neural network gradually reduces the scale of the feature map and finally generates the prediction with very low resolution, resulting in a rough segmentation result and prone to inaccurate boundaries. Later, UNet [14] based structure was proposed, which adopts a stepwise upsample learning to restore the feature map resolution while maintaining the relatively large receptive field of the neurons. At the same time, the skip connection is used to enhance the fusion of shallow and deep features to improve the original FCN, greatly improving the segmentation performance and boundary localization of the specific organs or diseased regions. SegNet [19] is similar to UNet, but utilizes the max pooling indices to achieve up-sample operation in the decoder branch. SFANet [3] incorporates a sharing encoder branch and two decoder branches to detect polyp regions and boundaries respectively, and includes a new boundary-sensitive loss to mutually improve both polyp region segmentation and boundary detection. In addition, by adopting the upward concatenation to fuse multi-level features and embedding the selective kernel module to learn multi-scale features, the model is further enhanced and achieves competitive results. However, most of the methods have not taken proper measures to deal with the shape and size variance of polyps regions.

In this paper, we propose the Adaptive Context Selection Network (ACSNet). Inspired by [4], we believe that the global context features are helpful for the segmentation of large polyps, while the local context information is crucial for the identification of small ones. Therefore, the intent of our designed network is to adaptively select context information as contrast learning and feature enhancement based on the size of the polyp region to be segmented. Specifically, our ACSNet is based on the encoder-decoder framework, with Local Context Attention (LCA) module, Global Context Module (GCM), and Adaptive Selection Module (ASM). LCAs and GCM are responsible for mining local and global context features and sending them to the ASM modules in each decoder layer. Through channel-wise attention, ASM well achieves adaptive feature fusion and selection. In summary, the contributions of this paper mainly include: (1) Our designed ACSNet can adaptively attend to different context information to better cope with the impact of the diversity of polyp size and shape on segmentation. (2) Our tailored LCA and GCM modules can achieve more consistent and accurate polyp segmentation through complementary selection of local features and cross-layer enhancement of global context. (3) ACSNet achieves new state-of-the-art results on two widely used public benchmark datasets.
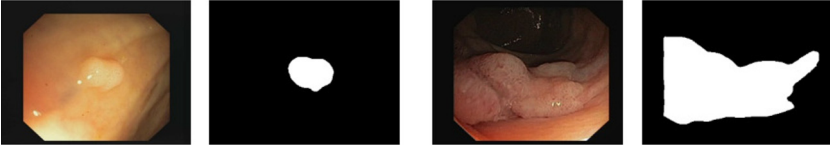
Fig. 1. Two examples of polyp segmentation

2   Method

The architecture of our ACSNet is shown in Fig. 2, which can be regarded as an enhanced UNet [14] or Feature Pyramid Network (FPN) [10]. We utilize ResNet34 [5] as our encoder, which contains five blocks in total. Accordingly, the decoder branch also has five blocks. Each decoder block is composed of two Conv-BN-ReLU combinations, and generates one prediction map with different resolution, which is supervised by the down-sampled ground truth respectively.

The GCM is placed on top of the encoder branch, which captures the global context information and densely concatenates to the ASM of each layer in the decoder path. At the same time, each skip-connection between the encoder and decoder paths of UNet [14] is replaced by the LCA module, which gives each positional feature column of every decoding layer a local context enhancement of different receptive field and at the same time delicately leverages the prediction confidence of the previous layer as a guidance to force the current layer to focus on harder regions. Finally, we utilize the ASM modules to integrate the features output from each previous decoder block, the LCA module and the GCM, based on a channel-wise attention scheme for context selection.

2.1   Local Context Attention Module (LCA)

LCA is designed as a kind of spatial attention scheme, which aims to incorporate hard sample mining when merging shallow features and pay more attention to the uncertain and more complex area to achieve layer-wise feature complementation and prediction refinement. As shown in Fig. 3, the attention map of each LCA module is determined by the prediction map generated from the upper layer of the decoder stream. Specifically, the attention map of the ith LCA module is denoted as $Att_i \in \mathbb{R}^{1 \times H_i \times W_i}$, in which H_i, W_i are the height and width of the attention map respectively. The value of position $j \in [1, 2, \cdots, H_i \times W_i]$, denoted as Att_i^j can be calculated as follows:

$$Att_i^j = 1 - \frac{\left| p_{i+1}^j - T \right|}{\max\left(T, 1-T\right)}, \tag{1}$$

where $P_{i+1}^j \in (0,1)$ is the jth location value of the prediction map $P_{i+1} \in \mathbb{R}^{1 \times H_i \times W_i}$ which is generated by the $(i+1)^{th}$ decoder block. T is the threshold to determine whether the specific position belongs to foreground or background.
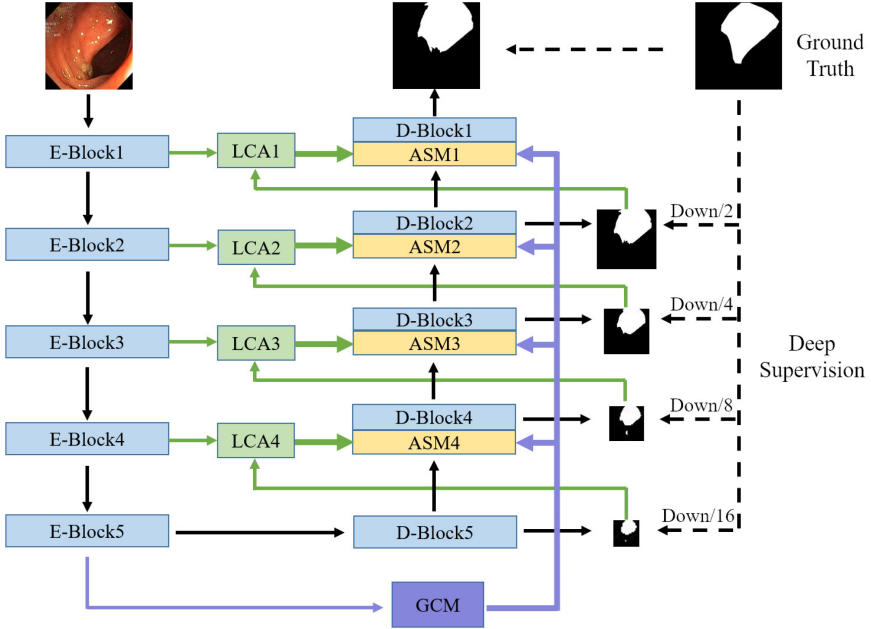
Fig. 2. Overview of our proposed ACSNet

We calculate the absolute difference between the prediction value and threshold T, and limit it to the range of 0 to 1 by dividing the maximum difference. We believe that the closer the predicted value is to the threshold T, the more uncertain the prediction of the corresponding position is, so it should be given a larger attention weight in the forwarding layer, and vice versa. Finally, we multiply the features by the attention values, and then sum with the original features to get the output of this module. For simplicity, T is set to 0.5 in our experiments.
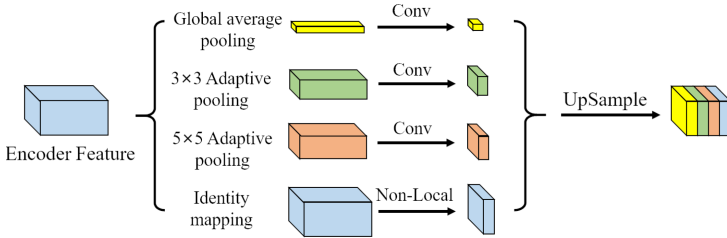
2.2 Global Context Module (GCM)

We borrow the idea from pyramid pooling [6,11,20] to design our GCM and also put it as an independent module for global context inferring on top of the encoder branch. Meanwhile, GCM forwards the output to each ASM module to compensate the global context which is gradually diluted during layer-wise refinement.

As shown in Fig. 4, GCM contains four branches to extract context features at different scales. Specifically, this module is composed of a global average pooling branch, two adaptive local average pooling branches, and outputs three feature maps of spatial size 1×1, 3×3, 5×5, respectively. It also contains an identity mapping branch with non local operation [18] to capture the long range dependency while maintaining the original resolution. We introduce a non-
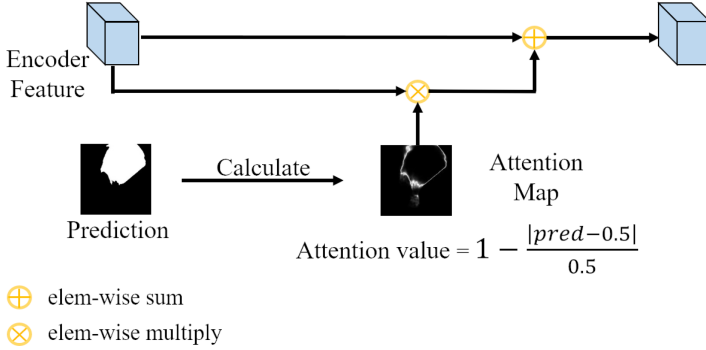
Fig. 3. Local Context Attention Module (LCA)

local operation based feature representation here to finely capture the global dependency of each positional feature to enhance the output of the encoder. In the end, we up-sample the above four feature maps and concatenate them to obtain the resulted global context feature of this module, which will be densely fed to each designed ASM module in the decoder stream.


Fig. 4. Global Context Module (GCM)

2.3 Adaptive Selection Module (ASM)

We believe that local context and global context have different reference values for the segmentation of polyp regions with different appearances, sizes, and feature contrasts. Therefore, we attach an adaptive context selection module (ASM) to each block in the decoder stream. Based on the local context features generated by the LCA, the global context features from the GCM, and the output features of previous decoder block as inputs, it learns to adaptively select context feature for aggregation in each block.

As shown in Fig. 5, we incorporate a "Squeeze-and-Excitation" block [7] to adaptively recalibrate channel-wise feature responses for feature selection. Specifically, ASM takes the concatenated feature as input, and employs global average

pooling to squeeze the feature map to a single vector which is further fed to a fully connected layer to learn the weight of each channel. After sigmoid operation, the attention weight is limited to the range of 0 to 1. Through multiplying the original feature maps with the attention values, some informative context features can be picked out while those not conducive to improving discrimination will be suppressed. Noted that we also apply non local operation [18] to the features output from previous decoder block before concatenation to enhance the decoder features with long range dependency.
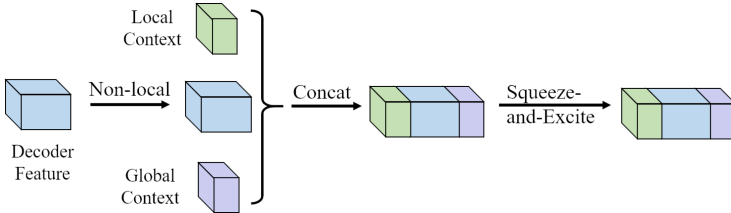

Fig. 5. Adaptive Selection Module (ASM)

3   Experiments

3.1   Datasets

We evaluate our proposed method on two benchmark colonoscopy image datasets, collected from the examination of colorectal cancer. The first is the EndoScene Dataset [17], which contains 912 images and each of which has at least one polyp region. It is divided into the training set, validation set and test set, with 547, 183, and 182 images respectively. For simplicity, we resize the images to 384×288 uniformly in our experiments. The second is Kvasir-SEG Dataset [8] containing 1000 images with polyp regions. We randomly use 60% of the dataset as training set, 20% as validation set, and the remaining 20% as test set. Since the image resolution of this dataset varies greatly, we refer to the setting of [8] and set all images to a fixed size of 320×320.

3.2   Implementation Details and Evaluation Metrics

In the training stage, we use data augmentation to enlarge the training set, including random horizontal and vertical flips, rotation, zoom and shift. All the images are randomly cropped to 224×224 as input. We set batch size to 4, and use SGD optimizer with a momentum of 0.9 and a weight decay of 0.0005 to optimize the model. A poly learning rate police is adopted to adjust the initial learning rate, which is $lr = init\_lr \times (1 - \frac{epoch}{nEpoch})^{power}$, where $init\_lr = 0.001$, $power = 0.9$, $nEpoch = 150$. We utilize the combination of a binary cross
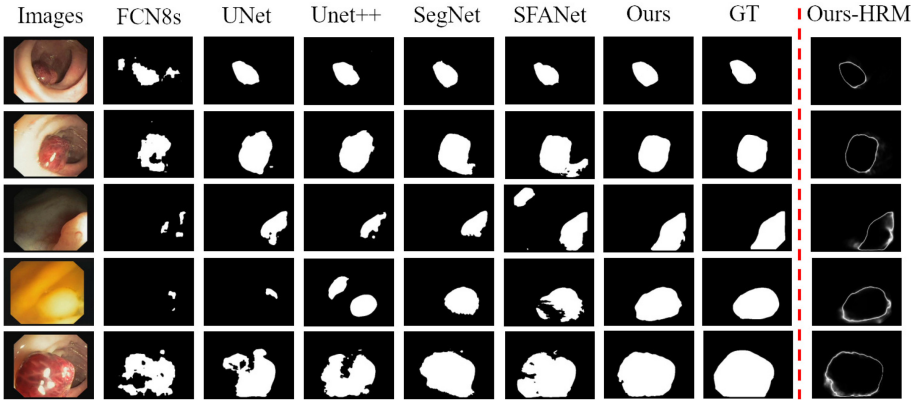
Fig. 6. Visual comparison of polyp region segmentation from state-of-the-art methods. The ground truth (GT) is shown in the penultimate column. Our proposed method consistently produces segmentation results closest to the ground truth. The hard region mining result is shown in the rightmost column.

entropy loss and a dice loss as the loss function. Our model is implemented using PyTorch [13] framework.

As in [3], we use eight metrics to evaluate the segmentation performance, including "Recall", "Specificity", "Precision", "Dice Score", "Intersection-over-Union for Polyp (IoUp)", "IoU for Background (IoUb)", "Mean IoU (mIoU)" and "Accuracy".

Table 1. Comparison with other state-of-the-art methods on the EndoScene dataset

Methods	Rec	$Spec$	$Prec$	$Dice$	$IoUp$	$IoUb$	$mIoU$	Acc
FCN8s [1]	60.21	98.60	79.59	61.23	48.38	93.45	70.92	93.77
UNet [14]	85.54	98.75	83.56	80.31	70.68	95.90	83.29	96.25
UNet++ [21]	78.90	99.15	86.17	77.38	68.00	95.48	81.74	95.78
SegNet [19]	86.48	99.04	86.54	82.67	74.41	96.33	85.37	96.62
SFANet [3]	85.51	98.94	86.81	82.93	75.00	96.33	85.66	96.61
Ours	**87.96**	**99.16**	**90.99**	**86.59**	**79.73**	**96.86**	**88.29**	**97.11**

3.3   Results on the EndoScene Dataset

We compare our ACSNet with FCN8s [1], UNet [14], UNet++ [21], SegNet [19] and SFANet [3] on the test set. As shown in Table 1, our method achieves the best performance over all metrics, with $Dice$ of 86.59%, a 3.66% improvement over

the second best algorithm. Some visualization results are shown in Fig. 6 (Col.1-8), as can be seen that our algorithm is very robust to some complex situations such as polyp region sizes and image brightness changes. At the same time, due to the introduction of the effective context selection module and especially the hard region mining (abbr.HRM) mechanism, the algorithm is significantly more accurate for polyp boundary positioning. In the rightmost column of Fig. 6, it can be observed that the hard regions mined by our method are usually located in the border area of polyps, which is worthy of attention during prediction refinement.

3.4   Results on the Kvasir-SEG Dataset

On this dataset, we compare our ACSNet with UNet [14], UNet++ [21], Seg-Net [19], ResUNet [8] and SFANet [3]. The results are listed in Table 2. Similarly, our method achieves the best performance and outperforms others by large margins, further demonstrating the robustness and effectiveness of our method.

Table 2. Comparison with other state-of-the-art methods and Ablation study on the Kvasir-SEG dataset

Methods	Rec	Spec	Prec	Dice	IoUp	IoUb	mIoU	Acc
UNet [14]	87.89	97.69	83.89	82.85	73.95	94.73	84.34	95.65
UNet++ [21]	88.67	97.49	83.17	82.80	73.74	94.49	84.11	95.42
ResUNet [8]	81.25	98.31	87.88	81.14	72.23	94.00	83.11	94.90
SFANet [3]	91.99	97.05	82.95	84.68	77.06	94.83	85.94	95.71
SegNet [19]	90.03	98.13	87.51	86.43	79.11	95.90	87.51	96.68
Ours	**93.14**	98.55	**91.59**	**91.30**	**85.80**	**97.00**	**91.40**	**97.64**
Baseline	89.53	98.63	90.32	88.21	81.59	96.27	88.93	96.99
Baseline+LCAs	91.79	98.39	89.15	89.00	82.47	96.41	89.44	97.15
Baseline+LCAs+GCM	92.18	**98.72**	90.90	90.28	84.35	96.88	90.62	97.52

3.5   Ablation Study

To validate the effectiveness and necessity of each of the three modules in our proposed method, we compare ACSNet with its three variants in Table 2. Specifically, the baseline model refers to the original U-shape encoder-decoder framework with skip-connections, and we gradually add LCAs, GCM, and ASMs to it, denoted as Baseline+LCAs, Baseline+LCAs+GCM and Ours, respectively. As shown in the table, with the progressive introduction of LCAs, GCM, and ASMs, our algorithm has witnessed a certain degree of performance improvement, boosting *Dice* by 0.79%, 1.28%, 1.02% respectively.

4   Conclusion

In this paper, we believe that an efficient perception of local and global context is essential to improve the performance of polyps region localization and segmentation. Based on this, we propose an adaptive context selection based encoder-decoder framework which contains the LCA module for hard region mining based local context extraction, the GCM module for global feature representation and enhancement in each decoder block, and the ASM component for contextual information aggregation and selection. Extensive experimental results and ablation studies have demonstrated the effectiveness and superiority of the proposed method.

References

1. Akbari, M., et al.: Polyp segmentation in colonoscopy images using fully convolutional network. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 69–72 (2018)
2. Brandao, P., et al.: Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134, p. 101340F. International Society for Optics and Photonics (2017)
3. Fang, Y., Chen, C., Yuan, Y., Tong, K.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 302–310. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_34
4. Fu, J., et al.: Adaptive context network for scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6748–6757 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y.: Non-local context encoder: robust biomedical image segmentation against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8417–8424 (2019)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
8. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
9. Li, G., Yu, Y.: Contrast-oriented deep neural networks for salient object detection. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 6038–6051 (2018)
10. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

11. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3917–3926 (2019)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
13. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8026–8037 (2019)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2020. CA: A Cancer J. Clin. **70**(1), 7–30 (2020)
16. Van Rijn, J.C., et al.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. **101**(2), 343–350 (2006)
17. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. **2017** (2017)
18. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
19. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. arXiv preprint arXiv:1807.10584 (2018)
20. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
21. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1