3D-ESPNet with Pyramidal Refinement for Volumetric Brain Tumor Image Segmentation

Nicholas Nuechterlein(✉) and Sachin Mehta

University of Washington, Seattle, WA 98195, USA
{nknuecht,sacmehta}@cs.washington.edu

Abstract. Automatic quantitative analysis of structural magnetic resonance (MR) images of brain tumors is critical to the clinical care of glioma patients, and for the future of advanced MR imaging research. In particular, automatic brain tumor segmentation can provide volumes of interest (VOIs) to scale the analysis of advanced MR imaging modalities such as perfusion-weighted imaging (PWI), diffusion-weighted imaging (DTI), and MR spectroscopy (MRS), which is currently hindered by the prohibitive cost and time of manual segmentations. However, automatic brain tumor segmentation is complicated by the high heterogeneity and dimensionality of MR data, and the relatively small size of available datasets. This paper extends ESPNet, a fast and efficient network designed for vanilla 2D semantic segmentation, to challenging 3D data in the medical imaging domain [11]. Even without substantive pre- and post-processing, our model achieves respectable brain tumor segmentation results, while learning only 3.8 million parameters. 3D-ESPNet achieves dice scores of 0.850, 0.665, and 0.782 on whole tumor, enhancing tumor, and tumor core classes on the test set of the 2018 BraTS challenge [1–4,12]. Our source code is open-source and available at https://github.com/sacmehta/3D-ESPNet.

Keywords: Glimoa · BraTS · ESPNet · CNN · Semantic segmentation

1 Introduction

Glioma is the most common primary brain tumor. Due to glioma's highly heterogeneous appearance, extent, and shape, segmentation of brain tumors in MR volumes is one of the most challenging tasks in neuroradiology [7]. This is compounded by the sparsity of data and the heterogeneity incurred by differing scanner models and manufacturers, imaging sites, variation in clinical standards and protocols, and the noise introduced by the movement of patients' heads during scans. At every clinical visit, glioma patients generally receive standard

N. Nuechterlein and S. Mehta—Equally contributed.

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11384, pp. 245–253, 2019.
https://doi.org/10.1007/978-3-030-11726-9_22

of care FLAIR, post-contrast T1-weighted (T1ce), T2, and T1 MR sequences, each of which is described by a distinct volume. These sequences give distinct and complementary information about the tumor extent and composition.

Automated brain tumor segmentation also ranks among the most difficult problems in medical image analysis. The notion that massive amounts of data are required to train deep networks is widely held. Not only are MR scans scarce, they are high dimensional (e.g. $240 \times 240 \times 155 \times 4$) and contain high class imbalances (e.g. $\geq$95% background class). Thus, naive models are predisposed to exhibit extreme background bias.

In similar biomedical domains, patchwise approaches have helped address problems of data shortages and dimensionality. Ciresan et al. proposed a sliding-window method to segment electron microscopic images of the brain, which both localized the problem and exaggerated the dataset [6,14]. Ronneberger et al.'s 2D encoder-decoder network, U-Net, outperformed Ciresan's method [14]. U-Net is a fully convolutional network (FCN) where the traditional pooling operations in the contracting (encoding) path are mirrored by upsampling operations in the symmetric expanding (decoding) path. Skip connections are passed from encoding blocks on the contracting path to same-level decoding blocks in the expanding path.

While some success has been reached using 2D FCNs, like U-Net, these models ignore crucial 3D spatial context, which is undesirable given that most clinical imaging data are volumetric. However, even among 3D FCNs such as DeepMedic, a previous winner of the BraTS competition, fine spatial information is discarded in pooling [9]. This motivates our interest in U-Net's skip connections and, in particular, the architecture of Milletari et al.'s 3D extension of U-Net, V-Net. V-Net benchmarked well on the "PROMISE2012" challenge, where it gave impressive segmentations of MR prostate scans after training on only 50 examples [13].

ESPNet is a faster, more efficient take on U-Net's encoder-decoder architecture [11]. In this paper, we seek to extend and benchmark ESPNet on 3D medical imaging data.

We outline our paper as follows. Section 2 describes our network architecture. We report our methods in Sect. 3. Experimental results are given in Sect. 4. Finally, we close with a discussion of limitations and future directions for our work in Sect. 5.

2 Network Architecture

Our network is an end-to-end system consisting of 3D-ESPNet followed by pyramidal refinement, as shown in Fig. 1. We describe the main building block of our architecture, the ESP module, and, later, 3D-ESPNet's segmentation architecture and pyramidal refinement.

2.1 ESP Module

The Efficient Spatial Pyramid (ESP) module, shown in Fig. 2, is an efficient convolutional module proposed in [11]. The module is based on the RSTM

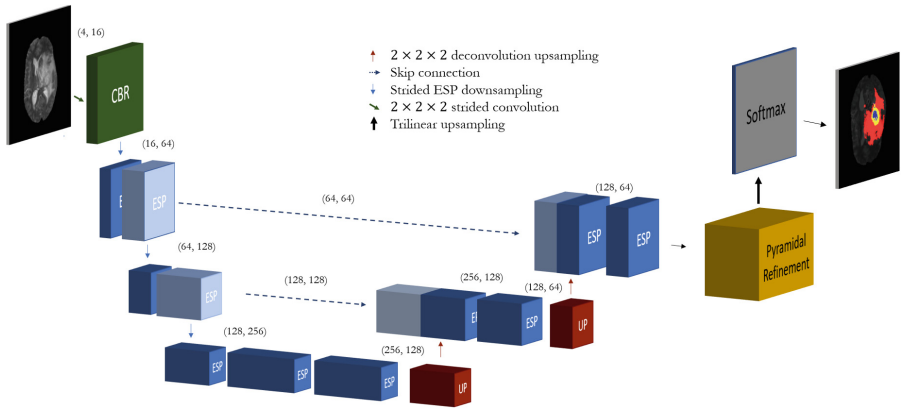


Fig. 1. 3D-ESPNet with pyramidal refinement. 3D-ESPNet's encoder is shown on the left; the decoder is shown on the right with pyramidal refinement. Parentheses give the channel dimensions of incoming and outgoing feature maps. The CBR block consists of a convolutional block followed by batch normalization and ReLU. Light-blue feature maps in the decoder indicate concatenation by long-range, skip connections. Light-blue feature maps in the encoder indicate strided ESP models for downsampling. Arrows are defined in the legend. (Color figure online)

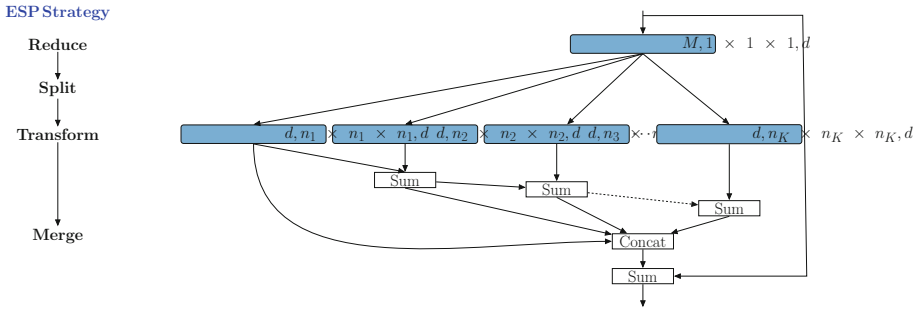


Fig. 2. The Efficient Spatial Pyramid (ESP) module. The blocks in blue represent 3D convolutional layers and are denoted as (# input channels, effective receptive field, # output channels). The ESP module takes an input feature map with M channels and produces an output feature map with N channels, where $d = \frac{N}{K}$ and K represents the number of parallel branches. (Color figure online)

(Reduce-Split-Transform-Merge) strategy and allows the aggregation of the information from a large effective receptive field while learning fewer parameters. We extend the ESP block by replacing its spatial 2D convolutions with volumetric 3D convolutions.

2.2 3D-ESPNet Structure

3D-ESPNet is an encoder-decoder network that extends U-Net [14]. The primary distinction between 3D-ESPNet and U-Net is that 3D-ESPNet employs efficient

convolutional blocks for aggregating features instead of stacking convolution layers (with or without residual connections) after the first layer.

In the encoder stage, the network learns feature representations by performing convolutional and downsampling operations. The encoder downsamples once with a strided convolutional layer and three subsequent times with strided ESP modules. In downsampling ESP modules, we use convolutions with $n_i \times n_i \times n_i$ sized kernels and stride of two, for $i \in \{1, \ldots, K\}$, as shown in Fig. 2. The combination of varying receptive fields allows 3D-ESPNet to learn feature representations at multiple scales.

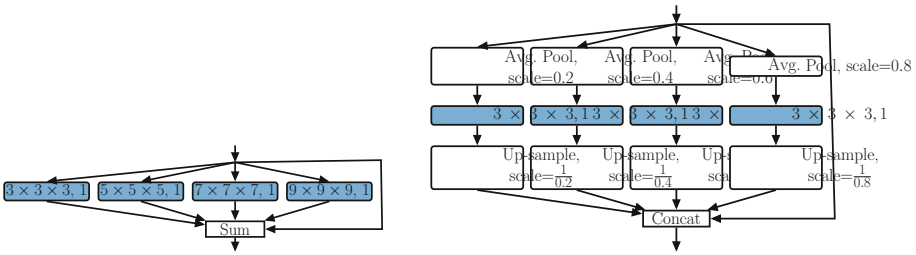
In the decoder stage, we share the feature maps in the encoder with same-level feature maps in the decoder via skip-connection concatenation. Skip-connections allow fine details lost in downsampling in the encoder to be recovered in the decoder, which gives the segmentation maps a granularity simple interpolation cannot achieve. The decoder uses $3 \times 3 \times 3$ deconvolution kernels to upsample the encoder output once, followed by trilinear upsampling layer to return to the resolution at the networks second level. The feature maps of the final ESP module in the decoder are passed into the pyramidal refinement module. The block diagram of 3D-ESPNet is shown in Fig. 1.

Pyramidal Refinement: Pyramid-based approaches sub-sample either the feature maps or the convolutional kernel to learn global contextual information. Inspired by the success of such approaches for segmenting complex 2D scenes, we extend these modules for volumetric data. We call this module pyramidal refinement. Our module combines both feature map-based and convolutional kernel-based pooling methods in a novel fashion.

Pyramidal refinement, shown in Fig. 4, consists of three layers:

- *Projection Layer:* This is a standard $3 \times 3 \times 3$ convolutional layer followed by batch normalization and ReLU that projects the feature maps from the previous ESP block to C-dimensional space, where C is the number of classes.
- *Spatial Pyramid Pooling (SPP) Block*: The input feature maps to this block are low dimensional ($C = 4$). We sub-sample them using convolutional kernels of different sizes and merge their output using sum operations. This is similar to the ASPP block except that we do not use dilated convolutions [5].
- *PSP Block:* A PSP block, sketched in Fig. 3b, is based on the principle of *split-pool-transform-upsample* [15]. *Split:* A PSP block distributes the input feature maps across four parallel branches. *Pool:* Each branch downsamples the feature maps using a different pooling rate. *Transform:* The downsampled feature maps are transformed using point-wise convolutions. *Upsample:* The transformed feature maps are upsampled to the same resolution as the input feature maps using bilinear interpolation. *Merge:* The upsampled feature maps are concatenated with the input feature maps to produce the output feature maps.

Pyramidal refinement is followed by a classification layer. This final layer pools the feature maps using another SPP block and then upsamples by a factor of two using trilinear interpolation. Two convolutional layers are stacked on top of the upsampled feature maps before a softmax.



SPP block. PSP block.

Fig. 3. Pooling modules used in a pyramidal refinement block. Here, a convolutional layer is represented as (kernel size, dilation rate).

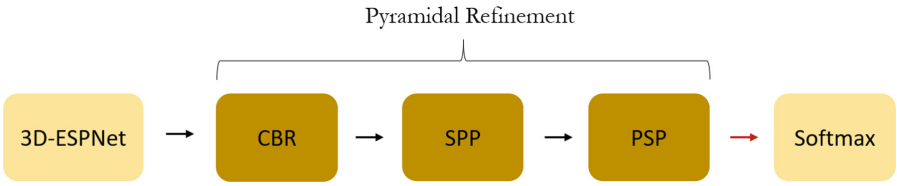


Fig. 4. Pyramidal-refinement. After the second upsampling operation in the 3D-ESPNet decoder, the feature maps are passed through a CBR block, a spatial pyramid pooling block (SPP), and a pyramid pooling module (PSP) at 1/4 resolution. We then upsample to input resolution using trilinear interpolation and compress and pass the feature maps through a softmax to obtain a prediction.

3 Methods

3.1 Data

We train on the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2018 training set, which provides 285 multi-institutional pre-operative multimodal MR tumor scans, each consisting of T1, post-contrast T1-weighted (T1ce), T2, and FLAIR volumes [1–4,12]. Each case is annotated with the following voxel labels: enhancing tumor, peritumoral edema, background, and necrotic core and non-enhancing tumor. Necrotic core and non-enhancing tumor share a single label. These data are co-registered to the standard MNI anatomical template, interpolated to the same resolution, and skull-stripped. Ground-truth segmentations are manually drawn and approved by neuroradiologists.

3.2 Preprocessing

We used minimal preprocessing. We performed min-max normalization. We also cropped each volume to remove any padding around the brain common in every modality; this allowed us to double our batch size to four, which stabilized training.

3.3 Training

To tune our model's hyperparameters, we randomly partitioned our dataset into a training set and a validation set using an 80:20 split (228:57). We selected the hyperparameters that maximized the mean intersection over union (mIOU) on the 57 withheld volumes in the validation set. We used mean intersection over union (mIOU) for our loss function instead of cross entropy for empirical reasons as we and others have observed [8]. We weight our mIOU loss to address the severe class imbalance. We used data augmentation heavily including scaling and random flips.

We implemented our model in PyTorch. We trained at full resolution on all modalities on an NVIDIA Titan X using a batch size of four. We trained for 300 epochs. Training took less than five hours; test time evaluation takes less than twenty seconds. We found that the optimizer Adam outperformed SGD with momentum [10]. We experimented with learning rate decay and settled on a learning rate of $10e^{-4}$, which we decreased to $10e^{-5}$ after 200 epochs. Code for this adaptation of ESPNet is available at https://github.com/sacmehta/3D-ESPNet.

4 Results

Results on the BraTS 2018 online test and validation sets are shown in Table 1. Visual inspection reveals out model's flexible performance on difficult cases such as gliomas that cross the corpus callosum–so-called butterfly gliomas–shown in Figs. 5 and 6. However, our method lacks some of the granularity present in

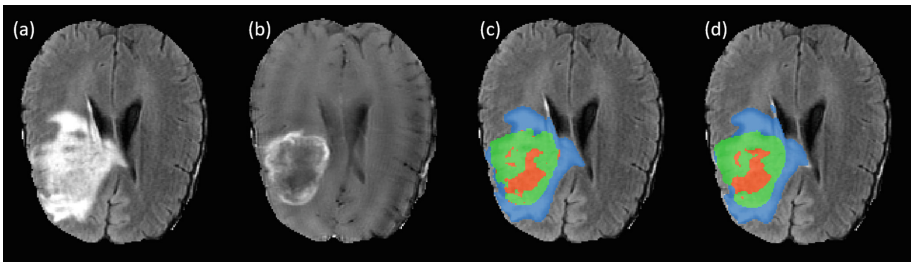


Fig. 5. A butterfly high-grade glioma. (a) FLAIR sequence; (b) T1ce sequence; (c) network prediction; (d) ground truth segmentation.

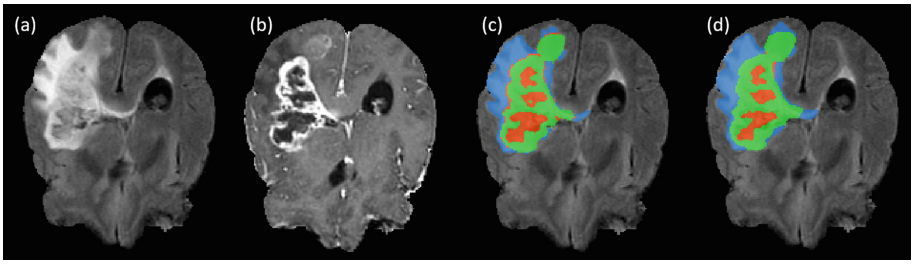


Fig. 6. A second butterfly high-grade glioma. (a) FLAIR sequence; (b) T1ce sequence; (c) network prediction; (d) ground truth segmentation.

the ground truth segmentation. It is clear in the examples provided that our network's predictions are too smooth, especially in Fig. 5, where the predicted non-enhancing and necrotic class is the correct size and in the correct position, but the segmentation does not follow the sharp contours of the gyri outlined in the ground truth. In Fig. 6, we notice that our network tends not to predict necrotic or non-enhancing tumor outside of the tumor-enhancing ring. However, our model is able to handle gaping holes inside tumors filled with cerebrospinal fluid (CSF) just as a resection cavity would appear. This is shown in Fig. 8. These cavities differ from a typical necrotic core on the T2 sequences of a tumor as CSF shows extreme hyperintensity. This robustness is crucial for segmenting post-operative scans which can contain large resection cavities (Figs. 7 and 9).

Table 1. Results obtained on BraTS 2018 online test set are shown in bold. Results obtained on BraTS 2018 online validation set are shown in parenthesis. Sensitivity and specificity results were not given for the online test set.

3D-ESPNet	Dice Score		Sensitivity		Specificity		Hausdorff95	
Whole tumor	**0.850**	(0.883)	-	(0.934)	-	(0.990)	**9.598**	(5.461)
Enhancing tumor	**0.665**	(0.737)	-	(0.831)	-	(0.997)	**5.497**	(5.295)
Tumor core	**0.782**	(0.814)	-	(0.821)	-	(0.997)	**8.668**	(7.850)

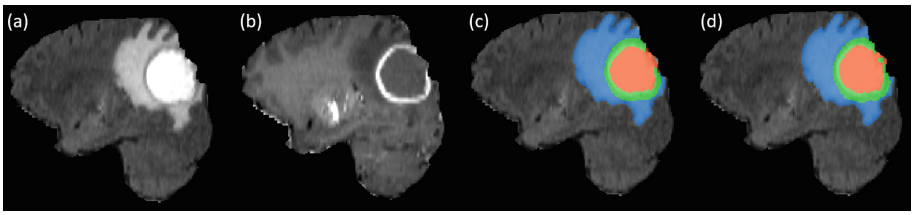


Fig. 7. A sagittal view of a high-grade glioma. (a) FLAIR sequence; (b) T1ce sequence; (c) network prediction; (d) ground truth segmentation.

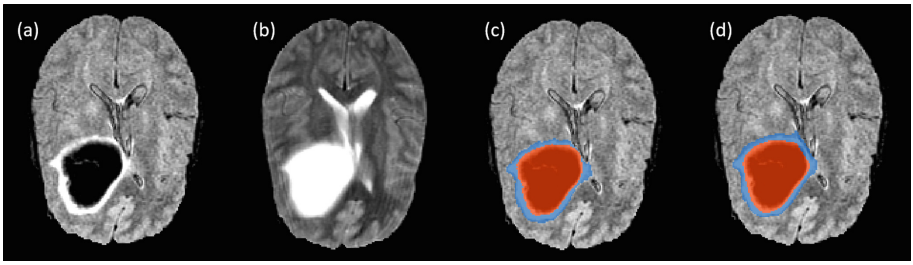


Fig. 8. Low-grade glioma showing bright CSF fluid in ventricles and tumor cavity on the T2 sequence. (a) FLAIR sequences; (b) T2 sequence; (c) network prediction; (d) ground truth segmentation.

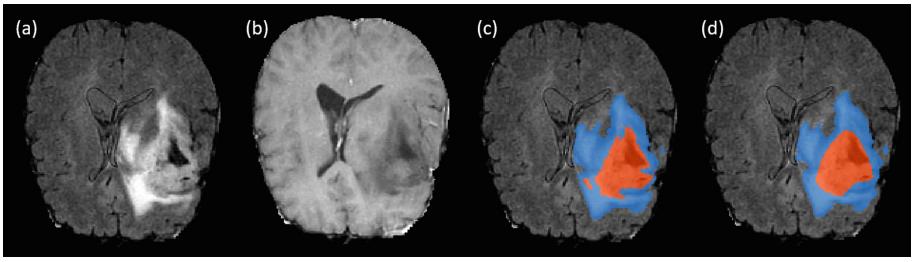


Fig. 9. Low-grade glioma. (a) FLAIR sequence; (b) T1 sequence; (c) network prediction; (d) ground truth segmentation.

5 Discussion

We propose a fast and efficient network for semantic brain tumor segmentation. 3D-ESPNet with pyramidal refinement achieves a respectable 0.850 dice score for whole tumor segmentation on the 2018 BraTS online test set without substantial pre- or post-processing, while learning only 3.8 million parameters.

Brain tumor segmentation has its place in clinic, though neuroradiologist and neuro-oncologists usually limit its use to quantifying volumetric changes in tissue types (edema, enhancing tissue, non-enhancing or necrotic tissue) between patient visits for evaluating tumor progression [7]. However, tumor segmentation is essential to the analysis of advanced MR imaging (DWI, DTI, MRSI). Because such segmentation is usually done manually, segmentation time and cost prevent advanced MR imaging studies from being done at scale. Automatic brain tumor segmentation will allow such advanced imaging studies to be done on massive datasets and, therefore, avail themselves of strong ML analysis and more definitive conclusions.

We plan to add pre- and post-processing techniques to our model. Histogram equalization and N4BiasFieldCorrection might better prepare the training data, and adding a conditional random field after the classifier may help eliminate spurious tumor predictions. We achieved a dice score of 0.850 on the whole

tumor class, but work remains to be done on the individual classes. Better hyperparameter tuning and non-linear data augmentation may also improve our performance.

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
2. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
3. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
4. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
6. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS, pp. 2852–2860 (2012)
7. Fink, J.R., Muzi, M., Peck, M., Krohn, K.A.: Multimodality brain tumor imaging: MR imaging, PET, and PET/MR imaging. J. Nucl. Med. **56**(10), 1554–1561 (2015)
8. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 450–462. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_38
9. Kamnitsas, K., et al.: Deepmedic for brain tumor segmentation. In: BrainLes@MICCAI (2016)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation. arXiv preprint arXiv:1803.06815 (2018)
12. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993 (2015)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
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. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890 (2017)