NAS-SCAM: Neural Architecture Search-Based Spatial and Channel Joint Attention Module for Nuclei Semantic Segmentation and Classification

Zuhao Liu[1], Huan Wang[1], Shaoting Zhang[1,2], Guotai Wang[1(✉)], and Jin Qi[1(✉)]

[1] University of Electronic Science and Technology of China, Chengdu, China
{guotai.wang,jqi}@uestc.edu.cn
[2] SenseTime Research, Shanghai, China

Abstract. The segmentation and classification of different types of nuclei plays an important role in discriminating and diagnosing of the initiation, development, invasion, metastasis and therapeutic response of tumors of various organs. Recently, deep learning method based on attention mechanism has achieved good results in nuclei semantic segmentation. However, the design of attention module architecture relies heavily on the experience of researchers and a large number of experiments. Therefore, in order to avoid this manual design and achieve better performance, we propose a new Neural Architecture Search-based Spatial and Channel joint Attention Module (NAS-SCAM) to obtain better spatial and channel weighting effect. To the best of our knowledge, this is the first time to apply NAS to the attention mechanism. At the same time, we also use synchronous search strategy to search architectures independently for different attention modules in the same network structure. We verify the superiority of our methods over the state-of-the-art attention modules and networks in public dataset of MoNuSAC 2020. We make our code and model available at https://github.com/ZuhaoLiu/NAS-SCAM.

Keywords: Neural architecture search · Nuclei segmentation · Attention mechanism

1 Introduction

The segmentation and classification of different types of nuclei plays a great role in discriminating and diagnosing of the initiation, development, invasion, metastasis and therapeutic response of tumors of various organs [1]. In recent years, deep learning methods are widely applied in nuclei segmentation and classification [2–6]. For example, Kang et al. [4] proposed stacking two U-Nets [7] and Yoo et al. [5] proposed a weakly supervised method for nuclei segmentation.

Z. Liu and H. Wang—Equal contribution.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12261, pp. 263–272, 2020.
https://doi.org/10.1007/978-3-030-59710-8_26

However, as an important step of nuclei instance segmentation, the results of semantic segmentation are greatly impacted by the imbalance of different types of nuclei. Therefore, based on deep neural network, attention module is applied to alleviate class imbalance problem because it can make network automatically enhance important features and suppress unimportant ones.

However, the choice of attention module architecture is based on prior knowledge or limited experiments, which has a lot of disadvantages. Firstly, attention modules with the same architecture may not be applicable to all datasets. However, due to the diversity of optional operations within the attention module, it takes a lot of experimental resources to find the optimal attention module architecture. Secondly, attention modules are often used in different positions of the network. Traditionally, in the same network, these are multiple attention modules with the same architecture. However, due to the influence of convolution and nonlinear operations in network, the spatial and channel information contained in the feature maps of different positions is often different. So, adopting attention modules with different architectures can lead to better use of information in different positions, but this will significantly increase the amount of trail in traditional methods. Therefore, it is time-consuming to try all the experiments to finish attention module selection.

In recent years, with the original intention of simplifying the difficulty of hyper parameter adjustment, Neural Architecture Search (NAS) is widely used in classification [8,9] and semantic segmentation [10]. It can automatically learn the optimal architecture of the network, thus greatly reducing the difficulty of model architecture selection. Therefore, based on NAS, our paper aims to explore the optimal architectures of attention modules. At the same time, we also propose a new searching strategy, i.e., synchronous search strategy, which can search out architectures of different attention modules in the same network.

Our contributions mainly include two folds: (1) We propose a new attention module, i.e., Neural Architecture Search-based Spatial and Channel joint Attention Module (NAS-SCAM), which can efficiently complete the automatic search of architecture to produce better space and channel weighting effect. As far as we know, this is the first application of NAS in the field of attention mechanism, which provides a new development direction for the application of attention module. (2) We propose a synchronous search strategy, which can make our attention module search out different architectures in different positions of the same network, and make the attention module more suitable for certain position of the network, so as to produce a better weighting effect. We have verified our results on the public dataset, MoNuSAC 2020 [1]. Compared with the state-of-the-art attention modules and networks, our method achieves better results in nuclei semantic segmentation and classification.

2 Methods

2.1 NAS-Based Spatial and Channel Joint Attention Module

NAS-SCAM is composed of NAS-based spatial attention module (NAS-SAM) and NAS-based channel attention module (NAS-CAM), which can generate spatial and channel weighting effect, respectively.

The architecture of NAS-SAM is shown in Fig. 1(a). Assuming that the input feature map is $M = [m_1, m_2, ..., m_c]$, which has width W, height H, and channel C. M is transformed into spatial weight map $n \in \mathbb{R}^{H \times W}$ by one or multiple convolutions and nonlinear operations $F_{NS}(.)$ in search space. Through the learning of parameters in search space, n contains the spatial weighting information. Finally, we use the multiplication operation to fuse spatial weight map n into the input feature map M and generate output feature map M' as equation Eq. (1).

$$M' = n \otimes M = [nm_1, nm_2, ..., nm_C] \tag{1}$$

The proper architecture selection of $F_{NS}(.)$ is the key operation and it can make a great difference in the weighting effect. However, because $F_{NS}(.)$ has numerous choices, it is difficult to find the optimal one. So, we propose to select the proper architecture of $F_{NS}(.)$ by using NAS.

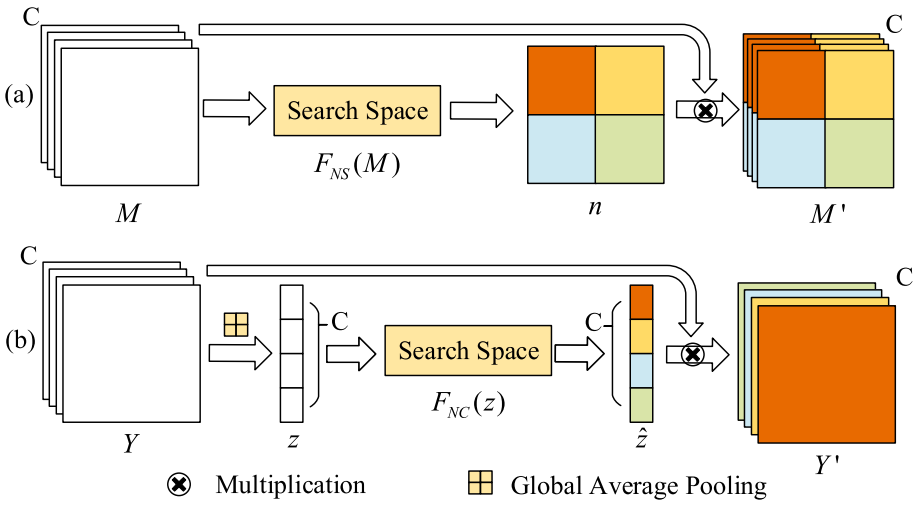

Fig. 1. Illustration of NAS attention modules, (a) is NAS spatial attention module and (b) is NAS channel attention module.

The architecture of NAS-CAM is shown in Fig. 1(b). In order to generate channel weighting effect on the premise of keeping spatial information unchanged. Assuming that the input feature map is $Y = [y_1, y_2, ..., y_c], y_i \in$

$\mathbb{R}^{H\times W}$. We firstly use a global average pooling along the spatial dimension as Eq. (2) and generate vector $z \in \mathbb{R}^{1\times 1\times C}$.

$$z_i = \text{Avgpool}(y_i) = \frac{1}{H \times W} \sum_{p=1}^{H} \sum_{q=1}^{W} y_i(p, q) \tag{2}$$

Then, one or multiple convolutions and nonlinear operations $F_{NC}(.)$ in search space are used to generate channel weight vector $\hat{z} \in \mathbb{R}^{1\times 1\times C}$ which contains channel-wise weighting information. Then, we use NAS to search for the optimal selection of $F_{NC}(.)$. Finally, the output feature map Y' is generated by recalibrating $\hat{z}$ to Y by Eq. (3).

$$Y' = \hat{z} \otimes Y = [\hat{z}_1 y_1, \hat{z}_2 y_2, ..., \hat{z}_C y_C] \tag{3}$$

NAS-SAM and NAS-CAM can be combined in series or in parallel, which are NAS-SCAM-P and NAS-SCAM-S, respectively. For NAS-SCAM-P, the input feature map is weighted along spatial dimension and channel dimension independently, and then element-wise maximum operation is used to fuse two output feature maps to retain more important weighting information. And for NAS-SCAM-S, the input feature map is weighted in the order of spatial dimension followed by channel dimension.

2.2 Search Space

Search space is a series of alternative neural network structures. In order to obtain the appropriate attention module architecture, the network will automatically select the appropriate operations in the search space to achieve the purpose of automatic search. Inspired by literatures [8,10], we define a new search space based on the characteristics of NAS-SAM and NAS-CAM, enabling the network to automatically learn different operations inside the search space.

The architecture of search space is showed in Fig. 2. The input feature map of search space has channel number C. Between each two nodes, there are multiple operations need to be selected and the input shape and output shape of each

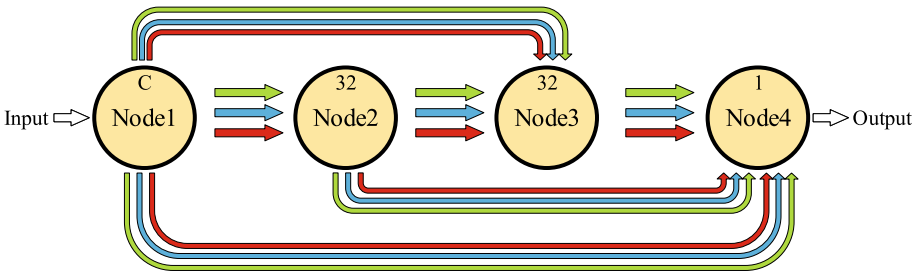

Fig. 2. Architecture of search space.

Table 1. Operations of NAS-SAM and NAS-CAM

NAS-SAM		NAS-CAM	
Zero (No connection)	Conv2D 5	Zero (No connection)	Conv1D 9
Conv2D 1	Atrous Conv2D 3	Conv1D 1	Conv1D 15
Conv2D 3	Atrous Conv2D 5	Conv1D 3	Atrous Conv1D 3
		Conv1D 5	Atrous Conv1D 5

operation are the same except channel dimension. In addition, in order to make attention module have better spatial and channel information learning ability, we set the channel number of Node2 and Node3 to 32. And the number of Node4 is set to 1 to generate weight map along spatial dimension or channel dimension. Every operation in search space is followed by batch normalization to normalize the output value. And activate function following every operation is ReLU except the operations connected to Node4, whose activate function is sigmoid to make output values between 0 and 1.

Because NAS-SAM and NAS-CAM have different architectures, so we choose different operations between each two nodes in their search space. So, the operations of NAS-SAM and NAS-CAM are listed in Table 1. For NAS-SAM, because we need to extract information from the spatial dimension, we use 2D convolutions with different filter sizes to extract information from receptive fields with different sizes. We also use dilated convolutions to increase long-ranged information learning ability. For NAS-SAM, because of the influence of global average pooling, we use 1D convolutions to extract channel information. Furthermore, zero operation is also applied to both NAS-SAM and NAS-CAM to represent no connection between two nodes.

We use continuous relaxation to learn the optimal operation between two nodes. Assuming that operation set between i-th node and j-th node is $O^{(i,j)}(.)$, and $o(.)$ is certain operation in $O^{(i,j)}(.)$. $u_o^{(i,j)}$ is continuous weight coefficient which reflects weight of each $o(.)$ and $x(i)$ is the output of i-th node, $O^{(i,j)}(.)$ is defined as Eq. (4).

$$O^{(i,j)}(x(i)) = \sum_{o \in O^{(i,j)}} u_o^{(i,j)} o(x(i)) \tag{4}$$

Continuous weight coefficient $u_o^{(i,j)}$ is generated from continuous variable $\alpha_o^{(i,j)}$ by softmax function, which is showed in Eq. (5).

$$u_o^{(i,j)} = \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o \in O^{(i,j)}} \exp(\alpha_o^{(i,j)})} \tag{5}$$

Consequently, updating $\alpha_o^{(i,j)}$ through backward propagation can change values of $u_o^{(i,j)}$, and different value of $u_o^{(i,j)}$ represents different importance of each operation. Finally, we choose the operation corresponding to the highest $u_o^{(i,j)}$ as the final operation between i-th node and j-th node.

In both NAS-SAM and NAS-CAM, every node is connected to all previous nodes by summing all output of previous nodes. The output of j-th node $x(j)$ is defined as Eq. (6).

$$x(j) = \sum_{x<j} O^{(x,j)}(x(i)) \tag{6}$$

2.3 Synchronous Search Strategy

In order to search optimal architectures of multiple attention modules in the same network, we propose to use synchronous search strategy to search each attention module independently. Conventionally, after designing an architecture of attention module, the module with the same architecture will be plugged into the end of every down and up sampling block [11,12]. However, the feature maps in different positions of network have large semantic difference because of the convolution and pooling. Therefore, searching different architectures of attention module can make them more suitable for different positions, i.e., down and up sampling blocks in network.

The synchronous search strategy initializes a unique attention module for each down and up sampling block and optimizes them independently. Because the adjustment of attention module architecture is through optimizing continuous variable α, different α in different attention module will have different gradient in the optimization process, so as to be optimized towards different direction and generate more suitable architecture for certain position.

3 Experiments

3.1 Dataset Description

Multi-organ nuclei segmentation and classification dataset in MoNuSAC 2020 is used to validate the performance of the proposed NAS attention module. This dataset contains 209 annotated H&E stained histopathology images and four types of nuclei including epithelial, lymphocytes, macrophages, and neutrophils. We randomly select 120 images as training set, 39 images as validation set and 50 images as testing set. We use Dice similarity coefficient (DSC) as evaluation metric which can calculate the similarity between prediction and ground truth.

3.2 Implementation Details

In order to augment the training data, firstly, we use overlapping crop to generate more training data, and the cropped image size is set to 256×256 and the total number of cropped images is 4803. Secondly, we use a series of augmentation methods including random rotate, random flip, Gaussian blur, median blur and elastic transformation. To normalize the data, we use standard color normalization to preprocess dataset. In addition, it is noticeable that each class containing the same amount of data in each training batch can potentially alleviate the class imbalance problem and achieve better results.

Table 2. DSC scores (%) of different approaches on multi-organ nuclei dataset

Methods	Average	Epithelial	Lymphocyte	Macrophage	Neutrophil
U-Net [7]	56.77 ± 3.40	77.58 ± 3.08	75.40 ± 2.95	27.35 ± 3.46	46.75 ± 4.86
Deeplab V3+ [14]	62.37 ± 3.03	78.91 ± 0.45	76.46 ± 0.85	33.61 ± 11.52	60.53 ± 0.98
CBAM [11]	59.40 ± 3.51	77.51 ± 2.14	77.11 ± 0.88	25.93 ± 6.09	57.07 ± 6.65
scSE [12]	56.86 ± 4.27	79.48 ± 0.46	77.59 ± 0.45	27.94 ± 8.30	42.44 ± 9.55
NAS-SCAM **without** synchronous search strategy					
NAS-SCAM-S	61.22 ± 3.19	78.30 ± 0.79	76.68 ± 0.15	31.39 ± 6.92	58.52 ± 5.45
NAS-SCAM-P	62.80 ± 1.77	79.58 ± 0.36	77.62 ± 0.33	37.75 ± 3.84	56.27 ± 8.44
NAS-SCAM **with** synchronous search strategy					
NAS-SCAM-S	62.47 ± 1.23	78.55 ± 0.78	76.64 ± 1.76	35.42 ± 6.53	59.28 ± 0.15
NAS-SCAM-P	65.01 ± 1.07	80.67 ± 0.66	77.43 ± 1.94	40.85 ± 6.06	61.10 ± 2.72

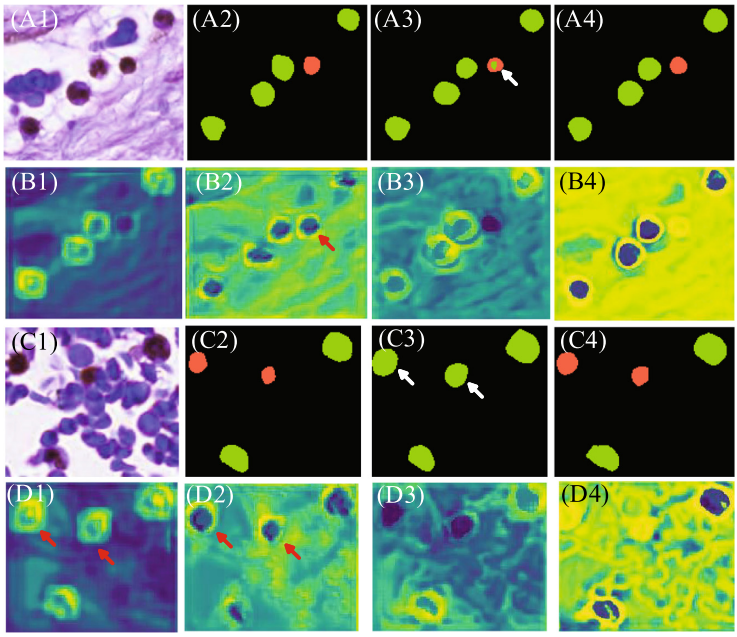

Fig. 3. Visualization of segmentation results. A1 and C1 are original images; A2 and C2 are ground truth; A3 and C3 are segmentation results of CBAM; A4 and C4 are segmentation results of NAS-SCAM-P; B1-2 and D1-2 are attention maps in CBAM; B3-4 and D3-4 are attention maps in NAS-SCAM-P. White arrows highlight the wrong classification in segmentation results and red arrows highlight misjudgment of attention maps.

We implemented our experiments in Python and machine learning framework Pytorch, and two Nvidia RTX 2080 GPU are used. Exponential logarithmic loss described in literature [13] and Adam optimizer are used in optimization process.

The baseline of the networks is U-Net which contains four down sampling and four up sampling blocks. For the networks with attention modules, attention module is plugged into the end of each down or up sampling block.

We use the first-order approximation of bilevel optimizer described in literature [8] to optimize our network. In searching process, the total epoch number is 120, and we only update network parameters in the first 40 epochs. The update of w and update of continuous variable α are implemented alternatively after first 40 epochs. The learning rate is 0.0001 when updating w and 0.001 when updating α. After learned an optimized architecture, we rebuild network based on learned α. In rebuilding process, the total epoch number is 300, and learning rate is 0.0001. We saved the model that performed best in validation set as the model in testing set. All experiments are implemented three times, and mean and deviation of DSC of each class are calculated in each experiment.

3.3 Experimental Results

In order to verify the effectiveness of our proposed NAS-SCAM and synchronous search strategy. Firstly, we compare the performances of NAS-SCAM with or without synchronous search strategy. Secondly, we compare the performances of NAS-SCAM with two state-of-the-art attention modules, convolutional block attention module (CBAM) [11] and spatial-channel squeeze & excitation (scSE) module [12]. Thirdly, we compared the performance of NAS-SCAM with two state-of-the-art networks: U-Net [7] and Deeplab V3+ [14]. Results are showed in Table 2.

From Table 2, The effectiveness of synchronous search strategy can be verified by the comparison between NAS-SCAM with this strategy and that without this strategy. For NAS-SCAM-S, synchronous search strategy can increase average DSC from 61.22% to 62.47%. For NAS-SCAM-P, synchronous search strategy has greater influence on results which increases average DSC from 62.80% to 65.01%. Results prove that searching more suitable architecture for each attention module can achieve better results than searching a unique architecture for all attention modules.

In addition, NAS-SCAM can achieve better results than existing attention modules and state-of-the-art networks. The average DSC of NAS-SCAM-S and NAS-SCAM-P are 62.47% and 65.01%, which are higher than CBAM and scSE whose average DSC are 59.40% and 56.86%. And the result of NAS-SCAM is also better than that of Deeplab V3+ and U-Net, whose average DSC are 62.37% and 56.77%, respectively. Moreover, NAS-SCAM-P can achieve the best results for all type of nuclei except lymphocyte. The results show the effectiveness of NAS-SCAM compared with state-of-the-art attention modules and networks.

3.4 Visualization

From Fig. 3, we visualize the segmentation results and spatial attention maps generated by the last two attention blocks in network from CBAM and NAS-SCAM-P. The red mask is lymphocyte and green mask is neutrophil. It is noted

that compared with NAS-SCAM-P, CBAM generates wrong classification results in A3 and C3 as white arrows point out. The misjudgment of attention maps has direct relationship with this result, which is showed in B2 and D1-2, where attention maps generated from CBAM give equal weights to two classes as red arrows point out. But in B3-4 and D3-4, attention maps generated from NAS-SCAM-P give different weights to different classes, so as to generate better results.

4 Conclusion

In this paper, we propose new attention module, NAS-SCAM, which is the first application of NAS in attention mechanism. This provides a new direction for the development of attention mechanism. We also propose a new search strategy, synchronous search strategy, which can make searched architecture of attention module better fit to the network. Our proposed methods can generate better spatial and channel weighting effect which is beneficial for network to distinguish different types of nuclei, so as to achieve better results in nuclei semantic segmentation and classification and can be better applied in clinical diagnosis.

Future work will focus on improving the search strategy and increasing the search space. We are dedicated to add channel number and activation function selection into our search space, and find more effective attention module.

Acknowledgements. This work is supported by Sichuan Jiuzhou electric Group Co. Ltd, Sichuan, Mianyang, 621000, China, and National Natural Science Foundation of China under grant no. 81771921, and Glasgow College, University of Electronic Science and Technology of China.

References

1. Verma, R., Kumar, N., Patil, A., et al.: Multi-organ Nuclei Segmentation and Classification Challenge (2020, unpublished). https://doi.org/10.13140/RG.2.2.12290.02244/1
2. Su, H., Shi, X., Cai, J., Yang, L.: Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 559–567. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_62
3. Saha, M., Chakraborty, C.: Her2Net: a deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. TIP **27**, 2189–2200 (2018)
4. Kang, Q., Lao, Q., Fevens, T.: Nuclei segmentation in histopathological images using two-stage learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 703–711. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_78
5. Yoo, I., Yoo, D., Paeng, K.: PseudoEdgeNet: nuclei segmentation only with point annotations. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 731–739. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_81

6. Qu, H., Yan, Z., Riedlinger, G.M., De, S., Metaxas, D.N.: Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 378–386. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_42
7. 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
8. Liu, H., Simonyan, K., Yang, Y.: Darts: differentiable architecture search. In: International Conference on Learning Representations (ICLR) (2019)
9. Real, E., Aggarwal, A. et al.: Regularized evolution for image classifier architecture search. In: AAAI, vol. 33, no. 01 (2019)
10. Liu, C. et al.: Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation. In: CVPR, pp. 82–92 (2019)
11. Woo, S., Park, J., Lee, J.Y., Kweon, I.N.: CBAM: convolutional block attention module. In: ECCV, pp. 3–19 (2018)
12. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_48
13. Wong, K.C.L., Moradi, M., Tang, H., Syeda-Mahmood, T.: 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 612–619. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_70
14. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, pp. 801–818 (2018)