electronics

Article

A Self-Spatial Adaptive Weighting Based U-Net for Image Segmentation

Choongsang Cho [1,†], Young Han Lee [1,*,†], Jongyoul Park [2] and Sangkeun Lee [3,†]

[1] Korea Electronics Technology Institute, Seongnam 13488, Korea; ideafisher@keti.re.kr
[2] Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul 01811, Korea; jongyoul@seoultech.ac.kr
[3] Smart Vision Global, Seoul 02024, Korea; sangkny@gmail.com
* Correspondence: yhlee@keti.re.kr; Tel.: +82-31-739-7458
† These authors contributed equally to this work.

Abstract: Semantic image segmentation has a wide range of applications. When it comes to medical image segmentation, its accuracy is even more important than those of other areas because the performance gives useful information directly applicable to disease diagnosis, surgical planning, and history monitoring. The state-of-the-art models in medical image segmentation are variants of encoder-decoder architecture, which is called U-Net. To effectively reflect the spatial features in feature maps in encoder-decoder architecture, we propose a spatially adaptive weighting scheme for medical image segmentation. Specifically, the spatial feature is estimated from the feature maps, and the learned weighting parameters are obtained from the computed map, since segmentation results are predicted from the feature map through a convolutional layer. Especially in the proposed networks, the convolutional block for extracting the feature map is replaced with the widely used convolutional frameworks: VGG, ResNet, and Bottleneck Resent structures. In addition, a bilinear up-sampling method replaces the up-convolutional layer to increase the resolution of the feature map. For the performance evaluation of the proposed architecture, we used three data sets covering different medical imaging modalities. Experimental results show that the network with the proposed self-spatial adaptive weighting block based on the ResNet framework gave the highest IoU and DICE scores in the three tasks compared to other methods. In particular, the segmentation network combining the proposed self-spatially adaptive block and ResNet framework recorded the highest 3.01% and 2.89% improvements in IoU and DICE scores, respectively, in the Nerve data set. Therefore, we believe that the proposed scheme can be a useful tool for image segmentation tasks based on the encoder-decoder architecture.

Keywords: deep learning; self-spatial weighting; adaptive weighting; medical image segmentation

check for updates

Citation: Cho, C.; Lee, Y.H.; Park, J.; Lee, S. A Self-Spatial Adaptive Weighting Based U-Net for Image Segmentation. *Electronics* **2021**, *10*, 348. https://doi.org/10.3390/electronics10030348

Academic Editor: Giovanni Dimauro
Received: 22 December 2020
Accepted: 25 January 2021
Published: 2 February 2021

1. Introduction

Over the past few years, deep convolutional neural networks have made a lot of progress in computer vision-based tasks, including image classification [1,2], object detection [3,4], semantic segmentation [5,6], human pose estimation [7,8], image captioning [9,10], and so on.

Semantic image segmentation has a wide range of applications in the fields of computer vision, robotics, medical, and computer graphics. Image segmentation in natural images is used to parse the scene, and its performance has improved so that it can be applicable to automatic driving and robot sensing, to name a few [6,11]. When it comes to medical image segmentation, accuracy is even more important than other areas because the result gives important information for disease diagnosis, surgical planning, and history monitoring [12].

State-of-the-art scene segmentation frameworks for natural images are based on the fully convolutional network (FCN) [13], and the state-of-the-art models for medical image segmentation are variants of the encoder-decoder architecture called U-Net [14,15]. Encoder-decoder networks for segmentation use a similar structure: Skip connections, and coarse-grained feature maps. The skip connection-based scheme has been used in many successful image segmentation [14,16] and classification [17] methods. An attention framework was used to highlight salient features that stand out in many computer vision tasks, including segmentation, to take into account the nature of the task in the feature maps [18–20]. Considering the goal of segmentation, which assigns a category label to each pixel in the image, the segmentation result is obtained from the last feature map via the convolutional layer, so the feature maps in the encoder-decoder architecture should reflect the spatial characteristics of the task.

In encoder-decoder architecture, we propose a spatial adaptive weighting method for encoder-decoder architecture to reflect the spatial characteristics of feature maps. Since the segmentation result is predicted from the feature map through the convolutional layer [11,14,18], we estimate the spatial characteristics from the feature map and get the weighting parameters learned from the computed map. The weighting parameters are multiplied and added to the feature maps of the architecture.

We propose a self-spatial adaptive weighting scheme in a U-Net architecture (SS-U-Net) and apply it to medical images. The convolution block for extracting feature maps from the proposed network is replaced by the widely used convolution frameworks VGG, ResNet, and Bottleneck Resent structures. The up-convolution layer to increase the feature map resolution is replaced by the bilinear up-sampling method. To evaluate the proposed scheme, we use three sets of medical imaging data to include different medical imaging modalities: microscopy and ultrasound. Our experiments show that the proposed method has the smallest model size compared to the standard U-Net, while improving performance on three data sets. In particular, the model with the bottleneck structure, U-Net(B), has the smallest size among the compared methods, and is only about 60% the size of a standard U-Net.

2. Related Work

For natural image segmentation, the fully convolutional network (FCN) was first introduced by Long et al. [13]. This approach estimates a coarse segmentation map for each fully connected layer and improves the map by combining the fine segmentation score maps. The pyramid scene parsing network (PSPnet) based on FCN was proposed by Zhao et al. [5,6]. Since the FCN method's receptive field is not sufficient for complex scene images, the fusion information of these receptive fields and other sub-areas is calculated by the pyramid pooling structure and used as global prior to segmentation. To consider the context aggregation problem in a semantic segmentation scheme, an object-contextual representation method that characterized the pixel by representing the corresponding object class was proposed by Yuan et al. [21]. A method to maintain high-resolution representations throughout the entire process was proposed by Wang et al. [22]. To maintain high-resolution representation, it proposed the high-to-low resolution convolution streams and fused the representations from multi-resolution streams. A hierarchical attention mechanism for image segmentation was proposed to predict relative weights between adjacent scales and combine multiscale predictions at the pixel level [23].

U-Net [14], an encoder-decoder architecture based on the FCN, has been used in state-of-the-art models for medical image segmentation methods. It has symmetric architecture, and the feature map of the encoder is transmitted to the decoder side through a skip-connection. Then, that feature map is concatenated with the up-sampled feature map in the decoder path of the next convolutional layer. In order to highlight salient features of the network, attention architecture has been applied to the U-Net structure in the study [18]. For 3D structure medical images, H-dense U-Net based on the architecture of DenseNet [1] was proposed for liver and liver tumor segmentation by Li et al. [15]. To reduce the semantic difference between the feature maps of the encoder and decoder sub-networks, a skip pathway method was proposed in the U-Net++ [11].

3. Our Approach

3.1. Image Segmentation Problem

Image segmentation can be interpreted as an optimization problem to find a segmented image U in a given image V [12]. Thus, the given image is categorized into a set of optimized classes and the classes, G, are defined by

$$G = \{g_i \in \mathbb{R} : i = 1, \dots, N_c\}, \tag{1}$$

where N_c is the number of predefined classes. $\mathbb{R}$ and g_i denote real and i-th class values, respectively.

In previous studies, before the deep-learning method emerged, minimization of the cost function for image segmentation was used to solve the optimization problem based on the Mumford–Shah function [12,24,25]. After major advances in computer vision technology based on deep convolutional networks, the problem has been solved using deep learning and large amounts of labeled data sets in many studies [13,14].

3.2. Self-Spatial Adaptive Weighting

In the image generation task based on Generative Adversarial Nets (GAN), a semantic segmentation mask was used as a condition for adjusting the appearance of images generated by image generation [26–29]. The given semantic segmentation mask is also used as a conditional guide for their normalization, and it improved the performance of image generation in previous research [29].

To consider a spatial weighting method in the image segmentation task without the given mask, we propose a spatial weighting scheme for image segmentation called self-spatial (SS) adaptive weighting, as shown in Figure 1.

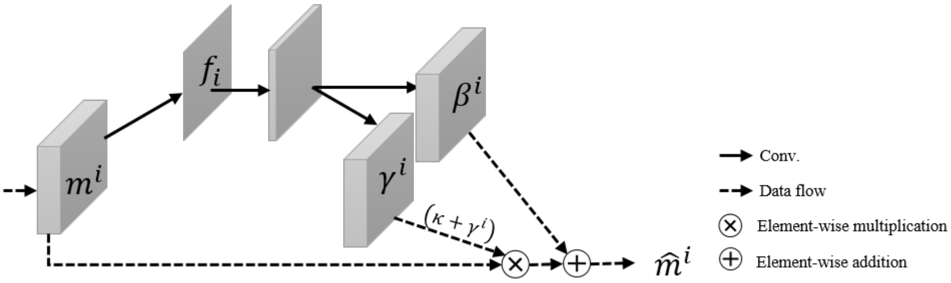


Figure 1. The proposed self-spatial adaptive weighting structure (SS block).

Let $m^{C^i \times H^i \times W^i}$ be the output feature map of the convolution block, and m^i denotes the feature map for the i-th convolution block in a segmentation network. Here, C^i is the number of channels in the convolution block, and H^i and W^i represent the height and width of those feature maps, respectively. In the i-th block, the spatial characteristics are estimated from the feature map of each convolution block site as

$$f^i = \nu(m^i). \tag{2}$$

A function $\nu(\cdot)$ is implemented by a single convolutional layer that converts m_i to f^i. That is, by this function, the feature map, m_i, is turned into a spatial feature, f^i, which has the number of division classes.

To consider the spatial characteristics for the segmentation, the spatial weighting parameter is obtained from the map computed according to Equation (3).

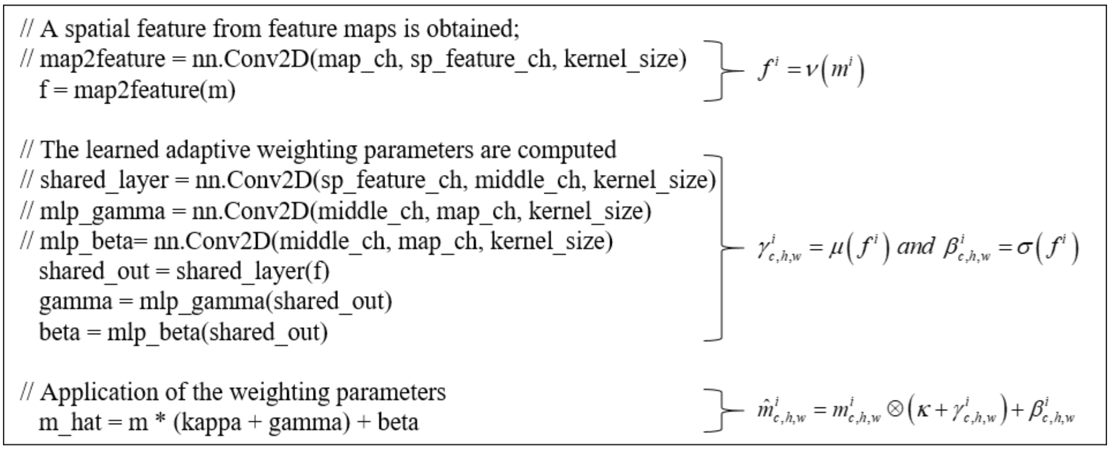
$$\gamma^i_{c,h,w} = \mu(f^i), \qquad \beta^i_{c,h,w} = \sigma(f^i), \tag{3}$$

where $\mu(\cdot), \sigma(\cdot)$ represent functions that convert f_i into the learned adaptive weighting parameters, $\gamma^i_{c,h,w}$ and $\beta^i_{c,h,w}$, respectively. The spatial weighting parameters, γ and β, are multiplied and added to the feature map of the i-th convolution blocks element by element, as

$$\hat{m}^{i}_{c,h,w} = m^{i}_{c,h,w} \otimes \left(\kappa + \gamma^{i}_{c,h,w}\right) + \beta^{i}_{c,h,w}. \tag{4}$$

The variables $\gamma^{i}_{c,h,w}$ and $\beta^{i}_{c,h,w}$ are the learned weighting parameters depending on the spatial map at the site $\left(c \in C^{i}, h \in H^{i}, w \in W^{i}\right)$, and κ is a predefined constant value.

The learned weight parameters, $\gamma^{i}_{c,h,w}$ and $\beta^{i}_{c,h,w}$, are calculated from the same spatial feature f^{i}, and the learning-based parameter computation is implemented using a two-layer convolutional network, as shown in Figure 2.



```
// A spatial feature from feature maps is obtained;
// map2feature = nn.Conv2D(map_ch, sp_feature_ch, kernel_size)
   f = map2feature(m)

// The learned adaptive weighting parameters are computed
// shared_layer = nn.Conv2D(sp_feature_ch, middle_ch, kernel_size)
// mlp_gamma = nn.Conv2D(middle_ch, map_ch, kernel_size)
// mlp_beta= nn.Conv2D(middle_ch, map_ch, kernel_size)
   shared_out = shared_layer(f)
   gamma = mlp_gamma(shared_out)
   beta = mlp_beta(shared_out)

// Application of the weighting parameters
   m_hat = m * (kappa + gamma) + beta
```

$$f^{i} = \nu\left(m^{i}\right)$$

$$\gamma^{i}_{c,h,w} = \mu\left(f^{i}\right) \text{ and } \beta^{i}_{c,h,w} = \sigma\left(f^{i}\right)$$

$$\hat{m}^{i}_{c,h,w} = m^{i}_{c,h,w} \otimes \left(\kappa + \gamma^{i}_{c,h,w}\right) + \beta^{i}_{c,h,w}$$

Figure 2. A pseudo-code of the proposed self-spatial adaptive weighing structure.

3.3. *Self-Spatial Adaptive Weighting-Based U-Net Structure for Image Segmentation*

In medical image segmentation, U-Net architecture consisting of convolution blocks, skip connection paths, and up-convolution steps has been widely used.

The proposed weighting scheme is integrated into the standard U-Net architecture to apply adaptive weights based on spatial characteristics to the feature map that is passed to the next convolutional block via downsampling and upsampling methods. In order to prevent an increase in the model size by applying the proposed technique, the up-convolution layer of the standard U-Net was implemented by using a bilinear upsampling method, which results in reducing the model size of the U-Net, as shown in Figure 3.

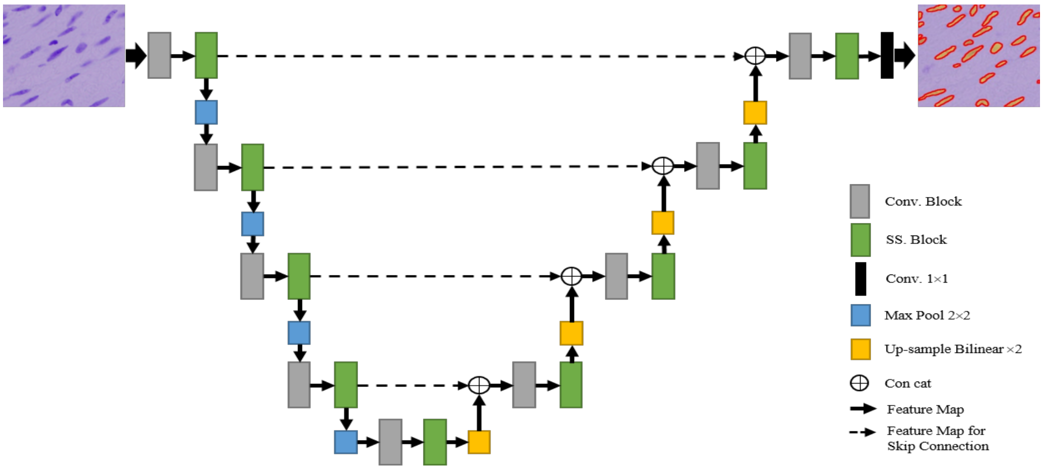


Figure 3. An overall architecture of the proposed self-spatial adaptive weighing-based U-Net (SS-U-Net) for image segmentation.

The proposed self-spatial adaptive weighting-based U-Net, SS-U-Net, is composed of three main blocks: a convolution block, a self-spatial adaptive weighting block, and an up/down sampling block. The feature map of the segmentation network is extracted from the convolutional block and weighted by the learned adaptive scales and biases calculated from the spatial features of the map in the SS block. In the encoding path of SS-U-Net, the weighted features transmitted through the skip connections and their resolutions are reduced in the down-sampling block implemented by max-pooling operations. On the other hand, the resolution of weighted features is increased by a bilinear upsampling method. The upsampled features and the ones passed through the skip connections are concatenated and propagated to the next convolutional block in the decoding path of the proposed network.

Deep convolutional neural networks for image classification have had a breakthrough method [17,30]. The VGG structure provides better performance with low complexity using a 3 × 3 convolutional kernel instead of a larger kernel, such as 5 × 5, or 7 × 7, and has a structure similar to that of a standard U-Net, but maintains the same spatial resolution at the input and output. The ResNet proposed skip-connection in depth so that the network stacked more layers compared to other networks. In the Bottleneck structure, the 1 × 1 convolutional layers were used to reduce the number of channels in the convolution block. Thus, it avoids increasing the complexity of the ResNet framework. These structures can be represented as shown in Figure 4. These convolutional frameworks have a similar purpose to the convolutional block of segmentation. In the proposed scheme, a widely used framework is used for the proposed structure, and in particular, a bottleneck structure is used to build a small model for segmentation. In this paper, a modified U-Net with bilinear upsampling is defined as U-Net $left(cdot right)$, and the framework used is indicated in parentheses.

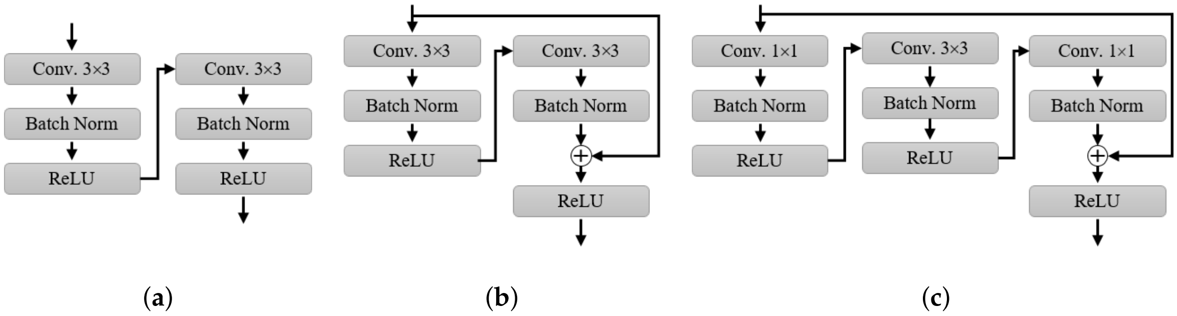


Figure 4. Convolution blocks applied to the proposed methods. From left to right: (**a**–**c**) are the VGG, ResNet, and Bottleneck structures, respectively.

4. Experimental Results

4.1. Datasets

As can be seen in Table 1, we cover a variety of medical imaging modalities using three medical imaging data sets for model evaluation. The first data set was obtained under a microscope to segment the cell area. The data set from the Data Science Bowl 2018 segmentation challenge consists of nuclei images from different modalities (brightfield and fluorescence) [11]. The other two data sets segmenting the fetal head consisted of 999 samples with no growth abnormalities [31], and the nerve regions were from an ultrasound imaging equipment. Given the resolution of the smallest image in the evaluation data sets, each image was scaled to 256 × 256 for our implementation. For the training, validation, and test sets, we split the data sets into training (80%), validation (10%), and test (10%) sets.

To evaluate the performance of our segmentation model, we calculate the intersection over Union (IoU), also known as the Jaccard index, which measures the area of intersection between the predicted segmentation and the ground truth divided by the area of union between them. In addition, we employ the Dice Coefficient, $F1$ score, which evaluates the value of 2× areas of intersection divided by the total number of pixels in the two images [32].

Table 1. Experimental data sets for image segmentation.

Data Set	# of Images	Input Size	Modality	Provider
Cell Nuclei [33]	670	320 × 256	Microscopy	Data Science Bowl competition
Fetal Head [31]	999	800 × 540	Ultrasound	HC18 Grand Challenge
Nerve [34]	5635	580 × 420	Ultrasound	Ultrasound Nerve competition Kaggle

4.2. Training Setup

We implemented the networks using Pytorch [35], an open-source machine learning library for Python. The Adam optimizer [36] was used to train network weights and biases using 400 epochs with an initial learning rate of 0.001 and batch size of 16. For the proposed method, κ was set to 1, and the input resolution of the segmentation networks was set to 256 × 256, taking into account the image resolution of the data set.

4.3. Performance Comparison

In order to evaluate the effect of the self-spatial adaptive weighting method and convolutional blocks in U-Net network, the three kinds of convolution blocks were individually applied to the proposed structure, and combined with the proposed SS block.

Table 2 compares the segmentation methods in terms of the model size and segmentation results that were measured by the IoU and DICE scores, respectively, for the three segmentation tasks. The model with the Bottleneck structure, U-Net(B), gives the smallest model size, and is about 60% of the size of standard U-Nets. In a similar manner, the model with the ResNet framework for convolutional blocks, U-Net(R), is about 95% of the standard size. Applying the SS block to a standard U-Net increases the U-Net's model size by 1.27 MB. Among the convolutional frameworks selected for the proposed method, we evaluate SS-U-Net(R), a ResNet convolutional block-based scheme with the best performance in the three segmentation tasks. In addition, the proposed method, SS-U-Net(R), produces the smallest model size, while the U-Net++ gives the largest model.

Table 2. Performance comparison for image segmentation using various convolution blocks with the SS scheme. The "B", "V", "R", and "SS" represent the Bottleneck structure, VGG blocks, Residual block, and Self-Spatial normalization, respectively. Intersection over union (IoU) and the Dice coefficient are used in terms of comparison metrics (%). Numbers in bold indicate the highest performance in each metric.

Method	Model Size	Cell Nuclei		Fetal Head		Nerve	
	Param (MB)	IoU	DICE	IoU	DICE	IoU	DICE
U-Net(B)	20.01	86.24	91.59	94.12	96.49	65.55	76.72
SS-U-Net(B)	21.29	86.38	91.65	95.26	97.24	68.56	79.61
U-Net(V)	29.96	86.05	91.34	95.26	97.24	68.32	79.20
SS-U-Net(V)	31.24	86.36	91.68	95.24	97.26	68.84	79.90
U-Net(R)	31.62	85.91	91.38	95.22	97.23	68.01	79.17
SS-U-Net(R)	32.89	**86.58**	**91.84**	**95.48**	**97.41**	**69.16**	**80.14**

As can be seen in Table 2, the network with the Bottleneck framework, U-Net(B), has the lowest IoU and DICE scores in the fetal head and nerve segmentation tasks, and the network with the VGG block, U-Net(V), has the lowest IoU and Dice scores in the cell segmentation. As the proposed SS block is applied to the segmentation, segmentation performance is improved in all three types of convolutional blocks. In particular, the segmentation network using the proposed SS block and Bottleneck framework improved the IoU and DICE scores by 3.01% and 2.89%, respectively. SS-U-Net(R), a network with the proposed SS block and ResNet framework, achieved the highest IoU and Dice scores in three tasks compared to other combinations. The network using the ResNet framework improved the segmentation performance by about 1.15% and 0.97% in the IoU and DICE scores, respectively, in the nerve segmentation.

The three task images are segmented by the network combined with the proposed SS block and a kind of convolutional framework, as shown in Figure 5. Figure 5a–i are the results of the first, second, and third tasks, respectively. The yellow region in the Figure indicates the ground truth, and the solid red line illustrates the contour of the results. It can be easily seen that the proposed networks, SS-U-Net(B), SS-U-Net(V), and SS-U-Net(R) have a closer shape to the ground truth compared to the other network without the SS Block.

The proposed method for evaluating the segmentation performance is compared with standard U-Net [14], attention U-Net [18], U-Net++ [11], and customized wide U-Net architectures, as the authors did in [11]. Wide-U-Net, a model extended from the standard U-Net, has a model size similar to the largest model among the compared networks.

Table 3 lists the experimental results and shows the effectiveness of the proposed scheme. The compared methods scored very high in fetal head segmentation, but the lowest performance in nerve segmentation. The proposed method in the three tasks scored the highest in both IoU and DICE. The U-Net with SS block, Att U-Net, and U-Net++ each had the second-highest performance in the three tasks. In addition, the performance of SS-U-Net was improved in all tasks compared to standard U-Net. The proposed method improved performance compared to the standard U-Net in three tasks, and in particular, it has a smaller model size.

Table 3. Performance comparison for medical image segmentation with various U-Net mechanisms. The Intersection over union (IoU) and Dice coefficients are employed as comparison metrics (%). Numbers in bold indicate the highest performance in each metric.

Segmentation Method	**Model Size**	**Cell Nuclei**		**Fetal Head**		**Nerve**	
	Param (MB)	**IoU**	**DICE**	**IoU**	**DICE**	**IoU**	**DICE**
U-Net [14]	32.95	86.09	91.39	95.31	97.29	68.35	79.34
U-Net + SS	34.22	86.14	91.63	95.37	97.33	68.90	79.73
Wide U-Net [11]	34.85	86.10	91.51	95.13	97.21	68.94	79.75
U-Net++ [11]	34.96	85.83	91.58	95.41	97.30	67.87	79.85
Att U-Net [18]	33.63	85.83	91.49	95.36	97.35	68.53	79.31
U-Net(R) + SS(our)	32.89	**86.58**	**91.84**	**95.48**	**97.41**	**69.14**	**80.14**

The three task images are segmented by standard U-Net, attention U-Net, U-Net++, wide-U-Net, and the proposed approaches, as shown in Figure 6. Figure 6a–i are the results of the first, second, and third tasks, respectively. The yellow area in the Figure illustrates the ground truth, and the solid red line represents the contour of the results. It can be easily seen that the result of the network with the proposed scheme has better segmentation performance than those of the compared methods.

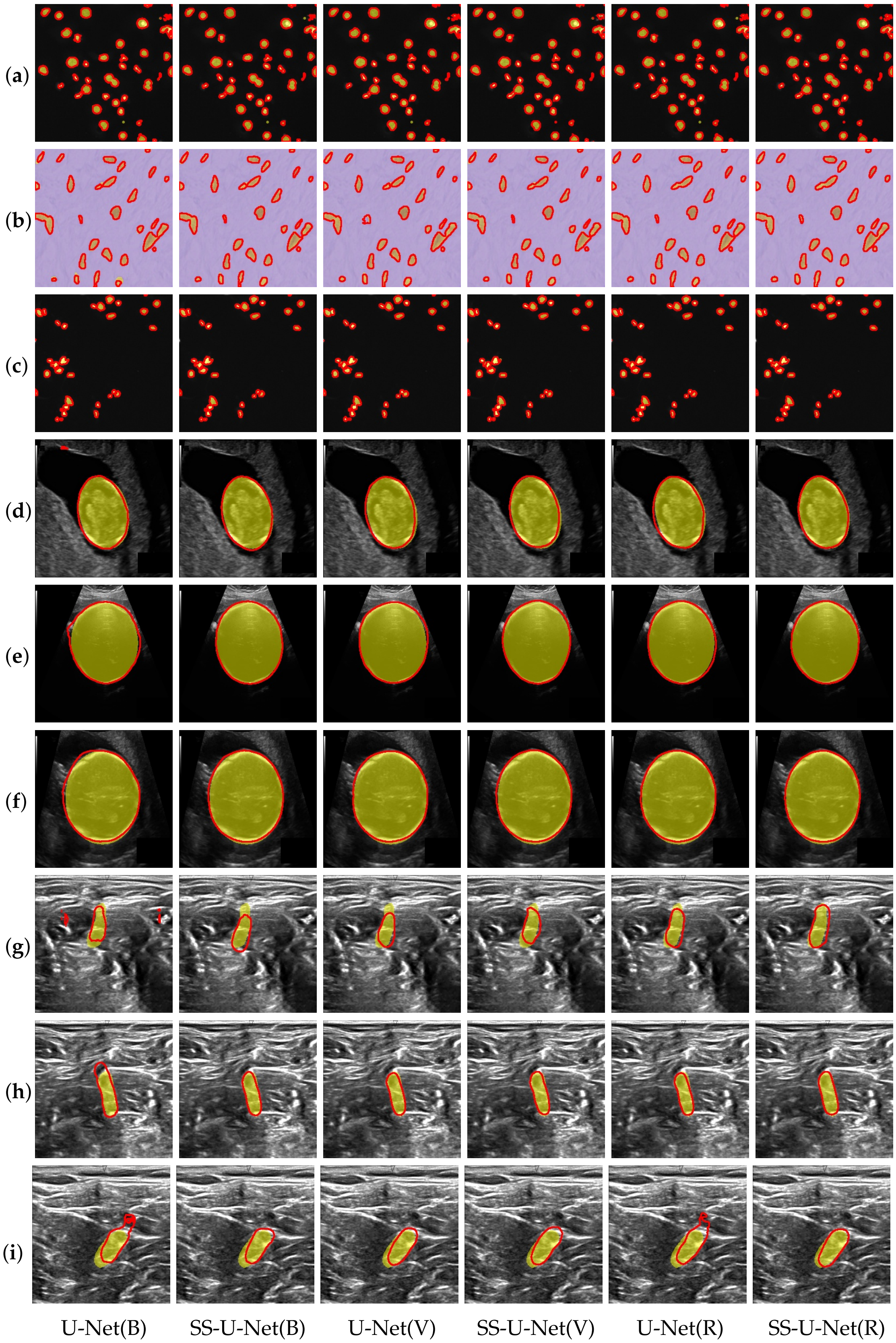


Figure 5. Performance comparison for real medical image segmentation. From top to bottom: each row represents one of the test data sets for three tasks—cell (**a**–**c**), fetal head (**d**–**f**), and nerve segmentation (**g**–**i**), respectively. From left to right: the results are segmented by U-Net(B), SS-U-Net(B), U-Net(V), SS-U-Net(V), U-Net(R), and SS-U-Net(R) approaches, respectively.

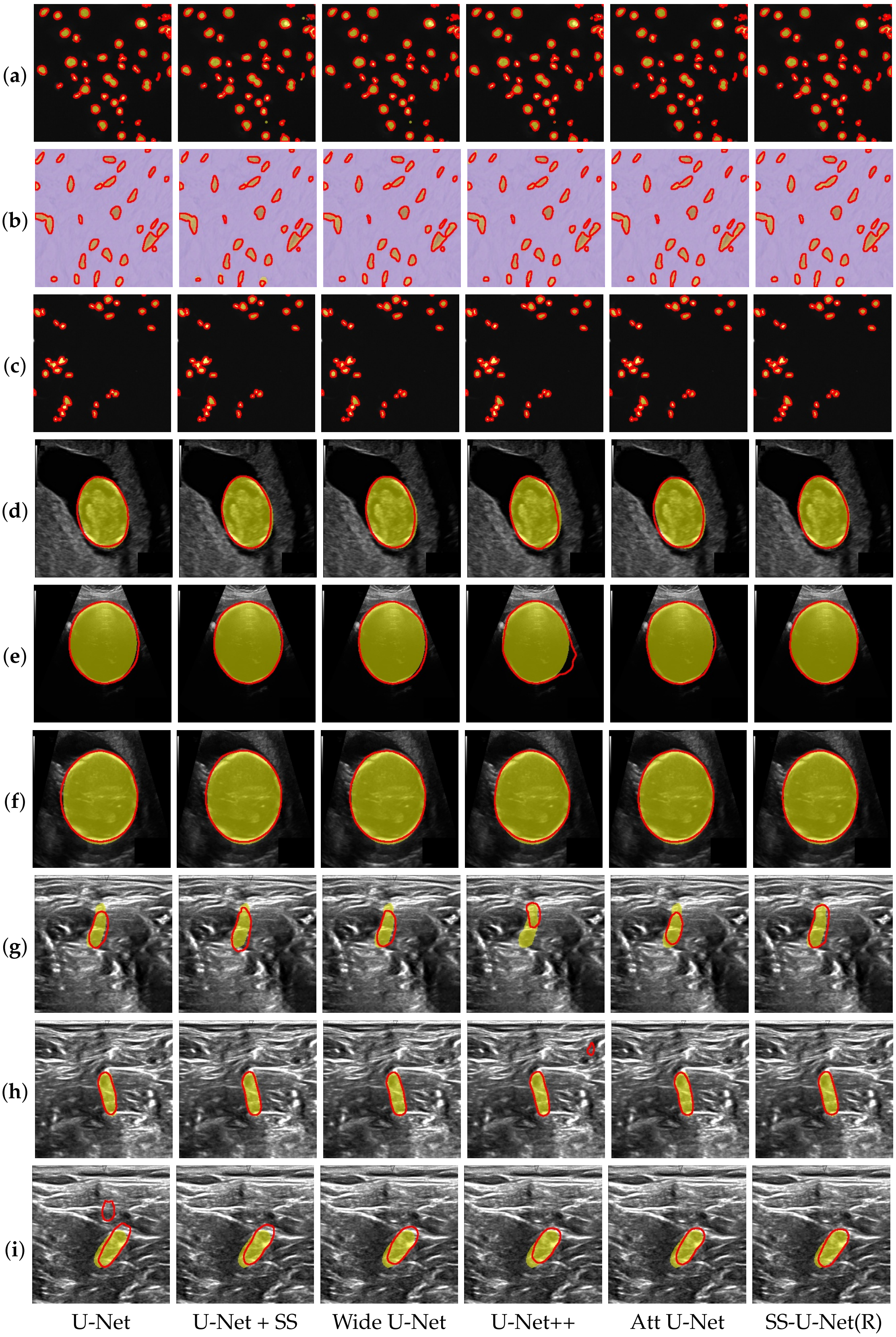


Figure 6. Performance comparison for real medical image segmentation. From top to bottom: the images are test images for three tasks—cell (**a**–**c**), fetal head (**d**–**f**), and nerve segmentation (**g**–**i**), respectively. From left to right: the results are segmented by U-Net, U-Net + SS, Wide U-Net, U-Net++, Att U-Net, and SS-U-Net(R) approaches, respectively.

5. Conclusions

In this paper, a self-spatial, adaptive, weighting-based U-Net for image segmentation was presented. The widely used convectional frameworks were employed for the proposed structure, the three kinds of convolution blocks were individually applied to the proposed structure, and their performances were compared. The experimental results showed that the proposed method could be effectively applied to the existing methods, and their performances were improved. In particular, it was verified that the proposed scheme, SS-U-Net, was efficient, and could provide the best segmentation result by combining the self-spatial adaptive weighting scheme and ResNet convolution block approach. In particular, the proposed approach with a bottleneck structure had the smallest model size among the compared methods, and they were improved in performance using the proposed block. Furthermore, the proposed method outperformed the compared methods with different segmentation targets and medical imaging modalities in terms of IoU and Dice metrics.

It is worth noting that the proposed scheme provided a compact model for SS-U-Net with the Bottlenet block structure while maintaining a similar performance to the standard U-Net. Therefore, we believe that the proposed scheme can be a useful tool for image segmentation.

Author Contributions: Conceptualization, C.C.; data curation, Y.H.L.; formal analysis, C.C.; investigation, J.P. and S.L.; methodology, C.C., Y.H.L. and S.L.; software, Y.H.L.; validation, C.C., J.P. and S.L.; visualization, C.C., Y.H.L. and J.P.; writing—original draft preparation, C.C. and Y.H.L.; writing—review and editing, C.C., Y.H.L. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-00056, Development of 3D high accuracy motion recognition SoC for automobile applications based on 140 GHz electromagnetic waves).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the policy of the institute.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
2. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Gao, M.; Yu, R.; Li, A.; Morariu, V.I.; Davis, L.S. Dynamic zoom-in network for fast object detection in large images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 21–26.
5. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
6. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Change Loy, C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
7. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
8. Luvizon, D.C.; Picard, D.; Tabia, H. 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5137–5146.
9. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.

10. Dai, B.; Lin, D. Contrastive Learning for Image Captioning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 898–907.
11. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
12. Cho, C.S.; Lee, S. Low-complexity topological derivative-based segmentation. *IEEE Trans. Image Process.* **2014**, *24*, 734–741.
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
15. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med Imaging* **2018**, *37*, 2663–2674. [CrossRef]
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
17. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
19. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
20. Li, W.; Liu, K.; Zhang, L.; Cheng, F. Object detection based on an adaptive attention mechanism. *Sci. Rep.* **2020**, *10*, 1–13. [CrossRef] [PubMed]
21. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.
22. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *arXiv* **2020**, arXiv:1908.07919.
23. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
24. Chan, T.F.; Vese, L.A. Active contours without edges. *IEEE Trans. Image Process.* **2001**, *10*, 266–277. [CrossRef] [PubMed]
25. Li, C.; Huang, R.; Ding, Z.; Gatenby, J.C.; Metaxas, D.N.; Gore, J.C. A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE Trans. Image Process.* **2011**, *20*, 2007–2016. [PubMed]
26. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
27. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Liu, G.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-video synthesis. *arXiv* **2018**, arXiv:1808.06601.
28. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
29. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Sobhaninia, Z.; Rafiei, S.; Emami, A.; Karimi, N.; Najarian, K.; Samavi, S.; Soroushmehr, S.R. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 6545–6548.
32. Zou, K.H.; Warfield, S.K.; Bharatha, A.; Tempany, C.M.; Kaus, M.R.; Haker, S.J.; Wells III, W.M.; Jolesz, F.A.; Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index1: Scientific reports. *Acad. Radiol.* **2004**, *11*, 178–189. [CrossRef]
33. Caicedo, J.C.; Goodman, A.; Karhohs, K.W.; Cimini, B.A.; Ackerman, J.; Haghighi, M.; Heng, C.; Becker, T.; Doan, M.; McQuin, C.; et al. Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl. *Nat. Methods* **2019**, *16*, 1247–1253. [CrossRef] [PubMed]
34. Ultrasound Nerve Segmentation Kaggle. 2016. Available online: https://www.kaggle.com/c/ultrasound-nerve-segmentation (accessed on 2 December 2020).
35. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8026–8037.
36. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980v9.