symmetry

MDPI

Article

Match Feature U-Net: Dynamic Receptive Field Networks for Biomedical Image Segmentation

Xiaofei Qin [1,2,3], Chengzi Wu [4], Hang Chang [5], Hao Lu [6] and Xuedian Zhang [1,2,3,7,*]

[1] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; xiaofei.qin@usst.edu.cn
[2] Shanghai Key Laboratory of Contemporary Optics System, Shanghai 200093, China
[3] Key Laboratory of Biomedical Optical Technology and Devices of Ministry of Education, Shanghai 200093, China
[4] School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 183780708@st.usst.edu.cn
[5] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; HChang@lbl.gov
[6] Guangxi Yuchai Machinery Co., Ltd., Nanning, Guangxi 530007, China; luhao@yuchai.cn
[7] Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China
* Correspondence: xdzhang@usst.edu.cn

Received: 2 July 2020; Accepted: 25 July 2020; Published: 26 July 2020

check for updates

Abstract: Medical image segmentation is a fundamental task in medical image analysis. Dynamic receptive field is very helpful for accurate medical image segmentation, which needs to be further studied and utilized. In this paper, we propose Match Feature U-Net, a novel, symmetric encoder–decoder architecture with dynamic receptive field for medical image segmentation. We modify the Selective Kernel convolution (a module proposed in Selective Kernel Networks) by inserting a newly proposed Match operation, which makes similar features in different convolution branches have corresponding positions, and then we replace the U-Net's convolution with the redesigned Selective Kernel convolution. This network is a combination of U-Net and improved Selective Kernel convolution. It inherits the advantages of simple structure and low parameter complexity of U-Net, and enhances the efficiency of dynamic receptive field in Selective Kernel convolution, making it an ideal model for medical image segmentation tasks which often have small training data and large changes in targets size. Compared with state-of-the-art segmentation methods, the number of parameters in Match Feature U-Net (2.65 M) is 34% of U-Net (7.76 M), 29% of UNet++ (9.04 M), and 9.1% of CE-Net (29 M). We evaluated the proposed architecture in four medical image segmentation tasks: nuclei segmentation in microscopy images, breast cancer cell segmentation, gland segmentation in colon histology images, and disc/cup segmentation. Our experimental results show that Match Feature U-Net achieves an average Mean Intersection over Union (MIoU) gain of 1.8, 1.45, and 2.82 points over U-Net, UNet++, and CE-Net, respectively.

Keywords: encoder-decoder; image segmentation; match feature; U-Net

1. Introduction

In the natural image task, various network structures have achieved satisfactory performance. Resnet [1], Mask-RCNN [2], and their derivative networks usually have hundreds of layers, or a large number of parameters. The continuous improvement of hardware performance and large data sets such as ImageNet [3] make these complex large-scale networks very effective. However, in medical image tasks, there is often sparse data. The typical medical image segmentation model U-Net [4], which has a modest number of parameters, was trained end-to-end from very few images and outperforms the previous best method [5] (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Medical image data are so precious that large-scale networks are hard to train on small sample data sets. Therefore, in the medical image segmentation task, the model should not have too many parameters. U-Net-based architecture is a mainstream solution.

In order to improve the accuracy of the model, the size of receptive field is very important [6]. In most existing convolution neural networks, once the model is trained, the parameters and the size of the receptive field are fixed [7]. However, the size of the targets in the pictures are different, e.g., optic disc and cup in REFUGE dataset [8]. For large objects, if the receptive field is too small it will lack the overall information, while for small objects, if the receptive field is too large it will lose the detailed information. Therefore, a dynamic and adjustable receptive field is of great benefit to the neural network. Benefiting from the low computational cost of grouped convolutions, Selective Kernel Networks [6] (SKNet) can dynamically adjust the receptive field without increasing the number of parameters. SKNet dynamically adjusts to a suitable kernel size by a triplet of operators: Split, Fuse, and Select. In SKNet, the order of features after the Split operator is often chaotic, that is to say, the output feature maps of different kernel size convolutions are often mismatched. The features that have the same channel index may not be similar, so adding them directly will cause further confusion and make the model difficult to train, as shown in Figure 1a. This study aims to improve the feature mismatch problem in SKNet and apply it to U-Net, which gives U-Net the ability to dynamically adjust the receptive field. We modify this architecture by inserting a Match block before the addition, which will sequence the features and shift similar features of different branches into the corresponding channel, as shown in Figure 1b. In order to introduce this adaptive receptive field mechanism to the medical image segmentation task, we replace the U-Net's convolution with the redesigned Selective Kernel convolution. This network inherits the advantage of the simple structure of U-Net, and enhances the efficiency of dynamic receptive field in SKNet, making it perform better for medical image segmentation tasks.

To summarize, we present Match Feature U-Net (MFUnet), a U-Net-based architecture with a dynamic receptive field for medical image segmentation. We modify and extend U-Net and SKNet architectures such that Match Feature U-Net acquires a dynamic receptive field and yields more precise segmentations. The source code of our method is available at https://github.com/WuChengziOrange/mfunet. Our main contributions are the two parts:

- We improved Selective Kernel Networks by inserting a Match operator, which matches features in different branches, and makes subsequent feature fusion more accurate.
- We embed an adaptive receptive field mechanism into U-Net, and achieve superior results in breast cancer cell segmentation, disc/cup segmentation, and gland segmentation.

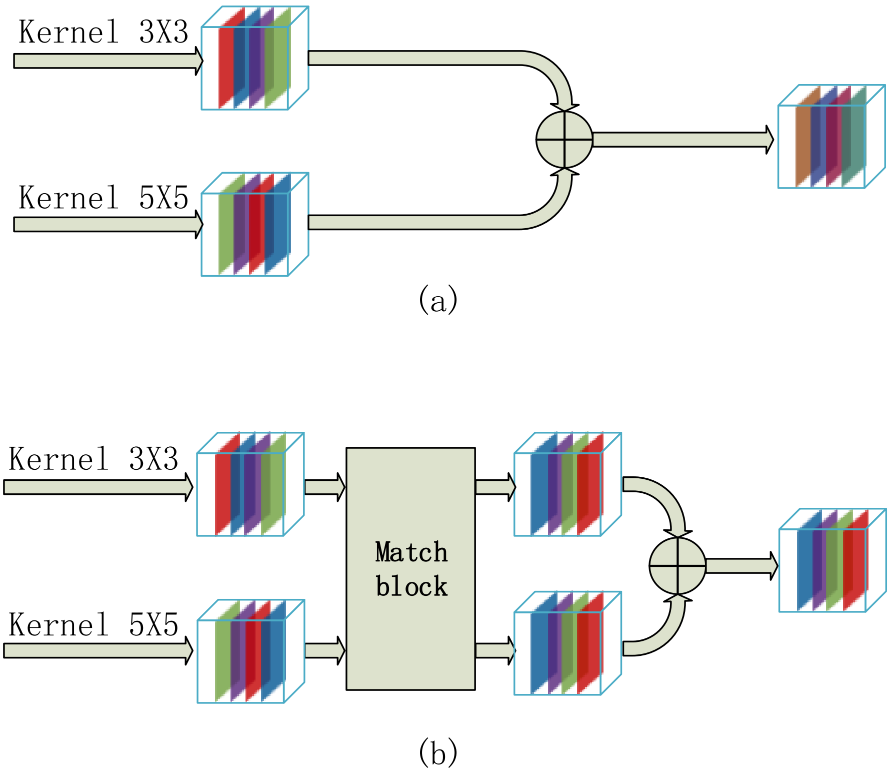


Figure 1. Illustration of Match block; different colors represent different kinds of features. SKNet adds features of different branches directly, which means that different kinds of features with the same channel index will be mixed (**a**), which will make the semantic information of the added features fuzzy and restrict the performance. We modify this architecture by inserting a Match block before the addition (**b**), which will shift the similar features of different branches to the matched channel.

2. Related Work

Encoder–decoder architecture can effectively promote the fusion of high-level semantic information and low-level spatial information, which has become the mainstream method in medical image segmentation [9]. The dynamic receptive field mechanism can make the convolutional neural network adapt to different sized targets, which has benefited many natural image segmentation tasks [6]. For the task of medical image segmentation, it is necessary to combine and make full use of their advantages.

2.1. Encoder–Decoder

Fully convolutional networks [10] refer to the pioneering work of deep learning in the field of semantic segmentation. Such a network uses convolution layers to replace the last fully connected layers, and deconvolution to up-sample the feature map and output the segmented image. Since then, a series of encoder–decoder architectures represented by U-Net has emerged. The encoder path extracts the feature information of the image while increasing the image receptive field through downsampling. The decoder path recovers the image size through deconvolution and other operations. U-Net introduces skip connections to fuse the shallow, precise, fine-grained features from downsampling and the deep, semantic, and coarse-grained features from upsampling. U-net shows great potential in the field of medical image segmentation. Many subsequent works are based on U-net, forming a huge U-net family. Drozdzal et al. [11] put forward that in the U-Net architecture, not only long jump connection structure, but also short jump connection structure can be used. Çiçek et al. [12] propose a 3D U-Net network structure, which can segment 3D images by inputting a continuous 2D slice sequence of 3D images. Milletari et al. [13] design V-Net, which is a 3D deformation structure of U-Net.

V-Net uses Dice coefficient loss function instead of traditional cross-entropy loss function, and reduces channel dimension through $1 \times 1 \times 1$ convolution kernel. Wang et al. [14] propose a wound image analysis system. First, U-Net is used to segment the wound image, and then a Support Vector Machine (SVM) classifier is used to classify the segmented wound image to determine whether the wound is infected. Finally, Gaussian Process (GP) regression algorithm is used to predict the wound healing time. Brosch et al. [15] use U-Net to segment white matter lesions in brain Magnetic Resonance Imaging (MRI), and add a jump connection structure between the first layer of convolution and the last layer of deconvolution of U-Net, making the network architecture acquire good segmentation results even with less training data. Inspired by DenseNet [16] and U-Net, Zhou et al. design UNet++ [9], which makes full use of different depths of downsampling, denser skip connection, and deep supervision to output more accurate segmentation results. Z Gu et al. propose CE-Net [17] to capture more high-level information and preserve spatial information for 2D medical image segmentation. ET-Net [18] embeds edge-attention representations to guide the segmentation network. In our work, we propose an alternative approach that makes full use of the information flow in the encoder–decoder architecture. Each convolution layer of our model can be dynamically adjusted according to the input image, which is more flexible and adaptive.

2.2. Receptive Field

The size of the receptive field plays a crucial role in Convolutional Neural Networks. Too large a receptive field will lose localization accuracy, and too small a receptive field will limit context information. Many networks [19–24] enrich the receptive field on the basis of the encoder–decoder structure. Inspired by spatial pyramid pooling [25], Pyramid Scene Parsing Network [19] and DeepLabv3 [20] use the Pyramid Pooling Module to obtain the feature maps of different receptive fields, taking into account the detailed features and global image-level features, and outputting more reasonable segmented images. DeepLabv3+ [21] introduces atrous separable convolution (dilated separable convolution) to enlarge the receptive field size and reduce computation complexity. SAC [22] relaxes the fixed size receptive field by predicting the position adaptive scale coefficient, thus easing the segmentation problem of invisible small targets and inconsistent large targets to some extent. Dynamic Filter [7] adaptively adjusts the parameters of the filter, but not the size of the kernel. Active convolution [26] uses offsets to increase the sampling field in convolution. These offsets are learned end-to-end, but become static in inference. Deformable Convolutional Networks [27] further learn offsets for each element of regular convolutional filters to augment the sampling locations with arbitrary form, which can extract geometric-invariant features. SKNet proposes an adaptive selection mechanism, which divides the common convolution operation into three steps: split, fuse, and select. Split operation allows different kernel size convolutions in multiple parallel branches to extract features with different receptive fields. Fuse operation sums the features of different receptive fields in an element-wise manner, and then applies global average pooling to generate channel-wise statistics. Finally, the select operation is carried out to select the appropriate receptive fields and features. In the task of focus segmentation, the deep learning algorithm needs to complete many tasks such as target recognition, organ segmentation, and tissue segmentation. Therefore, in the process of segmentation, the global information and local information of the image should be combined to achieve the accurate segmentation of the focus. Kamnitsas et al. [23] and Ghafoorian et al. [24] adapt multi-scale convolution to extract the global and local information.

3. Methods

3.1. Adaptive Receptive Field Convolution

As shown in Figure 2, to better fuse features from different branches, we insert a Match block between Split and Fuse operation in the Selective Kernel convolution [6].

We follow the symmetric two-branch structure of SKNet [6]. The two outputs, $U_a \in \mathbb{R}^{H \times W \times C}$ and $U_b \in \mathbb{R}^{H \times W \times C}$, of the Split operation are derived, respectively, from the 3×3 grouped convolution and the 5×5 grouped convolution (replaced by the dilated convolution with a 3×3 kernel and dilation size of 2). U_a and U_b contain similar features, but the order of these features on the channel is usually random: $0 < i \leq C, i \ \in \mathbb{N}, \exists U_{a(i)}, U_{b(i)}$ usually belongs to different kinds of features, where c represents the number of channels in the feature map and $U_{(i)}$ denotes the feature on the ith channel in the feature map. Adding them directly through fusion will lead to confusion. As stated above, our goal is to improve the feature mismatch problem in SKNet and apply it to U-Net, which gives U-Net the ability to dynamically adjust the receptive field. The main idea is to introduce a Match block to transfer similar features into the corresponding channel. To achieve this goal, we divide the Match block into two parts: Coarse tuning and Fine tuning.

In the coarse tuning part, we obtain two vectors, $(V_a, V_b \in \mathbb{R}^{1 \times c})$, by using global average pooling in the feature maps of two branches, and then arrange the elements of the vectors in descending order:

$$V_a = rank(F_{gap}(U_a)), V_b = rank(F_{gap}(U_b)) \tag{1}$$

where F_{gap} is the global average pooling and rank denotes the descending order arrangement. Further, the features in U_a and U_b are arranged in descending order according to V_a and V_b, so that we get the coarse sorted feature map $(\dot{U}_a, \dot{U}_b \in \mathbb{R}^{H \times W \times C})$:

$$\dot{U}_a = sort(U_a, V_a), \dot{U}_b = sort(U_b, V_b) \tag{2}$$

In the fine tuning part, the matched feature maps, $(\bar{U}_a, \bar{U}_b \in \mathbb{R}^{H \times W \times C})$, are created through the following mechanisms.

For the first branch, $\overline{U}_a$ is a direct copy of $\dot{U}_a$ and is considered as a feature map with the standard feature order (The standard feature order can comes from any branch, and we take the first branch as example.):

$$\overline{U}_a = \dot{U}_a \tag{3}$$

For the other branches (only the second branch in Figure 2), we calculate the correlation between each feature in $\dot{U}_a$ and the adjacent $2k + 1$ features in $\dot{U}_b$, and here we simply use the inner product to calculate the correlation. Then, we transfer the most relevant features into the corresponding channels where $\overline{U}_b$ is the stack of these features:

$$\overline{U}_{b(i)} = \underset{\dot{U}_{b(j)}}{\arg\max}(\dot{U}_{a(i)} \cdot \dot{U}_{b(j)}), i - k \leq j \leq i + k \tag{4}$$

$$\overline{U}_b = [\overline{U}_{b(1)}, \overline{U}_{b(2)}, \cdots, \overline{U}_{b(c)}] \tag{5}$$

where $i, j \in \mathbb{R}^c$, c denotes the channel size of the feature maps, and the fine tuning rate k is a hyperparameter to adjust the fine tuning range. The larger the value of k, the larger the matching range, and vice versa. Through experiments, we find that $k = 1$ is a suitable value. After Match block, the following operation is the same as SKNet, using the Fuse operator and Select operator to process the matched feature maps.

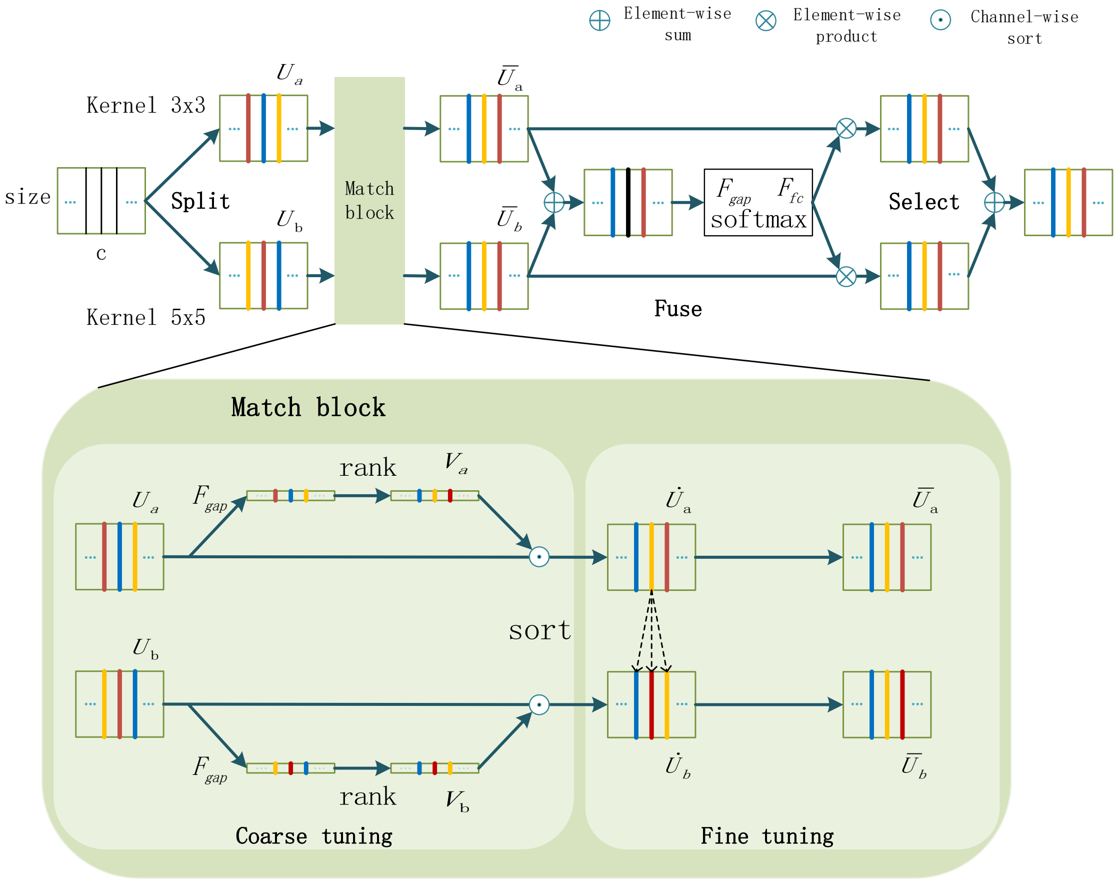


Figure 2. Illustration of Adaptive Receptive Field (ARF) convolution, different colors denote different kinds of features. The Match block consists of two parts: coarse tuning and fine tuning. In the coarse tuning part, we arrange the features in descending order according to the average pooling value. In the fine tuning part, we keep the 3 × 3 branch unchanged. For each feature in the 3 × 3 branch, we calculate the correlation with the nearest $2k + 1$ features (as shown by dashed arrows) in the 5 × 5 branch to match the most relevant features.

3.2. Network Architecture

As shown in Figure 3, the Dynamic Convolution Unit (DCU) is similar to the residual unit. Each DCU adopts a residual connection and consists of a sequence of 1 × 1 convolution, ARF convolution, and a further 1 × 1 convolution. In the lower branch of the DCU, the 1 × 1 convolution in front of the ARF convolution is used to reduce the number of channels in the feature maps, and thus reduce computational complexity. The 1 × 1 convolution in the back of the ARF convolution is used to increase the number of channels in the feature maps and improve the expression ability of the model. The 1 × 1 convolution in the upper branch of the DCU is used to change the number of channels of the feature maps so as to add it to the lower branch.

As shown in Figure 4, we replace the 3 × 3 convolutions in U-Net with DCU because U-Net is one of the state-of-the-art network architectures for medical image segmentation. This enables almost all layers to select the appropriate receptive field.

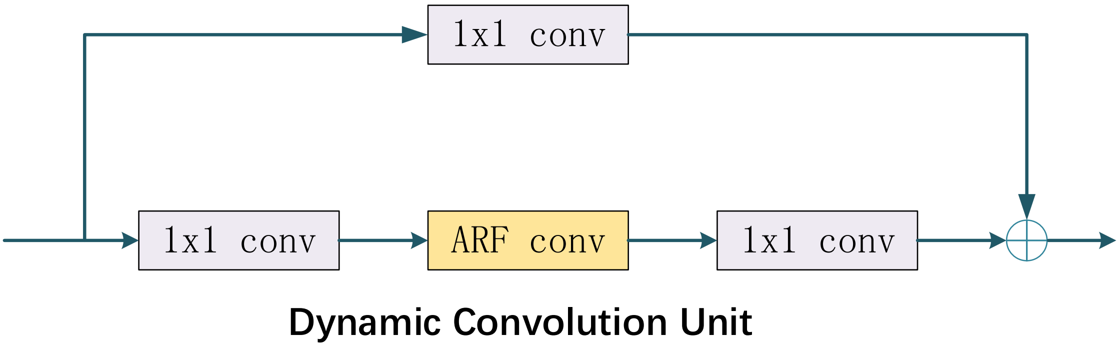


Figure 3. Illustration of the Dynamic Convolution Unit, which outputs the sum of a 1 × 1 convolution branch and a stack of a 1 × 1 convolution, ARF convolution, and a further 1 × 1 convolution. More details of the yellow block (Adaptive Receptive Field convolution) are shown in Figure 2.

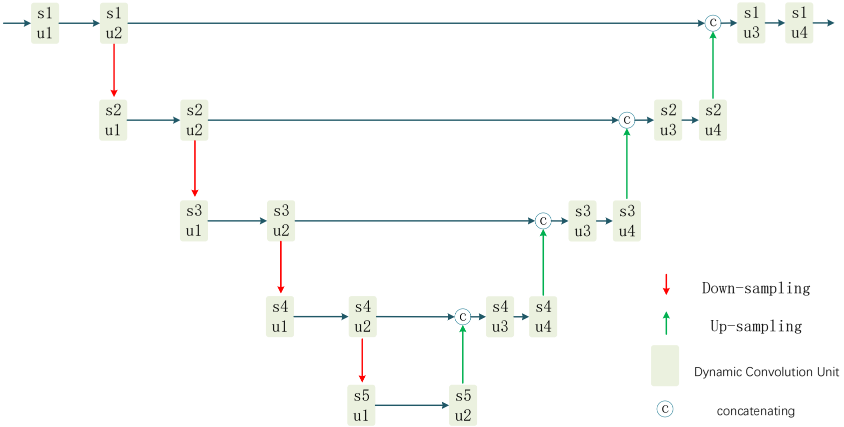


Figure 4. Overview of the MFUnet architecture. *si* and *uj* denote stage *i* and unit *j*, respectively. Each olive box represents a Dynamic Convolution Unit, red arrows represent downsampling (Max-pooling), and green arrows represent upsampling (Transposed convolution).

4. Experiments

4.1. Datasets

We use four datasets to evaluate the model, as shown in Table 1. The size of these images is not consistent, and the number of samples is very small, so we adopt standard data augmentation techniques including random horizontal flip, random scaling, and random crop. Horizontal flip doubles the dataset. We use two random scaling rates, so random scaling also doubles the dataset. Four random cropped images are selected for each original image, so random crop quadruples the dataset. In this way, each image in the original dataset is augmented to 2 × 2 × 4 = 16 images. Finally, the input resolution of the network during training is 512 × 512 (256 × 256 on the Cell nuclei dataset).

Table 1. The image segmentation datasets used in our implementation (For multiple resolution data sets, only the two most common resolutions are displayed.).

Dataset	Images	Size	Provider
Cell nuclei	670	256 × 256 256 × 320	https://www.kaggle.com/c/data-science-bowl-2018 Data Science Bowl 2018 [28]
Breast cancer cells	58	896 × 768 768 × 512	UCSB Bio-Segmentation Benchmark dataset [29]
Gland Segmentation	165	775 × 522 589 × 453	GlaS challenge at MICCAI 2015 [30]
REFUGE	800	2124 × 2056	REFUGE Challenge [8]

As shown in Figure 5, we use 20% of the augmented dataset as test data and the remaining 80% for 5-fold cross-validation. As shown in Table 2, five groups of experiments are conducted, and each group is repeated twice to obtain a total of 10 test results, and the mean and standard deviation of the results are calculated.

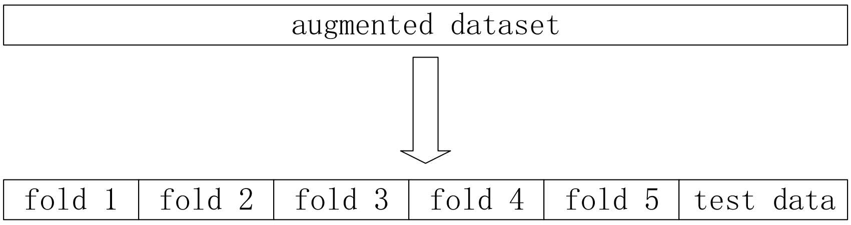


Figure 5. Data partition strategy for each dataset.

Table 2. Data distribution for five groups of experiments.

Group	Train Set	Validation Set	Test Set
1	fold 2, fold 3, fold 4, fold 5	fold 1	test data
2	fold 1, fold 3, fold 4, fold 5	fold 2	test data
3	fold 1, fold 2, fold 4, fold 5	fold 3	test data
4	fold 1, fold 2, fold 3, fold 5	fold 4	test data
5	fold 1, fold 2, fold 3, fold 4	fold 5	test data

4.2. Baseline Models

For contrast, we use U-Net, CE-Net, and UNet++, as U-Net is widely used in medical image segmentation, CE-Net is good at capturing high-level information, and UNet++ is one of the state-of-the-art network architectures for medical image segmentation.

4.3. Evaluation Metrics

To evaluate the performance, we adopt the Mean Intersection over Union (MIoU), which has been commonly used to evaluate the accuracy of semantic segmentation. Its definition is

$$MIoU = \frac{1}{m+1}\sum_{i=0}^{m}\frac{Area(A_i \cap B_i)}{Area(A_i \cup B_i)} \tag{6}$$

where m represents the number of categories (plus background for a total of $m+1$ categories), and A_i and B_i denote the predicted and the ground truth results for ith category, respectively. $MIoU \in [0,1]$, the closer its value is to 1, the better the model performance is, and the closer it is to 0, the worse the model performance is. The specific calculation method is shown in the following formula,

$$MIoU = \frac{1}{m+1}\sum_{i=0}^{m}\frac{n_{ii}}{\sum_{j=0}^{m}(n_{ij}+n_{ji}) - n_{ii}} \tag{7}$$

where m represents the number of categories (plus background for a total of $m+1$ categories); n_{ii} is the number of pixels whose predicted category and ground truth category are both ith category; n_{ij} means the number of pixels with the ground truth category is ith category and the predicted category is jth category; n_{ji} represents the number of pixels with the ground true category is jth category and the predicted category is ith category; n_{ii} is equivalent to truth positives (TP); and n_{ij} and n_{ji} are equivalent to false negatives (FN) and false positives (FP), respectively.

In our paper, the term "average MIoU gain" means the average value of our method's MIoU on four datasets subtracting that of other methods.

4.4. Implementation Details

Biomedical image segmentation is a pixel-wise classification problem and we expect to train the proposed framework to predict each pixel to be foreground or background. Commonly used loss function is cross-entropy. However, cross-entropy loss is not optimal for this problem because the objects in medical images, such as optic disc and cup, often occupy a small region in the image [17]. In our work, we use Dice coefficient loss [13,31] instead of cross-entropy loss. The Dice coefficient is a measure of overlap and its definition is shown in the following formula,

$$L_{dice} = 1 - \sum_{k=0}^{K} \frac{2w_k \sum_i^N p_{(k,i)} g_{(k,i)}}{\sum_i^N [p_{(k,i)}^2 + g_{(k,i)}^2]} \tag{8}$$

where K is the class number and N denotes the pixel number. $p_{(k,i)} \in [0,1]$ and $g_{(k,i)} \in 0,1$ are the predicted probability and ground truth label for class k, respectively. w_k denotes the weight for class k.

The final loss function is

$$L_{loss} = L_{dice} + L_{reg} \tag{9}$$

where L_{reg} denotes the regularization loss [32] used to avoid overfitting.

Our method is implemented in Keras, and uses TensorFlow [33] as the backend. During training, the parameters of all layers are randomly initialized and the *batchsize* is set to 8 with synchronized batch normalization. The $early - stop$ mechanism is also used to avoid overfitting. Our method is optimized using the Adam optimizer with initial learning rate 1×10^{-4}, weight decay of 0.0005, and momentum of 0.9.

Although many optimization methods show better performance, our model works best with Adam optimizer. The training losses of our model using different optimization methods are shown in Figure 6; the Adam optimizer has minimal training loss and fast convergence speed.

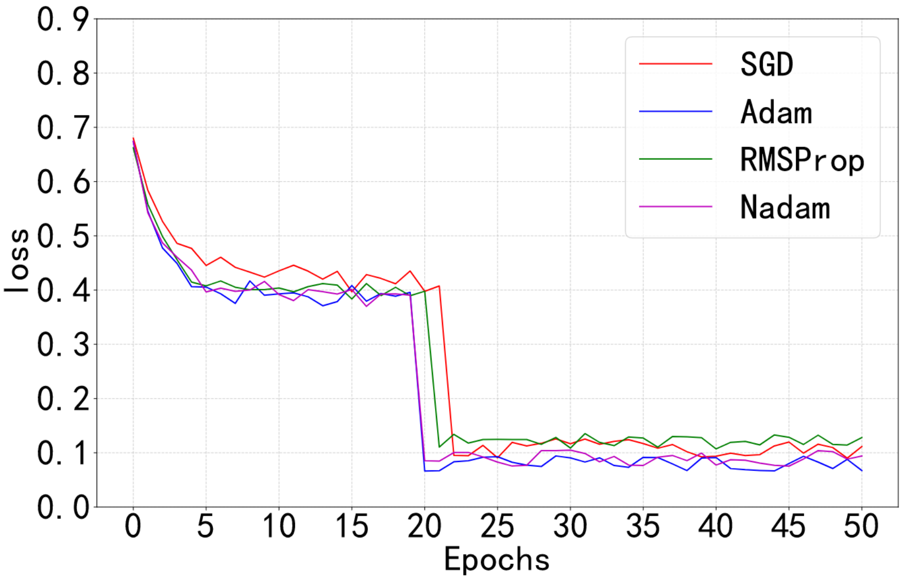


Figure 6. The training losses of our model using different optimization methods.

4.5. Results

As shown in Table 3, our method compares well with U-Net, UNet++, and CE-Net in terms of model size and Mean Intersection over Union (MIoU) for the tasks of cell nuclei segmentation, breast cancer cell segmentation, Gland segmentation, and disc/cup segmentation. Our method achieves an average MIoU gain of 1.8, 1.45, and 2.82 points over U-Net, UNet++, and CE-Net on four datasets, respectively. This improvement results from two causes: (1) our method can adjust the appropriate receptive field to achieve the best performance and (2) our model has much fewer parameters (9.1∼34% of other methods) and is easier to train on small datasets. Figure 7 shows a qualitative comparison between the results of U-Net, UNet++, CE-Net, and MFUnet (ours). On the task of cell nuclei segmentation, we achieve slightly worse performance (reduced by 0.69 points) than UNet++ because the cell nuclei objects that are to be segmented have similar size. For targets of similar size, the adaptive receptive field advantage of ARF unit can not be fully exploded; however, in medical images, where the target size is very random, our method still has an advantage.

Table 3. Comparison among different methods (mean ± standard deviation).

Architecture	Params	Cell Nuclei	Breast Cancer Cell	Gland	REFUGE	Average MIoU Gain
U-Net	7.76 M	90.77 ± 0.48	41.73 ± 0.43	83.45 ± 0.46	83.21 ± 0.49	1.8
UNet++	9.04 M	**92.52** ± **0.57**	39.92 ± 0.55	84.12 ± 0.53	83.97 ± 0.54	1.45
CE-Net	29 M	90.54 ± 0.45	37.95 ± 0.44	81.23 ± 0.47	85.36 ± 0.37	2.82
MFUnet (ours)	2.65 M	91.83 ± 0.39	**42.57** ± **0.37**	**85.29** ± **0.36**	**86.68** ± **0.31**	–

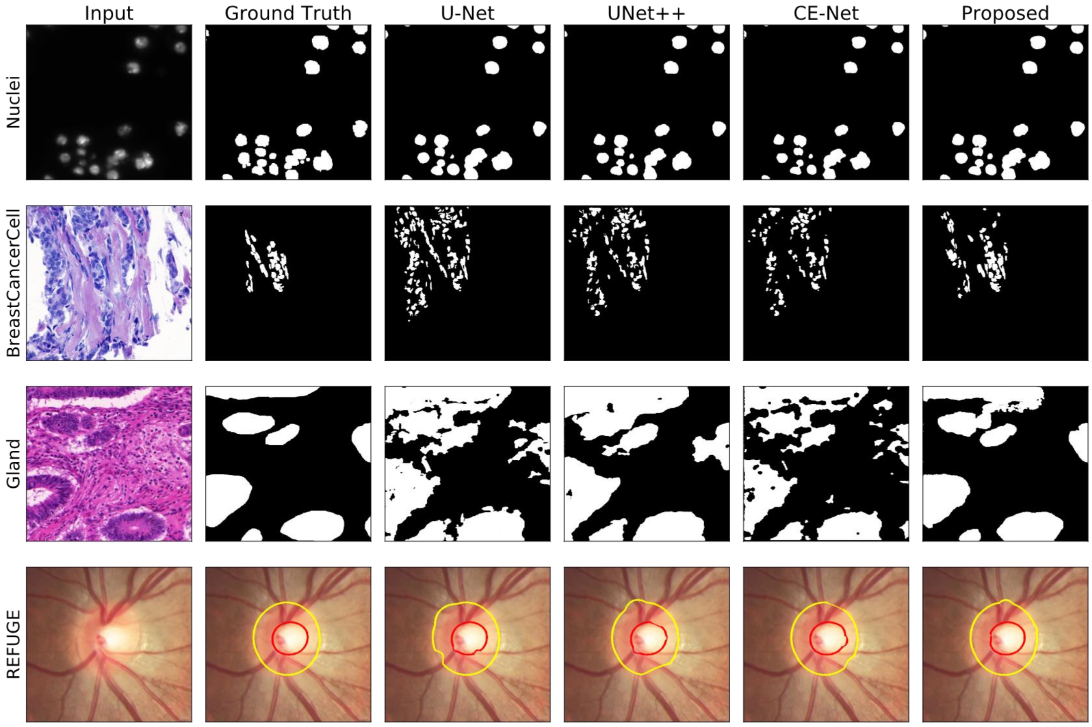


Figure 7. Examples of segmentation results with U-Net, UNet++, CE-Net, and MFUnet (ours).

4.6. Ablation Studies

Fine tuning rate k, combinations of different kernels, Match operation, and the adaptive receptive field mechanism are four important influence factors of model performance. In order to study their effectiveness, we evaluate model performance on the Cell nuclei, Breast cancer cell, and Gland Segmentation datasets.

Table 4 shows the performance of our method on three data sets for different k settings. With the increase of k, the fine tuning range gets larger. We find that $k = 1$ is a suitable value, and if we continue to increase k, the performances will decline (still better than the case of $k = 0$).

As shown in Table 5, the combination of 3×3 and 5×5 convolutional kernels achieves the best results (an average increase of 1.25 points over U-Net on three datasets). The results of the combination of 5×5 and 7×7 convolutional kernels are the worst, but they are still improved compared with U-Net (an average increase of 0.63 points on three datasets). This shows that our method is also suitable for other convolution kernel combinations, where the combination of 3 × 3 and 5 × 5 convolutional kernels is a suitable choice.

Table 4. Results (MIoU: %) of our method for different k settings (mean ± standard deviation).

k	Cell Nuclei	Breast Cancer Cell	Gland
0	91.54 ± 0.39	42.03 ± 0.33	84.82 ± 0.40
1	**91.83 ± 0.39**	**42.57 ± 0.37**	**85.29 ± 0.36**
2	91.67 ± 0.37	42.25 ± 0.36	84.95 ± 0.35
3	91.59 ± 0.38	42.23 ± 0.35	84.97 ± 0.36

Table 5. Results (MIoU: %) of MFUnet with different combinations of multiple kernels (mean ± standard deviation).

3 × 3	5 × 5	7 × 7	Cell Nuclei	Breast Cancer Cell	Gland
✓	✓	×	**91.83 ± 0.39**	**42.57 ± 0.37**	**85.29 ± 0.36**
✓	×	✓	91.17 ± 0.38	41.82 ± 0.36	84.52 ± 0.31
×	✓	✓	90.85 ± 0.40	41.79 ± 0.33	83.94 ± 0.40
✓	✓	✓	91.31 ± 0.45	41.98 ± 0.42	84.83 ± 0.39

Table 6 shows the performances of Match operator and adaptive receptive mechanism in U-Net, UNet++, and CE-Net. The results show that MFUet (U-Net with Match operator and adaptive receptive mechanism) can achieve the best results. MFUet achieves an average MIoU gain of 1.06, 0.84, and 1.84 points over original U-Net on nuclei segmentation, breast cancer cell segmentation, and gland segmentation, respectively. It is noted that the improvement is more marked for gland segmentation because the scale of the objects in the dataset is more variable and the information is more complex. Therefore, the Match operation plays a more important role. The features extracted by the convolution kernels of the two branches are more complex, so the matched feature maps fusing will further improve the results, and the improved performance of our model on this dataset is more obvious. We find that the performances of three networks (U-Net, UNet++, and CE-Net) using only adaptive receptive field mechanism (replace the convolution with the original Selective Kernel convolution) are worse than that of the original networks and then after introducing the Match operation, the performances are better than the original networks. The reason for performance degradation after introducing adaptive receptive field mechanism alone might be that medical images have lower level semantic information than natural images and have less dependency on the receptive field. Without the feature matching operation, the information flow in the network might get more chaotic. This shows that the proposed Match block is useful for our segmentation task.

Table 6. Results (MIoU: %) of Match operator and adaptive receptive mechanism in U-Net, UNet++, and CE-Net (mean ± standard deviation).

Architecture	Match Operator	Adaptive Receptive Field Mechanism	Cell Nuclei	Breast Cancer Cell	Gland
U-Net	×	✓	89.97 ± 0.42	41.85 ± 0.46	81.07 ± 0.39
U-Net	✓	✓	**91.83 ± 0.39**	**42.57 ± 0.37**	**85.29 ± 0.36**
UNet++	×	✓	91.03 ± 0.50	39.51 ± 0.53	82.29 ± 0.51
UNet++	✓	✓	91.79 ± 0.53	40.46 ± 0.49	84.81 ± 0.48
CE-Net	×	✓	89.88 ± 0.46	37.17 ± 0.38	80.03 ± 0.45
CE-Net	✓	✓	91.24 ± 0.44	39.58 ± 0.36	83.74 ± 0.43

4.7. Visualization of Features

To verify whether Match block works, we visualized the input and the output feature maps of the module, as shown in Figure 8. For clear observation (the semantic information with higher resolution is easier to understand, and the semantic information with lower resolution is more difficult to understand), we chose to observe the convolution layer (stage 1 unit 4). It consists of two branches: a 3 × 3 convolution kernel branch and a 5 × 5 convolution kernel branch. The feature maps of each branch contain 16 channels, with the same resolution as the input image: 512 × 512. We select the first three channels of feature map for comparison. Figure 8b is the input of the Match block. The above feature map comes from the 3 × 3 convolution kernel branch, and the following feature map comes from the 5 × 5 convolution kernel branch. It can be seen that the features of the corresponding channels are quite different. The activation values (global average pooling value) of each channel and the activation value difference between two branches are shown in Figure 9a,c. Figure 8c shows the feature maps processed by Match block. It can be seen that similar channels in the feature maps are matched, thus providing better conditions for fusion. Figure 9b shows the activation values of each channel of the two branch feature maps after matching, and Figure 9d shows the activation value difference between two branches after matching. It can be seen that the activation values of the 3 × 3 branch are arranged in descending order (which is consistent with the expectation of coarse tuning), while the 5 × 5 branch is fluctuating, but the overall arrangement is in descending order (This is the result of fine tuning, because the activation values between the most similar channels are not necessarily the closest.). This shows that the proposed Match block is able to make similar features in different convolution branches have corresponding positions.

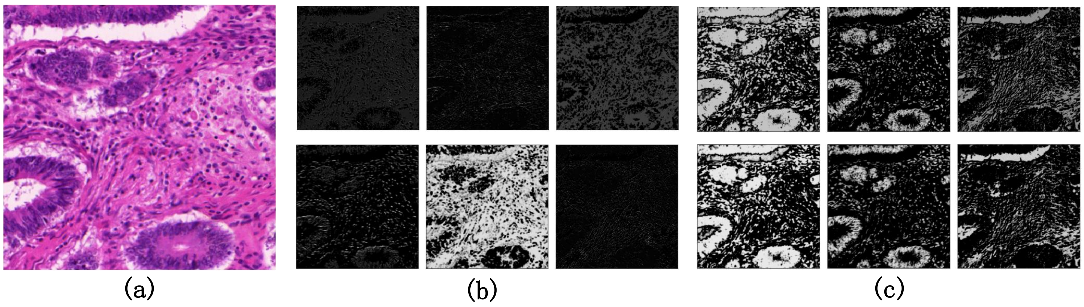
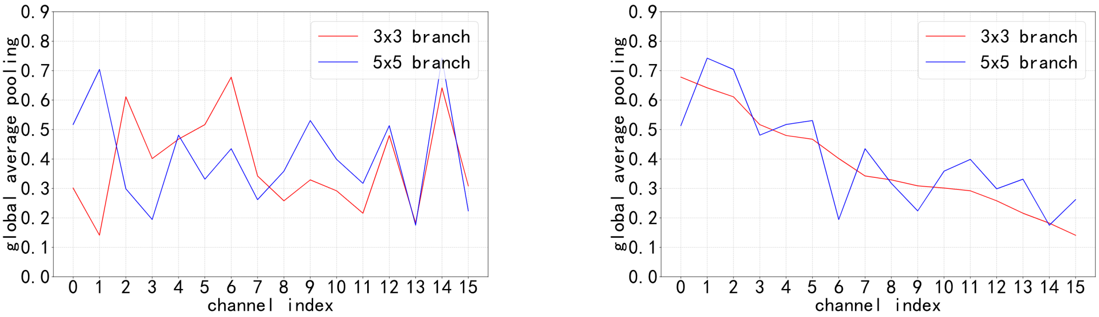


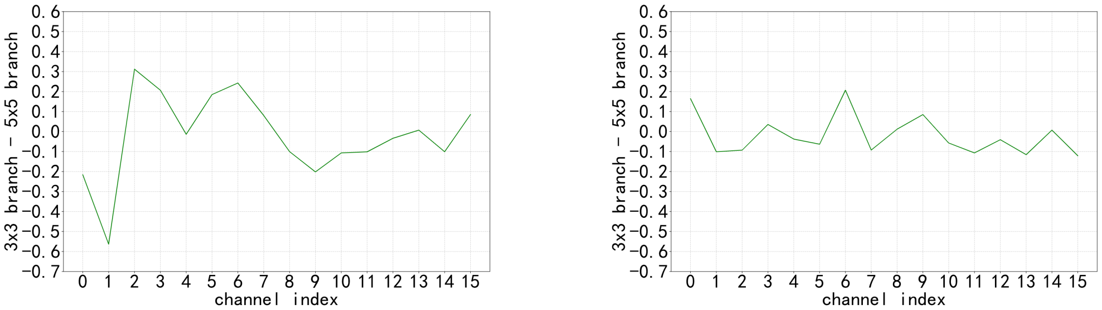
Figure 8. Visualization of feature maps (stage 1 unit 4). (**a**) Original input picture. (**b**) Feature maps before matching. (**c**) Feature maps after matching. The above feature map comes from the 3 × 3 convolution kernel branch, and the following feature map comes from the 5 × 5 convolution kernel branch.



(**a**) Activation values before matching

(**b**) Activation values after matching

Figure 9. *Cont.*



(**c**) Activation value difference before matching

(**d**) Activation value difference after matching

Figure 9. Illustration of the activation values (global average pooling value) of each channel. Activation value difference represents the activation values of the 3 × 3 branch and that of the 5 × 5 branch.

5. Conclusions

By inserting a Match block into SK convolution, we propose an Adaptive Receptive Field (ARF) convolution which improves the regulation of the receptive field. We also construct MFUnet from the original U-Net architecture by replacing the 3 × 3 convolution with the ARF convolution. We evaluate the MFUnet on four datasets, and the experimental results show that our proposed Match operation can enhance segmentation performance, especially on variable scale datasets.

Author Contributions: Conceptualization, X.Q. and C.W.; methodology, X.Q., H.C., and C.W.; software, C.W. and H.L.; validation, X.Z., X.Q., and H.C.; formal analysis, H.L. and H.C.; investigation, X.Q. and C.W.; resources, X.Q. and X.Z; data curation, X.Q. and C.W.; writing—original draft preparation, X.Q. and C.W.; writing—review and editing, X.Q., H.C., and C.W.; visualization, X.Q. and C.W.; supervision, X.Z. and X.Q; project administration, X.Q. and X.Z; funding acquisition, X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Artificial Intelligence Program of Shanghai grant number 2019-RGZN-01077 and the National Key Research and Development Program of China grant number 2016YFF0101400.

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

References

1. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. 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.
3. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
4. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
7. Jia, X.; De Brabandere, B.; Tuytelaars, T.; Gool, L.V. Dynamic filter networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 9 December 2016; pp. 667–675.

8. Orlando, J.I.; Fu, H.; Breda, J.B.; Van Keer, K.; Bathula, D.R.; Diazpinto, A.; Fang, R.; Heng, P.; Kim, J.; Lee, J.; et al. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. *Med. Image Anal.* **2020**, *59*, 101570. [CrossRef]
9. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Drozdzal, M.; Vorontsov, E.; Chartrand, G.; Kadoury, S.; Pal, C. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 179–187.
12. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 424–432.
13. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
14. Wang, C.; Yan, X.; Smith, M.; Kochhar, K.; Rubin, M.; Warren, S.M.; Wrobel, J.; Lee, H. A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 2415–2418.
15. Brosch, T.; Tang, L.Y.; Yoo, Y.; Li, D.K.; Traboulsee, A.; Tam, R. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Imaging* **2016**, *35*, 1229–1239. [CrossRef] [PubMed]
16. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
17. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
18. Zhang, Z.; Fu, H.; Dai, H.; Shen, J.; Pang, Y.; Shao, L. ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2019; pp. 442–450.
19. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
20. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
21. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
22. Zhang, R.; Tang, S.; Zhang, Y.; Li, J.; Yan, S. Scale-adaptive convolutions for scene parsing. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2031–2039.
23. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef] [PubMed]
24. Ghafoorian, M.; Karssemeijer, N.; Heskes, T.; Van Uder, I.; de Leeuw, F.E.; Marchiori, E.; van Ginneken, B.; Platel, B. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; pp. 1414–1417.
25. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2; pp. 2169–2178.

26. Jeon, Y.; Kim, J. Active convolution: Learning the shape of convolution for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4201–4209.
27. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
28. Kaggle 2018 Data Science Bowl. Available online: https://www.kaggle.com/c/data-science-bowl-2018 (accessed on 19 July 2020).
29. Gelasca, E.D.; Byun, J.; Obara, B.; Manjunath, B. Evaluation and benchmark for biological image segmentation. In Proceedings of the 2008 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 1816–1819.
30. Sirinukunwattana, K.; Pluim, J.P.; Chen, H.; Qi, X.; Heng, P.A.; Guo, Y.B.; Wang, L.Y.; Matuszewski, B.J.; Bruni, E.; Sanchez, U.; et al. Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Anal.* **2017**, *35*, 489–502. [CrossRef] [PubMed]
31. Crum, W.R.; Camara, O.; Hill, D.L. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans. Med. Imaging* **2006**, *25*, 1451–1461. [CrossRef] [PubMed]
32. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
33. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.