Attention-based Forward Passing for Enhancing UNet Segmentation on Microscopy Images

Xinyu Xu
518021910645
xuxinyu2000@sjtu.edu.cn

Chenfeng Bai
518021910850
353349591@qq.com

Chao Wang
518021910166
791972041@qq.com

Abstract

Recent year has witnessed a great success of deep learning applications including bioinformatics. Deep learning techniques have become powerful weapons in cell structure analysis. Automatic biological image segmentation is a difficult problem but also of great value in the subject. Conventional image segmentation methods always suffer from lack of data in biological field but lightweight UNet [10] and variants work effectively, attracting much interests. The key of its success is about the forward passing of feature using a skip connection from encoder to decoder. However, such simple method is far from preciseness. In this project, we applies attention mechanism to achieve selective forward passing and integrate it into UNet family. We conduct experiments and make detailed analysis to validate the effectiveness of our approach. Code is publicly available in https://github.com/AllenXuuu/CS420-Machine-Learning.

1 Introduction

Microscopy images analysis is a hard problem of huge expense in traditional biological experiments. Researcher have to carefully inspect the whole image to make conclusion about the cellular property. This is such an physically exhausting phase. An automatic analysis about microscopy images is necessary for boosting relevant researches.

Deep learning gains a rapid development in recent years. Starting from AlexNet [8] who introduces deep convolutional neural networks into image classification, it attracts tons of attention from leading edge researchers. Some variants likes GoogLeNet [12] and VGG [11] achieves better and better performance and validates the large potential of CNN. ResNet [4] is another break point in the development of deep learning. He [4] proposed the idea of skip connection to alleviate gradient vanish in extremely deep neural network, which has a great impact on following researches. These typical convolutional networks learn robust feature representation and show much superiority in many down-stream tasks. These contribution supports the large potential of deep learning applications.

Image segmentation is an important subject in image understanding. Deep learning is also an powerful weapon in this field. Some classic frame like Mask R-CNN [3] achieves great performance in real world scenario. However, when turn back to microscopy image segmentation, models suffer from the lack of biological data and can't support lightweight application. Ronneberger [10] et.al. proposed a light framework UNet [10] as a solution to this problem. It absorbs the advantages of skip connection in ResNet to hierarchically connects the encoder-decoder modules. Such lightweight design leverages both information from low-level patterns and high-level semantics thus achieves great effects within limited biological data. It also causes a larger impact on many following works. Some UNet variants [14] [15] turn out to be more and more effective and efficient. One of the most influential network among them is UNet++ [15], or can be named as Nested UNet. It proposes a densely nested UNet with up-sampling and skip connection at each hidden level. The hierarchical utilization of feature representation at different scales can lead to a more robust prediction. These are both effective tools for microscopy image segmentation and boost the process of biological experiments.

The main reason for the huge success of UNet based segmentation is the light skip connection. It forces model to memorize the original local information from the low level and helps the segmentation. But we may wonder, does all low level information helps to enhance high level segmentation? This is really hard to say. But generally some low-level patterns always contain slight noise or unrelated local information, which leads to small harms on segmentation performance. To improve the model, we should more precisely consider which part from low level is better for supplementing high level representation. Some noisy feature pattern should be carefully forgotten while other essence should be passed forward to the decoder. Such idea can be easily achieved by attention mechanism, which attracts many researchers in recent studies. Attention map usually plays an important role in precisely selecting feature and turns out to be effective in many tasks.

Our contribution is applying attention mechanism into electron microscopy image segmentation. From the standard architecture of UNet [10] and UNet++ [15], we modified the skip connected in forward passing by an attention-based selective passing. Details, encoder feature of each pixel goes to attention mechanism with activation to generate attention map and filters feature by multiplication. Then, such cleaned feature is better for understanding high level semantics and get used by decoder. In the given dataset for electron microscopy image segmentation, our simple attention-based forward passing module affects well and enhances the performance of both UNet and UNet++ [15].

2 Related Work

Biological Image Segmentation

Classic image segmentation methods [3] are usually trained on large-scale database such as ImageNet [2]. They performs well when applied in real world scenario. However, in the domain of biological images, data source are usually limited and hard to afford the learning of large-scale applications. The most famous solution to this problem is UNet [10], which achieves great performance in biomedical image segmentation. The key point of UNet is the hierarchical skip connection (absorbed from ResNet [4]) from encoder modules to decoder modules. This design tackles the problematic gradient vanish in neural network and makes the whole model easy to train. The UNet architecture attracts the attention from many researchers who propose many advanced version of UNet. One of the influential variant of UNet is UNet++ [15], proposed by Zhou et.al. It uses a nested feature forward passing and up-sampling to leverage the advantages of skip connection. Zhang et.al. [14] proposed MDU-Net to densely connect various feature at multiple level. Jin et.al. [6] proposed DUNet which use three branches with deformable convolution to encode the finer-grained information of biological images. And for more interesting work, nnU-Net [5] became the first segmentation method that is designed to deal with the dataset diversity found in the domain. Simon et.al. [7] introduced an other probabilistic view of UNet, to deal with the segmentation of ambiguous images.

3 Approach

In this section, I will review the fundamental architecture UNet [10] and UNet++ [15] in head. Then I will introduce our modification for improvement. Last, I will claim the detailed learning protocol.

3.1 Review: UNet and UNet++

We illustrate the main architecture of UNet and UNet++ in Figure 1 and 2, cited from their original paper [10] [15].

The architecture of UNet [10] is in the shape of "U", consisting of encoders (in the left) and decoders (in the right). Both encoder and decoder block is 2-layer convolutional neural network. In the encoder side, the original image pass to a block to obtain the feature map. Then, its width and height shrink to half for down sampling via a Maxpooling layer. Then each block at next level will increase the feature channels. This indicates the resolution of feature is shrink to half after each level but the dimensionality is doubled. This results in an aggregated bottleneck of small resolution but large dimensionality. It contains rich global information and it's computationally efficient due to the limited size. The decoder module in right part is the inverse operation of encoders. At each level, it use a transposed convolution for up sampling from feature to double the size of the origin as well as reduce the feature dimensionality to half. The cleverest design of UNet is the

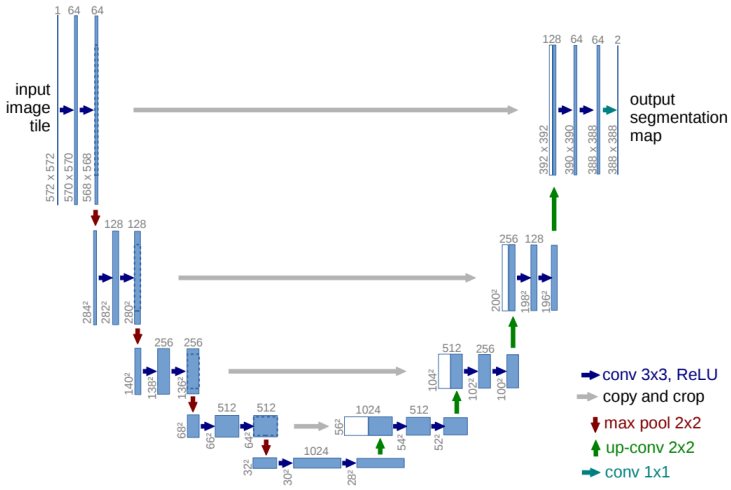

Figure 1: Architecture of UNet. Cite from original paper [10].

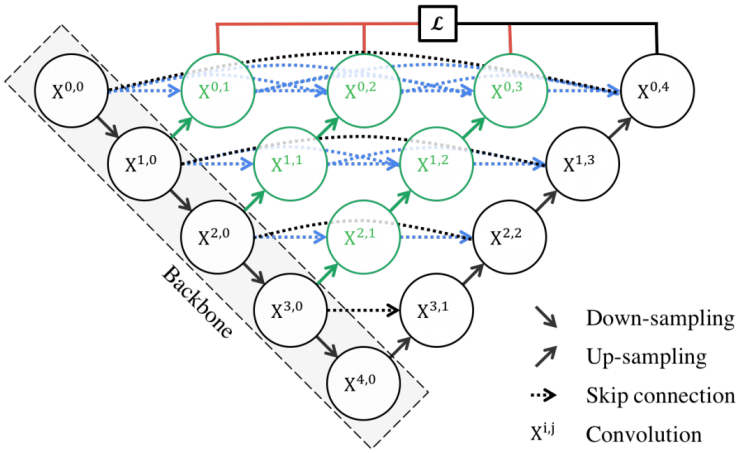

Figure 2: Architecture of UNet++. Cite from original paper [15].

skip connection, concatenating encoder feature before passing to the next upper level. It efficiently utilizes the original information without extra learnable parameters. Finally, the output feature of decoders is passed to a convolutional layer with unit kernel to classify the predicted labels. Generally, this "U" shaped network achieves great performance in bioinformatics.

UNet++ [15] is the advance of UNet [10]. The main success of UNet can be attributed to two reasons (1) Conversion between dispersed upper-layer and aggregated bottom-layer. (2) Forward passing of feature using skip connection. Inspired by these reasons, we consider to increase the upper-bottom interaction and skipped forward passing of feature. This is the main idea of UNet++ [15], illustrated in Figure 2. We use $X^{i,j}$ to denote the output feature at i-th level and j-th in forwarding path. We can find the most left or right feature at each level is the feature of original UNet, but we add multiple intermediate blocks. The source to each intermediate block is the up-sampling of previous block feature at lower level (left donw) with all previous feature at same level. For intermediate feature, it can be formulated as

$$Z^{i,j} = \text{concat}\left[\text{ up-sampling }(X_{i+1,j-1}), X_{i,j-1}, X_{i,j-2}, \cdots, X_{i,0}\right] \tag{1}$$

$$X^{i,j} = ConvBlock(Z^{i,j}) \tag{2}$$

$$j \geq 1 \text{ and } i + j \leq 3 \tag{3}$$

UNet++ use multiple classifiers on $X^{0,1}, X^{0,2}, X^{0,3}, X^{0,4}$ for final classification. Both the loss computation and the predication is averaged over these 4 feature maps. This gives a more robust estimation from various scaled feature.

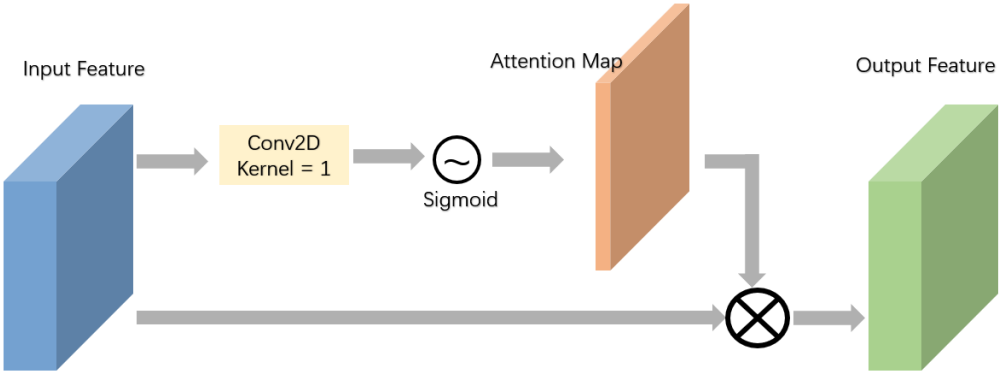

Figure 3: Illustration of the proposed attention mechanism.

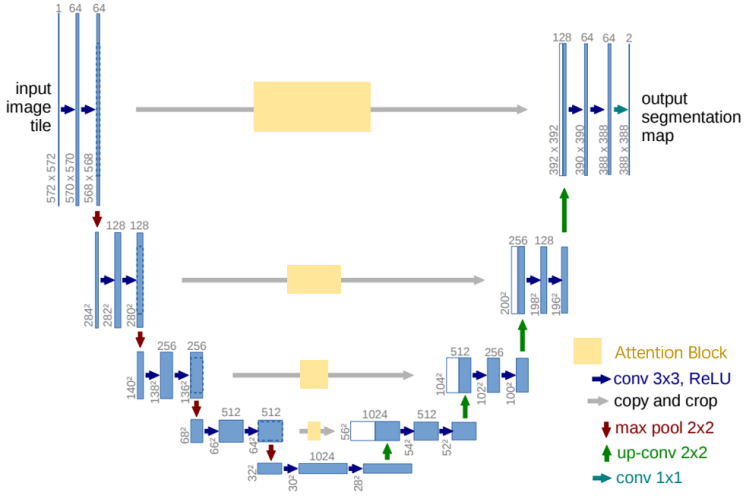

Figure 4: Integrate attention block into UNet. It can be integrated to UNet++ similarly.

3.2 Attention Mechanism in Forward Passing

Next, we introduce our improvement based on UNet [10] and UNet++ [15]. Our motivation is that, the original feature is less processed thus may be mixed with some noise. Then, directly forward passing the original feature to the decoder is not precise enough to support better performance. More than the simple skip connection, we want to use a more precise method to determine which feature should be passed to the decoder.

Recently, attention became one of the most important technique in deep learning community, which supports many applications such as Transformer [13]. Attention map is able to selectively choose what feature should be better passed forward, but what should be directly ignored. This helps decoder modules correctly focus on the valuable feature.

The detailed design of proposed attention mechanism is illustrated in Figure 3. On the upper branch, from the input feature, we use the convolutional layer of kernel size 1 followed by Sigmoid function to regress the attention map. It has same size with the input but only of single channel. Numerical values in the attention map is scaled to (0,1) due to the Sigmoid function. The attention map effectively determines which grid in the feature map is valuable for decoder modules. Then the attention map makes multiplication with the input feature to obtain the output, which will be passed to next

module in forward. Generally, attention mechanism filters feature of less importance but highlights valuable ones. It can greatly improve segmentation performance.

Our proposed attention mechanism can be easily integrated into UNet [10] and UNet++ [15]. We simply illustrate the attention integrated UNet in Figure 4. It only has small difference to the original UNet [10] in the middle forward passing phase. Besides, the attention mechanism can also be integrated into UNet++ [15], by densely replacing each skip connection with an attention mechanism.

3.3 Learning Protocol

Image Augmentation

Training network on the limited biological data still remains a difficult problem. Thus we consider to use some image augmentation technique to help the training. This is a valuable method to enrich the given data samples. It helps to avoid over-learning on the original limited data space and improve generalization ability on unseen data. The augmentation methods involved in our training are: (1) 90-degree rotation (2) horizontal/vertical flipping (3) Gaussian noise (4) image blurring (5) image distortion. First two methods (rotation and flipping) are about geometric operations on images. They lead to image with same semantics but different geometric form. Then, Gaussian noise and image blurring are used to produce polluted images. They help to enhance the robustness of network with less variance in training. Last one is about image distortion, which can be grid distortion. This can help the segmentation on the marginal pixels. Generally, image augmentation enlarges the space of image data and improves model performance.

Loss Function

We introduce the selection of loss functions from candidates in this part.

One of the most widely used loss functions is Cross Entropy loss L_{CE}, which can be generalized to any classification problems. It aims in minimizing the Cross Entropy between prediction and the ground truth. The formulation of Cross Entropy loss is in Eq. 4.

$$L_{CE}(\widehat{Y}, Y) = \frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \mathcal{C}_i} \left(Y_{i,c} \log(\widehat{Y_{i,c}}) - (1 - Y_{i,c}) \log(1 - \widehat{Y_{i,c}}) \right) \tag{4}$$

where $\widehat{Y}$ is the prediction and Y is the ground truth. N is the number of images in a minibatch and C_i is the set of pixels of i-th image.

Another widely used loss function in biomedical image segmentation is Dice loss [9], originally from VNet [9]. It views the network prediction and ground truth as two sets. And it minimize the discrepancy of these two set. It is formulated in Eq 5.

$$L_{Dice}(\widehat{Y}, Y) = 1 - \frac{2\,|\widehat{Y} \bigcap Y|}{|\widehat{Y}| + |Y|} \tag{5}$$

These two losses have their own advantages and differences in training. Dice gradient is directly towards the optimization direction of some evaluation metric like mIoU. But Cross Entropy is just a proxy in optimization. Thus Dice seems to be more suitable in our task. But definitely not, we found Cross Entropy loss works better in our experiment. Dice loss is more suffer from variance in training and we are hard to observe the convergence of the training process. Overall, we use Cross Entropy loss in the reported best model, but we make more analysis about the other design choice of loss function like Dice loss in the ablation study sections below.

Entire Learning Program

We claim the whole learning program of our method in Alg 1, in the pseudo-code format.

Algorithm 1: Entire Learning Program

Initialize $model \longleftarrow$ build up neural network.;
for $i = 1$ ***to*** max_epoch **do**
  for $j = 1$ ***to*** *data loader length* **do**
    $X, Y \longleftarrow$ sample a mini-batch of images and labels via data loader;
    $X \longleftarrow$ augment each image X;
    $\widehat{Y} \longleftarrow$ feedforward X in $model$ to obtain the final prediction;
    $loss \longleftarrow$ compute CE or Dice loss between $\widehat{Y}$ and Y;
    Back propagate $loss$;
    Using SGD optimizer to optimize $model$;

4 Experiment

4.1 Dataset

We use the given dataset of 30 electron microscopy images for following experiment. The whole dataset splits into 25 images for training and 5 for testing. Each image is a gray scale image of 512 * 512 resolution. The final labels for segmentation is of same size to the image and each pixel is classified into 2 classes.

4.2 Evaluation Metric

We implement four metrics for evaluation. They are pixel accuracy, mIoU, V^{rand} and V^{info} [1]. We use $\widehat{Y}$ and Y to denote the predicted class and ground truth. And let $p_{i,j}$ denote the probability of a random pixel is predicted in i-th class but of j-th class actually. We can get the marginal distribution $s_i = \sum_j p_{ij}$ and $t_j = \sum_i p_{ij}$. We take $\alpha = 0.5$ in both V^{rand} and V^{info} [1].

Pixel Accuracy directly measures number of correctly classified pixels. But it has strong bias in different distribution of ground truth categories. It is formulated as

$$Pixel\_Accuracy = \frac{1}{N \times |C|} \sum_{i=1}^{N} \sum_{c \in \mathcal{C}_i} \mathbf{1}[\widehat{Y}_{i,c} = Y_{i,c}] \tag{6}$$

mIoU is a less biased metric to ground distribution of each classes. For a specific category k, we compute the intersection over union between the set of pixels predicted as k and the set of pixels labeled k. Then take average value over all classes. For K classes in total, it's formulated as

$$mIoU = \frac{1}{K} \sum_{k=1}^{K} \frac{p_{i,i}}{s_i + t_i - p_{i,i}} \tag{7}$$

$\mathbf{V^{rand}}$ is used to measure two distributions via a randomly selected pixel. It is the harmonic mean of two component. One is the random split score but the other is the random merge score, constrained by marginal distributions.

$$V^{rand} = \frac{\sum_{i,j} p_{i,j}^2}{0.5 \sum_i s_i^2 + 0.5 \sum_j t_j^2} \tag{8}$$

$\mathbf{V^{info}}$ extends the randomly selection of pixels into an information theoretic view. It is measured by mutual information with constraints of marginal entropy.

$$V^{info} = \frac{\sum_{i,j} p_{i,j} \log(p_{i,j}) - \sum_i s_i \log(s_i) - \sum_j t_j \log(t_j)}{0.5\Big(- \sum_i s_i \log(s_i)\Big) + 0.5\Big(- \sum_j t_j \log(t_j)\Big)} \tag{9}$$

4.3 Implementation

We implement our framework in PyTorch and one Nvidia 1080Ti GPU is used in training. Our UNet [10] family networks have [64,128,256,512,1024] hidden channels at each level, while the resolution is power of halves to the original size. We defaultly use CrossEntropy loss in our main experiments due to its great performance and also report results with Dice loss [9] in ablation study. We use SGD optimizer to optimize the model with learning rate 1e-3. The batch size is 4 images for UNet but 1 image for UNet++ since the latter in nested architeture requires more GPU memory. We evaluate model after each epoch with 500 epoches in total, and report best score among them.

Network	Use Attention	Pixel Acc	mIoU	V^{rand}	V^{info}
UNet [10]	✗	0.9278	0.8028	0.9018	0.5153
UNet++ [15]	✗	0.9268	0.7976	0.9013	0.5071
UNet [10]	✓	**0.9282**	**0.8043**	**0.9021**	**0.5181**
UNet++ [15]	✓	0.9267	0.8009	0.9016	0.5118
Best		**0.9282**	**0.8043**	**0.9021**	**0.5181**

Table 1: Results of UNet [10] and UNet++ [15] w/w.o. attention.

	Pixel Acc	mIoU	V^{rand}	V^{info}
UNet + Attention	**0.9282**	**0.8043**	**0.9021**	**0.5181**
without Attention	0.9278	0.8028	0.9018	0.5153
without Augmentation	0.9247	0.7966	0.8976	0.5034
Dice Loss	0.9270	0.8014	0.9006	0.5125

Table 2: Results of ablation study

4.4 Result

We conduct detailed experiments on each network architecture setting and provides results in Tab. 1. Our methods turn out to work well according to the score. And we can observe some interesting phenomenon from Tab. 1.

First, we can easily get UNet [10] based networks perform better than UNet++ [15] based networks. This is mainly because UNet++ is densely nested with more parameters then UNet. The given data space is limited thus model is easily to get overfitting on the data space. Thus a lightweight model is better in this task.

Second, we validate the effectiveness of our proposed attention block. Both UNet and UNet++ with a attention module gains a slight performance improvement. It proves that we can use an extra attention mechanism to enhance the segmentation performance of UNet and variants.

Last, we make vertical comparison over different evaluation metrics. We can find that the metric of pixel accuracy is extremely high over others. This is because the it's strongly biased with distribution of the ground truth. Besides, the V^{info} is really a hard metric in scoring. The errors are usually scaled and expanded by $\log$ function, leading to a relatively low score.

4.5 Ablation Study

We conduct ablation study to verify each component. We start from the base model of UNet with attention, which is of best. Then we do some operations including attention mechanism, image augmentation and loss function.

First when removing the attention mechanism, the performance trends to drop a small margin. This is because the attention based forward passing is a more precise way than naive skip connection in passing feature from encoder to decoder.

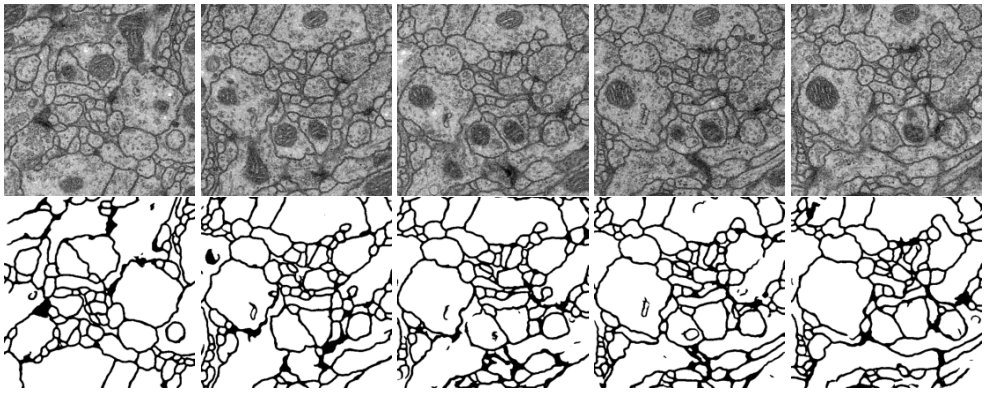

Figure 5: Visualization of segmentation maps of 5 test set images.

Second, when removing the image augmentation, the network also degrades a lot. This validates image augmentation technique enlarging the training data space. It helps model to be of less variance and performs robustly to unseen testing data.

Third, we we try to use different loss functions like Dice loss [9]. You can refer to Sec.3.3 for discussion about two loss functions. Unfortunately, it doesn't live up to our expectation, with a small performance degradation from Cross Entropy trained model. We find the scale of Dice loss is much smaller to the Cross Entropy loss and guess this is mainly because the small gradient blocks the model in a local optimum.

4.6 Visualization

We visualize the segmentation maps of 5 test set image in Figure 5.

5 Conclusion

In this project, we are dealing with the challenging electron microscopy image segmentation. Traditionally, UNet [10] and its variants work very well in this field. This is mainly because of the advantages of lightweight skip connection, which directly passes the encoder feature into decoders. However, this approach is not very precise, with some noise in the input feature also involved in forward passing. To tackle this problem, we are inspired by attention mechanism and adopt attention blocks to achieve selective forward passing. This method helps model focus on feature of more importance to improve the decoding phase. We do experiments on the given electron microscopy image dataset and give detailed analysis about the results. Fortunately, our proposed attention mechanism block works very well and enhance of original performance of UNet [10] and UNet++ [15].

References

[1] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018.

[6] Qiangguo Jin, Zhaopeng Meng, Tuan D. Pham, Qi Chen, Leyi Wei, and Ran Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019.

[7] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images, 2019.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[9] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[14] Jiawei Zhang, Yuzhen Jin, Jilan Xu, Xiaowei Xu, and Yanchun Zhang. Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation. *arXiv preprint arXiv:1812.00352*, 2018.

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.