Cell Image Segmentation by Feature Random Enhancement Module

No Author Given

Paper ID 25

Abstract. In order to perform semantic segmentation with high accuracy, it is important to extract good features using an encoder. Although loss function is optimized in training deep neural network. far layers from the layers for computing loss function are difficult to train. Skip connection is effective for this problem but there are still far layers from the loss function even if we use skip connection. In this paper, we propose the Feature Random Enhancement Module which enhances the features only in training. By emphasizing the features at far layers from loss function, we can train those layers works well and the accuracy was improved. In experiments, we evaluated the proposed module on two kinds of cell image datasets, and our module improved the segmentation accuracy without increasing computational cost in test phase.

Keywords: cell,Semantic Segmentation,U-Net

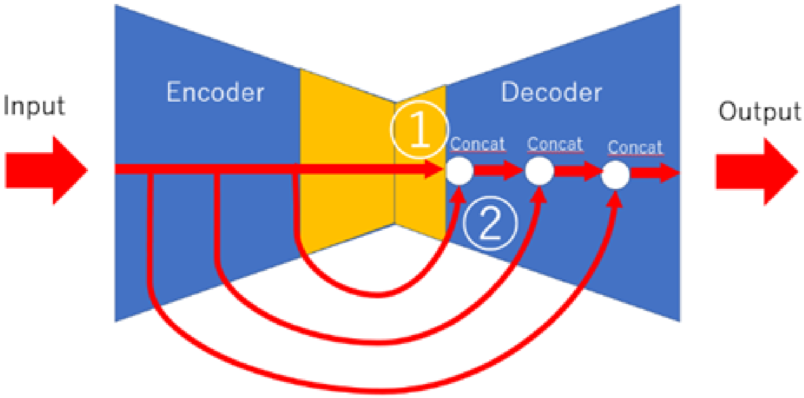


Fig. 1. U-Net and the problem

1 Introduction

In recent years, the development of deep learning technology has been remarkable, and there is a demand to use it in various situations. Among them, the segmentation of cell images obtained by microscopes, which is performed manually by human experts and tends to produce subjective results, requires objective results by the same criteria using deep learning technology [1]. However, the optimal network for segmentation using deep learning has not been established yet. Even if the accuracy is not so high, it is actually used in the field of cell biology to obtain objective results. Therefore, automatic segmentation method with high accuracy is desired. U-Net is still widely used for segmentation of microscope images because the computational cost is not high, it works well for small number of training images and high accuracy is obtained without adjusting hyperparameters. For this reason, many improvements of U-Net have been proposed for microscope images [6-8].

This study belongs to one of those variations and improves the accuracy of U-Net. While conventional improvement is done by deepening, the proposed method in this paper does not require any additional computational resources at all during inference. Therefore,it retains the advantage of U-Net in that it requires fewer computing resources. Therefore, it is a very significant proposal in the segmentation of medical images where there is a demand for lightweight and accurate models.

A neural network such as CNN basically uses backpropagation of loss for training. For this reason, there is a phenomenon that near layers to the layer for computing loss are more updated in comparison with far layers [3]. In order to solve the problem, ResNet [4] used skip connection and improved the accuracy. U-Net [2] is famous deep neural network for cell segmentation task. U-Net also has skip connection between encoder and decoder. It contributes to improve the accuracy. In general, it is well known that skip connection gives the information of location and fine objects which were lost in encoder to decoder. However, we consider that the same theory as ResNet is used in skip connection to improve the accuracy. By using skip connection, the loss is propagated to encoder well, and weights are successfully updated. This is also the reason why U-net improved the accuracy in comparison with standard Encoder-Decoder CNN.

Figure 1 shows the structure of U-net. We see that skip connection is effective to propagate the loss to encoder. However, the layers shown as yellow in the Figure are the farthest from the loss at the final layer. Therefore, in the case of U-net, the yellow layer in Figure 1 is the most difficult to train though the layer has semantic information. In this paper, we propose new module to train those layers effectively.

We consider to enhance the feature map at yellow layer which is the farthest layer from the loss function. Since the yellow layer is difficult to learn, network learns to decrease the feature values at the yellow layer not to affect the output. In order to alleviate the difficulty of learning, we select some feature maps randomly at yellow layer and increased the absolute value of the feature map multiplied by a large constant value. This allows the features at yellow layer to be used effectively for segmentation.

In experiments, we evaluated our method on two kinds of cell image datasets. Intersection over Union (IoU) is used as an evaluation measure. The effectiveness of the proposed module was shown in comparison with the conventional U-Net without our module and U-net with SuperVision that loss function is computed at yellow layer.

The structure of this paper is as follows. Section 2 describes related works. Section 3 describes the details of the proposed method. Experimental result on two kinds of cell image datasets are shown in section 4. Finally, we summarize our work and discusses future works in section 5.

2 Related works

U-Net is a kind of Encoder-Decoder CNN [5]. Unlike the PSPNet [9], the Encoder-Decoder CNN does not use features in parallel, but features are extracted in series. Thus, in Encoder-Decoder CNN, far layers from the layer for computing loss are not updated well. ResNet and U-net solved this problem by skip connection.

There is also a technique called Deep supervision proposed in Deeply-Supervised Nets [10] to address the problem. In deep supervision, loss is also computed at middle layer. Far layers from final layer are updated well by supervision. U-Net++ [11] also used this technique. However, forcing loss from the ground truth in the middle of U-Net may not obtain an intermediate representation for better inference. In addition, U-Net ++ has a structure in which the output image is restored by the decoder from various parts of the encoder, and the decoders are connected to each other. However, the advantage of U-Net, which is a small computational resource, is vanished. This is accompanied by a large increase in the number of parameters due to multiple decoders. In this paper, we propose new methods based on the merits and demerits of these previous studies.

There are some methods that we referred to consider a new method. In the proposed method, feature enhancement is performed on some feature maps during training. There are many techniques for weighting feature maps. SENet [12] proposed to weight important channels. Attention, which has been proposed in the field of natural languages [13], is also used in the field of images. In recent years, many methods have been proposed that focus on channels [14-16]. Attention-U-net used attention for skip connection [17].

Dropout [18] is also related to our approach. Dropout sets a part of the feature map to 0 in only training. This prevents overfitting by randomly removing elements only during training. Our method randomly enhances some feature maps at farthest layer from loss function. We do not set some elements to 0 and enlarge some feature maps. When some elements are set to 0, backpropagation from the element is stopped. In our method, features are enlarged randomly to use backpropagation effectively for the farthest layer.

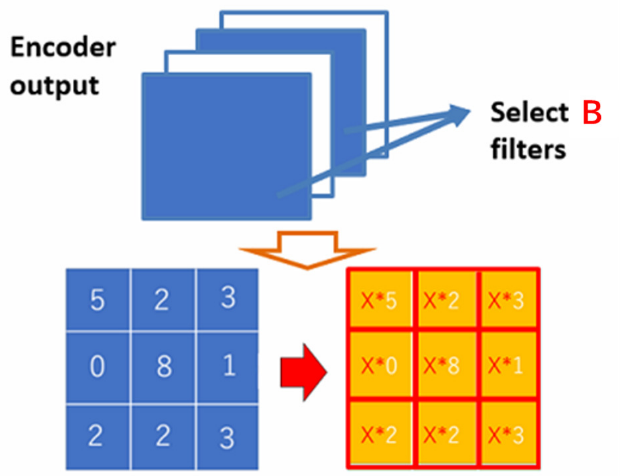


Fig. 2. Feature Random Enhancement Module

3 Proposed method

This section describes the proposed method. Section 3.1 gives the overview of the proposed method. Section 3.2 mentions the implementation details of our method.

3.1 Overview of the proposed method

When we obtain segmentation result by U-Net, the magnitude of features at yellow layer as shown in Fig. 1 is often smaller than that of features at skip connection from encoder to decoder. Fig.3 shows the fact when U-net is trained on two different datasets. Two lines in each graph show the average feature values at yellow layer in Fig.1 and those at skip connection from encoder to decoder. Note that both features are extracted after ReLU function. Since those two feature maps are concatenated in the U-net, the magnitude of features should be similar. Fig.3 shows that the encoder's output is obviously smaller than the features at skip connection. This demonstrates the yellow layer is not trained well because the layer is farthest from loss function.

Table 1. mIoU due to differences in enhancement at test phase

	Transgenic mouse cell dataset	Drosophila dataset
baseline(U-Net + SEblock)[%]	59.50	73.98
the random enhancement in test phase[%]	60.52	76.85
without random enhancement in test phase[%]	61.62	76.95

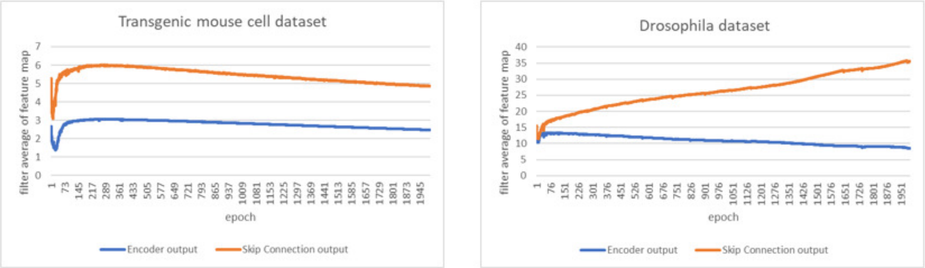


Fig. 3. Comparison of feature values at the yellow layer and skip connection

Both features at yellow layer and skip connection are concatenated and normalized by batch renormalization. This normalizes like to 0 mean and 1 variance. In addition, the perfect average of 0 means and 1 variance is not effective, so there are scale and shift parameters learned.

After normalization, the yellow layer still has small features in comparison with skip connection though yellow layer has semantic information.

Does this fact show that yellow layer is not required? Our answer is “NO”. This phenomenon is caused from that near layers to the layers for computing loss are updated well and far layers are not updated well. Yellow layer in Figure 1 is the farthest layer from loss function because encoder is updated through skip connection. Therefore, network learns he layers connected by a skip connection in comparison with the yellow layer because it is difficult to update the yellow layer. Thus, normalization do not work well if there is a difference in the ease of updating such as between near and far layers.

Therefore, the proposed method emphasizes the features at the farthest layer from loss function because it is not trained well by back-propagation. Main purpose of our proposed module is to train the farthest layer well. Thus, this enhancement is used in only training.

Although we show the result in section 4.4, the proposed module emphasizes the outputs of the selected filters but it affects the non-enhanced filters. The outputs of non-enhanced filters enlarged automatically. Furthermore, Table 1 in section 4 shows that the accuracy is improved when we do not enhance features in test phase. Surprisingly, the network without enhancing features in test phase is better than the network with enhancing features. This is because features at the farthest layer from the loss function are already enlarged by our module in training phase.

From this result, At test phase, we think that emphasizing all the parallel filters is fatal break the balance in training phase. Also, with the same random enhancement as in training, the output of the selected filter will be enhanced, but it will be more accurate if it is not enhanced during testing.Therefore, we think that the impact of enhancement during training is greater for the non-enhanced filter than for the selected filter.Thus, we do not need to use our module in test phase.

Thus, the computational cost of the network does not increase in test phase. However, the accuracy is improved without changing the inference time or computational resources. This is an advantage of the proposed method because many methods deepened the network to improve the accuracy.

To describe the proposed method for U-Net, the encoder's output is enhanced by multiplying feature maps selected randomly by X. The number of feature maps selected randomly is denoted as B. The feature maps are re-selected each epoch and the network weight is updated during training.

The closest approaches is Dropout. Similarly, dropout is used only during training, and some neurons are randomly set to 0. If there is an element set to 0, the backpropagation stops at that element. It is a method of learning in a small network and allowing for ensembles. The proposed method differs from Dropout. We use an adjustable magnification emphasis not setting to 0. This is to improve the case where there is a difference in the ease of updating between the near and far layers.

The proposed method can be implemented in addition to Dropout. However, this does not mean that Dropout will be replaced by the proposed method. In addition to the same proportion of parameters to be acted upon as in Dropout, there is a parameter that is difficult to learn. It is the magnification of the emphasis. Therefore, it is difficult to implement it in many places. Implementing it at the farthest layer from the loss function solves the problem presented in this paper and is the most effective.

Figure 2 shows the detailed description of the proposed method. In the proposed method, we multiply X by all values in the selected feature maps which are the end of encoder shown as yellow layer in Figure 1. This operation is performed only in training phase. The enhanced feature maps are selected randomly. Thus, all channels in encoder's output are not enhanced. We need to select hyperparameters X and B appropriately. Hyperparameters were searched by using the optimization with Tree-structured Parzen Estimator (TPE) [19] which is a new method in Bayesian optimization.

4 Experiments

This section shows the experimental results of the proposed method. Section 4.1 describes the dataset used in experiments, and the experimental results are shown in section 4.2. In section 4.3, additional experiments are conducted for considerations.

4.1 Dataset

In this paper, we conduct experiments on two kinds of cell image datasets. The first dataset includes only 50 fluorescent images of the liver of a transgenic mouse expressing a fluorescent marker on the cell membrane and nucleus [20]. The size of the image is 256 256 pixels and consists of three classes; cell membrane, cell nucleus, and background. We use 35 images for training, 5 images for validation, and 10 images for test.

The second dataset includes 20 Drosophila feather images [21]. The size of the image is 1024 x 1024 pixels and consists of four classes; cell membrane, mitochondria, synapse, and background. Training and inference were performed by cropping 16 images of 256 256 pixels from one image without overlap due to GPU memory size. Intersection over Union (IoU) and Mean IoU were used as evaluation measure for both datasets.

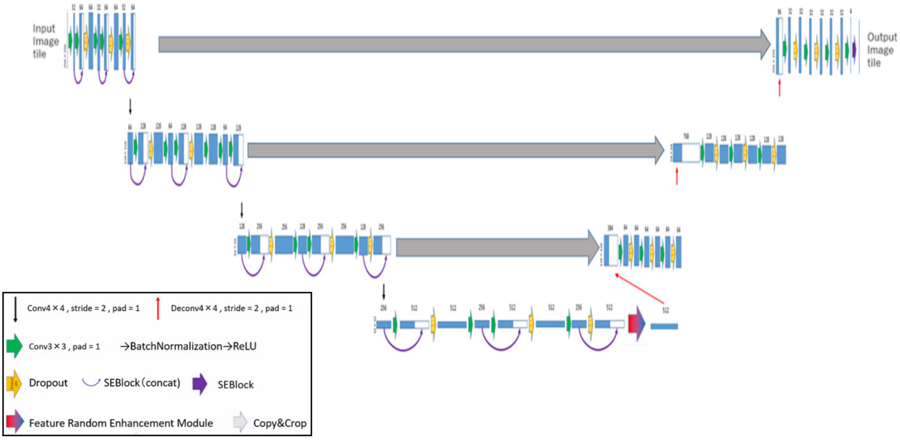


Fig. 4. U-Net with SE block

4.2 Implementation details

The proposed method introduces a module that randomly enhances the features at the final layer of encoder during training. We call it "Feature Random Enhancement Module". Fig. 3 shows the U-Net used in this paper. As shown in Fig. 3, the proposed module was implemented on a standard U-Net with SE block.

Some feature maps are selected from 512 feature maps at the farthest layer from the loss function which is shown as the bottom right in Fig. 4 at training phase, and the value in the feature map is multiplied by X.

4.3 Results

In all experiments, we trained all methods till 2000 epochs in which the learning converges sufficiently, and evaluation is done when the highest mIoU accuracy is obtained for validation images. We used softmax cross entropy. The hyperparameters B and X were searched 50 times using the Tree-structured Parzen Estimator algorithm (TPE), which seems to be a sufficient number.

For comparison, we evaluate U-Net with only SE block. This is the baseline. We also evaluate the U-net with SE block which uses SuperVision instead of the proposed module in order to show the effectiveness of Feature Enhancement module. SuperVision uses 1x1 convolution to change three or four channels at the end of encoder, and resize it to the size of input image and softmax cross entropy loss is computed.

When we use SuperVision, we must optimize two losses; the first loss is standard softmax cross entropy loss at the final layer and the second loss is for supervision. In general, the balancing weight for two losses should be optimized.

$$\text{Loss}=(1\text{- }\lambda)\text{*Loss.1}+\ \lambda\text{*Loss.2}(1)$$

where is λ the balancing weight. The parameter is also optimized by TPE. The search was performed 15 times to find the appropriate parameter. Since λ is a single parameter, the number of searches is smaller than that of the two parameters B and X in our method.

First, we describe the experimental results for the mouse cell dataset. Table 2 shows the results when we use B = 162 and X = 632 which gives the highest mean IoU for validation set. From Table 2, the accuracy of the proposed method is improved in all classes compared with the conventional U-Net with SE block, and 2.12% improvement on mean IoU is confirmed.

The accuracy of mean IoU is not improved by U-Net with SuperVision ($\lambda = 0.3257$ determined by TPE) even if loss is computed at the end of encoder that our module is used. Figure 5 shows the qualitative results. In Fig. 5, (a) is input image, (b) is ground truth, (c) and (d) are the results by the conventional U-Net with SE block and U-Net with SE block and SuperVision, respectively, and (e) is the result by the proposed method. We see that cell image is blurred and it is difficult for not experts to assign class labels. This is because cells are killed by strong light and images are captured with low illuminance.

In conventional method (c), there are many undetected and over detection of cell membrane or nucleus. In addition, in conventional method (d), furthermore, there are many undetected membranes. However, in the proposed method (e), more accurate segmentation results are obtained. This is because the proposed module enables to extract features from areas where training has not been done successfully in conventional method. In addition, the method using SuperVision gave lower accuracy than the proposed method. The encoder output is in the center of the network with lower resolution and many channels. I do not consider that forcing the loss from the middle output with ground truth image will always give an intermediate representation for good segmentation result.

Next, the experimental results are described for the Drosophila dataset. Table 3 shows the results when we use B =8 and X =250 when the highest mIoU is obtained for validation set. Table 3 shows that the accuracy of the proposed method is higher than that of the U-Net with SE block, and the mean IoU was improved by 2.97%. Furthermore, we see that the accuracy is improved in comparison with the U-Net with SE block and SuperVision($\lambda = 0.2781$ determined by TPE).

Figure 6 shows qualitative results. In Fig. 6, (a) is the input image, (b) is ground truth, (c) and (d) are the results by the U-Net with SE block and U-Net with SE block and SuperVision, and (e) is the results by the proposed method. In the Drosophila dataset, the image seems to contain enough information but it is difficult for ordinary people to assign correct class labels to each pixel. However, we confirmed that the proposed method (e) performs better segmentation for menbrane, nuclear in cropped image's edge, synapse with small area.

Figure 7 and 8 show the results of hyperparameter search using the TPE algorithm. The vertical and horizontal axes show the hyperparameters B and X in the proposed module. Red points mean high mean IoU for validation set, and the blue points mean low accuracy. We see that the TPE algorithm focuses on searching for places with high accuracy. Of course, optimal B and X depend on the dataset. However, we can find good hyperparameters by TPE.

Table 2. IoU of Transgenic mouse cell dataset

	menbrane[%]	nuclear[%]	background[%]	mIoU[%]
U-Net + SEblock	37.78	65.75	74.96	59.50
U-Net + SEblock + Supervision	39.23	64.99	73.34	59.19
Proposed method	40.53	67.58	76.75	61.62

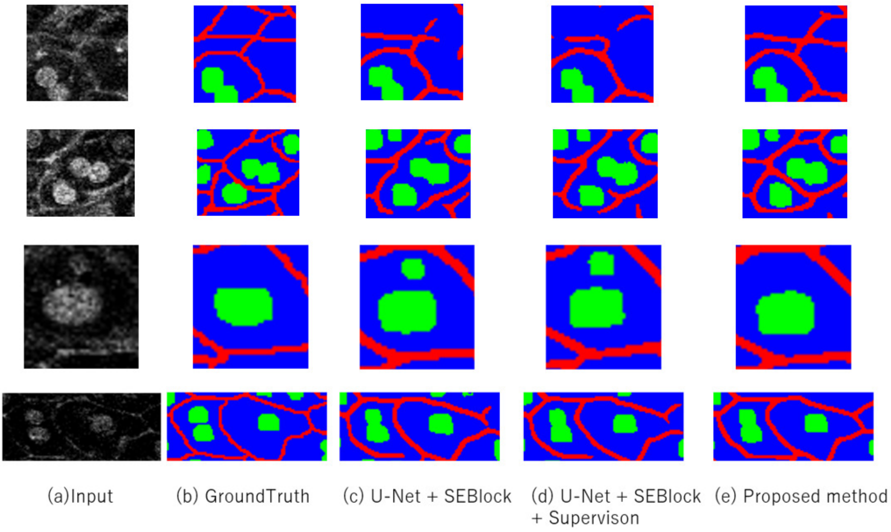


Fig. 5. Segmentation results on transgenic mouse cell images

Table 3. IoU of the Drosophila dataset

	menbrane[%]	nuclear[%]	background[%]	Synapus[%]	mIoU[%]
U-Net + SEblock	91.80	76.87	76.89	50.46	73.98
U-Net + SEblock + Supervision	92.39	77.77	78.24	52.21	75.15
Proposed method	92.93	78.71	78.02	58.14	76.95

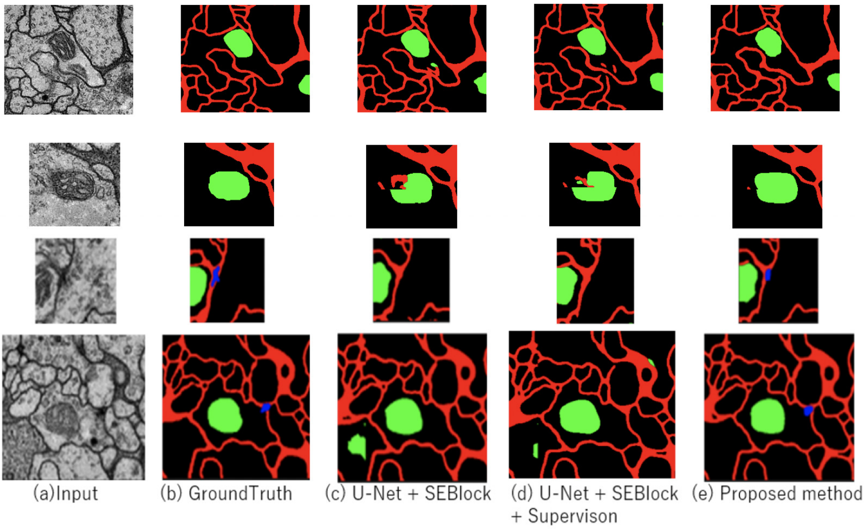


Fig. 6. Segmentation results on the Drosophila dataset

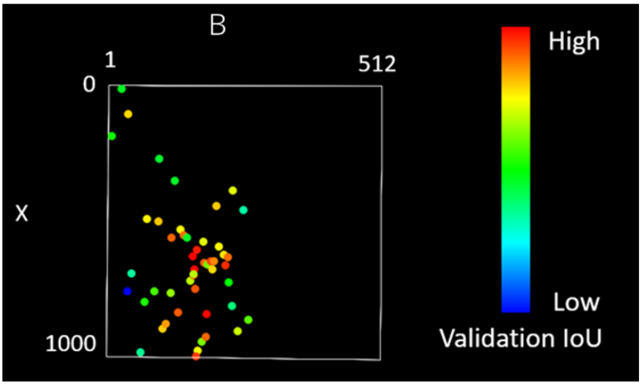


Fig. 7. TPE algorithm of Transgenic mouse cell

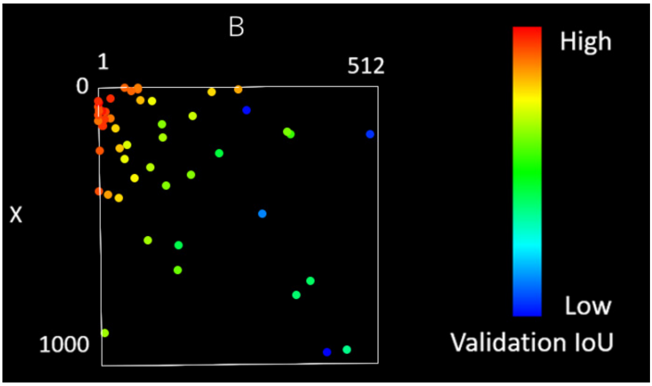


Fig. 8. TPE algorithm of Drosophila feather

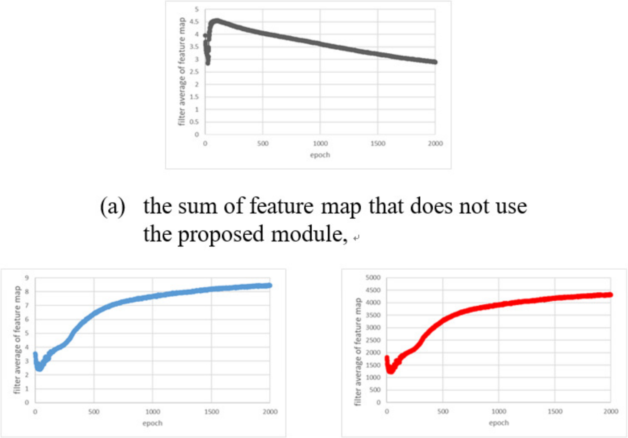
4.4 Additional Experiments

The proposed method emphasizes some feature maps randomly at each epoch to prevent over-fitting. However, as shown in Table 4, applying 10,000 enhancements to the ten filters fixed during training can improve IoU accuracy by about 1%. We observe the sum of the values in the feature map that ReLU function is used after convolution.

Figure 9 (b) shows the sum of emphasized feature map by the proposed module, and (c) shows the sum of feature map that is not emphasized though the proposed module is used. In (a), the sum of feature map gradually decreases and it is no longer used for output. On the other hand, in (b) and (c), the sum of feature map increased through training. This means that the feature maps at the end of encoder have large value automatically and those features are used to obtain segmentation results. The proposed method emphasizes some feature maps randomly at each epoch to prevent over-fitting. Therefore, from the change of the value of (c) compared with (a), it can be said that the proposed module has an effect on the feature maps that are not emphasized.

Table 4. For additional experiment,IoU of Transgenic mouse cell dataset

	menbrane[%]	nuclear[%]	background[%]	mIoU[%]
U-Net + SEblock	37.78	65.75	74.96	59.50
Proposed method(additional experiment)	40.43	67.34	73.71	60.49



(a) the sum of feature map that does not use the proposed module, ↵

(b) the sum of emphasized feature map by the proposed module ↵

(c) the sum of feature map that is not emphasized though the proposed module is used. ↵

Fig. 9. Sum of feature map with/without Feature Enhancement module

5 Conclusion

In this paper, we introduced the Feature Random Enhancement Module, which is enhanced feature map randomly only during training, and succeeded in improving the accuracy on cell image segmentation. We could propose the method for improving accuracy though the amount of computation during inference does not change.

A future task is to establish a method for deriving the parameters of the proposed module. Although TPE seems to be effective for parameter search from the results, it requires training for each parameter until the accuracy converges. Therefore, the computational cost for inference is fast but training takes longer time. Thus, we would like to study whether parameters can be determined faster without convergence.

References

1. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. Advances in Neural Information Processing Systems. pp.2852–2860 2012.

2. Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention, Springer, LNCS, Vol.9351: 234-241, 2015.
3. Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, Vol.5(2), pp.157–166.1994.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. Conference on Computer Vision and Pattern Recognition. 2016.
5. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia. Pyramid scene parsing network. Conference on Computer Vision and Pattern Recognition 2017.
6. .Zhang, Jiawei, et al. MDU-Net: Multi-scale Densely Connected U-Net for biomedical image segmentation. arXiv preprint arXiv:1812.00352 .2018.
7. Jin, Qiangguo, et al. DUNet: A deformable network for retinal vessel segmentation. Knowledge-Based Systems 178 149-162. 2019.
8. Jin, Qiangguo, et al. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. arXiv preprint arXiv:1811.01328. 2018.
9. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia. Pyramid scene parsing network. Conference on Computer Vision and Pattern Recognition. 2017.
10. Lee, Chen-Yu, et al. Deeply-supervised nets. Artificial intelligence and statistics. 2015.
11. Zhou, Zongwei, et al. ”Unet++: A nested u-net architecture for medical image segmentation.” Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp.3-11,2018.
12. Hu, Jie, Li Shen, and Gang Sun. Squeeze-and-excitation networks. Conference on Computer Vision and Pattern Recognition. 2018.
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017.
14. Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, Yun Fu. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. pp. 286-301, 2018.
15. Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3-19, 2018.
16. Yanting Hu, Jie Li, Yuanfei Huang, Xinbo Gao. Channel-wise and Spatial Feature Modulation Network for Single Image Super-Resolution. arXiv:1809.11130, 2018.
17. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, Daniel Rueckert. Attention U-Net: learning where to look for the pancreas. International Conference on Medical Imaging with Deep Learning . 2018.

18. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, Vol.15, No.1, pp.1929-1958, 2014.
19. Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. Algorithms for hyper-parameter optimization. Advances in neural information processing systems. 2011.
20. A. Imanishi, T. Murata, M. Sato, K. Hotta, I. Imayoshi, M. Matsuda, and K. Terai, "A Novel Morphological Marker for the Analysis of Molecular Activities at the Single-cell Level," Cell Structure and Function, Vol.43, No.2, pp.129-140, 2018.
21. Segmented anisotropic ssTEM dataset of neural tissue. Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, Richard Fetter. figshare. Retrieved 16:09, (GMT) Nov 20, 2013.