Cell Image Segmentation by Integrating Multiple CNNs

Yuki Hiramatsu[1], Kazuhiro Hotta[1], Ayako Imanishi [2], Michiyuki Matsuda[2] and Kenta Terai[2]
[1] Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan
[2]Kyoto University, Yoshida-Konoecho, Sakyo-ku, Kyoto 606-8501, Japan
140442113@ccalumni.meijo-u.ac.jp, kazuhotta@meijo-u.ac.jp, imanishi.ayako.38a@st.kyoto-u.ac.jp,
{matsuda.michiyuki.2c, terai.kenta.5m}@kyoto-u.ac.jp

Abstract

Convolutional Neural Network is valid for segmentation of objects in an image. In recent years, it is beginning to be applied to the field of medicine and cell biology. In semantic segmentation, the accuracy has been improved by using single deeper neural network. However, the accuracy is saturated for difficult segmentation tasks. In this paper, we propose a semantic segmentation method by integrating multiple CNNs adaptively. This method consists of a gating network and multiple expert networks. Expert network outputs the segmentation result for an input image. Gating network automatically divides the input image into several sub-problems and assigns them to expert networks. Thus, each expert network solves only the specific problem, and our proposed method is possible to learn more efficiently than single deep neural network. We evaluate the proposed method on the segmentation problem of cell membrane and nucleus. The proposed method improved the segmentation accuracy in comparison with single deep neural network.

1. Introduction

Convolutional Neural Network (CNN) [1, 2] achieved very high accuracy on various kinds of image recognition problems. In addition, semantic segmentation using CNN [1, 2] is beginning to be applied not only automatic driving [3, 14] but also medicine and cell biology [4, 6]. Semantic segmentation is to assign labels to all pixels in an image. For semantic segmentation, Fully Convolutional Neural Network (FCN) [5] and Encoder-Decoder structure such as U-Net [6] had been proposed. Especially the network structure of U-Net [6] is proposed as a segmentation method for cell images. However, they used single deep neural network [7, 8]. It is difficult for single deep neural network to improve the accuracy further because many devices are already done.

To solve this problem, we use the idea of Mixture of Experts (MoE) [9] as the base of the proposed method. MoE [9] consists of Gating network and Expert network. Gating network divides a complex problem into some sub-problems and assigns a sub-problem to each Expert network. By using this idea, it is possible to train a network more efficiently than single deep neural network. However, it is difficult to obtain high accuracy because MoE [9] uses a simple perceptron. In recent years, MoC-CNN [10] which used CNN instead of perceptron was proposed for crowd counting. MoC-CNN [10] improved the accuracy in comparison with the VGG-16 [11] which is one of the state-of-the-art single deep neural networks though the number of parameters of MoC-CNN [10] (170M) is much smaller than that of VGG 16 (1380M) [11]. This shows the effectiveness of the integration of multiple CNNs.

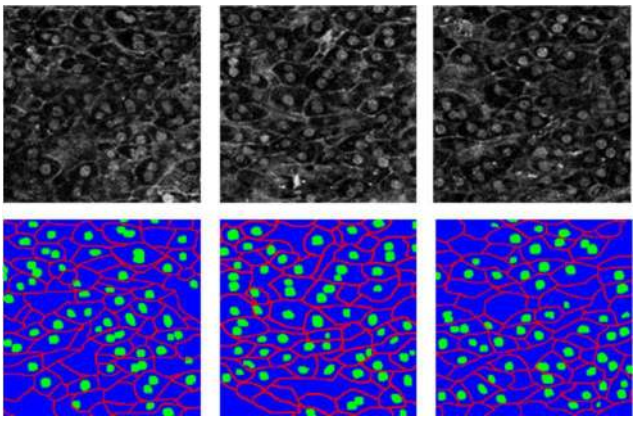

Figure 1. Examples of fluorescence images of the liver of transgenic mice that expressed fluorescent markers on the cell membrane and nucleus. The upper row shows input images and the lower row shows ground truth.

We propose a semantic segmentation method by integrating multiple CNNs based on the structure of MoE [9]. Figure 3 shows the overview of the proposed method. This method consists of Gating network and Expert networks. Both networks use Encoder-Decoder structure as U-Net [6]. There are multiple Expert networks whose outputs are segmentation results. In the Gating network, the feature maps obtained at the last convolutional layer of all Expert networks are used as input because Gating network should be higher-layer than Expert networks for assigning roles to Expert networks. Since Gating network assigns the weights to each pixel, the output of Gating network is the

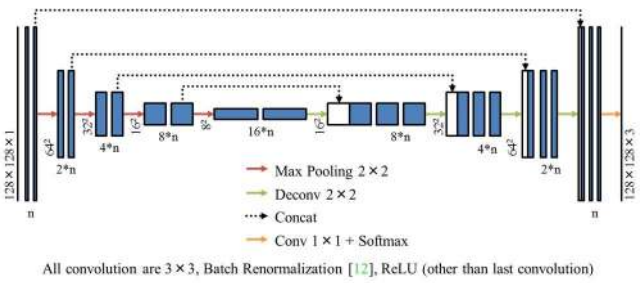

Figure 2. Overview of the conventional method. n indicates of the initial number of filters.

feature map with the same resolution as the segmentation results while the output of Gating network in MoC-CNN [10] corresponds to the number of Expert networks. We multiply the output of Gating network and the outputs of all Expert networks, and final segmentation result is obtained.

In experiments, we segment cell nucleus and membrane from microscopy images. Figure 1 shows the examples of them. We use IoU (Intersection over Union) and mean IoU as evaluation measures. We confirmed that the proposed method gave higher accuracy than that of single deep network. Furthermore, we investigated how Gating network automatically divides an input image into sub-problems and assigns roles to Expert networks.

This paper is organized as follows. In section 2, we describe related works. The details of the proposed method are described in section 3. In section 4, we evaluate our proposed method on segmentation of cell nucleus and membrane. Finally, we describe conclusion in section 5.

2. Related Works

This section describes related works. We explain Encoder-Decoder structure in Section 2.1. Mixture of Experts [9] is explained in Section 2.2. Finally, we explain Mixture of CNN [10] in Section 2.3.

2.1. Encoder-Decoder structure

U-Net [6] has been proposed as a segmentation method using CNN. U-Net [6] is often used for segmentation of biomedical images. The network structure of this method adopts encoder-decoder structure. In encoder, features in an image are extracted by convolution and pooling. In decoder, the segmentation result is constructed step by step while keeping their features. Furthermore, skip connection is introduced for each resolution. Skip connection in U-net [6] is to concatenate the feature map obtained by encoder with the feature map with the same resolution at decoder part. As a result, fine information which is lost at feature extraction process can be used effectively. In experiments, we use U-Net [6] as the conventional method.

Figure 2 shows the overview of the conventional method. Conventional method gave the best accuracy when initial number of filters "n" in Figure 2 is set to 32. Even if the number of layers increased, the accuracy did not improve while the number of parameters much increased. Therefore, we consider that this is the limit of a single deep neural network.

2.2. Mixture of Experts

Mixture of Experts (MoE) [9] was proposed by Jacobs et al. [9]. This is a model oriented toward a strategy of dividing a complex problem into sub-problems and solving simple sub-problems. MoE [9] consists of Expert networks E $\{E_1 \cdots E_m\}$ and Gating Network G whose output is m dimensional vectors. Input size and output size of all Expert networks must be the same but the structure of each Expert network does not need to be the same. The output y of MoE [9] is represented as

$$y = \sum_{i=1}^{m} G(x)_i E_i(x) \tag{1}$$

$$G(x) = \mathrm{Softmax}(x \cdot W_g) \tag{2}$$

where $\mathrm{G}(x)$ is the output of Gating network and $\mathrm{E}_i(x)$ is the output of the *i*-th Expert network. This is equal to assign weights to the outputs of each Expert network. By doing this, it is possible to train more efficiently than large single neural network. However, it is difficult to obtain high accuracy because MoE [9] used a simple perceptron for integrating networks.

2.3. Mixture of CNNs

Mixture of Counting CNN (MoC-CNN) [10] was proposed for crowd counting. This is a method which extends simple perceptron in MoE [9] to CNN. This method consists of a Gating CNN and multiple Expert CNNs. Loss function of Expert CNN is mean squared error and the loss of Gating CNN is softmax cross entropy. Furthermore, by introducing the absolute value of the counting error of each learning sample as a trade-off function, it made to select as many Expert CNNs as possible. MoC-CNN [10] obtained the comparable accuracy with VGG-16 [11] which is the state-of-the-art single deep neural network.

In this paper, we also use the structure of MoC-CNN [10] but we use encoder-decoder structure in expert networks because of semantic segmentation. The output of Gating CNN is the same resolution as the segmentation result. This enables to assign roles to Expert networks at each pixel. Thus, we cannot use MoC-CNN [10] for segmentation problem directly. Furthermore, we use softmax classifier instead of the mean squared error in Expert network.

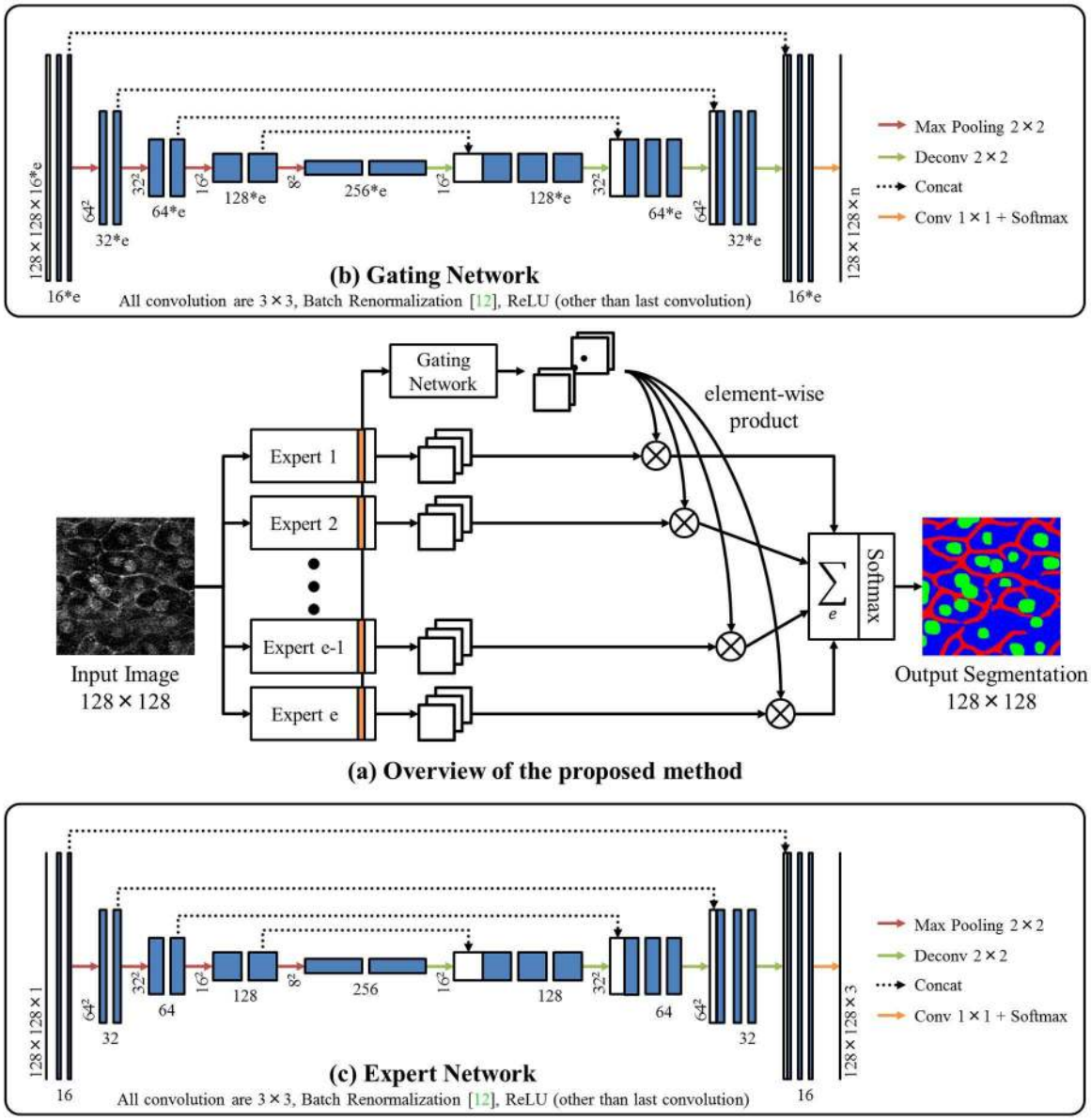

Figure 3. (a) Overview of the proposed method. (b) The detail of Gating network. (c) The detail of Expert network. e indicates the number of Expert networks. In this paper, we set the number of initial filters of Expert network as 16.

3. Proposed Method

This section describes the details of the proposed method. We explain Gating network and Expert network in Section 3.1. Divide-and-Assign by Gating network is explained in Section 3.2.

3.1. Network Details

Figure 3(b) shows the detail of Gating network and Figure 3(c) shows the detail of Expert network. Both networks use encoder-decoder structure except for the number of filters in each block. Feature maps obtained at the last convolutional layer of all Expert networks are fed into Gating network because Gating network should be a director of Expert networks. The first 5 blocks consist of 2 convolutional layers with 3×3 kernels and max pooling. The last 4 blocks consist of a deconvolution layer with 2×2 kernels [13], concatenation layer and 2 convolutional layers with 3×3 kernels. The final convolutional layer consists of a convolutional layer with 1×1 kernel and a softmax classifier. Final block is used to change the number of

feature maps to the number of classes. The output channel of Expert network corresponds to the number of classes while the output channel of Gate network corresponds to the number of Expert network.

3.2. Divide-and-Assign

Figure 3(a) shows how to integrate Expert networks by Gating network. Gating Network divides an input image into some sub-problems, and the pixel-wise role for each Expert network is made as a feature map. Since softmax layer is used at the final layer of Gating network, pixel-wise role is defined as the probability of each pixel. Therefore, after segmentation results are obtained from all Expert networks, the feature map of Gating network is multiplied with the results by Expert networks. Final segmentation result is obtained by the weighted sum of results of all Expert networks.

The pixel-wise weight is determined adaptively by Gating network according to the appearance of an input image. Of course, each Expert network trains the pixels that high probability is assigned by Gating network.

4. Experiments

This section shows evaluation results by the proposed method. We explain the dataset used in experiments in section 3.1. How to train the proposed method is explained in section 3.2. Finally, experimental results are shown in section 3.3.

4.1. Dataset

We evaluate our method on the fluorescence images of the liver of transgenic mice that expressed fluorescent markers on the cell membrane and nucleus. There are 50 images in total and the size of the image is 512×512 pixels. These images consist of 3 classes of cell membrane, cell nucleus and background. We divided those images into 35 training, 5 validation and 10 testing images. We augmented 35 training images to 280 images by rotating an image at the interval of 90 degrees and left-right mirroring.

We use intersection over union (IoU) and mean IoU (mIoU) as evaluation measures. They are computed as

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

$$mIoU = \frac{1}{c}\sum_{c}\frac{TP}{TP + FP + FN} \tag{4}$$

where TP, FP, and FN denote the true positive, false positive and false negative counts, respectively, and c denotes number of class.

4.2. Training

Since the classes are unbalanced in images, we use class weighting [14, 15] where the weight is assigned to each class in the cross-entropy loss function. The weight is defined as

$$w_c = median\_frequency\ /frequency(c) \tag{5}$$

where median frequency is the median of all class frequencies and frequency(c) is the number of pixels of class c in training images. In experiments, we assign the smallest weight to background class and cell nucleus is assigned to the largest weight.

In experiments, we evaluated 2-4 Expert networks to confirm the behavior of Gating network. We set the batch size to 4 in all process. The learning rate was set to 1e-3 and Adam optimizer was used for training.

4.3. Experimental Results

First, we confirmed the behavior of Gating network. Figure 4 shows feature maps at final layer of Gating network. (a) shows the case of 2 Expert networks, (b) shows the case of 3 Expert networks and (c) shows the case of 4 Expert networks. Red pixels mean high probability and blue pixels mean low probability. We can confirm that 3 Expert networks are the best for segmentation problem of cell membrane, nucleus and background. If we use 4 Expert networks, Gating network divided the roles into cell membrane, the outline of membrane, nucleus and background.

Table 1. Accuracy of the proposed method.

Method	Membrane	Nucleus	Background	mIoU
Our method: Expert(2)	**35.80**	61.82	70.76	56.13
Our method: Expert(3)	35.60	**62.52**	71.53	**56.55**
Our method: Expert(4)	35.73	61.22	**71.91**	56.29

Table 1 shows the accuracy of the proposed method when we change the number of Expert networks. From Table 1, we can confirm that 3 Expert networks are the best for both IoU and mIoU. Figure 4 and Table 1 demonstrated that 3 Expert networks are the most efficient in this problem.

Next, we change the output of Gating network. In the proposed method, the output of Gating network is the feature map with the same resolution as the segmentation result. Thus, we can assign a role to each pixel of Expert networks. However, in MoC-CNN [10], the scalar weight was used for integrating Expert networks. To investigate the effectiveness of our approach, we add an FC layer to the final layer of the Gating network like MoC-CNN [10]. The

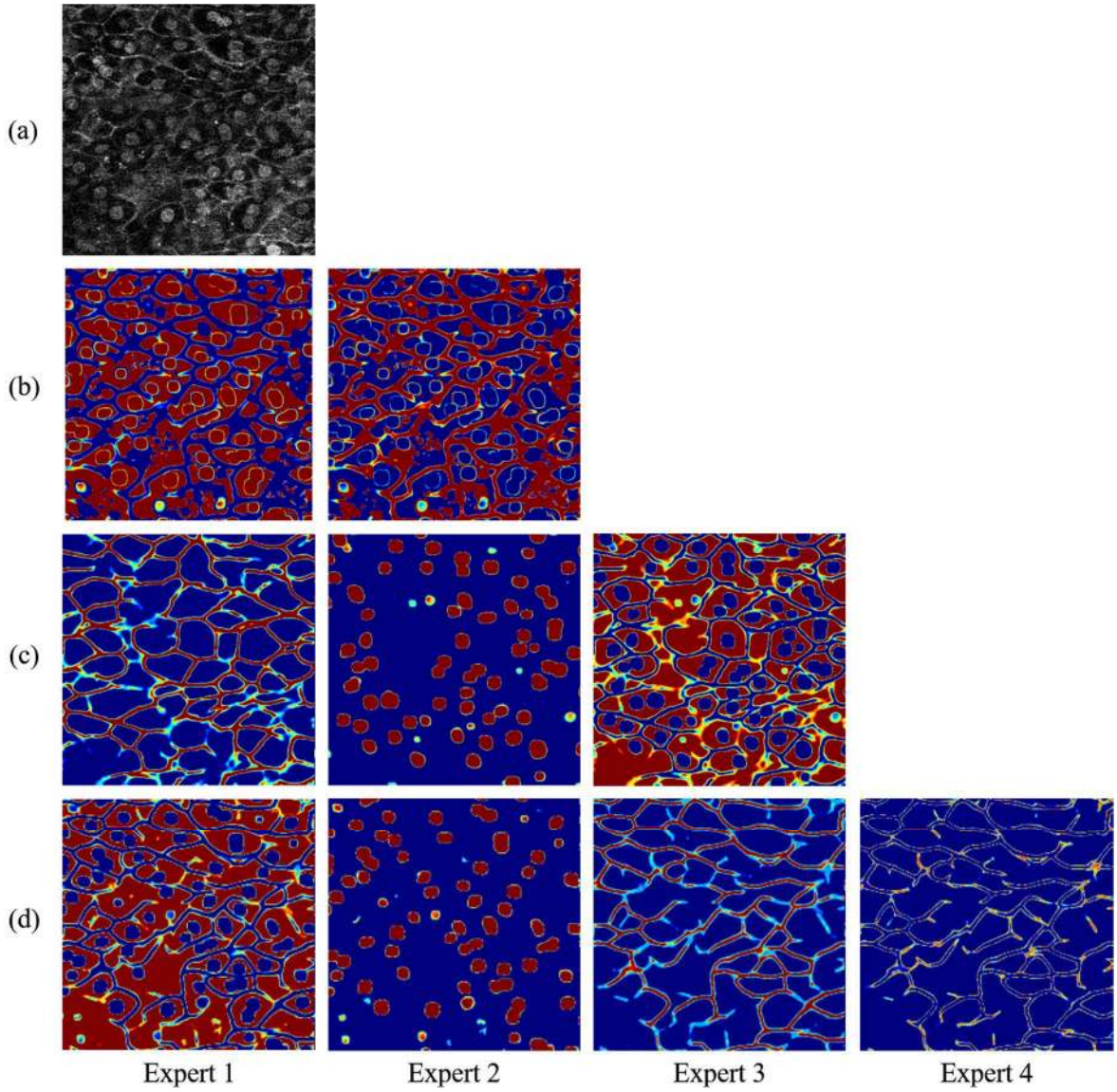

Figure 4. (a) Input image (b) The case of 2 Expert networks. (c) The case of 3 Expert networks. (d) The case of 4 Expert networks. Red pixels mean high probability and blue pixels mean low probability.

integration using scalar weights cannot assign a role to each pixel. Namely, the same integration weights are used at all pixels.

Table 2. Effectiveness of pixel-wise role assignment

Method	Membrane	Nucleus	Background	mIoU
Our method: vector	**35.60**	**62.52**	71.53	**56.55**
Our method: scalar	34.62	61.24	**72.10**	55.30

Table 2 shows the comparison result. We can confirm that the accuracy of our proposed Gating network is better for segmentation problem than the scalar weight. Therefore, pixel-wise role assignment is effective.

Table 3. Segmentation results (IoU and mIoU) in comparison with the conventional method.

Method	Membrane	Nucleus	Background	mIoU
U-Net [2] (n=16)	34.48	60.47	67.03	53.99
U-Net [2] (n=32)	35.32	60.76	69.08	55.05
U-Net [2] (n=64)	35.37	59.95	68.21	54.51
Our method: Expert(3)	**35.60**	**62.52**	**71.53**	**56.55**

Finally, we compare the proposed method with the U-Net [6]. Table 3 shows the comparison result. From

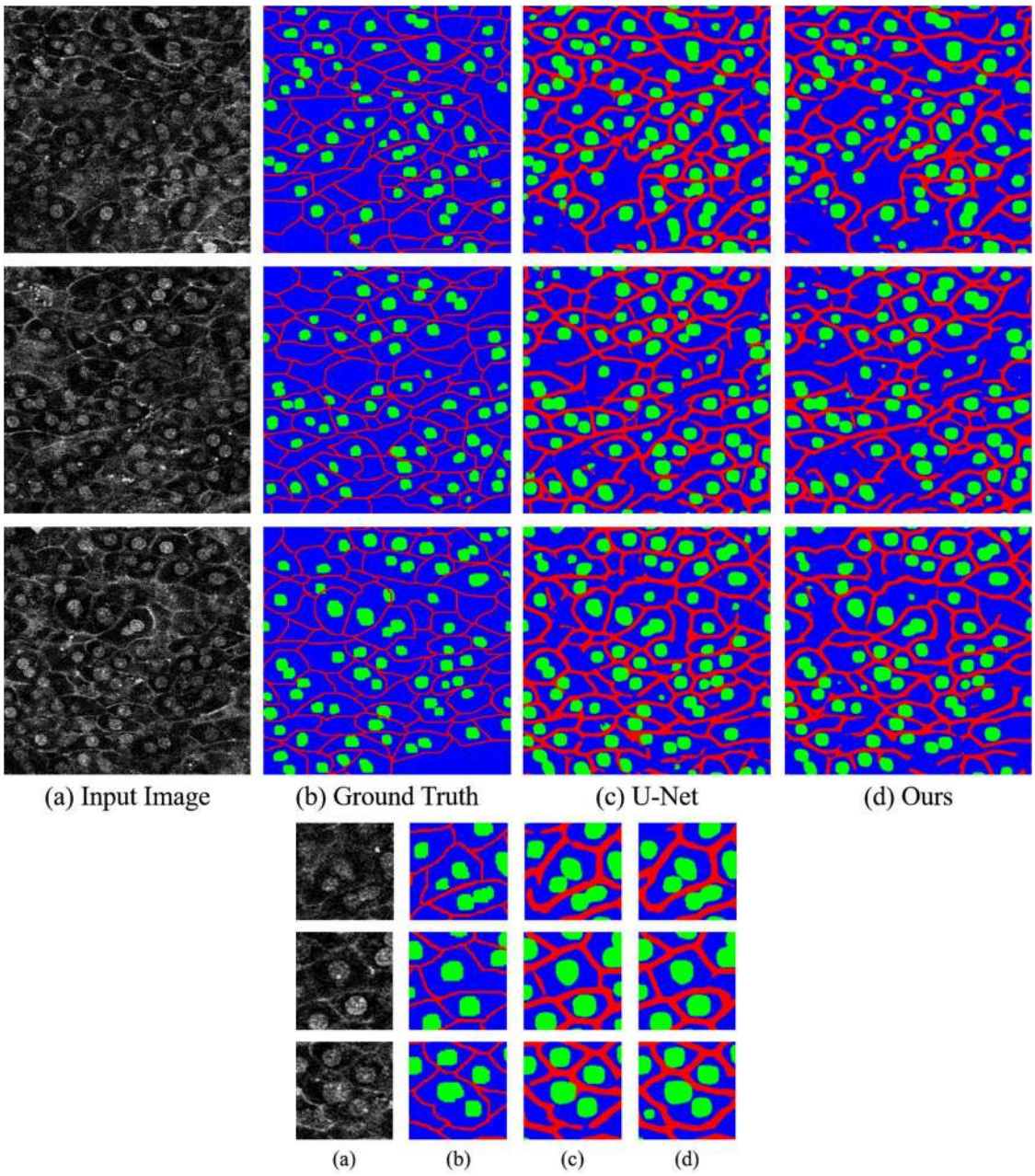

Figure 5. Segmentation results. (a) Input image (b) Ground truth (c) U-Net [6] (d) Our method

Table 3, we see that the proposed method improved IoU and mIoU in comparison with the conventional U-Net [6]. In detail, we confirmed that the proposed method improved mIoU accuracy 2.53% in comparison with the conventional method (n=16) which is the same structure as Expert network of our method.

In addition, the proposed method improved mIoU accuracy about 1% in comparison with the conventional method (n=32, 64) which uses larger number of filters than Expert network in the proposed method. Therefore, the proposed method achieved higher accuracy than that of single deep neural network with a large number of filters.

5. Conclusion

In this paper, we proposed a segmentation method by integrating multiple CNNs adaptively. The proposed method consists of Gating network and multiple Expert networks. Expert network outputs a segmentation result. On the other hand, Gating network divides the input image into sub-problems and assigns roles to each pixel of Expert

network according to the appearance. Thus, Expert networks train only the pixels assigned by Gating network. The accuracy of our method was improved in comparison with single deep neural network with a larger number of filters.

However, the accuracy of cell membrane was not so high yet because the distinction between the outline of membrane and background is ambiguous. In the future, we would like to improve the accuracy of the cell membrane by separately adding the weight to the outline of cell membrane.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE, vol.86, Issue.11, pp.2278-2324, 1998.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Proceedings of the International Conference on Neural Information Processing Systems, vol.1, pp.1097-1105, 2012.

[3] H. Zhao, J. Shi, X. Qi, Z. Wang and J. Jia, "Pyramid Scene Parsing Network", Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881-2890, 2017.

[4] F. Milletari, N. Navab, S. A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation", In the International Conference on 3D Vision, pp.565-571, 2016.

[5] J. Long, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

[6] O. Ronneberger, P. Fischer, T. Brox. "U-Net: Convolutional networks for biomedical image segmentation", Proceedings of the international Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[8] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning", In AAAI, pp. 4278-4284, 2017.

[9] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton, "Adaptive mixtures of local experts", Neural Computation, vol.3, Issue.1, pp.79–87, 1991.

[10] S. Kumagai, K. Hotta, T. Kurita, "Mixture of Counting CNNs: Adaptive Integration of CNNs Specialized to Specific Appearance for Crowd Counting", arXiv: 1703.09393, 2017.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", Proceedings of International Conference on Learning Representations, 2015.

[12] S. loffe, "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models", Advances in neural information processing systems, 2017.

[13] H. Noh, S. Hong, B.Han, "Learning deconvolution network for semantic segmentation", Proceedings of IEEE International Conference on Computer Vision, pp.1520-1528, 2015.

[14] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", Pattern Analysis and Machine Intelligence, vol.39, pp.2481-2495, 2017.

[15] K. L. Tseng, Y. L. Lin, W. Hsu, and C. Y. Huang, "Joint Sequence Learning and Cross-Modality Convolution for 3D Biomedical Segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6393-6400, 2017.