# X-net with Different Loss Functions for Cell Image Segmentation

Haruki Fujii[1], Hayato Tanaka[2], Momoko Ikeuchi[2] and Kazuhiro Hotta[1]
[1] Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan
[2]Niigata University, 8050 Ikarashi 2-noYoshida-Konoecho, Sakyo-ku, Kyoto 606-8501, Japan
170442134@ccalumni.meijo-u.ac.jp, kazuhotta@meijo-u.ac.jp,
ikeuchi@bio.sc.niigata-u.ac.jp, omenstresure4@gmail.com

## Abstract

*Convolutional neural network is valid for object segmentation. In recent years, it has been applied to the fields of medicine and cell biology. Each class has a different number of pixels in an image. Therefore, the accuracy of semantic segmentation varies drastically between objects with a large number of pixels and objects with a small number of pixels. In this paper, we propose X-Net that integrates two encoders and decoders to solve this problem. This has the advantage of extracting rich features from two encoders and using two decoders to complement the location information and small objects. By using different loss functions for each decoder, we can use the ensemble of two decoders with different viewpoints. We evaluated our method on the Arabidopsis cell images and Drosophila cell images. Experimental results show that our method achieved better accuracy than the conventional methods.*

## 1. Introduction

Convolutional neural network has been used in various image recognition problems such as image classification [3], object detection [4], image generation [5], and semantic segmentation [6]. In particular, semantic segmentation has been applied not only to autonomous driving [7,8] but also to medicine and cell biology [9,10]. Semantic segmentation is a task for assigning class labels to each pixel in an image.

Only one label is assigned to each image in image classification but all pixels in an image have class labels in semantic segmentation. Thus, class imbalance problem easily occurs in segmentation. The classes with a large number of pixels can achieve high accuracy, but the classes with a small number of pixels become low accuracy.

In this paper, we propose X-Net that integrates two encoders and decoders to solve the problem. This has the advantage of extracting rich features from two encoders and using two decoders to complement the location information and small objects in an image. By using two encoders, we will be able to obtain rich features that could not be extracted by a single network. For each decoder, we use different loss functions. We use Softmax Cross Entoropy loss for top decoder which classifies each pixel in an image. For bottom decoder, we use IoU loss which predicts on entire image by calculating the overlap ratio between the prediction result and ground truth in each class. By using different loss functions, each decoder can predict segmentation results from different point of views. Thus, we can use the ensemble of segmentation results obtained by two decoders with different viewpoints. This improves the accuracy further.

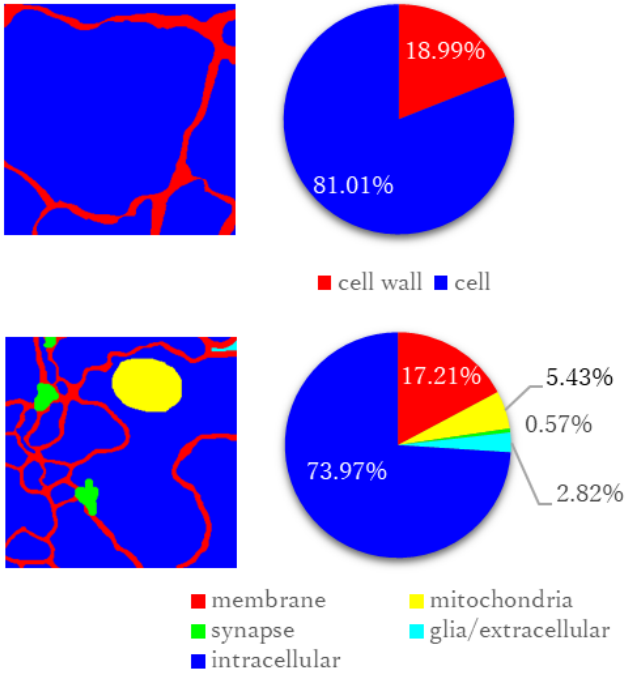


Fig. 1: Percentage of the number of pixels in each class. Top row shows *Arabidopsis* cell dataset. Bottom row shows *Drosophila* cell dataset.

In experiments, we evaluate the proposed method on two kinds of cell image datasets; the *Arabidopsis thaliana* (*Arabidopsis*) cell dataset [11] and the *Drosophila melanogaster* (*Drosophila*) cell dataset [12]. As a result of experiments, our method improved the segmentation accuracy of 2.21% on the *Arabidopsis* cell images and 2.18% on the *Drosophila* cell images.

The structure of this paper is as follows. Section 2 describes related works. Section 3 shows the details of the proposed X-Net. Section 4 shows the experimental results on two kinds of cell image datasets. Finally, conclusion and future works are described in section 5.

## 2. Related Work

Recent semantic segmentation methods are based on fully convolutional network (FCN) [14]. Since FCN does not use fully connected layers and consists of only convolutional layers, it is no longer necessary to fix the size of the input image. Encoder-decoder structures are often used for semantic segmentation. SegNet [15] memorizes the position at the pooling layer in encoder, and the original position is used at upsampling layer in decoder. This complements the location information and makes the memory more efficient.

U-Net [10] was proposed for medical image segmentation, and it is one of the most famous CNN models. In our implementation, max pooling is used for down-sampling and deconvolution is used for upsampling. The most important characteristic of U-Net is skip connection between encoder and decoder. The feature map in the encoder is concatenated to the restored feature map in the decoder. Therefore, the position information is complemented, and class label can be more accurately assigned to each pixel.

There are the methods using two U-nets to improve the accuracy. Double U-Net [16] has the structure that two U-Nets are connected in a series. By feeding the output of the first U-Net into another U-Net, more semantic information can be obtained efficiently. Ensemble U-Net [17] used the ensemble of outputs of two independent U-Nets to obtain a final segmentation result. Authors reported the accuracy improvement by the ensemble.

In contrast to those methods, our proposed X-Net consists of two encoders and two decoders. This allows us to extract rich features which cannot be extracted by a single network. After aggregating the features obtained by two encoders, we use two decoders to obtain two segmentation results independently. Finally, we integrate two outputs in two decoders and obtain a final segmentation result. By training two decoders with different loss functions, the effect of ensemble is enhanced and the accuracy will be improved further.

## 3. X-Net

We explain our proposed X-net in this section. At first, we explain the architecture of X-net which consists of two encoders and two decoders in section 3.1. Section 3.2 describes the loss functions for decoders.

### 3.1. Architecture

Figure 2 shows the network architecture of the proposed method. This method consists of two encoders and two decoders, and has three outputs. The input image is fed into two encoders, and two feature maps at the final layer in two encoders are then aggregated by concatenation. This allows us to extract rich features which could not be extracted by a single network, and it contributes to improve the segmentation accuracy.

The aggregated feature map is fed into two decoders simultaneously. The decoders use skip connections to provide the feature maps in each encoder with the same resolution to complement the positional information. The output (output1 and output2) of each decoder is the feature maps with the same resolution as the input image, and the number of channels is converted to the number of classes by 1x1 convolution. Although X-net has three outputs, two of the three outputs are from two decoders and the third output (output3) is the ensemble result of two outputs by 1x1 convolution after concatenation.

The reason for the ensemble of two outputs is to complement each other. The reason to prepare multiple outputs is to use different loss functions for each decoder in order to obtain segmentation results from different viewpoints. We explain loss functions in the next section.

### 3.2. Loss function

In semantic segmentation, Softmax Cross Entropy (SCE) loss is the loss function for classifying each pixel in an image. On the other hand, Intersection over Union (IoU) loss computes the overlap ratio between the prediction result and ground truth at each class. This means that it predicts on the entire image. If we use different loss functions for each decoder in the X-net, each decoder can predict the segmentation result from different viewpoints. We expect to enhance the ensemble effect at the output3 in Figure 2 by using different loss functions for each decoder. Therefore, X-Net is trained so that top decoder is trained with SCE loss and bottom decoder is trained with IoU loss.

SCE loss is defined as

$$\text{SCE loss} = -\sum_{i}\sum_{c} p_c^i \log q_c^i \tag{1}$$

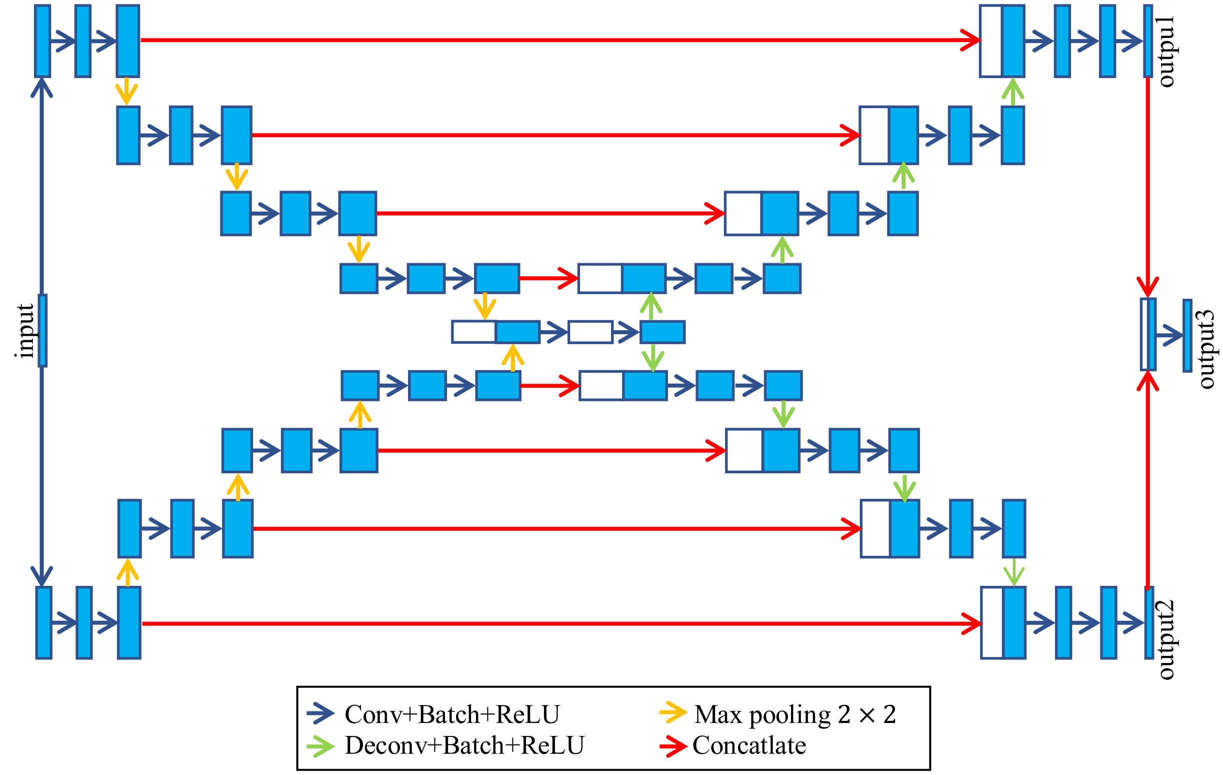


Fig.2: Architecture of X-Net

where $i$ means the i-th sample in training data, c means the c-th class, $p_c^i$ is one hot vector of ground truth, $q_c^i$ is the probability of class c for the i-th sample.

IoU loss is defined as

$$IoU\ loss = 1 - \frac{p_c^i q_c^i + \gamma}{p_c^i + q_c^i - p_c^i q_c^i + \gamma} \tag{2}$$

where $\gamma = 1.0^{-5}$ so that the function is not indefinite even when $p_c^i = q_c^i = 0$.

Dice loss is defined as

$$Dice\ loss = 1 - \frac{2p_c^i q_c^i + \gamma}{p_c^i + q_c^i + \gamma} \tag{3}$$

Let the loss functions of output1, output 2, and output3 in Figure 2 be Loss1, Loss2, and Loss3, respectively. As described previously, we used SCE loss as Loss1 and IoU or Dice loss as Loss2 to enhance the ensemble effect. Since output 3 is the integration of output 1 and 2, the loss function was also set to the sum of Loss1 and Loss2.

$$Loss1 = SCE\ loss \tag{4}$$

$$Loss2 = IoU\ loss \tag{5}$$

$$Loss3 = SCE\ Loss + IoU\ loss \tag{6}$$

## 4. Experiments

### 4.1. Datasets and evaluate measure

We use two datasets with very different image properties. The first one is the Arabidopsis cell dataset [11] as shown in left two columns of Figure 3. This dataset consists of tissue sections stained with toluidine blue staining and photographed with light microscopy. The dataset consists of 2 classes; cells and cell walls. The sizes of the original images are large (e.g. 3118×2261 pixels). Thus, we cropped the regions of 512 x 512 pixels and resized them to 256 x 256 pixels. There is no overlap for cropping areas, and the total number of regions is 50. We used 30 regions for training, 10 for validation and 10 for test. We evaluate our method with 5 fold cross- validation.

The second is one the *Drosophila* cell image dataset [12] as shown in two right columns of Figure 3. This dataset is the *Drosophila melanogaster* third instar larva ventral nerve cord taken at serial section Transmission Electron Microscopy (ssTEM). The dataset consists of 5 classes;

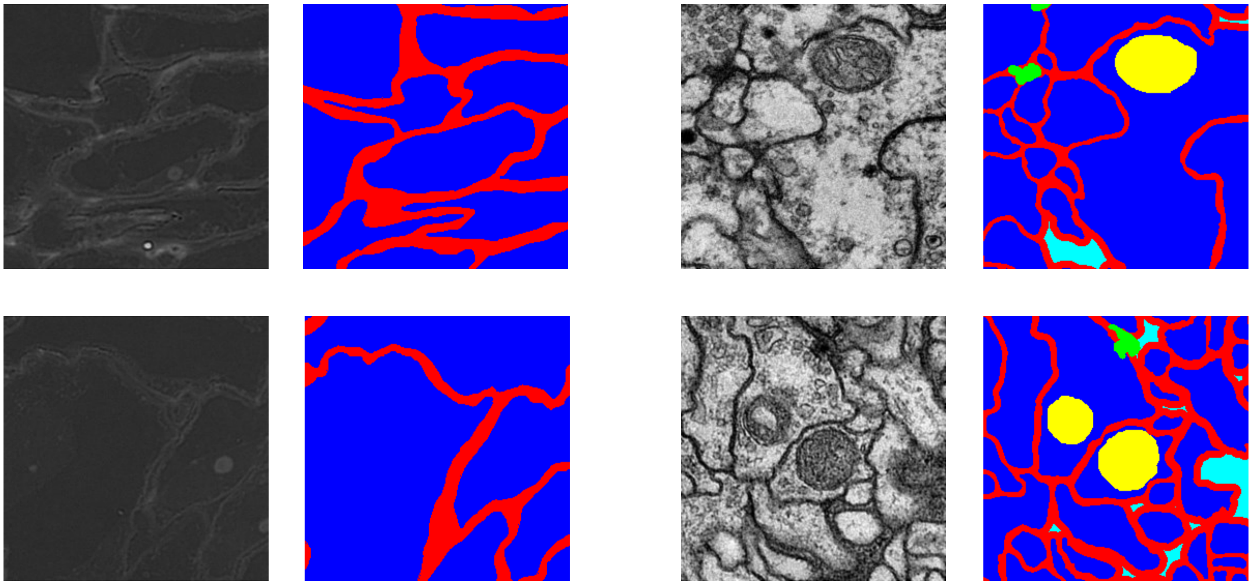


Fig. 3: Examples of datasets. Left two columns show the *Arabidopsis* cell dataset which consists of cell and cell wall. Right two columns show the *Drosophila* cell image dataset which consists of cytoplasm, cell membrane, mitochondria, and synapses.

membrane, mitochondria, synapses, glia/extracellular and intracellular. Since the original size is 1024×1024 pixels, we cropped regions of 256×256 pixels from original images due to the size of GPU memory. There is no overlap for cropping regions, and the total number of cropped regions is 320. We used 192 regions for training, 64 for validation and 64 for test. We evaluate our method with 5 fold cross-validation

We use Intersection over Union (IoU) as evaluation measure. IoU is the overlap ratio between segmentation result and ground truth labels. In this paper, we use IoU of each class and mean IoU which is the average IoU of all classes.

### 4.2. Implementation details

In this paper, we used the Pytorch library and trained our method using the Adam for 1,500 epochs. The learning rate was set to 0.001 as the initial value and 0.0001 after 1,000 epochs. Batch size was set to 8. Furthermore, class weight was used in SCE loss to address class imbalanced problem.

We evaluated the following 8 methods; U-Net using the sum of SCE loss and IoU loss, U-Net using the sum of SCE loss and Dice loss, Double U-Net using the sum of SCE loss and IoU loss, Ensemble U-Net using the sum of SCE loss and IoU loss, X-Net using the sum of SCE loss and IoU loss, X-Net using the sum of SCE loss and Dice loss, X-Net that top network is trained with SCE loss and bottom network is trained with IoU loss, and X-Net that top network is trained with SCE loss and bottom network is trained with Dice loss. The best model was selected by using the accuracy for validation set.

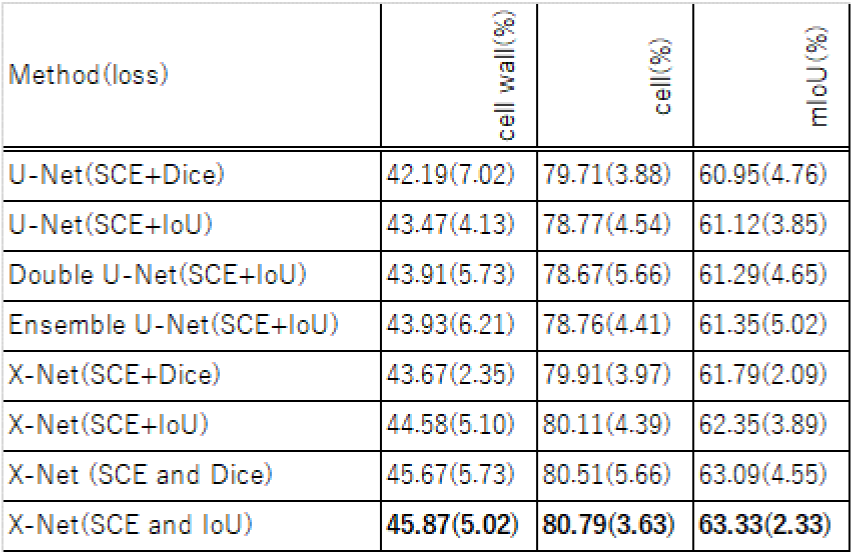
### 4.3. Comparison with Another Method

Table 1 shows the evaluation result on the *Arabidopsis* cell dataset. SCE+Dice in the Table indicates that the sum of SCE loss and Dice loss is used as the loss function, and “SCE and Dice” indicates that SCE loss is used for the upper network of X-Net and Dice loss is used for the lower network. The numbers in brackets are the standard deviations of the accuracies in 5-fold cross validation. The best accuracy was obtained by the proposed X-Net that top network is trained with SCE loss and bottom network is trained on IoU loss. The accuracy of our method was 63.33% on mean IoU which is 2.21% higher than the U-net with sum of SCE loss and IoU loss. In addition, the standard deviation was smaller than conventional methods.

The accuracy was improved when we use X-net with different loss functions. This shows the effectiveness of ensemble of two decoders trained by different losses. By comparison with conventional methods using two U-nets; Double U-net and Ensemble U-net, we demonstrated the effectiveness of architecture of our X-net using two encoders and two decoders.

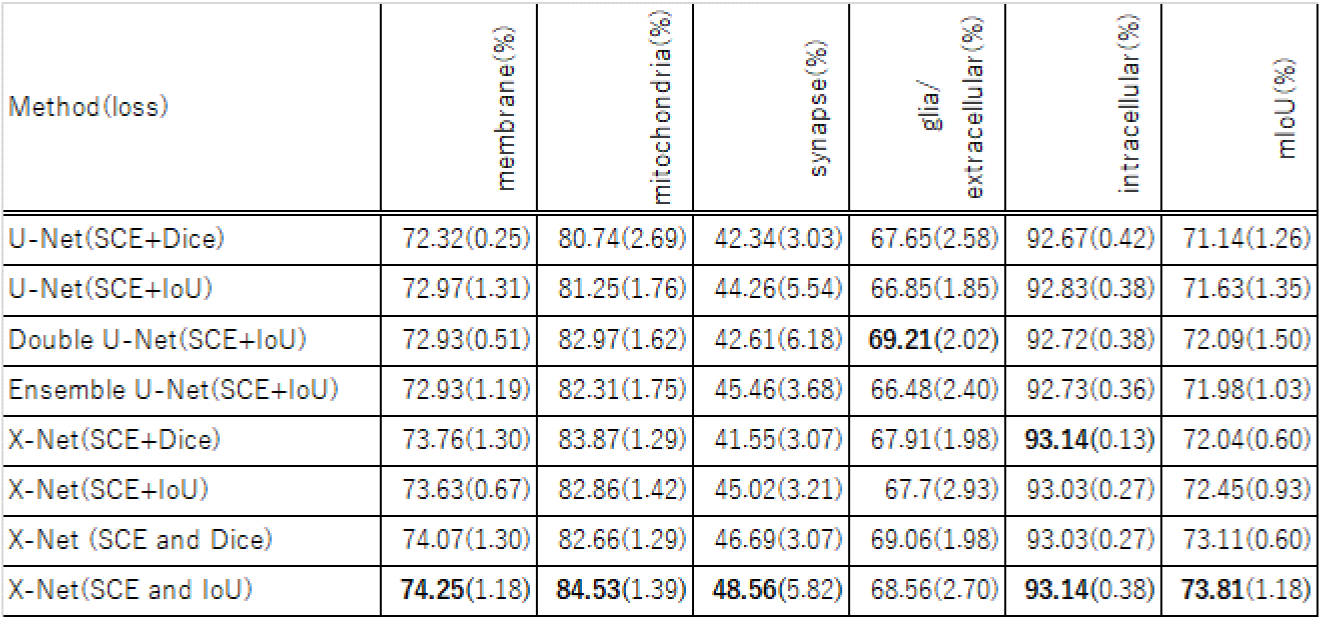
Table 2 shows evaluation results on the *Drosophila* cell image dataset. The proposed method achieved 73.81% on

Table 1: Comparison result on the *Arabidopsis* cell dataset



| Method(loss) | cell wall(%) | cell(%) | mIoU(%) |
|---|---|---|---|
| U-Net(SCE+Dice) | 42.19(7.02) | 79.71(3.88) | 60.95(4.76) |
| U-Net(SCE+IoU) | 43.47(4.13) | 78.77(4.54) | 61.12(3.85) |
| Double U-Net(SCE+IoU) | 43.91(5.73) | 78.67(5.66) | 61.29(4.65) |
| Ensemble U-Net(SCE+IoU) | 43.93(6.21) | 78.76(4.41) | 61.35(5.02) |
| X-Net(SCE+Dice) | 43.67(2.35) | 79.91(3.97) | 61.79(2.09) |
| X-Net(SCE+IoU) | 44.58(5.10) | 80.11(4.39) | 62.35(3.89) |
| X-Net (SCE and Dice) | 45.67(5.73) | 80.51(5.66) | 63.09(4.55) |
| X-Net(SCE and IoU) | **45.87(5.02)** | **80.79(3.63)** | **63.33(2.33)** |

Table 2: Comparison result on the *Drosophila* cell image dataset.



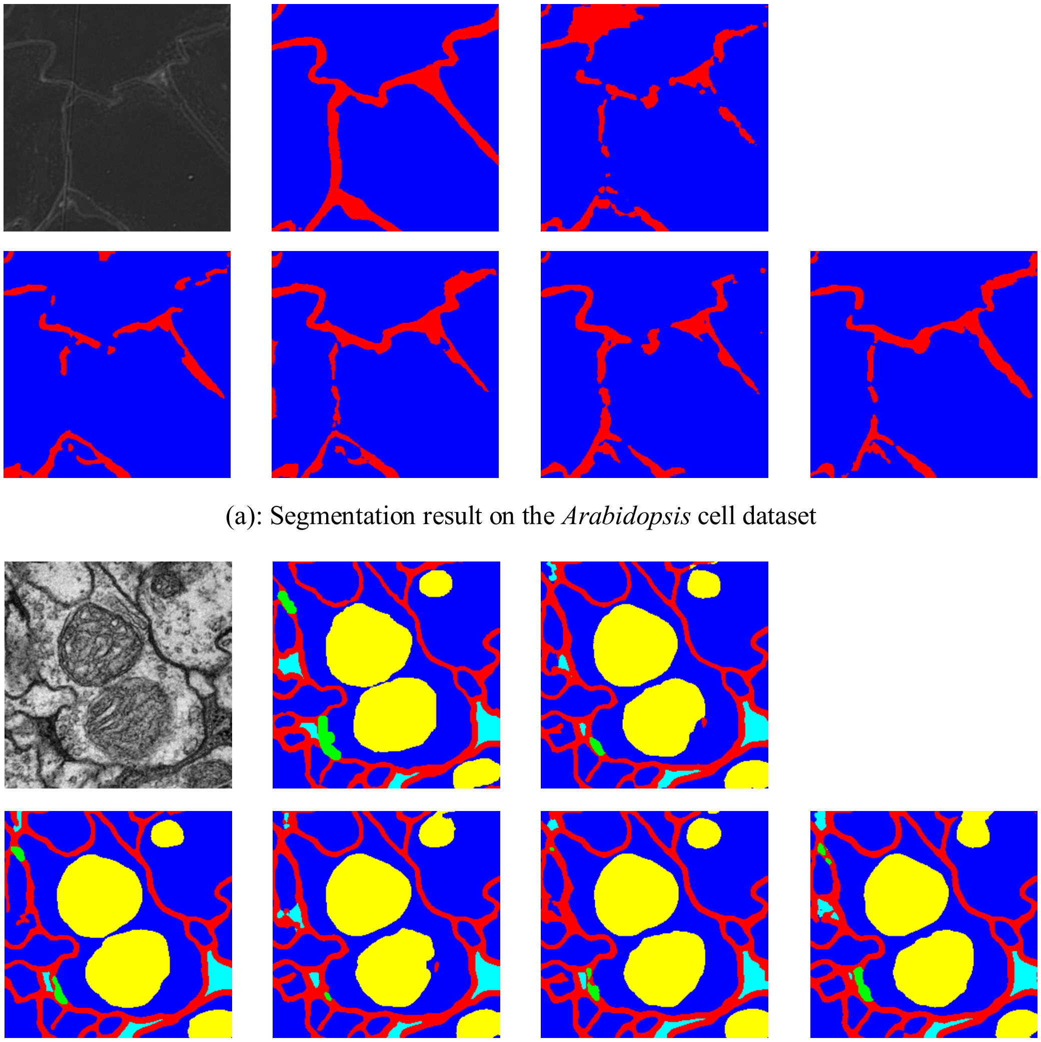
| Method(loss) | membrane(%) | mitochondria(%) | synapse(%) | glia/ extracellular(%) | intracellular(%) | mIoU(%) |
|---|---|---|---|---|---|---|
| U-Net(SCE+Dice) | 72.32(0.25) | 80.74(2.69) | 42.34(3.03) | 67.65(2.58) | 92.67(0.42) | 71.14(1.26) |
| U-Net(SCE+IoU) | 72.97(1.31) | 81.25(1.76) | 44.26(5.54) | 66.85(1.85) | 92.83(0.38) | 71.63(1.35) |
| Double U-Net(SCE+IoU) | 72.93(0.51) | 82.97(1.62) | 42.61(6.18) | **69.21**(2.02) | 92.72(0.38) | 72.09(1.50) |
| Ensemble U-Net(SCE+IoU) | 72.93(1.19) | 82.31(1.75) | 45.46(3.68) | 66.48(2.40) | 92.73(0.36) | 71.98(1.03) |
| X-Net(SCE+Dice) | 73.76(1.30) | 83.87(1.29) | 41.55(3.07) | 67.91(1.98) | **93.14**(0.13) | 72.04(0.60) |
| X-Net(SCE+IoU) | 73.63(0.67) | 82.86(1.42) | 45.02(3.21) | 67.7(2.93) | 93.03(0.27) | 72.45(0.93) |
| X-Net (SCE and Dice) | 74.07(1.30) | 82.66(1.29) | 46.69(3.07) | 69.06(1.98) | 93.03(0.27) | 73.11(0.60) |
| X-Net(SCE and IoU) | **74.25**(1.18) | **84.53**(1.39) | **48.56**(5.82) | 68.56(2.70) | **93.14**(0.38) | **73.81**(1.18) |

mean IoU. The accuracy was 2.18% higher than the U-net. Our method was also better than conventional Double U-net and Ensemble U-net.

Table 1 and 2 show that the usage of IoU loss achieved higher accuracy than that of Dice loss. This is because IoU loss is the same as the evaluation measure. When we check the accuracy of each class, there is no significant change in accuracy for cells in the *Arabidopsis* dataset or intracellular in the *Drosophila* dataset. However, there is a significant increase in accuracy for cell walls and synapses which have small number of pixels. This means that the accuracies of difficult classes were improved by the proposed X-net.

### 4.4. Qualitative Results

Figure 4 shows the segmentation results on two cell image datasets. In the case of the *Arabidopsis* dataset, we see that the proposed method was better than other methods



(a): Segmentation result on the *Arabidopsis* cell dataset

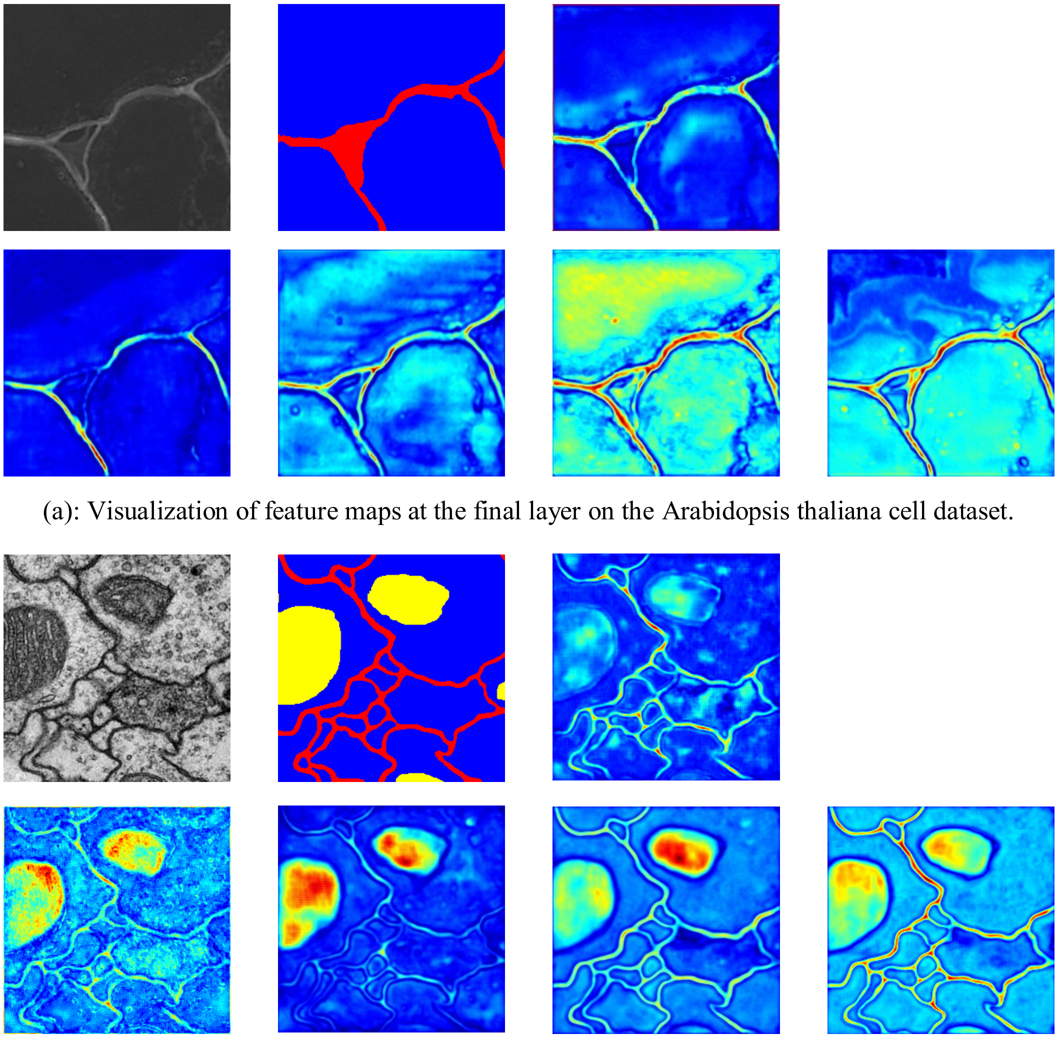
(b): Segmentation result on the *Drosophila* cell image dataset

Fig. 4: Segmentation results on two cell image datasets. (a) shows the segmentation results on the *Arabidopsis* cell dataset and (b) shows the results on the *Drosophila* cell image dataset. From left to right images in top row show input image, ground truth, the result by U-Net using the sum of SCE loss and IoU loss. From left to right images in bottom row show the result by Double U-Net using the sum of SCE loss and IoU loss, the result by Ensemble U-Net using the sum of SCE loss and IoU loss, the result by X-Net using the sum of SCE loss and IoU loss, and the result by X-Net that top network with SCE loss and bottom network with IoU loss.

in the cell wall class. The cell walls were broken in the result of U-net and Double U-Net, but the number of the connected cell walls increased by our method. In the *Drosophila* datasets, other methods could not discriminate between membranes, synapses, and glia/extracellular cells, but X-Net was able to segment them successfully. These results demonstrated that X-Net can segment the classes with fewer pixels well in comparison with the other methods.

Figure 5 shows the visualization results of the feature map on two cell image datasets. For the U-Net, Double U-Net, and Ensemble U-Net, we visualized the convolutional layer just before the output. For the X-Net, we visualized the convolutional layer just before output3. The feature map for visualization was created by averaging all feature maps. We normalized the feature map from 0 to 1 and painted red to the pixels that are close to 1 and blue to the pixels that are close to 0.

The feature map of X-Net, that top decoder is trained with SCE loss and bottom decoder is trained with IoU loss, has more reddish mitochondria and membranes than the other methods. This indicates that X-Net is better at



(a): Visualization of feature maps at the final layer on the Arabidopsis thaliana cell dataset.

(b): Visualization of feature maps at the final layer on the Drosophila cell image dataset.

Fig. 5: Visualization of feature maps at the final layer on two cell image datasets. (a) shows the feature map on the *Arabidopsis* cell dataset and (b) shows that on the *Drosophila* cell image dataset. From left to right images in top row show input image, ground truth, the feature map in U-Net using the sum of SCE loss and IoU loss. From left to right images in bottom row show the feature map in Double U-Net using the sum of SCE loss and IoU loss, that in Ensemble U-Net using the sum of SCE loss and IoU loss, that in X-Net using the sum of SCE loss and IoU loss, and that in X-Net that top network with SCE loss and bottom network with IoU loss.

identifying mitochondria and membranes than the other methods.

## 5. Conclusion

In this paper, we proposed the X-Net which consisted of two encoders and two decoders. This made it possible to extract rich features from two encoders and obtain superior segmentation results by two decoders from the aggregated features of two encoders. In addition, we trained top decoder with SCE loss for classifying each pixel in an image, and we trained bottom decoder with IoU loss or Dice loss which predicts on the entire image. By using different losses, we obtained better discrimination ability.

In addition, this paper evaluates the accuracy using two datasets with very different image properties, and found that it is effective for both. This indicates that X-Net is a highly versatile analysis method.

X-Net improved segmentation accuracy but used multiple encoders and decoders. This increases the computational cost and memory required. It is necessary to reduce them. By using pointwise convolution or summing the connections, we may be able to achieve the same accuracy with lower computational cost. This is a subject for future works.

### Acknowledgements

This paper is partially supported by JSPS KAKENHI Grant Number 20H05427.

### Reference

[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”, Advances in Neural Information Processing Systems, pp.1097-1105, 2012.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1-9. 2015.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580-587. 2014.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, pp. 2672-2680. 2014.

[6] 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.

[7] 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.

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

[9] F. Milletari, N. Navab, S. A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”, Proceedings of International Conference on 3D Vision, pp.565-571, 2016.

[10] O. Ronneberger, P. Fischer, T. Brox. “U-Net: Convolutional networks for biomedical image segmentation”, Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241, 2015.

[11] M. Ikeuchi, A. Iwase, B. Rymen, A. Lambolez, M. Kojima, Y. Takebayashi, J. Heyman, S. Watanabe, M. Seo, L. De Veylder, H. Sakakibara, K. Sugimoto “Wounding triggers callus formation via dynamic hormonal and transcriptional changes,” Plant Physiology, Vol.175, No.3 pp.1158-1174, 2017.

[12] G. Stephan, F. Jan, M. Julien, C. Albert, F. Richard. “Segmented anisotropic ssTEM dataset of neural tissue,” figshare. Retrieved 16:09, 2013.

[13] 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.

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

[15] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.39, pp.2481-2495, 2017.

[16] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Double U-net: A deep convolutional neural network for medical image segmentation,” IEEE International Symposium on Computer-Based Medical Systems, pp. 558-564. 2020.

[17] Z. Fatemeh, S. Nicola, K. Satheesh, and U. Eranga, “Ensemble U-net based method for fully automated detection and segmentation of renal masses on computed tomography images,” Medical Physics, Vol.47, No.9, pp.4032-4044. 2020.