HUMAN PERCEPTION-BASED EVALUATION CRITERION FOR ULTRA-HIGH RESOLUTION CELL MEMBRANE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Computer vision technology is widely used in biological and medical data analysis and understanding. However, there are still two major bottlenecks in the field of cell membrane segmentation, which seriously hinder further research: lack of sufficient high-quality data and lack of suitable evaluation criteria. In order to solve these two problems, this paper first proposes an Ultra-high Resolution Image Segmentation dataset for the Cell membrane, called U-RISC, the largest annotated Electron Microscopy (EM) dataset for the Cell membrane with multiple iterative annotations and uncompressed high-resolution raw data. During the analysis process of the U-RISC, we found that the current popular segmentation evaluation criteria are inconsistent with human perception. This interesting phenomenon is confirmed by a subjective experiment involving twenty people. Furthermore, to resolve this inconsistency, we propose a new evaluation criterion called Perceptual Hausdorff Distance (PHD) to measure the quality of cell membrane segmentation results. Detailed performance comparison and discussion of classic segmentation methods along with two iterative manual annotation results under existing evaluation criteria and PHD is given.

1 INTRODUCTION

Electron Microscopy (EM) is a powerful tool to explore ultra-fine structures in biological tissues, which has been widely used in the research areas of medicine and biology (ERLANDSON (2009); Curry et al. (2006); Harris et al. (2006)). In recent years, EM techniques have pioneered an emerging field called "Connectomics" (Lichtman et al. (2014)), which aims to scan and reconstruct the whole brain circuitry at the nanoscale. "Connectomics" has played a key role in several ambitious projects, including the BRAIN Initiative (Insel et al. (2013)) and MICrONS (Gleeson & Sawyer (2018)) in the U.S., Brain/MINDS in Japan (Dando (2020)), and the China Brain Project (Poo et al. (2016)). Because EM scans brain slices at the nanoscale, it produces massive images with ultra-high resolution and inevitably leads to the explosion of data. However, compared to the advances of EM, techniques of data analysis fall far behind. In particular, how to automatically extract information from massive raw data to reconstruct the circuitry map has growingly become the bottleneck of EM applications.

One critical step in automatic EM data analysis is Membrane segmentation. With the introduction of deep learning techniques, significant improvements have been achieved in several public available EM datasets ISBI 2012 and SNEMI3D (ISBI 2012 (2012); ISBI 2013 (2013)). One of the earliest works (Ciresan et al. (2012) used a succession of max-pooling convolutional networks as a pixel classifier, which estimated the probability of a pixel is a membrane. Ronneberger et al. (2015) presented a U-net structure with contracting paths, which captures multi-contextual information. Fully convolutional networks (FCNs) proposed by Long et al. (2015) led to a breakthrough in semantic segmentation. Follow-up works based on U-net and FCN structure (Xie & Tu (2015); Drozdzal et al. (2016); Hu et al. (2018); Zhou et al. (2018); Chaurasia & Culurciello (2017); Yu et al. (2017); Chen et al. (2019)) have also achieved outstanding results near-human performance.

Despite much progress that has been made in cell membrane segmentation for EM data thanks to deep learning, one risk to these state-of-the-arts (SOTA) methods is the dataset such as ISBI 2012

and SNEMI3D, which employed compressed images instead of the original ultra-high resolution images. Considering that image compression generally loses many texture details, how can these classic deep learning based segmentation methods work on EM images with the original resolution? Moreover, how robust if these methods are compared to human performance on original EM images?

To expand the research of membrane segmentation on more comprehensive EM data, we first established a dataset "U-RISC" containing images with original resolution (10000 $\times$ 10000, Fig. 1). To ensure the quality of annotation, it also costs us over 10,000 labor hours to label and double-check the data. To the best of our knowledge, U-RISC is the largest annotated EM dataset, and the only uncompressed annotated EM dataset today. Next, we tested several classic deep learning based segmentation methods on U-RISC and compared the results to human performance. We found the human performance is far superior to these methods. To understand why human perception is better than the SOTA segmentation methods, we examined in detail the Membrane segmentation results by these SOTA segmentation methods. Surprisingly, we found there was a certain inconsistency between current evaluation criteria for segmentation(e.g. F1 score, IoU) and human perception: while some figures were rated significantly lower in F1 score or IoU, they were "perceived" better by humans (Fig. 4). Such inconsistency motivated us to propose a human-perception based criterion, Perceptual Hausdorff Distance (PHD) to evaluate image qualities. Further, we set up a subjective experiment to collect human perception about membrane segmentation, and we found the PHD criteria is more consistent with human choices than traditional evaluation criteria. Finally, we found the SOTA segmentation methods need to be revisited with PHD criteria.

Overall, our contribution in this work lies mainly in the following two parts: (1) we established the largest, original image resolution-based EM dataset for training and testing; (2) we proposed a human-perception based evaluation criterion, PHD, and verified the superiority of PHD by subjective experiments. The dataset we contributed and the PHD criterion we proposed may help researchers to gain insights into the difference between human perception and conventional evaluation criteria, thus motivate the further design of the segmentation method to catch up with the human performance on original EM images.

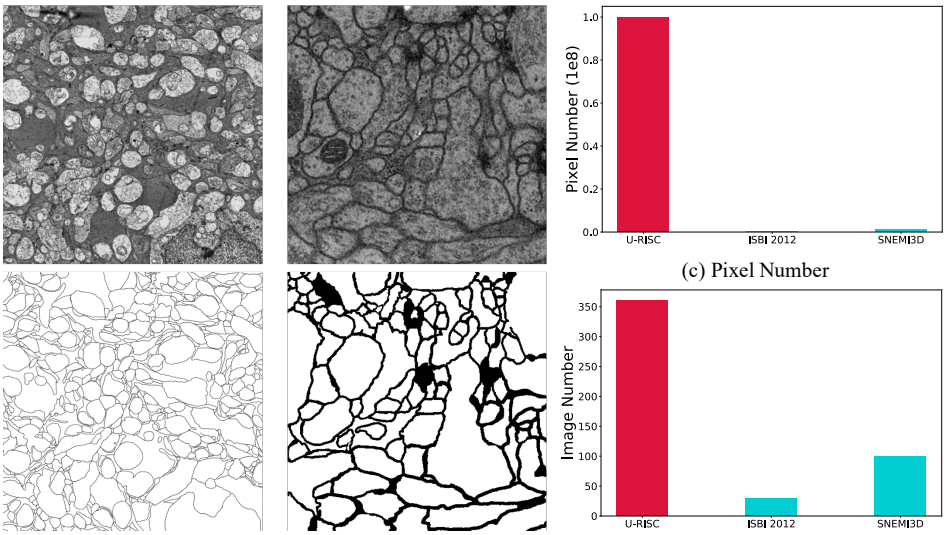
2 U-RISC: Ultra-high Resolution Image Segmentation dataset for Cell membrane

Supervised learning methods rely heavily on high-quality datasets. To alleviate the lack of training data for cell membrane segmentation, we proposed an Ultra-high Resolution Image Segmentation dataset for Cell membrane, called U-RISC. The dataset was annotated upon RC1, a large scale retinal serial section transmission electron microscopic (TEM) dataset provided by Marc's lab (Anderson et al. (2011)). The raw data represents a 0.25mm diameter retinal area acquired at an x-y resolution of 2.18nm/pixel and 70nm thickness in z-axis. In this work, we cut out 120 images in the size of 10000 pixel $\times$ 10000 pixel from different layers of original TEM Images. Then, we manually labeled the cell membranes in an iterative procedure. Since the human labeling process is very valuable for uncovering the human leaning process, we reserved the intermediate results. The U-RISC dataset will be released on `https://Anonymous.com` on acceptance.

2.1 Comparison With Other Datasets

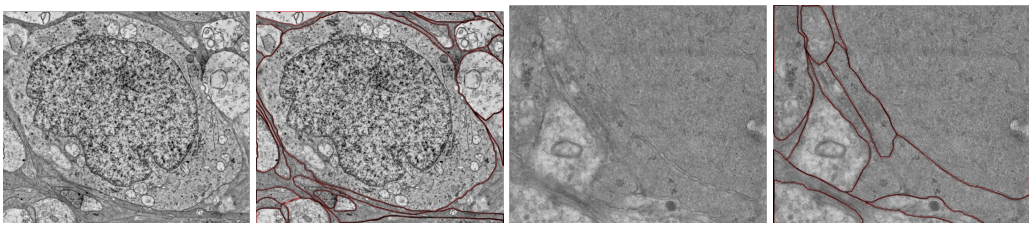
U-RISC includes many advantages compared to previous datasets. Based on the RC1, which was acquired under 2nm/pixel, U-RISC provides much more unambiguous information to identify the subcellular structures compared to ISBI 2012 (Fig. 1 (a) and (b)).

Fig. 1 (c) shows that the pixel number of each image in U-RISC is 400 times the image size of the ISBI 2012 dataset while 100 times of the image size of the SNEMI3D dataset. Besides that, U-RICS produced 3 sets of annotations with increasing labeling accuracy, all of which can be applied in developing deep learning based segmentation methods according to various demands. Thus U-RISC exceeds the previous dataset a lot in data size (Fig. 1 (d)). Overall, U-RISC is a much more challenging and suitable dataset in exploiting cell segmentation algorithms. An example of the image with its label is shown in the Supplementary. Due to the limitation of the size of the supplementary material, we only uploaded a quarter (5000 pixel $\times$ 5000 pixel) size of the original image with its label.



(a) U-RISC Example (b) ISBI 2012 Example (c) Pixel Number (d) Number of images

Figure 1: Dataset examples and statistical comparisons. (a) and (b) are examples of U-RISC and ISBI2012. The first line shows the original image, and the second line shows its annotation. (c) shows the pixel number of U-RISC, ISBI2012, and SNEMI3D datasets. (d) shows the image number of the three datasets.



(a) Example 1 (b) Example 2

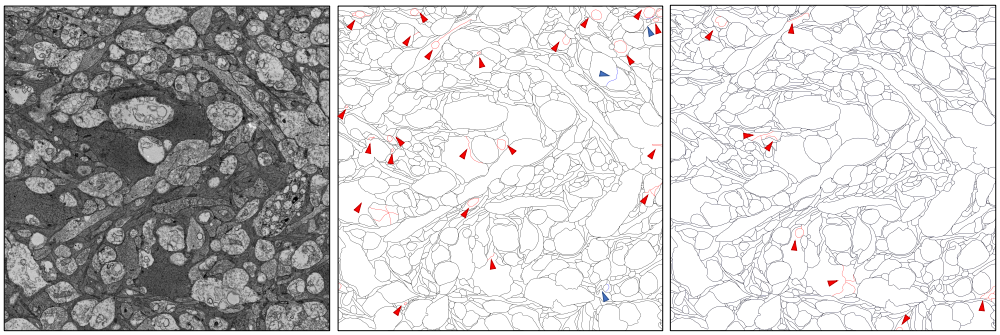
Figure 2: Examples of images with their labels. The left of (a) (b) are two original image parts, and the right are the images covered by their labels (red lines).

2.2 Triple Labeling Process

The character of high resolution in TEM image can display a much more detailed sub-cellular structure, which requests more patience to label out the cell (Fig. 2(a)). Besides, the imaging quality can be affected by many factors, such as section thickness or sample staining (Fig. 2(b)). And low imaging quality also requests more labeling efforts. Therefore, increasing labeling efforts is essential to completely annotate U-RISC. To guarantee the labeling accuracy, we set up an iterative correction mechanism in the labeling process (Fig. 3). Before starting the annotation, labeling rules were introduced to all annotators. 58 qualified annotators were allowed to participate in the final labeling process. After the first round annotation, 5 experienced lab staff with sufficient background knowledge were responsible to point out labeling errors pixel by pixel during the second and the third rounds of annotation. Finally, the third round annotation results were regarded as the final "ground truth". And previous two rounds of manual annotations are also saved for later analysis. Fig. 3 shows an example of the two inspection processes. We can see that there are quiet a few mislabeled and missed labeled cell membranes in each round. Therefore, the iterative correction mechanism is very necessary.

3 Perception-based Evaluation

In the analysis of EM data, membrane segmentation is generally an indispensable key step. However, in the field of cell membrane segmentation, most of the previous studies, such as Zhou et al. (2018); Chaurasia & Culurciello (2017); Drozdzal et al. (2016), were not specifically designed for high resolution datasets such as U-RISC. In addition, few researchers consider whether the existing evaluation criteria for general image segmentation are suitable for cell membrane segmentation.



(a) Original Image (b) First Iterative Correction (c) Second Iterative Correction

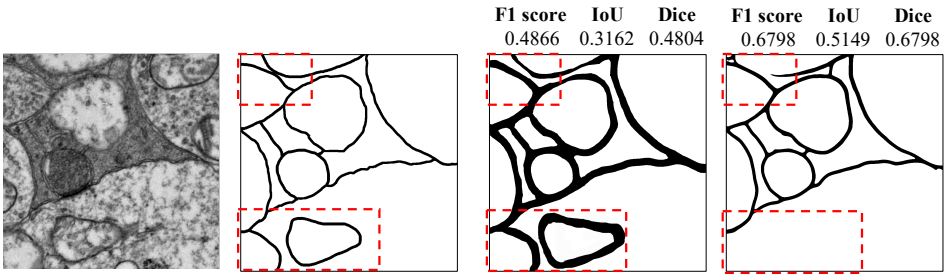
Figure 3: Example of iterative labeling. (a) shows the original EM image. (b) and (c) shows the first and second round of annotation processes. The red arrows point to the deleted parts during the inspection processes, and the blue arrows point to the added parts.

By comparing the segmentation results of the state-of-the-art segmentation methods, we found that the widely used evaluation criteria of segmentation were inconsistent with human perception in some cases, which is further discussed through the perceptual consistency experiment (details in Sec. 3.2). To address this issue, we proposed a new evaluation criterion called Perceptual Hausdorff Distance (PHD). The experimental results showed that it was more consistent with human perception.

3.1 Inconsistency between Existing Evaluation Criteria and Perception

In recent works, most of the segmentation methods evaluate the results (Ronneberger et al. (2015); Zhou et al. (2018); Chaurasia & Culurciello (2017); Yu et al. (2017); Chen et al. (2019)) following the evaluation criteria on general image segmentation, such as F1 score(Sasaki et al. (2007)), IoU(Kosub (2019)), and Dice Coefficient(Dice (1945)). All of the evaluation criteria above are based on the statistics of the degree to which pixels are classified correctly.

However, such statistics may not be consistent with human perception in cell membrane segmentation tasks. In the process of segmentation experiments, some interesting phenomena were found. Fig. 4 shows an example of the original image with its manual annotation and segmentation results by two methods GLNet (Chen et al. (2019)) and U-Net (Ronneberger et al. (2015)). The scores indicated that (d) was more similar to (b) than (c). It should be noted that when we looked at the area surrounded by the red dashed lines in the images, an natural first instinct was that (c) was a better prediction, because (d) missed some edges. The reason for the three scores of (c) are lower was that the predicted cell membrane of (c) was thicker than manual labeling. Therefore, it can be inferred that the existing evaluation criteria were not sufficiently robust to variations in the thickness and structures of the membrane, and the evaluation result was inconsistent with human perception.



(a) Original Image (b) Manual Annotation (c) GLNet (d) U-Net

Figure 4: Examples of some segmentation results. (a) is the original EM image. (b) is the manual annotation of cell membrane. (c) and (d) are segmentation results from GLNet and U-Net.

3.2 Perceptual Consistency Experiments

In order to verify the above conjecture, a subjective experiment was designed to explore the consistency with the existing evaluation criteria and human subjective perception. Six state-of-the-art segmentation methods were used to generate cell membrane segmentation results on U-RISC: U-net (Ronneberger et al. (2015)), LinkNet(Chaurasia & Culurciello (2017)), CASENet (Yu et al.

(2017)), SENet (Hu et al. (2018)), U-Net++ (Zhou et al. (2018)),and GLNet (Chen et al. (2019)). Using these segmentation results, 200 groups of images were randomly selected. Each group contained 3 images: the final manual annotation (ground truth) and 2 segmentation results for the same input cell image.

20 subjects were recruited to participate in the experiments. They had either a biological background or experience in cell membrane segmentation and reconstruction. For each group, each of the 20 subjects had three choices. If the subject could tell which segmentation result is more similar to the ground truth, he or she could choose which one. Otherwise, the subject could choose "Difficult to choose". The experiment interface is shown in the Appendix I.

Before the experiment, the subjects were trained on the purpose and source of the images. During the experiment, 200 groups of images were divided into four groups on average in order to prevent the subjects from choosing randomly due to fatigue. For each batch of groups, the subjects needed to complete the judgment continuously without interruption.

After the experiment, for each group, if there were more than 10 votes of the same number, it was called a valid group. Otherwise, it was invalid and discarded. There were a total of 113 valid groups. Then, based on these valid groups, the consistency of the F1 score, IoU, and Dice with human choices was calculated.

According to our experimental results, the consistency of F1 score, IoU, and Dice with human choice was only 34.51%, 35.40%, and 34.51%, respectively. Therefore, it can be inferred that the three criteria are not consistent with human subjective perception in most cases. More results are shown in the Appendix II.

3.3 Perceptual Hausdorff Distance

Based on the subjective experimental results, it was verified that the widely used evaluation criteria for general segmentation were inconsistent with human perception of cell membrane segmentation. This paper proposes a new evaluation standard based on human perception, namely, **Perceptual Hausdorff Distance (PHD for short)**, considering the structure but ignoring the thickness of cell membrane.

An Overview of PHD.

As Fig. 4 shows, from the perspective of neuronal reconstruction, the thickness of the cell membrane is not the key for evaluation. In fact, humans are more sensitive to structure changes, instead of thickness changes. Hence, when measuring the similarity of two cell membrane segmentation results, in order to eliminate the influence of thickness, the segmentation results of two cell membranes were skeletonized, and then the distance between two skeletons was calculated to measure the difference. Since the skeleton is a collection of different points, and Hausdorff distance is a common distance to calculate the difference between two sets of points, the proposed PHD is built upon Hausdorff distance.

On the other hand, through subjective experiments, it was found that people tend to ignore the slight offset between the membrane. Therefore, based on the above two considerations, the Perceptual Hausdorff Distance (PHD) based on Hausdorff distance(Huttenlocher et al. (1993); Aspert et al. (2002); Rachasingho & Tasena (2020)) with modification was designed. Fig. 5 shows the overview of PHD. The details are as follows.

Step 1. Skeletonize the cell membrane. Zhang-Suen thinning algorithm is used (Zhang & Suen (1984); Saudagar & Mohammed (2016)) and re-implemented to obtain the skeleton of the cell membrane.

Step 2. Calculate the distance between skeletons. Hausdorff distance is a common distance used to calculate the difference between two point sets. Consider two unordered nonempty sets of points $\mathbb{X}$ and $\mathbb{Y}$ and the Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y})$ between two point sets. The Hausdorff distance between $\mathbb{X}$ and $\mathbb{Y}$ is defined as

$$d_{\mathrm{H}}(\mathbb{X}, \mathbb{Y}) = \max\left\{d_{\mathbb{X},\,\mathbb{Y}}, d_{\mathbb{Y},\,\mathbb{X}}\right] = \max\left\{\max_{\boldsymbol{x}\,\in\,\mathbb{X}}\{\min_{\boldsymbol{y}\,\in\,\mathbb{Y}} d(\boldsymbol{x}, \boldsymbol{y})\},\ \max_{\boldsymbol{y}\,\in\,\mathbb{Y}}\{\min_{\boldsymbol{x}\,\in\,\mathbb{X}} d(\boldsymbol{x}, \boldsymbol{y})\}\right\}, \tag{1}$$

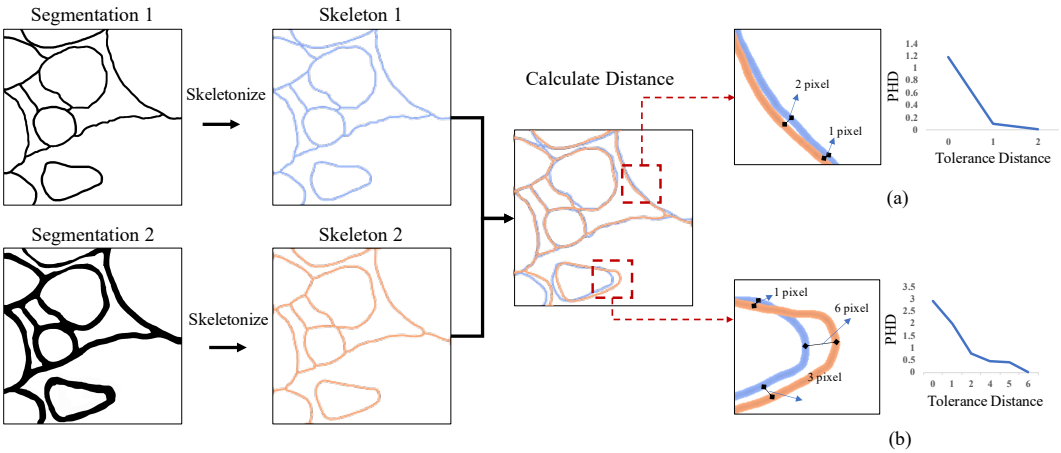


Figure 5: Overview of PHD. The PHD evaluation criterion takes two segmentation results as input. Then, the two inputs are skeletonized. Finally, a distance can be calculated between two skeletons with different tolerance distances. (a) and (b) are two toy cases for intuitively understanding the influence of tolerance distance in PHD.

which can be understood as the maximum value of the shortest distance from a point set to another point set. It is easy to prove that the Hausdorff distance is a metric (Choi (2019)).

In the task of cell membrane segmentation, we should pay attention to the global distance between two point sets, while Hausdorff distance is sensitive to outliers in two point sets. Therefore, the average distance of the two point sets is obtained naturally by using the average operation instead of all the max operations.

Furthermore, it was found that people have tolerance for the small offset between segmentation results. Specifically, if the distance between two points is very small, people tend to ignore it. Therefore, a concept called **Tolerance Distance** t is defined, which represents human tolerance for small errors.

The Perceptual Hausdorff Distance (PHD) is defined as Eq. 2.

$$d_{\mathrm{PHD}}(\mathbb{X},\mathbb{Y}) = \frac{1}{|\mathbb{X}|}\sum_{\boldsymbol{x}\in\mathbb{X}}\min_{\boldsymbol{y}\in\mathbb{Y}} d^*(\boldsymbol{x},\boldsymbol{y}) + \frac{1}{|\mathbb{Y}|}\sum_{\boldsymbol{y}\in\mathbb{Y}}\min_{\boldsymbol{x}\in\mathbb{X}} d^*(\boldsymbol{x},\boldsymbol{y}), \tag{2}$$

$$d^*(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \|\boldsymbol{x}-\boldsymbol{y}\|, & \|\boldsymbol{x}-\boldsymbol{y}\| > t \\ 0, & \|\boldsymbol{x}-\boldsymbol{y}\| \le t \end{cases} \tag{3}$$

To intuitively understand the influence of tolerance distance in PHD, toy cases (a) and (b) as shown in Fig. 5 are taken as examples. In case (a), the blue skeleton scored 19 points while the orange one scored 18 points. Two skeletons are close in the Euclidean Space but do not coincide. Among all the Euclidean distance $d(\boldsymbol{x},\boldsymbol{y})$ of $\boldsymbol{x}\in\mathbb{X}$ and $\boldsymbol{y}\in\mathbb{Y}$, the max distance is 2 pixels, and the most common distance is 1.

When $t=0$, which means no mistake can be tolerated, and the PHD is high. If $t=1$, the PHD value drops a lot. When the $t=2$, PHD becomes 0. In case (b), there is a large offset between two skeletons. When the t is set to $[2,4]$, the decline of PHD value is slow. When $t=6$, it drops to 0, which is the max distance between two point sets of skeletons. The Different settings of t represent the degree of tolerance to the distance between the two skeletons. In practical applications, different tolerance distances can be adopted according to different situations.

Consistency between PHD and human perception.

The consistency of PHD results in the subjective experiment with human perception was also calculated (as described in Section 3.2). The result showed that compared with F1 score, IoU and Dice, PHD with appropriate tolerance distance was more consistent with human perception.

As shown in Fig.6, while tolerance distance t of PHD increasing from 0 to 800, its consistency to human perception rose first and then dropped slowly to 0, suggesting human vision does have tolerance for certain offset. Specifically, the maximum value can be reached at 65.48%, when tolerance distance t was set to 3, suggesting that our perceptions prefer to tolerant small perturbations. It was

worthy to note that the optimal PHD score (65.48%) was nearly double of the consistency scores obtained by F1 score, IoU, or Dice.

Can F1 score, IoU, or Dice be improved by skeletonizing the segmentation before evaluation? Our experiment refutes this point of view. In Fig. 6, as shown by yellow, light green, and light blue bars, the consistency of the F1 score, IoU, and Dice with human perception calculated based on image skeletons can only reach 44.25%, 44.25%, 34.51%, respectively. They are still far from PHD performance with $t = 10$. Therefore, only extracting image skeleton can not eliminate the shortcomings of existing metrics.

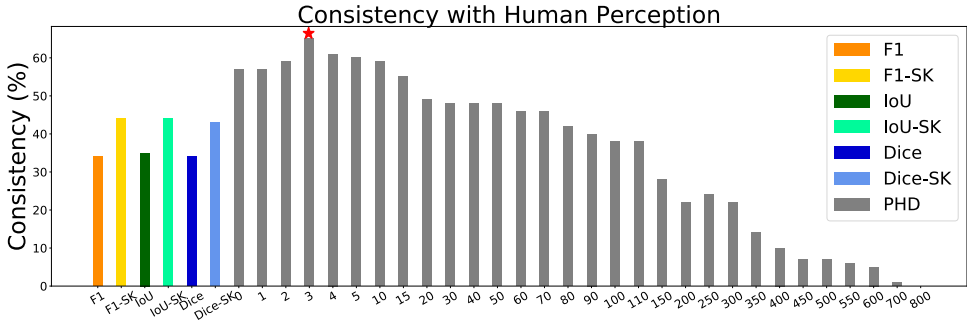


Figure 6: Consistency with human perception. The left 6 bars show the consistency results of F1 score, IoU, and Dice without and with skeletonization (-SK). The gray bars show the consistency result of PHD. The numbers on X-axis mean the tolerance distance settings.

4 Re-examining PHD on Classic Deep Learning Based Segmentation Methods with U-RISC

In the previous two sections, we proposed a new ultra-high resolution cell membrane segmentation dataset U-RISC and a new perceptual criteria PHD to help solve the two bottlenecks in the field of cell membrane segmentation. The subjective experiment on a small-scale dataset demonstrated that PHD is more consistent with human perception for the evaluation of cell membrane segmentation than some widely used criteria.

In order to understand the performance of deep learning methods on the U-RISC dataset, we conducted an in depth investigation on U-RISC with representative deep learning based segmentation methods and different evaluation criteria. To be specific, we chose 6 representative algorithms (U-net (Ronneberger et al. (2015)), LinkNet(Chaurasia & Culurciello (2017)), CASENet (Yu et al. (2017)), SENet (Hu et al. (2018)), U-Net++ (Zhou et al. (2018)),and GLNet (Chen et al. (2019))) and re-implemented them on U-RISC dataset. Then four evaluation criteria were used to compare the segmentation results: F1 score, IoU, Dice, and PHD.

As mentioned in Sec. 2, the results of the first two rounds of manual labeling results are retained. Therefore, the results of manual annotation under different evaluation criteria will also be analyzed.

Experiment Settings. The experiment settings of all segmentation algorithms were the same.

In the training stage, 60% of the dataset was used as the training data, and then the original image was randomly cut into 1024 $\times$ 1024 patches to generate 50,000 training images and 20,000 validation images.

Random flipping and clipping were used for data augmentation. The training parameters were consistent with those reported by original authors. Four V100 GPUs were used to train each algorithm. In the testing stage, the original image was cut into the same size of training image, and the patch was tested. These patches were eventually spliced back to the original size for evaluation.

Experiments Results. The experiment results were shown in Table. 1. The table showed the scores of different evaluation criteria on the first two rounds of manual annotation results and six segmentation results with the ground truth.

Our first finding was that U-RISC was a challenging dataset in the field of cell membrane segmentation. As shown in Table. 1,the performance of deep learning based methods gained around 0.6 in

Table 1: Experiments on U-RISC Dataset. This table shows the four different evaluation results of the first two rounds of human annotations and six segmentation results: U-Net (Ronneberger et al. (2015)), LinkNet(Chaurasia & Culurciello (2017)), CASENet (Yu et al. (2017)), SENet (Hu et al. (2018)), U-Net++ (Zhou et al. (2018)),and GLNet (Chen et al. (2019)). PHD-t means the PHD score with tolerance distance t.Note that the ground truth is the third round of human annotation.

Methods	**F1** ↑	**IoU** ↑	**Dice** ↑	**PHD-0** ↓	**PHD-1** ↓	**PHD-3** ↓	**PHD-5** ↓
Label 1st	0.9840	0.8616	0.9212	0.2102	0.0576	0.0456	0.0419
Label 2nd	0.9953	0.9901	0.9933	0.0209	0.0145	0.0112	0.0084
GLNet	0.5123	0.3233	0.4883	1.7646	1.6311	1.2905	1.0018
U-Net	0.5224	0.3541	0.5212	1.7533	1.5892	1.2009	0.9101
SENet	0.5810	0.4107	0.5810	1.7404	1.5744	1.1694	0.8643
CASENet	0.6007	0.4307	0.6007	1.7252	**1.5495**	**1.1345**	0.8207
U-Net++	0.6030	0.4329	0.6030	**1.7251**	1.5535	1.1448	0.8351
LinkNet	**0.6070**	**0.4371**	**0.6070**	1.7254	1.5508	1.1357	**0.8175**

F1-scores, far below the human level (0.98-0.99) on U-RISC dataset, by contrast, they all exceeded 0.95 on the ISBI 2012. Despite possible improvements by parameter tuning, to such ultra-high resolution images, there was clearly a huge gap between the STOA segmentation methods and human performance.

Our second finding was that evaluation rankings for F1-score, IoU, and Dice were more consistent with each other, but different from PHD-based rankings. According to the subjective experimental results of Sec. 3.2, PHD was much closer to human perception. Therefore, the change of ranking led by PHD may also inspire researchers to re-consider the evaluation criteria for cell membrane segmentation algorithms. It also provides a new perspective for promoting the development of segmentation algorithms.

Discussion. Based on the results of these six algorithms, it can be seen that LinkNet and CASENet are better than other methods. From the perspective of network design, LinkNet makes full use of the low-level local information and directly connects the low-level encoder to the decoder of corresponding size. This design pays more attention to the capture of local information, which leads to a more accurate local prediction. CASENet takes full account of the continuity of the edge and makes the low-level features strengthen the high-level semantic information by jumping links between the low-level feature and high-level feature, which pays more attention to structural information. Therefore, the design of LinkNet might be preferred by the traditional evaluation criteria, while CASENet might be preferred by the PHD. This also explains why the two methods rank differently under these two types of evaluation criteria.

More local segmentation results of different algorithms are shown in the Appendix III.

5 Conclusion

This paper aims to solve the two bottlenecks in the development of cell membrane segmentation. Firstly, we proposed U-RISC,Ultra-high Resolution Image Segmentation dataset for Cell membrane, the largest annotated EM dataset for the Cell membrane so far. To our best knowledge, U-RISC is the only uncompressed annotated EM dataset with multiple iterative annotations and uncompressed high-resolution raw image data. During the analysis process of the U-RISC, we found a certain inconsistency between current evaluation criteria for segmentation (e.g. F1 score, IoU) and human perception. Therefore, this article secondly proposed a human-perception based evaluation criterion, called Perceptual Hausdorff Distance (PHD). Through a subjective experiment on a small-scale dataset, experiments results demonstrated that the new criterion is more consistent with human perception for the evaluation of cell membrane segmentation. In addition, the evaluation criteria of PHD and existing classic deep learning segmentation methods are re-examined. In future research, we will consider how to improve deep learning segmentation methods from the perspective of cell membrane structure and apply PHD criterion for connectomics research.

REFERENCES

James R Anderson, Bryan W Jones, Carl B Watt, Margaret V Shaw, Jia-Hui Yang, David DeMill, James S Lauritzen, Yanhua Lin, Kevin D Rapp, David Mastronarde, et al. Exploring the retinal connectome. *Molecular vision*, 17:355–79, 2011.

Nicolas Aspert, Diego Santa-Cruz, and Touradj Ebrahimi. Mesh: Measuring errors between surfaces using the hausdorff distance. In *Proceedings. IEEE international conference on multimedia and expo*, volume 1, pp. 705–708. IEEE, 2002.

Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4. IEEE, 2017.

Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8924–8933, 2019.

Jiwon Choi. Gromov-hausdorff distance between metric graphs, 2019.

Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pp. 2843–2851, 2012.

Alan Curry, Hazel Appleton, and Barry Dowsett. Application of transmission electron microscopy to the clinical study of viral and bacterial infections: present and future. *Micron*, 37(2):91–106, 2006.

Malcolm R Dando. Japan's brain/minds project. In *Neuroscience and the Problem of Dual Use*, pp. 127–147. Springer, 2020.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187. Springer, 2016.

ROBERT A ERLANDSON. Role of electron microscopy in modern diagnostic surgical pathology. *Modern Surgical Pathology*, pp. 71, 2009.

Alfie Gleeson and Abigail Sawyer. Mapping a brain, 2018.

Kristen M Harris, Elizabeth Perry, Jennifer Bourne, Marcia Feinberg, Linnaea Ostroff, and Jamie Hurlburt. Uniform serial sectioning for transmission electron microscopy. *Journal of Neuroscience*, 26(47):12101–12103, 2006.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9): 850–863, 1993.

Thomas R Insel, Story C Landis, and Francis S Collins. The nih brain initiative. *Science*, 340(6133): 687–688, 2013.

ISBI 2012. Segmentation of neuronal structures in em stacks challenge - ISBI 2012. `http://brainiac2.mit.edu/isbi_challenge/`, 2012.

ISBI 2013. Segmentation of neuronal structures in em stacks challenge - ISBI 2013. `http://brainiac2.mit.edu/SNEMI3D/`, 2013.

Sven Kosub. A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, 120:36–38, 2019.

Jeff W Lichtman, Hanspeter Pfister, and Nir Shavit. The big data challenges of connectomics. *Nature neuroscience*, 17(11):1448–1454, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Mu-ming Poo, Jiu-lin Du, Nancy Y Ip, Zhi-Qi Xiong, Bo Xu, and Tieniu Tan. China brain project: basic neuroscience, brain diseases, and brain-inspired computing. *Neuron*, 92(3):591–596, 2016.

Jumpol Rachasingho and Santi Tasena. A metric space of subcopulas—an approach via hausdorff distance. *Fuzzy Sets and Systems*, 378:144–156, 2020.

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*, pp. 234–241. Springer, 2015.

Yutaka Sasaki et al. The truth of the f-measure. 2007, 2007.

Abdul Khader Jilani Saudagar and Habeeb Vulla Mohammed. Opencv based implementation of zhang-suen thinning algorithm using java for arabic text recognition. In *Information Systems Design and Intelligent Applications*, pp. 265–271. Springer, 2016.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5964–5973, 2017.

TY Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

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*, pp. 3–11. Springer, 2018.

A APPENDIX

I. Fig. 7 is the interface of perceptual consistency experiments.

II. Fig. 8 and Fig. 9 are some examples of subjective experiment images.

III. Fig. 10 and Fig. 11 are some examples of segmentation results of different algorithms.

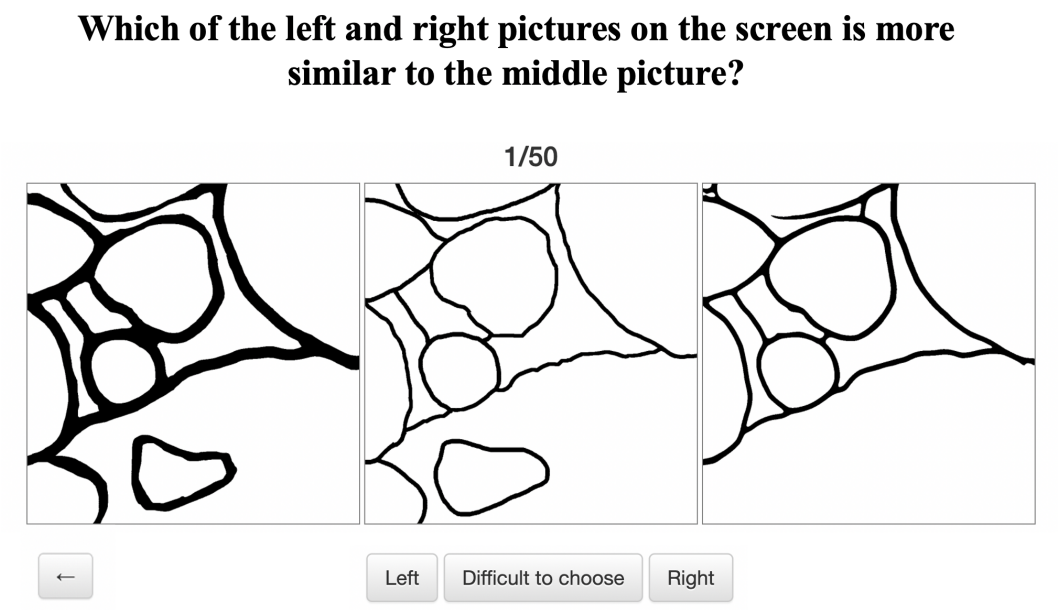


Figure 7: Interface of perceptual consistency experiments.

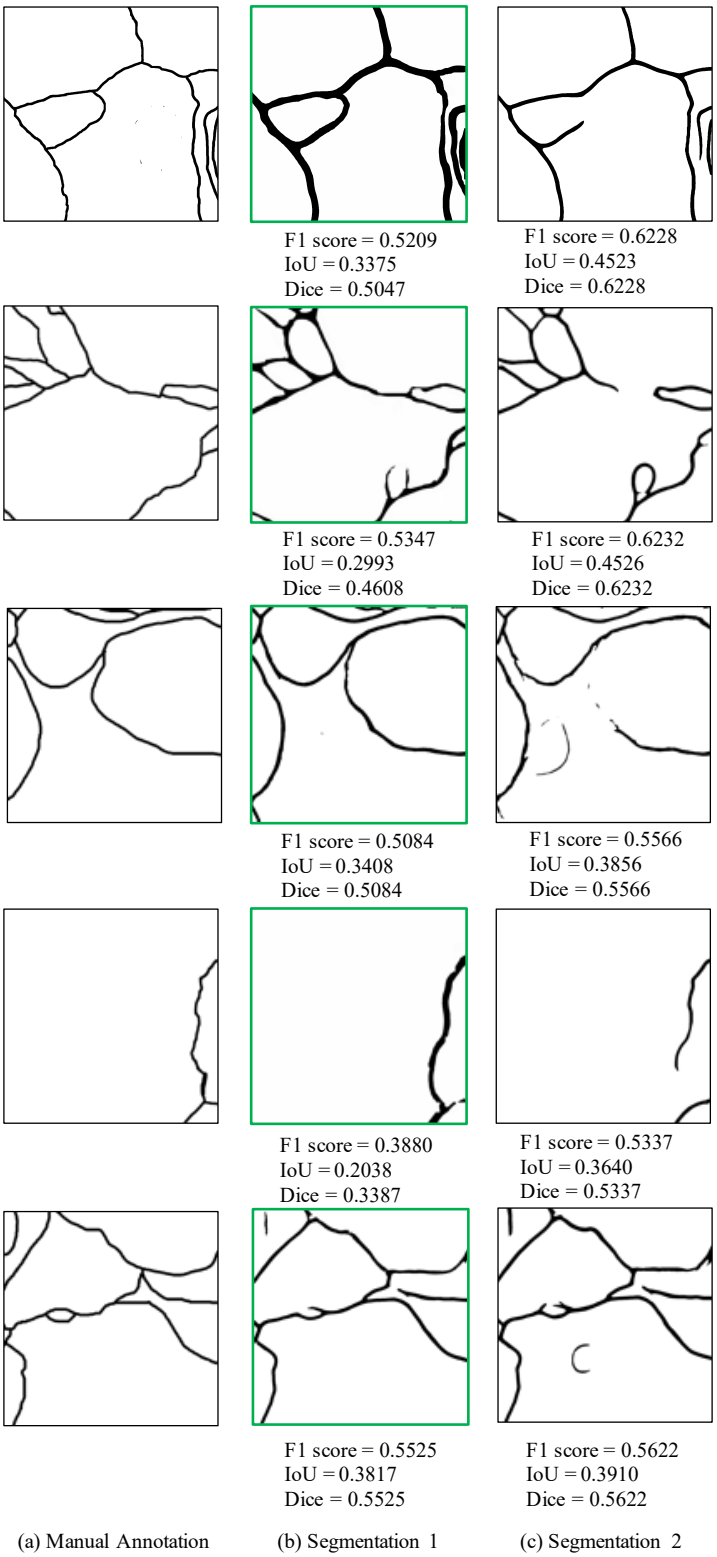


Figure 8: Examples of subjective experiment images (1). The figure in green box is the choice of most subjects. The scores of F1 score, IOU and Dice are only used to illustrate the inconsistency between the three criteria and human perception, which are not shown to the subjects during the subjective experiments.

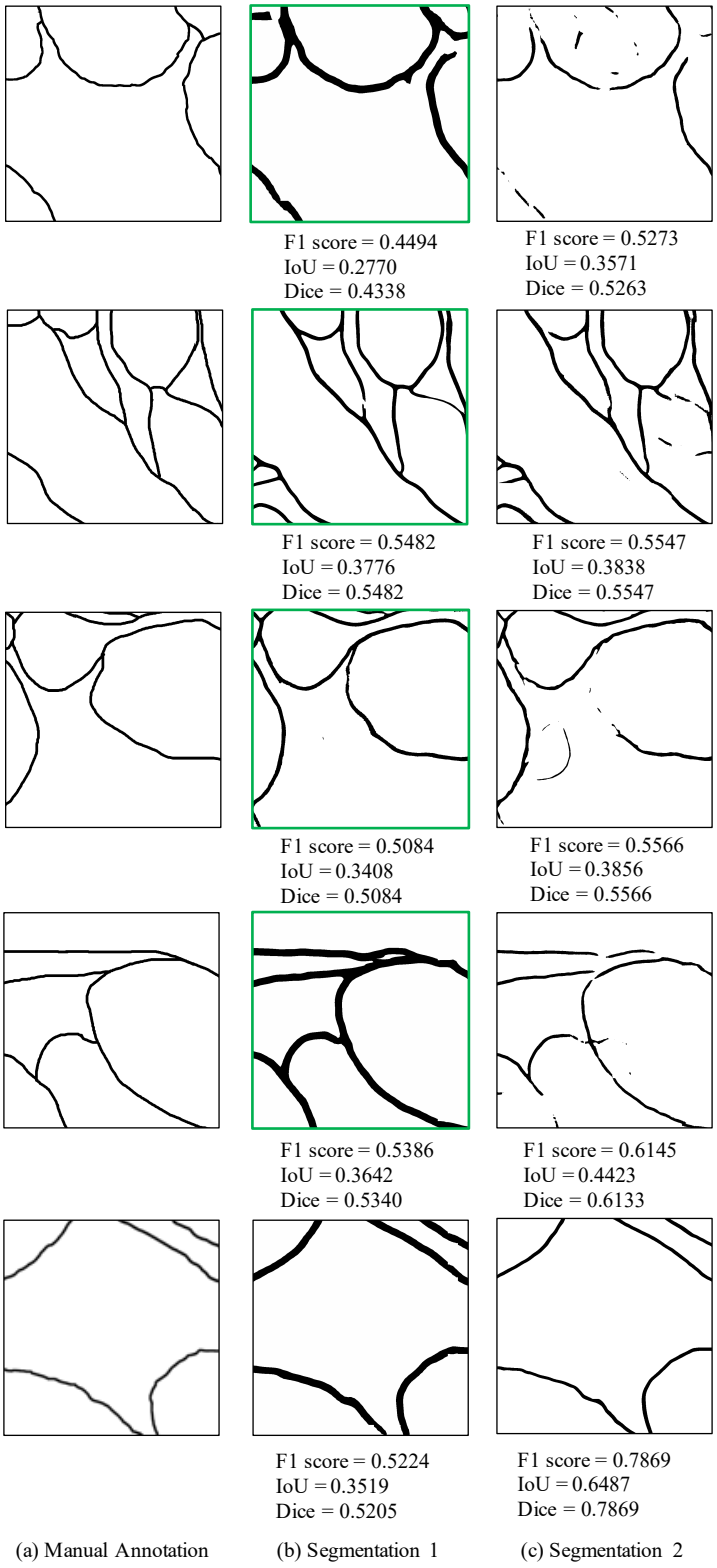


Figure 9: Examples of subjective experiment images (2). The figure in green box is the choice of most subjects. The scores of F1 score, IOU and Dice are only used to illustrate the inconsistency between the three criteria and human perception, which are not shown to the subjects during the subjective experiments. For the example in the last line, most of the subjects chose "Difficult to choose".

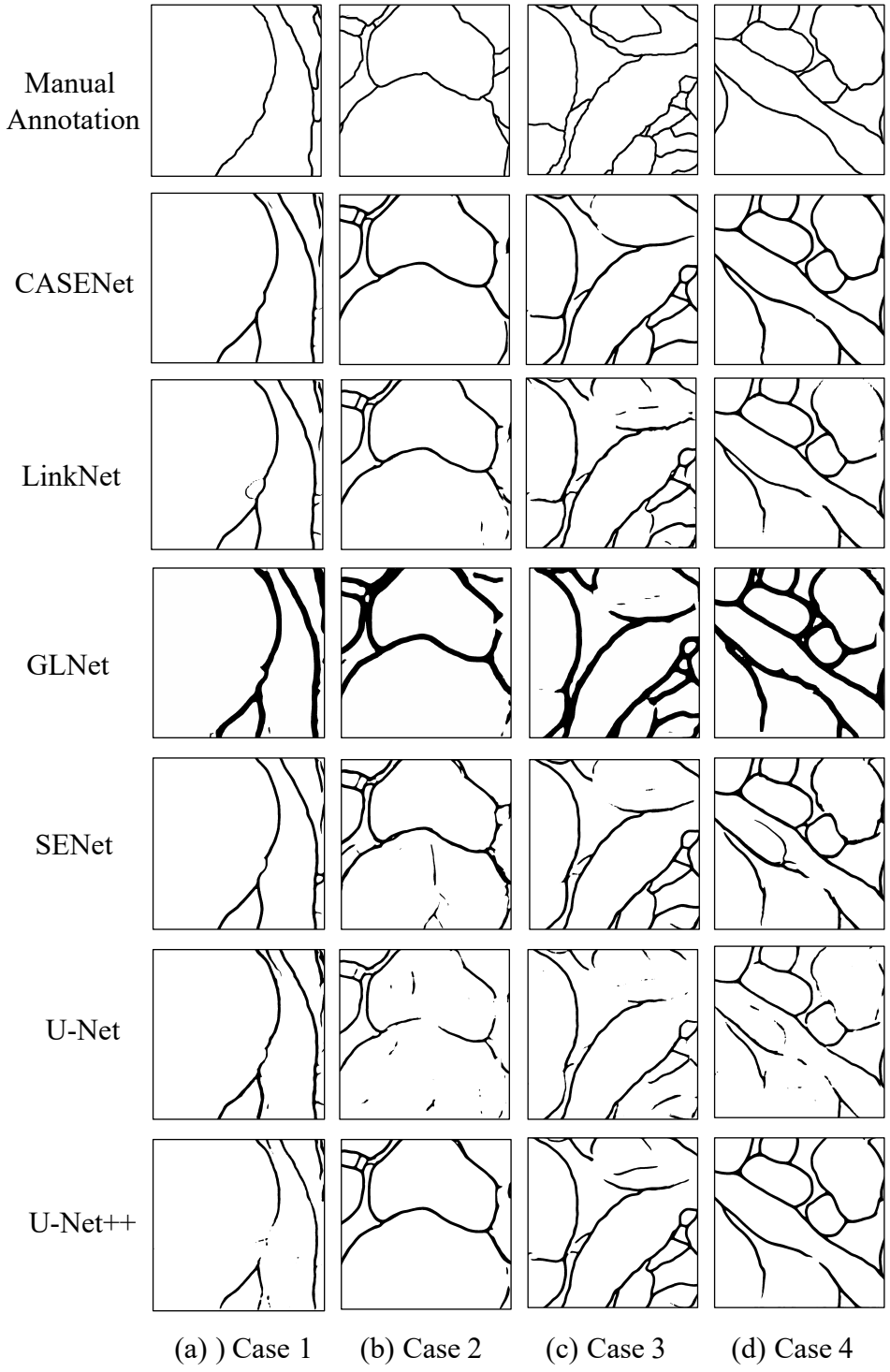


Figure 10: Examples of segmentation results (1).

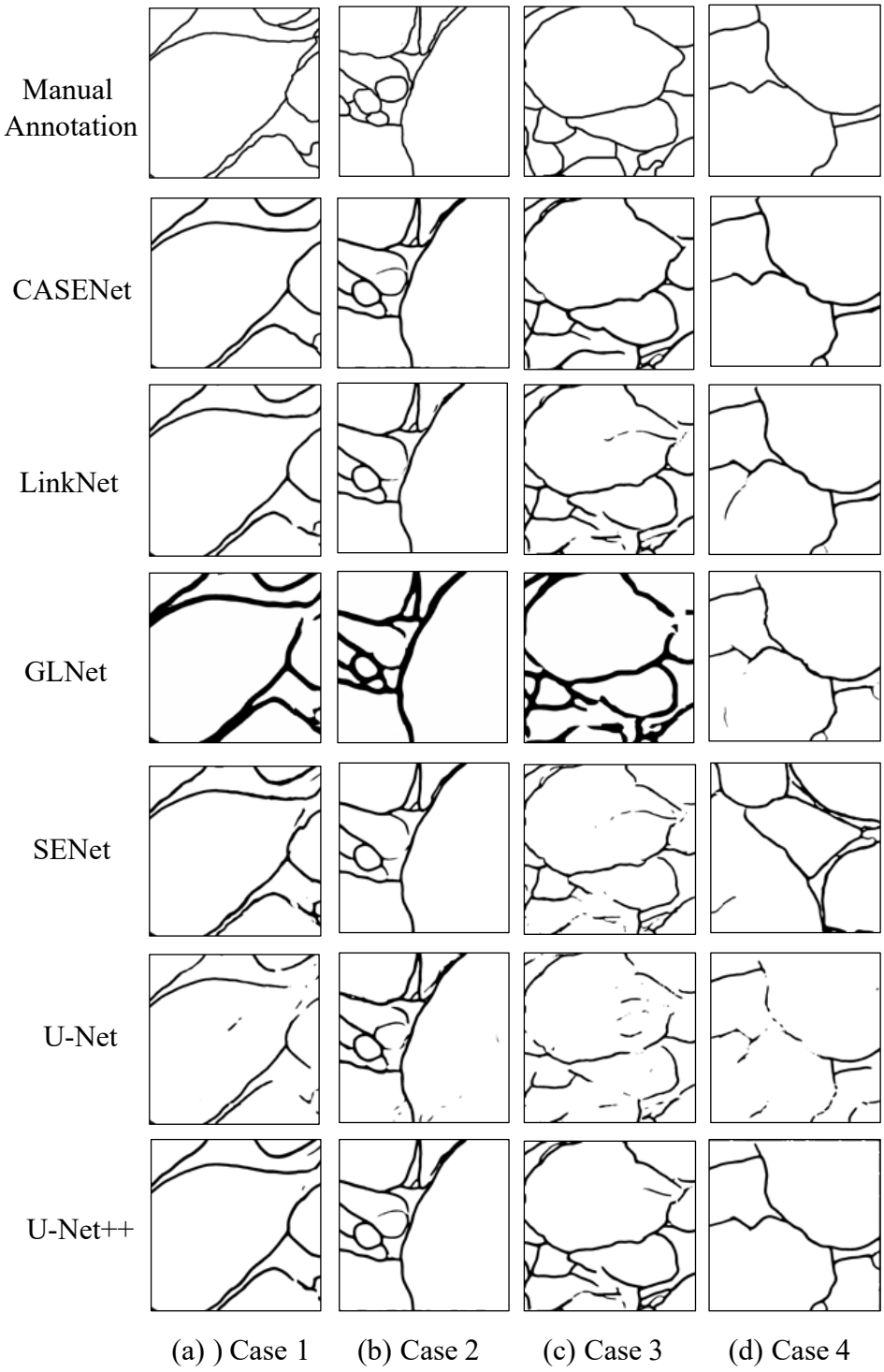


Figure 11: Examples of segmentation results (2).