ULTRA-HIGH RESOLUTION IMAGE SEGMENTATION DATAET FOR CELL MEMBRANE WITH PERCEPTUAL HAUSDORFF DISTANCE EVALUATION CRITERION

Anonymous authors
Paper under double-blind review

ABSTRACT

Computer vision technology is widely used to assist the analysis and understanding of biological and medical data. However, there are two bottlenecks in the development of supervised learning: the lack of enough training data and appropriate evaluation criteria. To alleviate the two problems, in this paper, we propose an Ultra-high Resolution Image Segmentation dataset for Cell membrane, called U-RISC. In particular, we retain the results of the two manual checks and the final annotation. In the process of using the state-of-the-art deep learning methods to segment our dataset, we find a phenomenon that the general segmentation evaluation criteria are inconsistent with human perception. Through subjective experiments, we prove the universality of this phenomenon. Therefore, we propose a Perceptual Hausdorff Distance (PHD) evaluation criterion to measure the quality of cell membrane segmentation task. The results of two manual checks and different segmentation methods on different evaluation criteria are also tested.

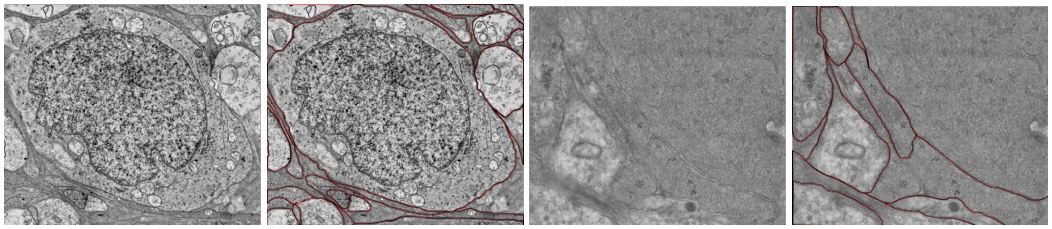
1 INTRODUCTION

A central theme of neuroscience is to understand how a brain's functions are related to its neuronal structure. Neuroscience calls the study of neuron distribution and connection "connectomics". Advancement of microscopic technologies established significant progress in our understanding of it and imaging of large volumes of neuronal tissue at high resolution, allowing neuroscience experts to reconstruct neuronal circuits and study the interconnections of neurons.

Segmentation is an active area of research in biological and medical image analysis. Membrane segmentation is widely used to study the morphology and connection of neurons, which is the essential first step of the reconstruction pipeline.

With the introduction of Convolutional Neural Networks (CNNs) and the rapid development of deep learning, significant improvements in performance have been achieved in many standard datasets. Recent deep learning-based neuron segmentation methods are well suited to deal with ambiguous neuronal boundaries in EM images. 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. With deep supervision and side outputs on FCN, a holistically-nested edge detector (HED) was then proposed by Xie & Tu (2015) to solve the ambiguity in edge and object boundary detection. Since then, a lot of works (Drozdzal et al. (2016),Hu et al. (2018),Zhou et al. (2018),Chaurasia & Culurciello (2017),Yu et al. (2017),Chen et al. (2019)) has been proposed based on U-net and FCN structure and achieved outstanding performance in neuron membrane segmentation.

The method based on machine learning has high requirements for accurate data annotation. Due to the high resolution of the neuron image, there are a large number of neurons in an image, so each contour of the neuron needs to be labeled manually, which is time-consuming and expensive. Several datasets for the segmentation of neuronal neuron membrane has been proposed. ISBI 2012 (ISBI 2012 (2012)) provided a full stack of EM slices is used to train machine learning algorithms for automatic segmentation of neural structures. ISBI 2013 (ISBI 2013 (2013))provided a larger



(a) Example 1. (b) Example 2.

Figure 1: Examples of images with their label. The left images of (a) (b) are two original image parts, the right iamges are the image covered by their labels (red lines).

dataset SNEMI3D to expand it into 3D. The evaluation they used is 1 - the maximal F-score of the Rand index, which calculates pixel error of the segmentation results.

However, the improvement of segmentation quality is limited by two aspects. (1) Lack of high-precision labeling training data. It is very difficult to label the neuron membrane accurately. The number, size, and resolution of previous several available training datasets are small and low, which can not meet the demand for research. (2) Methods and evaluation based on pixel-level might not consistent of human perception. According to our observation, many segmentation results are not consistent with subjective judgment. The different thickness will have a great impact on the evaluation results. However, the structural information of the neuron membrane which we are most concerned about is ignored.

In this paper, in alleviating the above two problems, we first propose a new super high-resolution neuron membrane segmentation dataset in section 2, including accurate manual annotation, which is more than 100 times the size of the previous dataset. Then, in view of the shortcomings of pixel-level evaluation criterion, a new evaluation criterion based on human perception is proposed in section 3. To verify the superiority of these criterion, we show some subjective experiments in the same section. In the end, the newly proposed evaluation method is applied to the new dataset for some widely used segmentation methods of the neuron membrane. We also compared human labeling results in the different inspection processes. The results are presented in section 4 and discussed in the last section.

2 DATASET

Supervised learning methods rely heavily on high-quality datasets. To alleviate the lack of training data for cell membrane segmentation, we propose an Ultra-high Resolution Image Segmentation dataet for Cell membrane, called U-RISC. The dataset is labeled on the published ssTEM Images Anderson et al. (2009) of the rabbit retinal GC layer Marc (1999) taken from Dutch Belted rabbits (Oregon Rabbitry). In this work, we cut out 120 images in the size of 10000×10000 from different layers of original ssTEM Images. Then, we directly manually labeled the cell membranes on the ultra-high resolution images. Since the human labeling process is very valuable for research, we also keep the intermediate results of 3 check and correction processes. Now the dataset is released on `https://Anonymous.com`.

2.1 LABELING PROCESS

Due to the high resolution of the images in our dataset, it's arduous to label all the cell membranes accurately. In the process of labeling, the situations of a lot of small organelle membranes (Fig.1(a)) or blur image quality (Fig.1(b)) is common, which will lead to the difficulty of labeling. As shown in Fig.6, (a)(b) are two examples of labeling processes. The area shown in the Fig.6 only accounts for less than 1% of the whole image. Therefore, In the process of long-time and high-intensity focused annotation, people are easy to get tired and ignore the indistinguishable cell membrane, or confuse organelle membranes with cell membranes.

In order to avoid such mistakes and correct them, we set up a correction mechanism in the labeling process. First, before the formal start of labeling, some annotators are trained to manually label some images guided by 5 experts. After training, 58 qualified annotators are allowed to participate in the final labeling process. Then, Five experts are responsible for checking whether the first labeling result is correct pixel by pixel. This cross-check was repeated three times. Finally, the labeling results after two checks are saved as the final labeling results. At the same time, the results of the previoustwo manual annotations are also saved for later analysis. Fig.2 shows the 2 examples of inspection process. The area of white dashed boxes shows that the wrong annotation is corrected at the second inspection, while the wrong in the area of yellow dashed boxes is corrected at the first inspection.

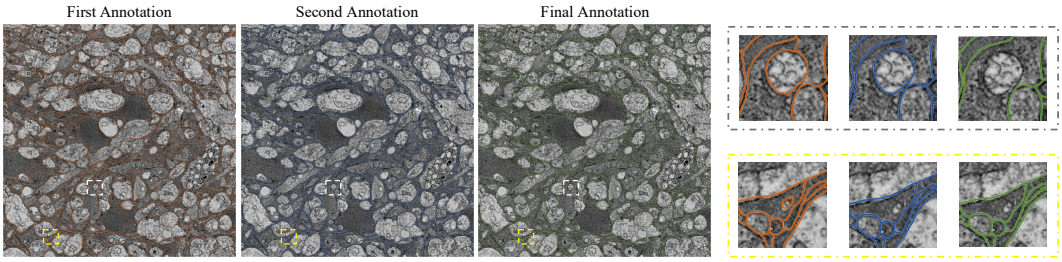
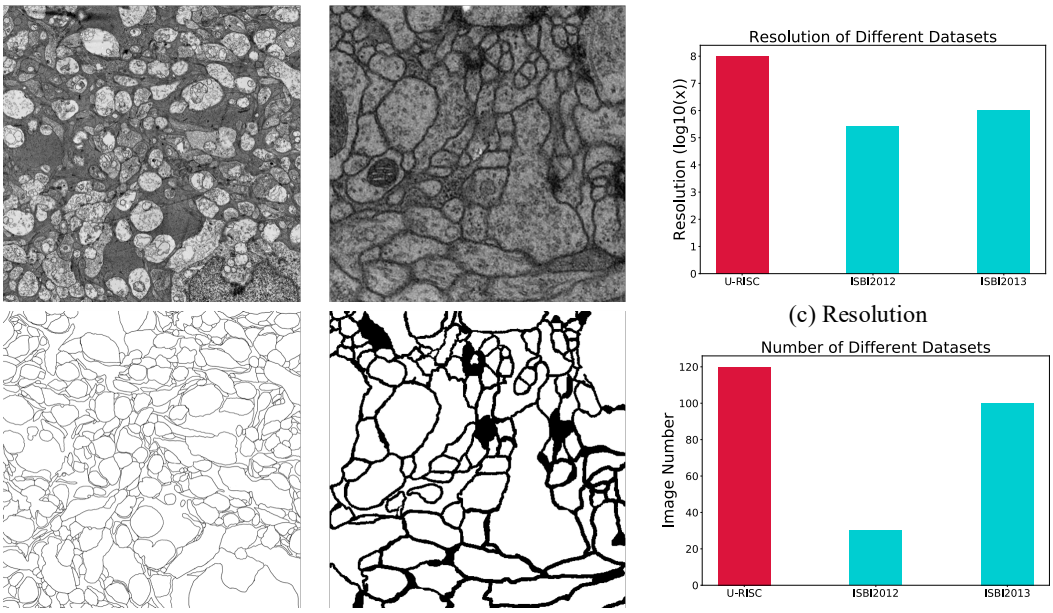


Figure 2: Example of three manual annotation modification results. The left three images are the first, second, and final manual annotation results. The small figures in the gray and yellow dashed boxes are details of the white and yellow boxes in the left three images, respectively.

2.2 Comparison with other datasets



(a) U-RISC (b) ISBI2012 (c) Resolution (d) Number of images

Figure 3: Dataset examples and statistical comparison. (a) U-RISC. (b) ISBI2012. (c) The resolution of U-RISC, ISBI2012 and ISBI2012 datasets. (d) The image number of the three datasets.

Although there are some published cell membrane segmentation datasets (ISBI 2012 (2012),(ISBI 2013 (2013))), they are small in size and have low resolution. As shown in Fig. 3, where (a) is an example of U-RISC, and (b) comes from ISBI 2012. We can see that U-RISC dataset contains

more cells in a single image than ISBI 2012, and the details are more abundant. Figure. 3 (a) shows that the resolution of each image in U-RISC is 400 times the size of the ISBI 2012 dataset while 100 times of the ISBI 2013 dataset. What's more, the number of published images is also higher than the ISBI 2012 and 2013. Therefore, U-RISC is more challenging in the field of cell membrane segmentation.

3 EVALUATION

In Section 2, we propose an annotated ultra-high resolution dataset for cell membrane segmentation. With the development of deep learning, there are many segmentation methods and criteria to predict the cell membrane and evaluate the results (Zhou et al. (2018),Chaurasia & Culurciello (2017),Yu et al. (2017),Chen et al. (2019))).

When using the state-of-the-art segmentation methods to obtain some segmentation results, we found that the general evaluation criteria are inconsistent with human perception in some cases, which is confirmed by some perceptual consistency experiments. Therefore, we propose a new evaluation criterion called Perceptual Hausdorff Distance (PHD). The experimental results show that it is more consistent with human perception.

3.1 MOTIVATION

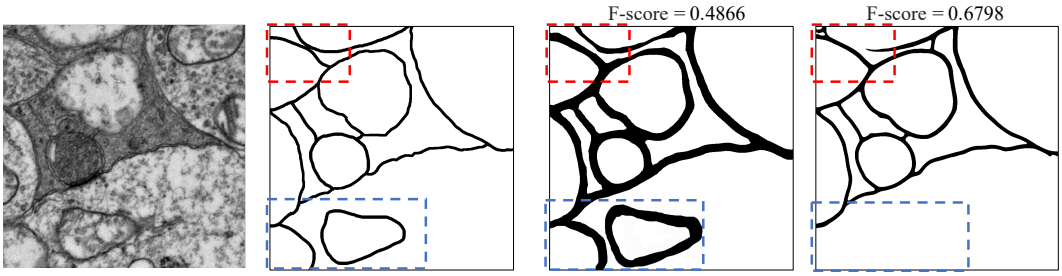
In recent work, most of the segmentation methods (Ronneberger et al. (2015), Zhou et al. (2018),Chaurasia & Culurciello (2017),Yu et al. (2017),Chen et al. (2019))) follow the general image segmentation evaluation criteria to evaluate the similarity of two images, such as F1 score, IoU, and Dice Coefficient. F1 score is the harmonic mean of precision and recall. It considers the accuracy and recall rate of the classification model, which can be regarded as a weighted average of accuracy and recall rate. IoU (Intersection-Over-Union) calculates the ratio of the area of overlap between two images, which is first used to measure the accuracy of an object detector at the beginning and later extended to segmentation tasks. Dice also calculates the size of the overlap of the two images but divided by the total size of the two objects. All of the evaluation criteria above are the statistics of whether the pixels are classified correctly.

However, such statistics might not be consistent with human perception. As F1 score is the most commonly used evaluation criterion for cell membrane segmentation, we take it as an example to give an illustration. Fig.3.1 shows an example of the original image with its manual annotation and two segmentation results from methods GLNet (Chen et al. (2019)) and U-Net (Ronneberger et al. (2015)). Take the F1 score as an evaluation criterion, the scores above Fig.3.1 (c) and (d) are the F1 score between themselves and (b). The scores indicate that (d) is more similar to (b) than (c). However, when we look at the area surrounded by the red and blue dotted lines in the images, our first instinct is that (c) is a better prediction, because (d) misses a lot of structures. The reason why the F1 score of (d) is higher is that the thickness of (d) is closer to that of manual annotation. So we deem that the F1 score is not sufficiently robust to variations in the thickness and structures of the membrane, and the evaluation result is inconsistent with human perception.

3.2 PERCEPTUAL CONSISTENCY EXCERIMENTS

To prove that the above inconsistency is a common phenomenon, we design a subjective experiment to explore the consistency between F1 score and human subjective perception.

The purpose of the experiment is to let the subjects choose which of the 2 cell segmentation results is more similar to the manual annotation. First, we generated cell membrane segmentations of the proposed U-RISC dataset by six state-of-the-art segmentation methods: 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)). Then, for each sub experiment, the subjects are given a group of three images, which contains 2 segmentation results and their corresponding manual annotation. The 2 segmentation results are randomly selected from the six methods. The experiment interface is shown in the appendix I.



(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.

We invited 20 subjects to participate in this subjective experiment. They are with biological background or participating in cell membrane segmentation and reconstruction works. Before the experiment started, the subjects are told about the purpose and image source of the experiment.

During the experiment, subjects are not allowed to interrupt the experiment for any reasons. To prevent the subjects from random choices due to feeling tired, the total of 200 groups of images are equally divided into four batchs.

After the experiment, we regard the experiments with more than half of the unanimous choices as valid experiments, and others as invalid ones. There are 113 valid experiments. Then we calculated the consistency of the F1 score with human choices in these valid experiments.

Consistency between F1 score and human perception. According to our experimental results, the consistency between the human choice and F1 score is only 34.51%, which means F1 score is not consistent with human subjective perception in most cases. Fig.6(a) shows two examples. The first row shows the manual annotation images. The second and third row show two segmentation results. And the F1 score above the images is caculated by themselves and its manual annotation. Fig.6(a) shows that the thickness and slight deviation of cell membrane dominate the judgment of F1 score, but the information of structure is not paid attention to.

3.3 Perceptual Hausdorff Distance

Based on the above phenomenon, in this section, we propose a new evaluation criterion based on human perception, called **Perceptual Hausdorff Distance (PHD)**, considering the structure of the cell membrane while ignoring the thickness.

As Fig.3.1 shows, the thickness of the cell membrane and slight offset between lines are not our major concern. In fact, the lack of structure can arouse people's attention. To eliminate the influence of thickness, we firstly skeletonize the results of cell membrane segmentation, and then calculate the distance between the two skeleton point sets to measure the difference of the two cell membrane segmentation. Hausdorff distance (Huttenlocher et al. (1993),Huttenlocher et al. (1993),Huttenlocher et al. (1993)) is a good metric for comparison of structural differences. In addition, from the perspective of cellal reconstruction, the slight skeleton deviation does not have a significant impact. So considering the overall distance of skeletons and tolerance of slight offset, we design the Perceptual Hausdorff Distance (PHD) based on Hausdorff distance. The details are as follows.

Fig.5 shows the overview of PHD. When we input two cell membrane segmentation images, we first skeletonize the two images to extract the skeleton of the cell membrane. Then, the Perceptual Hausdorff Distance is calculated between skeletons.

Step1. Skeletonize the cell membrane. Zhang-Suen thinning algorithm (Zhang & Suen (1984),Saudagar & Mohammed (2016)) is an algorithm used to skeletonize a black and white i.e. one bit per pixel images. So the input images need to be normalized to $\{0,1\}$ before the skeletonization process. Here we re-implemented this algorithm to obtain the skeleton of the cell membrane.

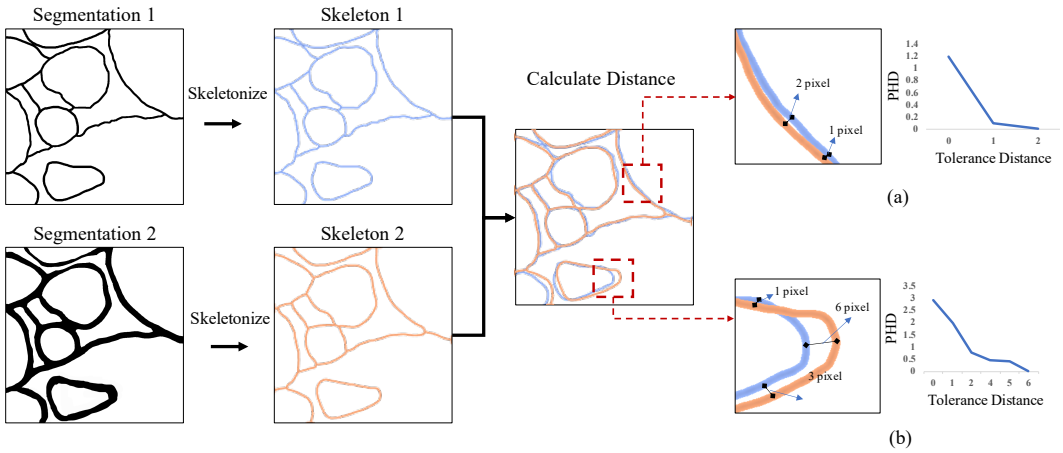


Figure 5: Overview of the evaluation criterion. First, two segmentation results are used as inputs. Then, the two skeletons are obtained by skeletonizing the two images. Finally, the PHD can be calculated between two skeletons with different tolerance distances. (a) and (b) are two toy cases for intuitively understanding the influence of tolerance distance on PHD.

Step2. Calculate the distance between skeletons. As mentioned above, the evaluation criteria based on pixel classification error is unreasonable. What we care about is the structural distance between the two skeletons. Therefore, based on the Hausdorff distance, we propose the Perceptual Hausdorff Distance (PHD) to measure the distance between two skeletons.

First, we briefly review the Hausdorff distance. Consider two unordered nonempty sets of points $\mathbb{X}$ and $\mathbb{Y}$ and a distance metric $d(\boldsymbol{x}, \boldsymbol{y})$ between two point sets $\boldsymbol{x} \in \mathbb{X}$ and $\boldsymbol{y} \in \mathbb{Y}$. The function $d(\cdot, \cdot)$could be any metric. In our case we use the Euclidean distance. The sets $\mathbb{X}$ and $\mathbb{Y}$ may have different number of points. Let $\Omega \subset \mathbb{R}^2$ be the space of all possible points. In its general form, the Hausdorff distance between $\mathbb{X} \subset \Omega$ and $\mathbb{Y} \subset \Omega$ is defined as

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

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)).

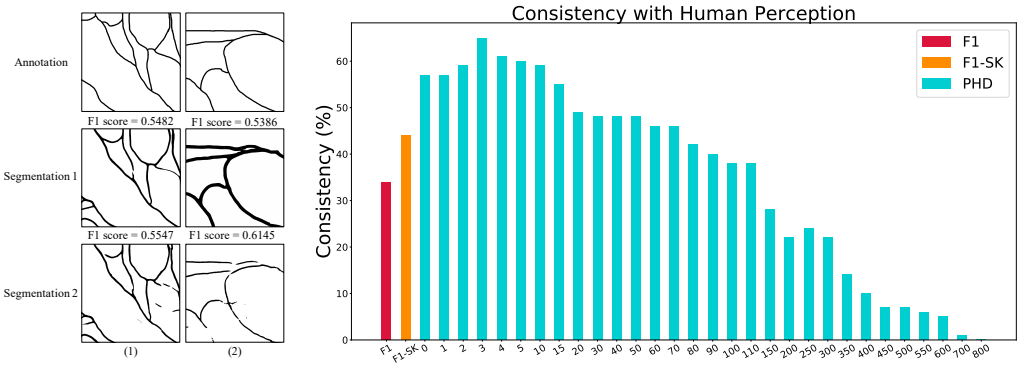
In our case, what we care about is the overall distance between two point sets, while Hausdorff distance is sensitive to outliers in two point sets. So naturally, we replace all max operations with average operations to obtain the mean distance of two point sets.

Furthermore, we found that humans are tolerant of slight deviations between cell membranes. If the distance between two points is small, we tend to ignore it and think that it is approximately similar. So we propose a concept called **Tolerance Distance** t, which represents human tolerance for slight skeleton excursions. Our Perceptual Hausdorff Distance (PHD) is designed as equation 2.

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

$$d(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \|\boldsymbol{x} - \boldsymbol{y}\|, & \|\boldsymbol{x} - \boldsymbol{y}\| > t \\ 0, & \|\boldsymbol{x} - \boldsymbol{y}\| \leq t \end{cases} \quad (3)$$

To intuitively understand the influence of tolerance distance in PHD, we take two toy cases (a) and (b) as shown in Fig.5 as an example.



(a) Some examples.

(b) Consistency of different evaluation criteria.

Figure 6: (a) Examples of inconsistency between F1 score and subjective choice. (b) Consistency of different evaluation criteria. The red and orange bar shows the consistency results of F1 score and the skeleton F1 score. The green bars shows the consistency result of PHD. The numbers in X-axis means the tolerance distance settings.

In case (a), the blue skeleton has 19 points while the orange one has 18. Two skeletons are close but not coincide. The farthest distance of two point sets is 2 pixels, and the most common distance is 1. When we set $t = 0$, which means we do not tolerate any offset distance, the PHD is high. If we set $t = 1$, the PHD drops a lot. When the $t = 2$, PHD becomes 0. In case (b), there is a big offset between two skeletons. When the t is set to $[2, 4]$, the PHD drops slowly. Different settings of t represent the degree of tolerance to the distance between the two skeletons. In practical application, different tolerance distances can be applicable to according to different situations.

Consistency between PHD and human perception. We also calculated the consistency between the PHD results with human perception in the subjective experiment of Section 3.2. Experiments show that the PHD with appropriate tolerance distances are more consistent to human perception than F1 score.

As the green bars show in Fig.6(b), when the tolerance distance is set to 3, the consistency reaches the maximum value of 65.48%, which is much higher than the result of F1 score. With the tolerance distance changes, the consistency increases first and then decreases. And it drops to 0 when the tolerance distance is set to 800. This trend of results shows that when judging these groups of images, most people can tolerate an offset of a small distance. If the distance increases to a certain extent, it will no longer be tolerated.

Some people may say that it is possible to improve the F1 score by only skeletonizing the segmentations. Our experiment refutes this point of view. As the orange bar shows, the consistency of the F1 score calculated from image skeletons can only reach 44.25%, which is also very low and with great randomness. Therefore, the disadvantages of F1 score can not be eliminated only by extracting image skeleton.

4 Further Experiments on U-RISC Dataset

In the previous two sections, we propose a new ultra-high resolution cell membrane segmentation dataset and a new evaluation criteria PHD. Through a subjective experiment on a small-scale dataset, we proved that PHD is consistent with human perception for the evaluation of cell membrane segmentation.

To observe the performance of different segmentation methods and evaluation criteria on the whole dataset, we choose six widely used 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-implement them on U-RISC dataset. As mentioned in section 2, we retain the results of two manual labeling results. So we will also analyze the results of manual labeling under different evaluation criteria.

Table 1: Experiments on U-RISC Dataset. This table shows the different evaluation results on human annotations and some sgementation results. The sgementation results are predicted by supervised learning methods: 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.

Methods	**F1** ↑	**IoU** ↑	**Dice** ↑	**PHD-0** ↓	**PHD-1** ↓	**PHD-3** ↓	**PHD-5** ↓
Label 1th	0.9840	0.8616	0.9212	0.2102	0.0576	0.0456	0.0419
Label 2th	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**

Experiments Settings The experiment settings are the same for each algorithm. For each training process, we first use 60% of the dataset as training data, and randomly cut the original image into 1024×1024 patches to generate 5,0000 training images and 2,0000 validation images. Random flipping and clipping are used for data augmentation. The training parameters are consistent with those reported by the original author. 4 V100 GPU are used for training each algorithm. During testing, 1 V100 is used to test the patches of the same size in rest images, and splice them back together.

Experiments Results The experiments results are shown in Table. 1. We calculate different evaluation criteria scores on final annotation with two labeling results and six segmentation results: F1 score, IoU, Dice, and PHD-t, $t = 0, 1, 3, 5$. PHD-t means the PHD score with tolerance distance t.

The first part shows the results of two labeling process. The results show that with the increase of inspection rounds, the quality of manual labeling has been improved. It is worth noting that F1 score increased by only 1.15%, while PHD-5 increased by 79.95%, which can also indirectly prove that PHD is more sensitive to structural changes.

The second part shows the results of six segmentation methods. From the scores we can see that F1 score, IoU and Dice have the same ranking for the six segmentation results. While PHD-1, PHD-3 and PHD-5 have different rankings. It's hard to say which ranking is the most reasonable, but it does provide researchers with a new idea to solve the problem of cell membrane segmentation. In the appendix II, we will show some local segmentation results of different algorithms.

5 Discussion

In this paper we propose an ultra-high resolution image segmentation dataet for cell membrane, called U-RISC and give a new evaluation criteria. Through a subjective experiment on a small-scale dataset, we proved that the new criterion is more consistent with human perception for the evaluation of cell membrane segmentation. Also, we show the performance of different segmentation methods and evaluation criteria, which provide researchers with a new idea to solve the problem of cell membrane segmentation. In the future research, we will continue to consider how to make better use of deep learning segmentation from the perspective of structure.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

James R Anderson, Bryan W Jones, Jia-Hui Yang, Marguerite V Shaw, Carl B Watt, Pavel Koshevoy, Joel Spaltenstein, Elizabeth Jurrus, UV Kannan, Ross T Whitaker, et al. A computational framework for ultrastructural mapping of neural circuitry. *PLoS Biol*, 7(3):e1000074, 2009.

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.

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.

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.

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.

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.

Robert E Marc. Mapping glutamatergic drive in the vertebrate retina with a channel-permeant organic cation. *Journal of Comparative Neurology*, 407(1):47–64, 1999.

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.

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. Interface of perceptual consistency exceriments.

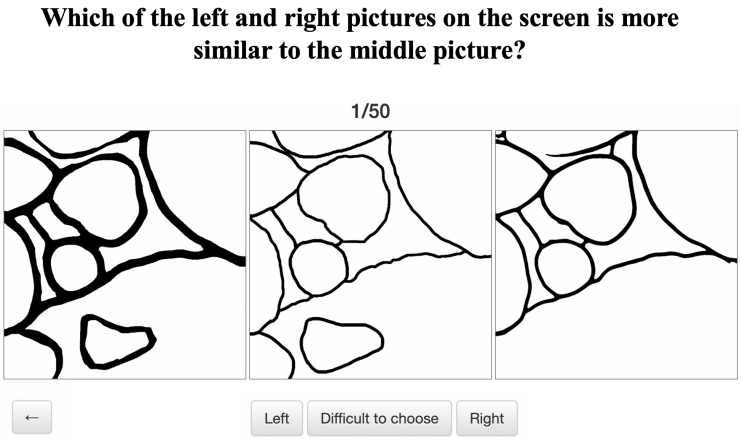


Figure 7: Interface of perceptual consistency exceriments.

II. Segmentation results of different algorithms.