Current Progress and Challenges in Large-scale 3D Mitochondria Instance Segmentation

Daniel Franco-Barranco, Zudi Lin, Won-Dong Jang, Xueying Wang, Qijia Shen, Wenjie Yin, Yutian Fan, Mingxing Li, Chang Chen, Zhiwei Xiong, Rui Xin, Hao Liu, Huai Chen, Zhili Li, Jie Zhao, Xuejin Chen, Constantin Pape, Ryan Conrad, Luke Nightingale, Joost de Folter, Martin L. Jones, Yanling Liu, Dorsa Ziaei, Stephan Huschauer, Ignacio Arganda-Carreras, Hanspeter Pfister and Donglai Wei

***Abstract*— In this paper, we present the results of the MitoEM challenge on mitochondria 3D instance segmentation from electron microscopy images, organized in conjunction with the IEEE-ISBI 2021 conference. Our benchmark dataset consists of two large-scale 3D volumes, one from human and one from rat cortex tissue, which are 3,600 times larger than previously used datasets. At the time of paper submission, 257 participants had registered for the challenge, 14 teams had submitted their results, and six teams participated in the challenge workshop. Here, we present eight top-performing approaches from the challenge participants, along with our own baseline strategies. Posterior to the challenge, annotation errors in the ground truth were corrected without altering the final ranking. Although several of the top methods are compared favorably to our own baselines, substantial errors remain unsolved for mitochondria with challenging morphologies. Thus, the challenge remains open for submission and automatic evaluation, with all volumes available for download. Additionally, we present a retrospective evaluation of the scoring system performed using TIMISE, our novel open-source evaluation toolbox, which revealed that (1) the challenge metric was permissive with the false positive predictions and (2) the size-based grouping of instances did not correctly categorize mitochondria of interest. Thus, we propose a new scoring system that better reflects the correctness of the segmentation results.**

***Index Terms*— Mitochondria, Electron Microscopy, 3D Instance Segmentation, Connectomics, Brain.**

Manuscript received xxx; accepted xxx. Date of publication xxx; date of current version xxx. Donglai Wei is the corresponding author.

D. Franco-Barranco and I. Arganda-Carreras are with the Department of Computer Science and Artificial Intelligence, University of the Basque Country, Donostia-San Sebastian, Spain, and the Donostia International Physics Center (DIPC), San Sebastian, Spain. I. Arganda-Carreras is also with Ikerbasque, Basque Foundation for Science, Bilbao, Spain and with Biofisika Institute (CSIC, UPV/EHU), Bilbao, Spain.

Z. Lin, W-D. Jang, and H. Pfister are with the John A. Paulson School of Engineering and Applied Sciences (SEAS), Harvard University, Allston, MA, USA.

X. Wang, W. Yin, and Y. Fan are with the Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA.

Q. Shen is with the Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom.

M. Li, C. Cheng and Z. Xiong are with the Dept. Electronic Engineering and Information Science (EEIS), University of Science and Technology of China, Anhui, China.

R. Xin, H. Liu and H. Chen are with the Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

Z. Li, J. Zhao, X. Chen are with the National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China, Anhui, China.

C. Pape is with the Georg-August University Goettingen, Germany. He has contributed to the challenge during his previous affiliation with the European Molecular Biology Laboratory (EMBL).

R. Conrad is with the Center for Molecular Microscopy, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, USA, and the Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, USA.

L. Nightingale, J. de Folter, M. L. Jones are with the Francis Crick Institute, London, United Kingdom.

Y. Liu and D. Ziaei are with the Advanced Biomedical Computational Science group, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

S. Huschauer is a non-affiliated contributor.

D. Wei is with the Computer Science Department, Boston College, Chestnut Hill, MA, USA. Contact email: weidf@bc.edu

I. Introduction

MITOCHONDRIA are the primary energy providers for cell activities, thus essential for metabolism. Quantification of the size and geometry of mitochondria is not only crucial to basic neuroscience research, *e.g.*, neuron type identification [1], but also informative to clinical studies, *e.g.*, bipolar disorder [2] and diabetes [3]. High-resolution imaging technologies like electron microscopy (EM) have been utilized to reveal their detailed 3D geometry at the nanometer level with the terabyte data scale [4]. Consequently, to enable an in-depth biological analysis, we need high-throughput and robust 3D mitochondria instance segmentation methods. Publicly accessible datasets that can exemplify the challenges are also of essential importance for understanding the empirical gain of segmentation approaches in this field.

The goal of this study is to (1) analyze the current progress in the mitochondria segmentation task based on the results of the Large-scale 3D Mitochondria Instance Segmentation challenge (MitoEM) [1], at the IEEE International Symposium on Biomedical Imaging (ISBI) 2021, and (2) present TIMISE [2], a novel open-source toolbox for identifying mitochondria instance segmentation errors, that reveals the difficulties of the current approaches and can be used as a guide for the creation of the next generation mitochondria segmentation models. To the best of our knowledge, MitoEM was the first open comparison of mitochondria instance segmentation algorithms on

[1]Challenge website: `https://mitoem.grand-challenge.org`
[2]`https://github.com/danifranco/TIMISE`

EM volumes. Moreover, we describe the associated annotated dataset of two 3D EM image stacks at the scale of $(30\mu m)^3$, which are freely available from the challenge website, and are two of a few large-scale 3D image volumes suitable for testing instance segmentation algorithms.

A. Previous Works

Mitochondria segmentation datasets. The *de facto* benchmark dataset for evaluating methods of mitochondria segmentation from EM images is the EPFL Hippocampus dataset [5], referred to as the Lucchi dataset in this paper. This dataset includes two EM image volumes along with corresponding binary segmentation masks. Subsequently, Kasthuri *et al.* [6] provided annotation for mitochondria masks for selected regions within the 3-cylinder volume. Additionally, Casser *et al.* [7] improved the annotation quality for both datasets through the implementation of a consistent annotation protocol for mask boundaries. Despite these efforts, the datasets remain small in size, less than 0.3 Gigavoxels and $(5\ \mu\text{m})^3$ physically, which does not adequately capture the complexity of mitochondria morphology. Furthermore, the provided binary masks are not easily converted into instance segmentation masks, which are necessary for detailed biological analysis as the instances of mitochondria can be connected to each other.

Instance segmentation evaluation metrics. The evaluation of instance segmentation results can be done at either the pixel level or the instance level. The pixel-level metric assumes high-quality ground truth instance masks and measures the correctness of the pixel grouping with a clustering-based criterion, such as the Rand index [8]. However, as dataset sizes increase, it becomes increasingly difficult to manually refine all masks for pixel-level accuracy. As a result, instance-level metrics are more commonly used for large-scale datasets. For each predicted instance mask, if its intersection-over-union (IoU) score with a ground truth mask is higher than a predefined threshold, it is considered a true positive (TP). Similarly, predictions that fall below the IoU threshold are considered false positives (FP), while ground truth predictions without a match with the TP prediction are considered false negatives (FN). For biomedical image datasets, metrics based on TP, FP, and FN rates, such as. accuracy=$\frac{TP}{TP+FP+FN}$ are widely used in the literature [9]–[11]. In the case of natural 2D images, popular methods like Mask R-CNN-based approaches, typically predict the confidence for each instance detection, and the average precision (AP) metric is used to average results over different detection thresholds [12], [13]. In addition, instances are usually divided into small/medium/large groups for separate evaluations. Wei *et al.* [14] provided an efficient implementation of the AP metric for instances inside 3D volumes. To further break down the analysis of instance matching results, Ka *et al.* [15] proposed association metrics, categorizing them into one-to-one, over-segmentation, under-segmentation, many-to-many, missing, and background.

Machine learning methods. Despite the advances in large-scale instance segmentation for neurons from EM images [16], [17], similar efforts for mitochondria have been largely overlooked in the field. The lack of a large-scale, public dataset has led to the majority of recent mitochondria (semantic) segmentation methods being benchmarked on the Lucchi dataset [5], where mitochondria instances are small in number, simple in morphology, and relatively sparse in distribution. Even in non-public datasets [18], [19], the complexity of mitochondrial shapes is limited by the small size of the dataset and the use of non-mammalian tissue. In the field of mitochondria semantic segmentation, previous studies have employed a variety of techniques to segment the Lucchi dataset. Early works have leveraged traditional image processing and machine learning techniques [20]–[23], while recent methods utilized 2D or 3D deep learning architectures for mitochondria segmentation [7], [24]–[26]. Furthermore, Liu *et al.* [27] proposed an instance segmentation approach utilizing a modified Mask R-CNN [28], while Xiao *et al.* [29] achieved instance segmentation through a tracking approach. However, it remains uncertain how the performance of these methods, developed on small datasets, would extend to larger datasets (*e.g.*, $(30\ \mu m)^3$ cube) for neuroscience analysis, where mitochondria exhibit more complex variations in appearance and shape.

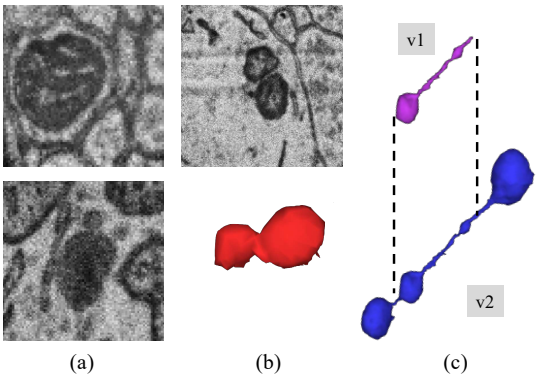

Fig. 1. Common annotation errors in the initial MitoEM dataset [14] (v1): (a) false positives of organelles that look similar to mitochondria, (b) false merges of mitochondrion, and (c) incomplete segmentation. Those errors were fixed after another around of expert proofreading (v2).

II. MitoEM Challenge

A. Dataset

We base the challenge on our released large-scale 3D mitochondria instance segmentation benchmark, the MitoEM dataset [14]. The MitoEM dataset consists of two $(30\mu m)^3$ 3D EM image stacks, one from an adult rat brain tissue (MitoEM-R) and one from an adult human brain tissue (MitoEM-H). The physical sample size of the MitoEM dataset is **3,600**$\times$ larger than the previous Lucchi benchmark [5] by the physical sample size. For information regarding the dataset acquisition and annotation strategy, we refer readers to Wei *et al.* [14].

Improved Annotation (V2). After the initial release of the MitoEM dataset, we identified three consistent categories of annotation errors (as depicted in Fig. 1). These errors include instances of organelles with a similar dark appearance that were mistakenly labeled as mitochondria, instances of neighboring mitochondria that were falsely merged into a single mitochondrion, and instances of *mitochondria-on-a-string* (MOAS) [30] that were occasionally incomplete due

to their thin microtubule connections. To address these errors, we enlisted three neuroscience experts who are familiar with EM images and mitochondria morphology to independently proofread the previous annotation. We then consolidated the changes and resolved any discrepancies among the experts through discussion. As a result of these efforts, the number of confirmed instances in the MitoEM-H dataset was reduced from 24.5K to 19K, and the number of confirmed instances in the MitoEM-R dataset was reduced from 14.4K to 10.8K.

In light of these changes, the ground truth was updated and uploaded to the grand-challenge website in December 2021, and all methods were re-evaluated accordingly. Despite these changes, the leaderboard rankings remained largely unaltered.

B. Evaluation Metric

In our initial release of the challenge, we used the evaluation metric proposed by Wei *et al.* [14], which computes the AP-75 score for small/medium/large groups of instances based on the instance size. However, upon conducting an analysis of the errors in the challenge submissions, we recognized the need to make certain improvements to the evaluation metric.

Improved metric: from AP to accuracy. We found that the AP-based metrics that were originally designed for top-down instance segmentation methods, such as Mask RCNN [28], are not well-suited for our challenge. In our case, most submission methods employed a bottom-up approach for instance segmentation, in which there is no estimation of the confidence score for each instance. To address this issue, Wei *et al.* [14] approximated the confidence score with the size of the instance, which can lead to unintuitive evaluation results, as discussed in Section IV. After careful consideration, we decided to adopt the popular accuracy metric [10] for evaluating the challenge submissions. This metric matches prediction instances with ground truth instances, providing a more intuitive evaluation of the methods' performance.

Improved instance grouping: from volume to cable length. In our initial release of the challenge, we utilized a splitting rule based on the volume to categorize mitochondria instances into small, medium, and large groups. However, we noticed that this approach was not effective for correctly categorizing complex mitochondria instances, such as the MOAS instances. For that reason, in this paper we have opted for the cable length[3] instead, using length thresholds of 1 μm and 4 μm to split the mitochondria into three groups: small, medium and large (as in the original MitoEM release). Under this new categorization, the number of small, medium, and large mitochondria instances are respectively: 5106, 3608, and 164 in MitoEM-H and 1292, 3832 and 524 in MitoEM-R. A visualization of the mitochondria of each new split is depicted in Fig. 2, where now all MOAS lie into the same (large) category, as previously expected. A fast inspection reveals that (1) the human tissue contains many more small mitochondria than the rat tissue, and (2) the large mitochondria from the human tissue are notably thinner than those of the rat tissue.

[3]Cable length is defined as the skeleton length of the instance taking into account each axes resolution, *e.g.*, $8 \times 8 \times 40$ for (x, y, z) axes in MitoEM.

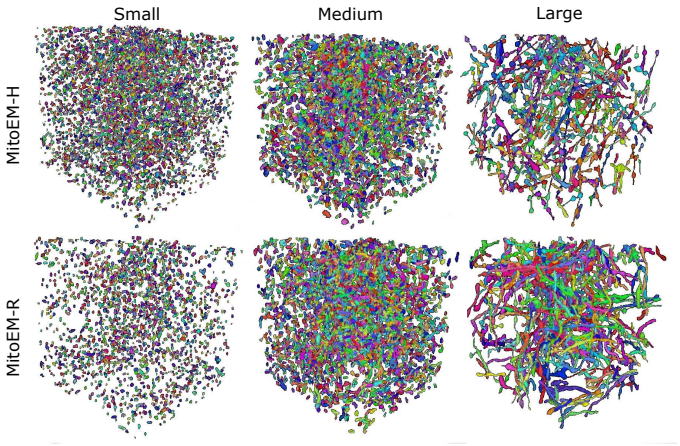

Fig. 2. Visualization of MitoEM-H and MitoEM-R datasets splitting categories based on cable length. From left to right: 3D meshes of small (length $\leq 1\mu$m), medium (1μm $<$ length $< 4\mu$m), and large (length $\geq 4\mu$m) mitochondria of human (top) and rat tissue (bottom).

All these changes became effective in July 2022 in the Grand Challenge platform, producing significant alterations in the leaderboard as described in Section V.

C. Open-source Baseline Methods

In order to increase the accessibility of our challenge, we have released two open-source baseline pipelines, each accompanied by a reproducible tutorial. These pipelines are designed to work on 2D and the 3D input image, respectively. Both baseline models utilize a U-Net [31] based architecture to predict both binary foreground segmentation masks and instance contours masks (referred to as *BC*). The final step in both models is to fuse the binary foreground mask and instance contour channel outputs to create mitochondria instance seeds, together with a foreground mask, which is then used as inputs for the marker-controlled watershed (MW) [32] algorithm.

U2D-BC[4]**.** The model was trained using an input size of 256×256. Data augmentation techniques such as flips, random rotations, variations in brightness and contrast, and elastic transformations were applied during the training process. The model was optimized until convergence, approximately 180 epochs, over a reduced version of the dataset (20% of training data) with stochastic gradient descent (SGD) using a learning rate of 0.002. We consider one epoch when visiting all training samples. We further applied median filtering in y-z axes to improve the network output predictions. This model was implemented based on the networks developed in [26].

U3D-BC[5]**.** The model was trained with an input size of $225 \times 225 \times 17$ for x, y, and z axes considering the anisotropy of the datasets, which contains a mixture of 2D and 3D convolutions. Besides applying common pixel and spatial augmentations, we also used misalignment augmentation to make the model more robust to the misalignment problem

[4]https://biapy.readthedocs.io/en/latest/tutorials/mitoem.html

[5]https://connectomics.readthedocs.io/en/latest/tutorials/mito.html

introduced in dataset acquisition. The model was optimized for 150K iterations with an initial learning rate of 0.04 and cosine learning-rate scheduling. We also applied Gaussian blending and test-time augmentations (self-ensemble) to improve the prediction quality. The model was implemented with PyTorch Connectomics [33] and can be reproduced based on the tutorial provided by the challenge organizers.

In comparison to our previous work [14], we made improvements to the implementation details in order to achieve superior results. Specifically, we have incorporated a number of additional data augmentation techniques, including misalignment, CutBlur [34], CutNoise, and motion-blur, in addition to the brightness, flip, elastic transform, and missing parts augmentations used in the original MitoEM paper [14]. Furthermore, we increased the probability and intensity of all augmentations to enhance the robustness of the model. Additionally, we updated the optimization technique used, switching from ADAM optimizer [35] to SGD [36], [37], following the recent findings of Zhou et al. [38], which indicate that ADAM-alike adaptive optimization algorithms do not generalize as well as SGD. We also implemented the *cosine learning rate decay* policy [39] to update the learning rate, while maintaining the number of training iterations and the initial learning rate as in the original model.

D. Organization

The challenge was accepted to ISBI 2021 in October 2020 and officially announced in November 2020. This announcement was accompanied by the creation of a dedicated website and the preparation of an evaluation system. The two image volumes, MitoEM-R and MitoEM-H, were made immediately available to participants to enable them to begin developing their methods. From the $1,000$ consecutive 3D slices of each stack, ground-truth mitochondria instance labels were provided for the first 500 slices and split into training (400 slices) and validation (100 slices) subsets. The annotations of the remaining 500 slices of each volume were kept private and used as the test set. Participants performed the segmentation on their own computers. The challenge was widely advertised and was open to any interested participants. A total of 257 individuals registered for the challenge and 14 teams submitted their results. For comparison, we also used two "internal submissions" corresponding to our 2D and 3D baseline methods (as described in Section III). To lower the barrier of entry for the challenge, an initial version of the code of our 3D baseline was made publicly available. The teams were also asked to submit a description of their method. Eight teams were invited to a workshop on April 13, during the ISBI 2021 conference, and to participate in the writing of this article. The winners of the challenge were announced at this workshop.

Some of the teams that participated in the challenge did not register for the conference or participate in the workshop. However, six teams did submit short papers and presented their methods. The results announced at the workshop (ranked using the AP-75 metric) are given in Table VII. Those results may be based on updated submissions. After the workshop, the challenge remained open to submissions and all image volumes, as well as their ground-truth labels, are available for download. The testing labels are kept confidential.

III. Summary of Segmentation Methods

In this section, we described the evaluated segmentation methods from the eight teams who successfully completed the challenge, together with our baseline methods. See Table I for an overview. Specific details of all algorithms are provided in the respective manuscripts submitted by participants as per the MitoEM challenge policies and are available at the challenge webpage under the "manuscripts" tab.

A. Participants' Methods

The following methods by the participant teams produced successful results that were submitted to the challenge. Notice that the method names used here may differ from the team names found on the MitoEM webpage. See Table VII for the link to the code, documentation and manuscript (if available) for each method.

- **VIDAR** (USTC)[6]: Two specialized networks, Res-UNet-R and Res-UNet-H, predict both the instance boundaries and the semantic masks of mitochondria. Both architectures are inspired by the 3D U-Net [40] and contain residual blocks where the initial convolution is performed only in 2D to address the anisotropic resolution of the input data. While in the Res-UNet-R, the decoder outputs the semantic mask and the instance boundary simultaneously, the Res-UNet-H contains two decoders, one for each output. Moreover, they used a weighted binary cross-entropy loss function to compensate for the class imbalance and deployed a multi-scale training strategy to train the network in two stages with progressively larger input images. For pre-processing, denoising was performed with their own image restoration network [41]. Finally, the semantic masks and instance boundaries are used to create a seed map to perform hierarchical agglomeration [42] and extract individual instances.
  You can find more detailed information about the method in the later work presented by the authors on [43].
- **IIPPR** (SJTU)[7]: The submissions were mostly based on the U3D-BC+MW baseline method provided by the challenge organizers, as described in the next section. Different random seeds, as well as training and decoding hyper-parameters, can result in different predictions from those generated by the baseline configurations.
- **VGG** (NEL-BITA)[8]: A contrastive learning [44], [45] framework is proposed using a representative point sampling strategy, and a loss function combining a point-wise similarity term (to increase the similarity of points from the same class and the separability of points from different classes) and an inter-frame consistency term to enhance the sensitivity of the 3D model to changes in image content from frame to frame. A classic 3D

[6]M. Li, C. Chen, Z. Xiong
[7]R. Xin, H. Liu, H. Chen
[8]Z. Li, J. Zhao, X. Chen

TABLE I

OVERVIEW OF THE MITOEM PARTICIPANT METHODS. CE — CROSS-ENTROPY, WBCE — WEIGHTED BINARY CROSS-ENTROPY, MSE — MEAN SQUARED ERROR, WMSE — WEIGHTED MSE, SGD — STOCHASTIC GRADIENT DESCENT, HA — HIERARCHICAL AGGLOMERATION, MCWS — MARKER-CONTROLLED WATERSHED, MWSMC — MUTEX WATERSHED AND MULTICUT, CC — CONNECTED COMPONENTS, HUA — HUNGARIAN ALGORITHM. (*) REUSE U3D-BC+MW CODE.

Method	Available code	Model architecture	Input shape	Loss function	Optimizer	Connectivity method	Pre-/post-processing
VIDAR	✓	Residual U-Net	3D	WBCE	Adam	HA	Denoising as pre-processing
IIPPR	✓(*)	Residual U-Net	3D	BCE+Dice	Adam	MCWS	Ensemble+Blending inference
U3D-BC+MW*	✓	Residual U-Net	3D	BCE+Dice	SGD	MCWS	Aggressive DA
U2D-BC+MW*	✓	U-Net	2D	BCE	SGD	MCWS	Aggressive DA+YZ-Filtering
VGG		U-Net	3D	CE+Contrastive	SGD	MCWS	None
EMBL	✓	U-Net	3D	Dice	Adam	MWSMC	None
CEM-PDL	✓	Panoptic-DeepLab	2D	WBCE+MSE	AdamW	HUA	CEM500K pretraining, Z-filtering...
FCI	✓	U-Net	3D	Dice	Adam	MCWS	Tri-axis prediction
ABCS		U-Net	3D	Dice	Adam	CC	Ensemble
H2RNet		Hybrid-HRNet	2D	WMSE	Adam	MCWS	Morphological closing, size filtering

U-Net [40] is used as a backbone network to output binary masks and boundary maps, and marker-controlled watershed [32] is applied to extract the final instances. Feature maps are extracted from the last two layers of the backbone decoder to extract point features and build positive and negative pairs according to their classes. Thanks to this, contrastive learning can be used to maximize the similarity between feature vectors of the same class, while minimizing those of two different classes. Similarly, the consistency loss term is designed to enhance the feature similarity between points belonging to the same class at the same position in adjacent slices and contrastively decrease the similarity of points from different classes.
You can find more detailed information about the method in the later work presented by the authors on [46].

- **EMBL** (Heidelberg)[9]: Foreground probabilities and long-range affinity maps [47] are predicted using a 3D U-Net [40] without pooling across the z-axis in the first 2 pooling layers to address the anisotropy of the dataset. Next, Mutex Watershed [48] is applied in parallel on the predictions of subvolumes of the whole dataset, and the final whole-volume instance segmentation is obtained by means of solving a Multicut clustering problem [49].
- **CEM-PDL** (NIH)[10]: A Panoptic-DeepLab model [50] with a ResNet50 [51] backbone is trained to perform instance segmentation in 2D slices. More specifically, the model has three outputs: semantic masks, instance centers, and instance center regressions (offset from each pixel to its corresponding center). Instances are obtained by simple post-processing (assigning each pixel to the closest predicted center). The backbone network uses weights pre-trained on CEM500K [52], a large dataset of EM images. Training is performed on 512×512 patches and the inference is applied to the full-size image 4096×4096. Several post-processing methods are used including Z-filtering, 2D watershed to split false mergers, and the Hungarian algorithm and the Intersection-over-Area merging strategy to merge false splits.

[9]C. Pape
[10]R. Conrad

This method has been further developed since submission to the MitoEM challenge into an open-source model called MitoNet [53].

- **FCI** (London)[11]: Four separate convolutional neural networks (CNNs) were trained to predict semantic masks and boundaries of mitochondria on each MitoEM-H and MitoEM-R, respectively. All networks follow the same architecture based on a 3D U-Net [40] with Inception-like blocks [54], and were trained using a smoothed dice coefficient (or F-measure) loss function. Weights were initialized using a nuclear envelope segmentation model trained on crowd-sourced citizen science annotations [55]. Boundary predictions were improved by means of combining predictions from all three views of the volumes [55], and individual instances are extracted using marker-controlled watershed [32].
You can find more detailed information about the method in the later work presented by the authors on [56].
- **ABCS** (FNL)[12]: Two simple 3D versions of the original U-Net architecture [31] were trained to simulate different fields of views with input sizes of $64 \times 128 \times 128$ voxels and $64 \times 256 \times 256$, respectively. Basic data augmentation was used with flipping in all three axes, and ensemble prediction with patch overlap was applied at inference time. Averaging the prediction of the two networks improved both their individual scores.
- **H2RNet** (Zurich)[13]: A modified HRNet [57] network is used with two heads that respectively predict the energy surface and the curvature of mitochondria in 2D slices. Both heads are fused into a final prediction. A weighted mean-square-error loss function is used, with weighting based on the frequency of a given value for the energy head, and weighting based on bending loss [58] for the contour head. Watershed post-processing in 2D is not needed since a cut-off from the surface energy learned is used as the hyper-parameter in the prediction. Due to computing resource limitations, 2D predictions were downsampled to apply marker-controlled watershed [32]

[11]L. Nightingale, J. de Folter, M. L. Jones
[12]Y. Liu, D. Ziaei
[13]S. Huschauer

in 3D to connect regions across sections and later upsampled using nearest-neighbor interpolation.

IV. TIMISE: Error Analysis Toolbox

In addition to setting up the challenge and compiling results, we aimed to build a toolbox to facilitate error analysis to inspire novel approaches, similar to the TIDE toolbox [59] for 2D instance segmentation. To this end, we introduce TIMISE[14]: an open-source **T**oolbox for **I**dentifying **M**itochondria **I**nstance **S**egmentation **E**rrors. TIMISE enables (1) a compact summary of error types, (2) a 3D visual comparison of instance segmentation results, and (3) comparisons across datasets and object attributes for deeper analysis.

A. Compact Summary of Error Types

Instead of using a single set of metrics, we aim to select from commonly used metrics to create a compact and informative error report to debug 3D instance segmentation methods.

Design choices. There are three commonly used sets of metrics for the 3D instance segmentation task: average precision (AP)-based, matching-based, and association-based. (1) No AP-based metric. The AP-based metric requires sorting predictions by confidence, which is not provided by most bottom-up segmentation approaches. Wei *et al.* [14] heuristically used the segment size as the prediction confidence, which can lead to undesirable biases for method ranking. (2) Matching-based metric for ranking. The matching-based metric focuses on the number of segments that are predicted correctly without worrying about the type of segmentation errors. It turns the segmentation result into the object detection result by thresholding the IoU of the matched prediction and ground truth masks. Then, the accuracy metric can combine informative statistics, *i.e.*, TP/FP/FN, into a single value to rank the methods. (3) Association error for break-down analysis. Many segmentation methods need to set hyper-parameters to control the ratio between false-split and false-merge errors. Thus, the pie chart displaying the proportion of different types of segmentation association error [15] is critical for a more interpretable result understanding. The association errors are defined as follows:

- *One-to-one*, if it is an exact match.
- *Over-segmentation*, when one instance in the ground truth is divided into two or more in the prediction.
- *Under-segmentation*, when two or more instances in the ground truth are merged in the prediction.
- *Many-to-many*, when two or more instances in the ground truth are associated with two or more in the prediction, which is the most complex case.
- *Missing*, for instances of the ground truth that are not captured in the prediction.
- *Background*, for instances associated with the background, i.e. false positives.

Validating design choices. To better understand the pros and cons of each metric implemented in TIMISE, we created a toy example with a ground truth of mitochondria of different sizes and a realistic model prediction (see Fig. 3). The ground truth volume contains only six instances: one large (MOAS type), three medium, and two small mitochondria based on their cable lengths. The prediction presents an over-segmentation of small and medium instances, a merger of two mitochondria, and several split errors in the MOAS. The corresponding AP-75, association, and accuracy values are shown in Table II.

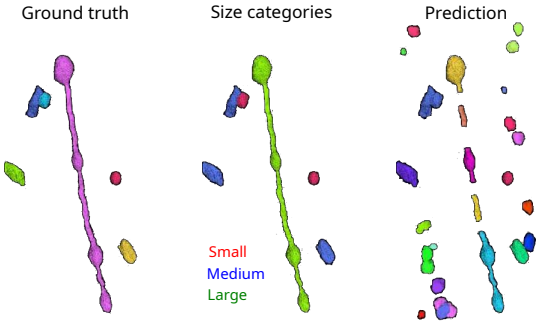

Fig. 3. Synthetic example of mitochondria instance segmentation. Left to right: ground truth instances, same instances color-coded by size, and model prediction.

TABLE II
AP-75, ASSOCIATION AND MATCHING METRICS EVALUATION OF THE SYNTHETIC EXAMPLE OF FIG.3 PERFORMED WITH THE TIMISE TOOLBOX. ASSOCIATION METRICS ARE EXPRESSED IN %.

		Small	Medium	Large	Total
AP-based	AP-75 ↑	0.51	0.44	0.00	0.22
Matching metrics	Precision ↑	0.06	0.67	0.00	0.12
	Recall ↑	0.50	0.67	0.00	0.50
	Accuracy ↑	0.05	0.50	0.00	0.10
Association metrics	Correct ↑	50.0	66.7	0.00	50.0
	Missing ↓	0.00	0.00	0.00	0.00
	Over ↓	0.00	0.00	100	16.7
	Under ↓	50.0	33.3	0.00	33.3
	Many ↓	0.00	0.00	0.00	0.00
	Back ↓	-	-	-	65.4

AP-75 overvalues small-size instances. In the AP-75 calculation [14], the mitochondria segments are sorted by size, resulting in small segments having the lowest confidence values. Therefore, when a small instance is merged with a medium one in the prediction, the small instance is considered an FN. Additionally, the large instance in the ground truth is split into several segments that do not reach the minimum IoU of 75% with the ground truth, so most of those segments are considered as medium FPs. This means the large mitochondrion is only matched with the blue instance that represents its bottom part (since it is the largest among all pieces). Although the prediction contains several small FPs, as well as more small and medium FPs considering the rest of the MOAS pieces not matched with it (e.g. all but the blue instance), the AP-75 values for small segments are still high.

Accuracy metric provides a good overall evaluation. As shown in Table II, the association metrics are useful for understanding the fate of the ground truth segments in the prediction but do not provide information on the overlap between the prediction and the ground truth. On the other

[14]https://github.com/danifranco/TIMISE

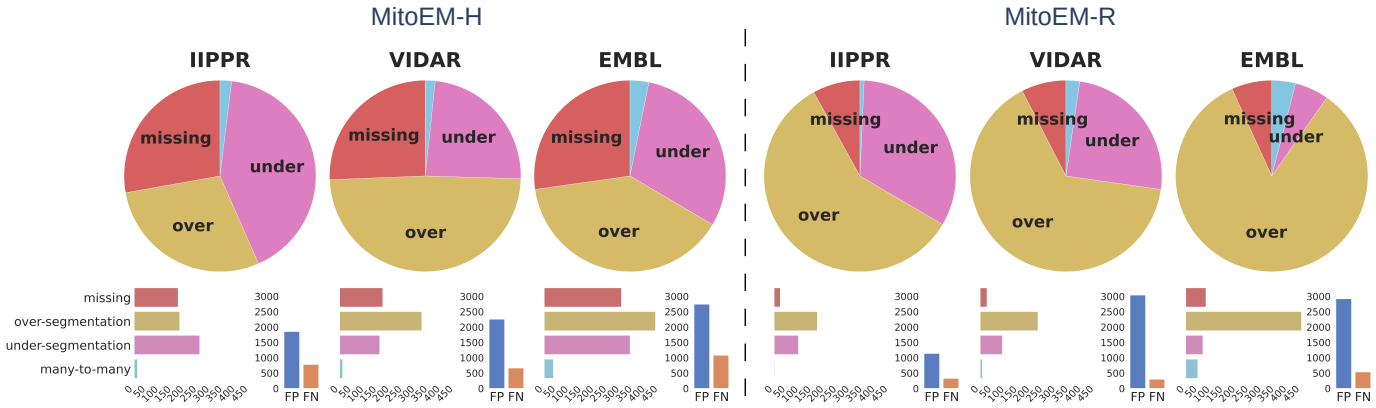

Fig. 4. Summary of errors in MitoEM for the top three methods. The pie chart illustrates the proportion of each type of association error, while the bar plots below show the absolute number of false positives (FP) and false negatives (FN) for each method. These values are used to calculate the accuracy metric.

hand, the matching metrics do provide this information by considering a prediction as a TP if the IoU with ground truth is greater than 75% (following [14]). However, the association metrics have multiple values, rather than just a single value, which complicates the direct comparison of the performance of different methods. For example, it is not clear whether a low *under-segmentation* rate is better or worse than a low *over-segmentation* rate, or whether *many-to-many* is worse than the previous two. These questions depend on the specific task at hand. Therefore, it is useful to have a single number, such as accuracy, to enable easy comparison of the performance of different models. In the toy example, there are many small FPs in the prediction, as previously mentioned, which results in low values for all matching metrics except recall. For medium instances, only the one merged with the small instance is not considered a TP due to its low IoU (< 0.75).

Association metrics provide a detailed breakdown of errors. Examining the association metrics helps us to understand where and how the prediction fails. A *missing* value of zero in all cases indicates that all ground truth instances have been captured by the prediction. More specifically, out of the two small mitochondria in the ground truth, one has been correctly predicted and is labeled as *correct*. The other one was merged with a medium mitochondrion, resulting in both small and medium being labeled as *under-segmentation*. The remaining three medium mitochondria are also labeled as *correct*. Also, the ground truth MOAS that was divided into medium-sized pieces in the prediction is labeled as *over-segmentation*.

B. 3D Mesh Visualization

TIMISE offers different plotting options based on the measured statistics and metrics with just a function call. This way, the user can (1) find correlations between one or more morphological measures and segmentation errors, as depicted in Fig 7 with the cable length and association metrics; (2) gather metrics of different methods to compare their performance in a single chart (as depicted in Fig. 9); (3) create neuroglancer [15] visualization scripts with just a function call to generate images such as Fig. 2 and 5; and (4) create interactive 3D plots[16] fusing more measurements and metrics.

[15] https://github.com/google/neuroglancer

TABLE III
MATCHING STATISTICS OF ALL METHODS ON THE MITOEM CHALLENGE LEADERBOARD. THE RANKINGS PRESENTED IN THIS TABLE DIFFER FROM THE ORIGINAL CHALLENGE LEADERBOARD AS A RESULT OF THE ALTERATION OF THE EVALUATION METRIC, AS DISCUSSED IN THE MANUSCRIPT. BASELINE METHODS FROM CHALLENGE ORGANIZERS (MARKED WITH *) ARE SHOWN BUT NOT INCLUDED IN THE RANKING. THE BEST SCORES ARE SHOWN IN BOLD.

Method	Rank	MitoEM-H			MitoEM-R			
		Prec.↑	Rec.↑	Acc.↑	Prec.↑	Rec.↑	Acc.↑	Total Acc.
IIPPR	1	**0.814**	0.913	**0.755**	**0.824**	0.943	**0.785**	**0.770**
VIDAR	2	0.785	**0.926**	0.739	0.638	**0.948**	0.616	0.678
U3D-BC*		0.706	0.916	0.663	0.663	0.913	0.623	0.643
EMBL	3	0.740	0.879	0.672	0.637	0.906	0.597	0.635
VGG	4	0.658	0.911	0.619	0.697	0.905	0.649	0.634
CEM-PDL	5	0.734	0.794	0.617	0.721	0.860	0.645	0.631
FCI	6	0.741	0.754	0.596	0.669	0.771	0.558	0.577
H2RNet	7	0.636	0.698	0.499	0.709	0.811	0.608	0.554
ABCS	8	0.628	0.766	0.526	0.675	0.694	0.520	0.523
U2D-BC*		0.435	0.925	0.420	0.354	0.911	0.342	0.381

V. Analysis of Current Progress on MitoEM

A. Overall Performance

To have an overview of each model error we created Fig. 4 with our proposed TIMISE toolbox, following the same plot types proposed in [59].

Matching based evaluation. To further analyze the performance of the methods, we present their corresponding matching metric values in Table III. The IIPPR method performs better than VIDAR in most cases, except in terms of recall. A similar pattern is observed in other cases, where high recall is achieved at the expense of precision (U3D-BC, VGG, and U2D-BC), due to a higher number of FPs and therefore higher *over-segmentation* values. This conclusion can also be drawn

[16] Find them in our prepared notebooks here: https://github.com/danifranco/TIMISE/tree/main/examples

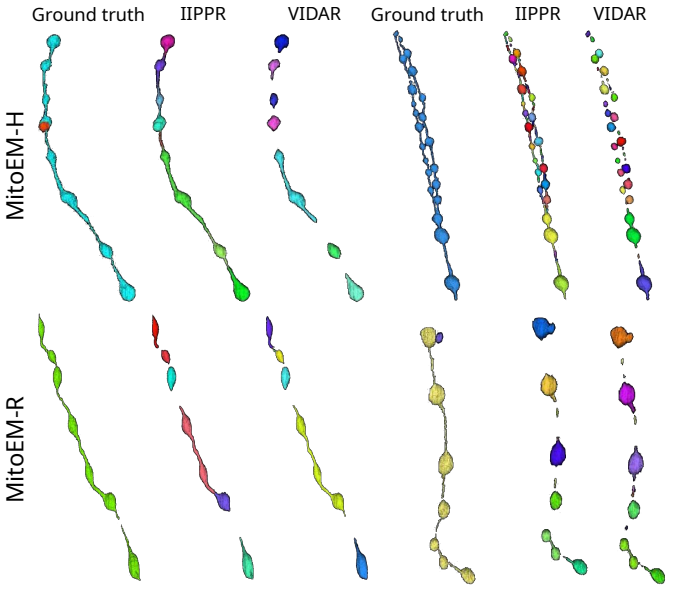

Fig. 5. 3D visualization of MOAS instances with the TIMISE toolbox for error inspection. We show the ground truth and segmentation results from the two top-performing models (IIPPR and VIDAR) in one MOAS instance per dataset. Different colors represent different instances.

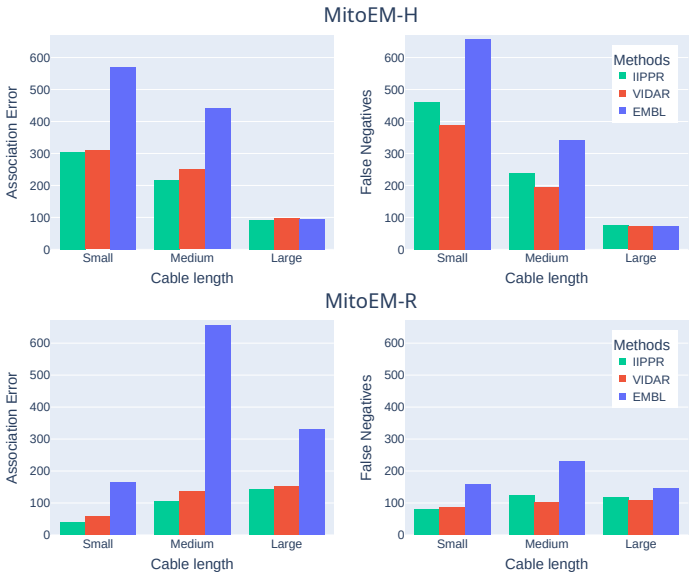

Fig. 6. Summary of the absolute number of error types per instance category for the three top-performing methods. The errors shown include cumulative association errors (on the left) and false negatives (on the right) for each method.

by examining the FP, FN and *missing* values of each method in Fig. 4. For a more detailed analysis, see Table VII.

Association based evaluation. To facilitate the understanding of the types of errors made by each competing method, we have created pie charts in Fig. 4. These pie charts provide a relative overview of the errors, but we also need to consider absolute magnitudes in order to compare between methods, leading to the creation of horizontal bar plots. In general, Fig.4 shows that the relative magnitude of missing instances is similar among the top three methods for both tissues. However, the absolute magnitudes indicate better performance for IIPPR and VIDAR compared to EMBL. The top methods tend to exhibit *over-segmentation* rather than *under-segmentation* (with the exception of IIPPR in human tissue). This demonstrates the failure of these methods in the most challenging instances of MitoEM, which are the MOAS-type mitochondria. A visual example of the *over-segmentation* problem is shown in Fig. 5. For all association error percentages, see Table VI, and for a more detailed breakdown analysis, see Fig. 9.

B. Comparison Across Skeleton Length

As previously mentioned, MOAS-type mitochondria pose a significant challenge in MitoEM due to their thin connections, which can result in *over-segmentation*, as illustrated in Fig. 5. While we have identified this issue, we have not yet considered the number of associations that each error involves. For example, an *over-segmentation* association may involve one ground truth instance that has been split into twenty predicted instances, while an *under-segmentation* may only involve two ground truth instances that have been merged into one predicted instance. It is therefore of interest to compare *over-segmentation*, *under-segmentation*, and *many-to-many* associations to determine which is the most detrimental.

For that comparison, we generated Fig. 7 using the TIMISE toolbox. This figure presents the number of associations per skeleton length bin. The values were calculated by adding the number of predicted *over-segmented* instances, subtracting the number of *under-segmented* instances from the ground truth, and calculating the ratio of predicted instances to those in the ground truth for the *many-to-many* case (with a positive value when there are more predicted instances than ground truth ones in the association, and negative otherwise). For example, a value of zero in a bin indicates that all associations sum to zero, as in the case of an *over-segmentation* of a GT instance that has been divided into two (resulting in a net increase of two instances), or an *under-segmentation* case where a predicted instance actually corresponds to two GT instances (resulting in a net decrease of two instances). On the other hand, a value of ten in a bin indicates that there are more cases of *over-segmentation* present than *under-segmentation*.

It can be observed that there are more *over-segmentation* associations in both tissues compared to other types of associations. Additionally, the number of associations appears to increase with cable length. The observed trend can be attributed to the MOAS-type mitochondria, as the size of these structures tends to correspond with a higher quantity of smaller constituent elements. It is also noted that the MOAS in human tissue are more *over-segmented* than in rat tissue. This difference is likely due to the thicker connections present in rat MOAS, as depicted in the middle of Fig. 7, which make them easier to segment in 3D.

Fig. 6 shows a breakdown analysis to identify the types of mitochondria that demonstrated the highest failure rates. In terms of the absolute number of cumulative association errors, the results follow the same ranking as in Table III, with IIPPR performing the best, followed by VIDAR, and

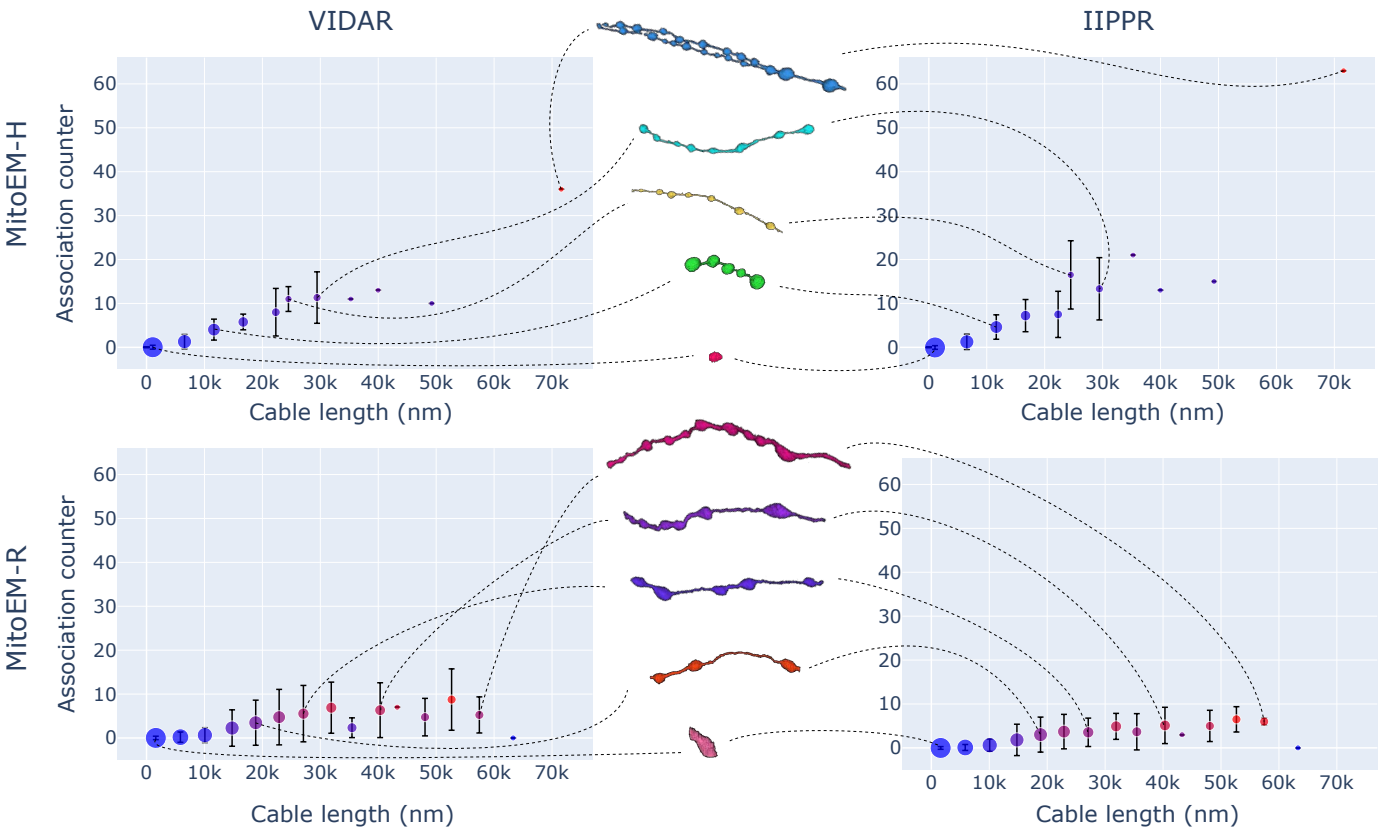

Fig. 7. Comparison across object scales. Histograms of association errors of the two top-performing methods (VIDAR and IIPPR) on the MitoEM-H (top) and MitoEM-R (bottom) test sets. The points show the association counts for mitochondria of each length, while the standard deviation is represented by the vertical black lines. The point size is proportional to the number of instances inside that bin. Some representative instances of different cable lengths are displayed in the middle and connected by dashed lines to their corresponding bins.

finally EMBL. In contrast, VIDAR performs better in terms of false negatives compared to IIPPR. This result aligns with the findings presented in Section V-A, which indicate that VIDAR is able to identify more instances, albeit at the cost of higher FP rates and lower precision.

VI. Discussion on Remaining Challenges

Despite the improved results during the competition, the following challenges remain for the community to tackle.

Model challenge. In the current full-supervised learning setting, *i.e.*, 40-10-50% data split, the IIPPR method serves as a strong baseline, achieving an overall 0.770 accuracy score. However, for practical large-scale deployment to recent petabyte scale datasets [17], the instance segmentation methods need to achieve at least 0.9 (accuracy) to make the saturated proofreading feasible. In addition to the challenges mentioned in the paper, *e.g.*, complex geometries and crowded instances, there is still an open challenge on the "large" segments (especially MOAS instances with super thin connections) as they often split those mitochondria into smaller pieces producing an *over-segmentation*. To address this issue, the VIDAR team at USTC's lab has proposed the use of knowledge distillation training [60] as a potential solution.

Limited label challenge. In practice, the annotation budget is often around 5-10% of the whole volume. Therefore, it is critical to developing data-efficient methods, *e.g.*, new data augmentation methods [61], unsupervised [62], semi-supervised, and active learning methods, that can achieve 0.9 (accuracy) with a limited amount of annotation.

Proofreading challenge. Regarding the suitability of a scoring system based on accuracy, one should assess the purpose of the segmentation result and its subsequent processing. In particular, for large datasets such as MitoEM, the current strategy assumes a proofreading phase after automatic segmentation. In that sense, a metric that does not penalize false positive predictions as much as false negative ones may be the most appropriate. In fact, eliminating false positives is proven much faster than correcting false negatives when proofreading 3D segments [63]. In a more general framework, the association and matching metrics provided in TIMISE help us complete the big picture in terms of evaluation.

VII. Conclusion

In this paper, we present the results of the ISBI 2021 challenge on MitoEM, the first large-scale instance mitochondria segmentation challenge that thoroughly benchmarks state-of-the-art methods in the field. To gain insight into the common errors of the proposed methods and identify current challenges that remain unresolved, we analyze the performance of the methods using various types of evaluation metrics. To assist

the research community in this endeavor, we have developed TIMISE, an open-source and user-friendly toolbox for identifying errors in mitochondria instance segmentation based on a wide range of segmentation metrics.

The release of MitoEM had the dual goal of attracting new computer vision researchers to the problem of EM mitochondria segmentation and pushing the state-of-the-art forward. We believe that the challenge was successful in this regard, as the participants improve over our own initial baseline methods. Furthermore, the competition received a very positive reaction from the community and had good attendance at its corresponding workshop at ISBI 2021.

After conducting a detailed analysis of the challenge results, we identified consistent annotation errors and released an updated version of the ground truth labels (V2). Additionally, using the TIMISE toolbox, we identified issues with the evaluation system based on the AP-75 metric and updated the challenge and method ranking using a more robust metric that is more sensitive to false negatives and *over-segmentations*, such as accuracy. However, the current accuracy values are still insufficient for fully automatic segmentation, therefore the challenge remains open for submissions.

Finally, we believe that our large-scale annotated dataset, similar to ImageNet for natural images, has the potential to be useful for a variety of applications beyond its original purpose. Some examples include using the dataset for deep feature pre-training, performing 3D shape analysis, and testing novel approaches such as active learning or domain adaptation.

APPENDIX I

TABLE IV
ADDITIONAL INFORMATION OF THE PRESENTED METHODS AT THE MITOEM CHALLENGE. (*) REUSE U3D-BC CODE.

Method	Accuracy Rank	Code	Documentation	Publication
IIPPR	1	Link (*)	-	
VIDAR	2	Link	-	[43]
U3D-BC*		Link	Link	[33]
EMBL	3	Link	Link	
VGG	4	-	-	[46]
CEM-PDL	5	Link	-	[53]
FCI	6	Link	-	[56]
H2RNet	7	-	-	
ABCS	8	-	-	
U2D-BC*		Link	Link	[26]

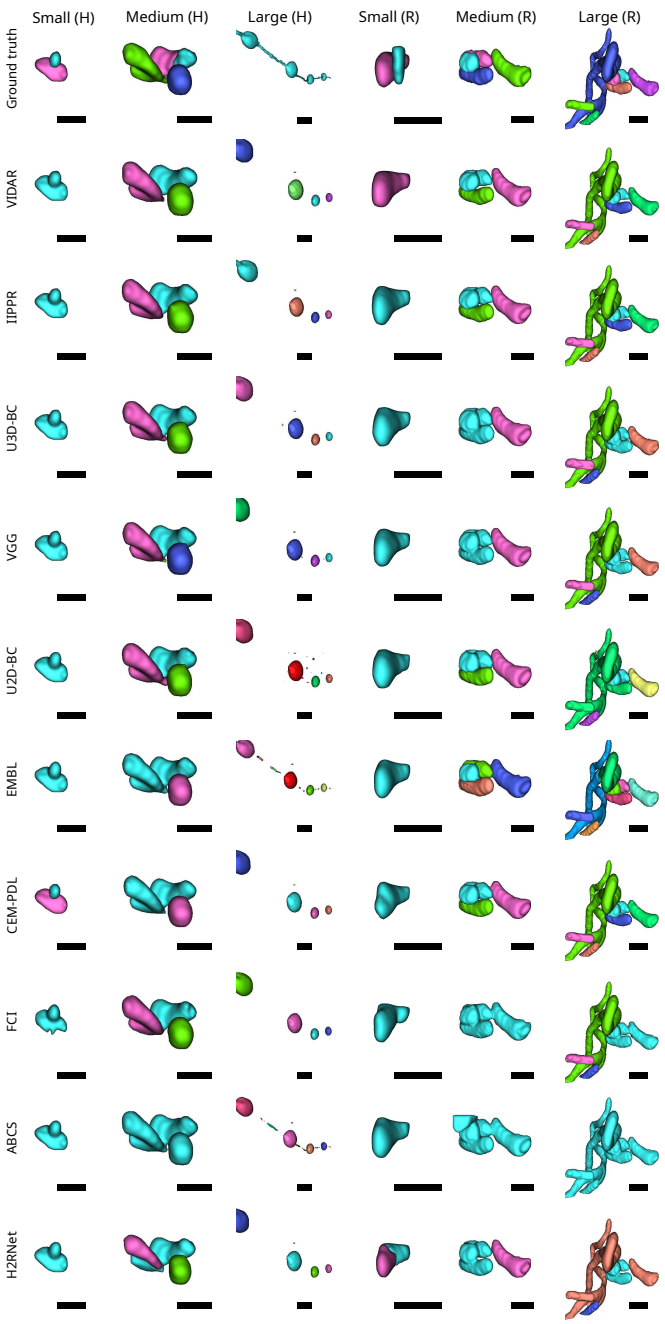

Fig. 8. Some examples of common segmentation errors by the analyzed methods in small, medium and large mitochondria of MitoEM-H and MitoEM-R tissue from the test set. Every instance is given a unique color. The scale bar represents 0.5 μm.

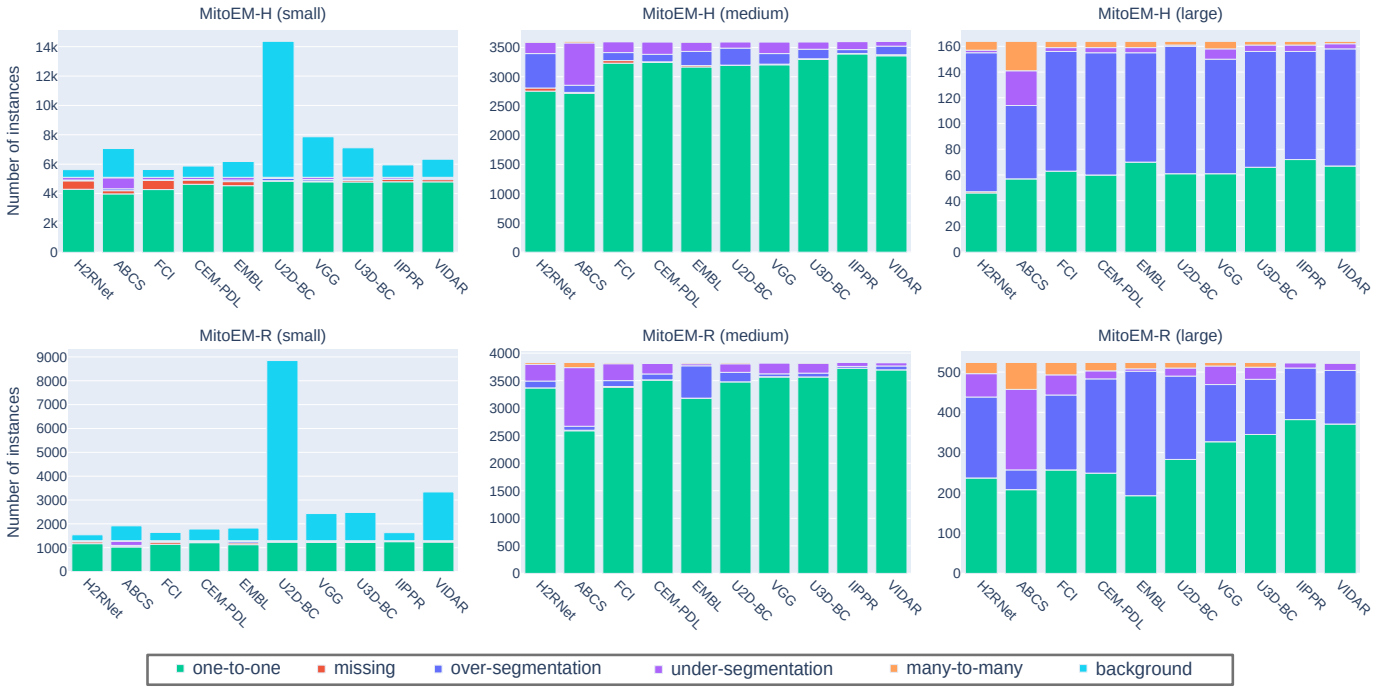

Fig. 9. Distribution of association errors of all methods on the MitoEM-H (top) and MitoEM-R (bottom) test sets for small (left), medium (center) and large (right) mitochondria. The methods are ordered from left to right by lowest-to-highest value of AP-75. A remarkable difference in performance can be appreciated for all methods on large mitochondria compared to small and medium mitochondria.

TABLE V

MATCHING BASED METRICS OF ALL METHODS ON THE MITOEM CHALLENGE LEADERBOARD PER CATEGORY. BASELINE METHODS FROM CHALLENGE ORGANIZERS (MARKED WITH *) ARE SHOWN BUT NOT INCLUDED IN THE RANKING. **BOLD** AND UNDERLINED NUMBERS DENOTE THE 1ST AND 2ND SCORES, RESPECTIVELY.

Method	Accuracy Rank	MitoEM-H				MitoEM-R		
		Category	Precision ↑	Recall ↑	Accuracy ↑	Precision ↑	Recall ↑	Accuracy ↑
IIPPR	1	large	0.148	0.543	0.132	**0.397**	0.779	**0.357**
		medium	**0.930**	0.934	**0.872**	**0.969**	0.967	**0.938**
		small	**0.811**	0.910	**0.750**	**0.753**	**0.938**	**0.717**
VIDAR	2	large	**0.172**	**0.549**	**0.151**	0.353	**0.794**	0.323
		medium	0.916	**0.946**	0.870	0.961	**0.974**	0.936
		small	0.758	0.924	0.713	0.362	0.933	0.353
U3D-BC*		large	0.131	0.537	0.118	0.268	0.763	0.247
		medium	0.890	0.933	0.836	0.936	0.934	0.878
		small	0.662	0.917	0.625	0.479	0.910	0.457
EMBL	3	large	0.108	0.549	0.099	0.231	0.725	0.212
		medium	0.853	0.905	0.783	0.786	0.940	0.748
		small	0.756	0.871	0.680	0.629	0.878	0.578
VGG	4	large	0.109	0.506	0.099	0.323	0.708	0.285
		medium	0.884	0.915	0.818	0.938	0.927	0.873
		small	0.603	0.922	0.574	0.494	0.919	0.473
CEM-PDL	5	large	0.164	0.427	0.134	0.232	0.519	0.191
		medium	0.834	0.849	0.727	0.900	0.901	0.819
		small	0.711	0.768	0.585	0.652	0.878	0.597
FCI	6	large	0.167	0.470	0.141	0.229	0.540	0.191
		medium	0.821	0.831	0.703	0.833	0.813	0.699
		small	0.733	0.709	0.564	0.624	0.741	0.513
H2RNet	7	large	0.117	0.341	0.096	0.214	0.515	0.178
		medium	0.580	0.692	0.461	0.865	0.852	0.752
		small	0.734	0.715	0.568	0.733	0.812	0.626
ABCS	8	large	0.134	0.537	0.120	0.381	0.548	0.290
		medium	0.776	0.787	0.641	0.808	0.684	0.588
		small	0.592	0.758	0.498	0.556	0.784	0.482
U2D-BC*		large	0.166	0.530	0.145	0.213	0.708	0.196
		medium	0.865	0.937	0.818	0.906	0.930	0.848
		small	0.328	**0.929**	0.320	0.137	0.936	0.135

TABLE VI

ASSOCIATION BASED METRICS (IN %) OF ALL METHODS ON THE MITOEM CHALLENGE LEADERBOARD. BASELINE METHODS FROM CHALLENGE ORGANIZERS (MARKED WITH *) ARE SHOWN BUT NOT INCLUDED IN THE RANKING. 'CORRECT', 'MISSING', 'OVER', 'UNDER' AND 'MANY' STANDS FOR 'ONE-TO-ONE', 'MISSING', 'OVER-SEGMENTATION', 'UNDER-SEGMENTATION', 'MANY-TO-MANY' AND 'BACKGROUND' ASSOCIATIONS, RESPECTIVELY. BEST SCORES ARE SHOWN IN BOLD. ALL BUT BACKGROUND ASSOCIATIONS SUM ONE FOR EACH TISSUE, I.E. MITOEM-H AND MITOEM-R. BACKGROUND PERCENTAGE HAS BEEN CALCULATED TAKEN INTO ACCOUNT ALL PREDICTED INSTANCES.

Method	Accuracy Rank	MitoEM-H						MitoEM-R					
		Correct↑	Missing↓	Over↓	Under↓	Many↓	Back↓	Correct↑	Missing↓	Over↓	Under↓	Many↓	Back↓
IIPPR	1	**93.07**	1.93	**1.99**	2.87	0.14	8.48	**94.92**	0.41	2.97	1.66	**0.04**	5.25
VIDAR	2	92.61	1.89	3.62	**1.76**	**0.12**	11.77	93.89	0.46	3.98	1.52	0.14	24.39
U3D-BC*		91.82	1.22	4.10	2.65	0.21	17.43	90.95	0.51	4.97	4.07	0.50	15.26
EMBL	3	87.55	3.39	4.89	3.77	0.39	10.22	79.62	1.38	17.0	**1.17**	0.81	6.62
VGG	4	90.82	0.78	4.16	3.86	0.38	22.46	90.67	0.55	3.67	4.75	0.37	15.54
CEM-PDL	5	89.4	3.32	2.96	4.05	0.26	7.98	88.05	0.62	6.32	4.21	0.80	7.28
FCI	6	85.23	7.51	3.09	3.91	0.26	5.87	84.86	1.84	5.49	6.94	1.04	5.36
H2RNet	7	79.88	7.07	8.49	4.10	0.45	**5.35**	84.54	1.43	5.81	7.03	1.19	**3.85**
ABCS	8	76.02	2.65	3.47	16.9	0.97	18.1	67.76	0.74	**2.76**	25.62	3.12	10.76
U2D-BC*		91.32	**0.32**	6.10	2.05	0.21	49.05	88.6	**0.09**	7.15	3.43	0.73	52.02

TABLE VII

THE MITOEM CHALLENGE LEADERBOARD AS ANNOUNCED AT THE WORKSHOP. THE METHODS ARE RANKED ACCORDING TO THEIR AP-75 SCORES, WITH THE HIGHEST SCORES DISPLAYED IN BOLD. THE RANKINGS PRESENTED IN THIS TABLE ALIGN WITH THE ORIGINAL CHALLENGE LEADERBOARD, BUT DEVIATE FROM THOSE PRESENTED IN THE PRESENT MANUSCRIPT DUE TO THE MODIFICATION OF THE EVALUATION METRIC. THE BASELINE METHODS FROM THE CHALLENGE ORGANIZERS (MARKED WITH *) ARE DISPLAYED BUT WERE NOT INCLUDED IN THE RANKING.

Method	AP-75 Rank	MitoEM-H				MitoEM-R				Overall
		Small	Medium	Large	All	Small	Medium	Large	All	
VIDAR	1	**0.835**	**0.905**	**0.420**	**0.827**	0.727	**0.955**	**0.550**	**0.850**	**0.839**
IIPPR	2	0.807	0.884	0.328	0.796	**0.815**	0.941	0.517	0.842	0.819
U3D-BC*		0.799	0.885	0.331	0.790	0.780	0.896	0.505	0.811	0.801
VGG	3	0.794	0.854	0.333	0.786	0.788	0.885	0.425	0.790	0.788
EMBL	4	0.783	0.837	0.389	0.762	0.773	0.896	0.444	0.779	0.771
U2D-BC*		0.741	0.885	0.349	0.779	0.623	0.879	0.433	0.751	0.765
CEM-PDL	5	0.642	0.742	0.249	0.644	0.730	0.834	0.194	0.674	0.659
ABCS	7	0.655	0.669	0.295	0.636	0.709	0.586	0.304	0.572	0.604
FCI	6	0.610	0.745	0.345	0.620	0.598	0.710	0.270	0.582	0.601
H2RNet	8	0.574	0.541	0.216	0.474	0.656	0.764	0.260	0.605	0.540

ACKNOWLEDGEMENTS

We gratefully thank the Grand Challenge team for providing the platform that enables public access, challenge organization, and automatic evaluation. This work has been partially supported by NSF award IIS-1835231, NIH award 5U54CA225088-03, by the University of the Basque Country UPV/EHU grant GIU19/027, by Ministerio de Ciencia, Innovación y Universidades, AEI, under grant PID2021-126701OB-I00, by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001999), the UK Medical Research Council (FC001999), and the Wellcome Trust (FC001999), by the Frederick National Laboratory for Cancer Research under NIH Contract No. 75N91019D00024.

REFERENCES

[1] P. J. Schubert, S. Dorkenwald, M. Januszewski, V. Jain, and J. Kornfeld, "Learning cellular morphology with neural networks," *Nature communications*, 2019.

[2] T. Kasahara, A. Takata, T. Kato, M. Kubota-Sakashita, T. Sawada, A. Kakita, H. Mizukami, D. Kaneda, K. Ozawa, and T. Kato, "Depression-like episodes in mice harboring mtDNA deletions in paraventricular thalamus," *Molecular psychiatry*, 2016.

[3] M. Zeviani and S. Di Donato, "Mitochondrial disorders," *Brain*, vol. 127, no. 10, 2004.

[4] A. Motta, M. Berning, K. M. Boergens, B. Staffler, M. Beining, S. Loomba, P. Hennig, H. Wissler, and M. Helmstaedter, "Dense connectomic reconstruction in layer 4 of the somatosensory cortex," *Science*, vol. 366, no. 6469, 2019.

[5] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua, "Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 474–486, 2011.

[6] N. Kasthuri, K. J. Hayworth, D. R. Berger, R. L. Schalek, J. A. Conchello, S. Knowles-Barley, D. Lee, A. Vázquez-Reina, V. Kaynig, T. R. Jones, M. Roberts, J. L. Morgan, J. C. Tapia, H. S. Seung, W. G. Roncal, J. T. Vogelstein, R. Burns, D. L. Sussman, C. E. Priebe, H. Pfister, and J. W. Lichtman, "Saturated reconstruction of a volume of neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, Jul. 2015. [Online]. Available: https://doi.org/10.1016/j.cell.2015.06.054

[7] V. Casser, K. Kang, H. Pfister, and D. Haehn, "Fast mitochondria detection for connectomics," in *Medical Imaging with Deep Learning*, 2020.

[8] SNEMI3D EM Segmentation Challenge and Dataset. [Online]. Available: http://brainiac2.mit.edu/SNEMI3D/home

[9] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018, pp. 265–273.

[10] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers, "Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[11] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[14] D. Wei, Z. Lin, D. Franco-Barranco, N. Wendt, X. Liu, W. Yin, X. Huang, A. Gupta, W.-D. Jang, X. Wang *et al.*, "MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2020, pp. 66–76.

[15] A. Kar, M. Petit, Y. Refahi, G. Cerutti, C. Godin, and J. Traas, "Benchmarking of deep learning algorithms for 3D instance segmentation of confocal image datasets," *PLoS computational biology*, vol. 18, no. 4, p. e1009879, 2022.

[16] M. Januszewski, J. Kornfeld, P. H. Li, A. Pope, T. Blakely, L. Lindsey, J. Maitin-Shepard, M. Tyka, W. Denk, and V. Jain, "High-precision automated reconstruction of neurons with flood-filling networks," *Nature Methods*, 2018.

[17] A. Shapson-Coe, M. Januszewski, D. R. Berger, A. Pope, Y. Wu, T. Blakely, R. L. Schalek, P. Li, S. Wang, J. Maitin-Shepard *et al.*, "A connectomic study of a petascale fragment of human cerebral cortex," *bioRxiv*, 2021.

[18] Ariadne.ai, *Automated segmentation of mitochondria and ER in cortical cells*, 2018 (accessed February 1, 2023), https://ariadne.ai/case/segmentation/organelles/CorticalCells/.

[19] S. Dorkenwald, P. J. Schubert, M. F. Killinger, G. Urban, S. Mikula, F. Svara, and J. Kornfeld, "Automated synaptic connectivity inference for volume electron microscopy," *Nature methods*, vol. 14, no. 4, pp. 435–442, 2017.

[20] K. Smith, A. Carleton, and V. Lepetit, "Fast ray features for learning irregular shapes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[21] A. Vazquez-Reina, M. Gelbart, D. Huang, J. Lichtman, E. Miller, and H. Pfister, "Segmentation fusion for connectomics," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[22] A. Lucchi, Y. Li, K. Smith, and P. Fua, "Structured image segmentation using kernelized features," in *ECCV*. Springer, 2012.

[23] A. Lucchi, P. Márquez-Neila, C. Becker, Y. Li, K. Smith, G. Knott, and P. Fua, "Learning structured models for segmentation of 2-D and 3-D imagery," *IEEE Transactions on Medical Imaging*, vol. 34, no. 5, pp. 1096–1110, 2014.

[24] I. Oztel, G. Yolcu, I. Ersoy, T. White, and F. Bunyak, "Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network," in *Bioinformatics and Biomedicine*, 2017.

[25] H.-C. Cheng and A. Varshney, "Volume segmentation using convolutional neural networks with limited training data," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 590–594.

[26] D. Franco-Barranco, A. Muñoz-Barrutia, and I. Arganda-Carreras, "Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes," *Neuroinformatics*, 2021.

[27] J. Liu, W. Li, C. Xiao, B. Hong, Q. Xie, and H. Han, "Automatic detection and segmentation of mitochondria from SEM images using deep neural network," in *EMBC*. IEEE, 2018.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2961–2969.

[29] C. Xiao, X. Chen, W. Li, L. Li, L. Wang, Q. Xie, and H. Han, "Automatic mitochondria segmentation for EM data using a 3D supervised convolutional network," *Frontiers in neuroanatomy*, vol. 12, p. 92, 2018.

[30] L. Zhang, S. Trushin, T. A. Christensen, B. V. Bachmeier, B. Gateno, A. Schroeder, J. Yao, K. Itoh, H. Sesaki, W. W. Poon, and K. Gylys, "Altered brain energetics induces mitochondrial fission arrest in Alzheimer's Disease," *Scientific reports*, vol. 6, p. 18725, 2016.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[32] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, no. 1, pp. 113–125, 1994.

[33] Z. Lin, D. Wei, J. Lichtman, and H. Pfister, "PyTorch Connectomics: A Scalable and Flexible Segmentation Framework for EM Connectomics," *arXiv preprint arXiv:2112.05754*, 2021.

[34] J. Yoo, N. Ahn, and K.-A. Sohn, "Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8375–8384.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[37] L. Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, p. 12, 1991.

[38] P. Zhou, J. Feng, C. Ma, C. Xiong, S. HOI *et al.*, "Towards theoretically understanding why SGD generalizes better than ADAM in deep learning," *arXiv preprint arXiv:2010.05627*, 2020.

[39] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[40] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016, pp. 424–432.

[41] S. Deng, W. Huang, C. Chen, X. Fu, and Z. Xiong, "A Unified Deep Learning Framework for ssTEM Image Restoration," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3734–3746, 2022.

[42] J. Funke, F. Tschopp, W. Grisaitis, A. Sheridan, C. Singh, S. Saalfeld, and S. C. Turaga, "Large scale image segmentation with structured loss based deep learning for connectome reconstruction," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 7, pp. 1669–1680, 2018.

[43] M. Li, C. Chen, X. Liu, W. Huang, Y. Zhang, and Z. Xiong, "Advanced deep networks for 3D mitochondria instance segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[44] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[46] Z. Li, X. Chen, J. Zhao, and Z. Xiong, "Contrastive learning for mitochondria segmentation," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3496–3500.

[47] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung, "Superhuman accuracy on the SNEMI3D connectomics challenge," *arXiv:1706.00120*, 2017.

[48] S. Wolf, A. Bailoni, C. Pape, N. Rahaman, A. Kreshuk, U. Köthe, and F. A. Hamprecht, "The mutex watershed and its objective: Efficient, parameter-free graph partitioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[49] C. Pape, T. Beier, P. Li, V. Jain, D. D. Bock, and A. Kreshuk, "Solving large multicut problems for connectomics via domain decomposition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1–10.

[50] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 475–12 485.

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

[52] R. Conrad and K. Narayan, "CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning," *Elife*, vol. 10, p. e65894, 2021.

[53] ——, "Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset," *Cell Systems*, vol. 14, no. 1, pp. 58–71, 2023.

[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[55] H. Spiers, H. Songhurst, L. Nightingale, J. De Folter, Z. V. Community, R. Hutchings, C. J. Peddie, A. Weston, A. Strange, S. Hindmarsh *et al.*, "Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations," *Traffic*, vol. 22, no. 7, pp. 240–253, 2021.

[56] L. Nightingale, J. de Folter, H. Spiers, A. Strange, L. M. Collinson, and M. L. Jones, "Automatic instance segmentation of mitochondria in electron microscopy data," *BioRxiv*, pp. 2021–05, 2021.
[57] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[58] H. Wang, M. Xian, and A. Vakanski, "Bending loss regularized network for nuclei segmentation in histopathology images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1–5.
[59] D. Bolya, S. Foley, J. Hays, and J. Hoffman, "Tide: A general toolbox for identifying object detection errors," in *European Conference on Computer Vision*. Springer, 2020, pp. 558–573.
[60] X. Liu, B. Hu, W. Huang, Y. Zhang, and Z. Xiong, "Efficient biomedical instance segmentation via knowledge distillation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2022, pp. 14–24.
[61] Q. Chen, M. Li, J. Li, B. Hu, and Z. Xiong, "Mask rearranging data augmentation for 3D mitochondria segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2022, pp. 36–46.
[62] W. Huang, X. Liu, Z. Cheng, Y. Zhang, and Z. Xiong, "Domain adaptive mitochondria segmentation via enforcing inter-section consistency," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2022, pp. 89–98.
[63] S. Dorkenwald, C. E. McKellar, T. Macrina, N. Kemnitz, K. Lee, R. Lu, J. Wu, S. Popovych, E. Mitchell, B. Nehoran *et al.*, "Flywire: online community for whole-brain connectomics," *Nature Methods*, vol. 19, no. 1, pp. 119–128, 2022.