3D Mitochondria Instance Segmentation with Spatio-Temporal Transformers

Omkar Thawakar[1(✉)], Rao Muhammad Anwer[1,2], Jorma Laaksonen[2], Orly Reiner[3], Mubarak Shah[4], and Fahad Shahbaz Khan[1,5]

[1] MBZUAI, Masdar City, UAE
omkar.thawakar@mbzuai.ac.ae
[2] Aalto University, Espoo, Finland
[3] Weizmann Institute of Science, Rehovot, Israel
[4] University of Central Florida, Orlando, USA
[5] Linköping University, Linköping, Sweden

Abstract. Accurate 3D mitochondria instance segmentation in electron microscopy (EM) is a challenging problem and serves as a prerequisite to empirically analyze their distributions and morphology. Most existing approaches employ 3D convolutions to obtain representative features. However, these convolution-based approaches struggle to effectively capture long-range dependencies in the volume mitochondria data, due to their limited local receptive field. To address this, we propose a hybrid encoder-decoder framework based on a split spatio-temporal attention module that efficiently computes spatial and temporal self-attentions in parallel, which are later fused through a deformable convolution. Further, we introduce a semantic foreground-background adversarial loss during training that aids in delineating the region of mitochondria instances from the background clutter. Our extensive experiments on three benchmarks, Lucchi, MitoEM-R and MitoEM-H, reveal the benefits of the proposed contributions achieving state-of-the-art results on all three datasets. Our code and models are available at https://github.com/OmkarThawakar/STT-UNET.

Keywords: Electron Microscopy · Mitochondria instance segmentation · Spatio-Temporal Transformer · Hybrid CNN-Transformers

1 Introduction

Mitochondria are membrane-bound organelles that generate the primary energy required to power the cell activities, thereby crucial for metabolism. Mitochondrial dysfunction, which occurs when mitochondria are not functioning properly

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_59.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 613–623, 2023.
https://doi.org/10.1007/978-3-031-43993-3_59

has been witnessed as a major factor in numerous diseases, including noncommunicable chronic diseases (*e.g*, cardiovascular and cancer), metabolic (*e.g*, obesity) and neurodegenerative (*e.g*, Alzheimer and Parkinson) disorders [23,25]. Electron microscopy (EM) images are typically utilized to reveal the corresponding 3D geometry and size of mitochondria at a nanometer scale, thereby facilitating basic biological research at finer scales. Therefore, automatic instance segmentation of mitochondria is desired, since manually segmenting from a large amount of data is particularly laborious and demanding. However, automatic 3D mitochondria instance segmentation is a challenging task, since complete shape of mitochondria can be sophisticated and multiple instances can also experience entanglement with each other resulting in unclear boundaries. Here, we look into the problem of accurate 3D mitochondria instance segmentation.

Earlier works on mitochondria segmentation employ standard image processing and machine learning methods [20,21,33]. Recent approaches address [4, 15,26] this problem by leveraging either 2D or 3D deep convolutional neural network (CNNs) architectures. These existing CNN-based approaches can be roughly categorized [36] into bottom-up [3,4,14,15,28] and top-down [12]. In case of bottom-up mitochondria instance segmentation approaches, a binary segmentation mask, an affinity map or a binary mask with boundary instances is computed typically using a 3D U-Net [5], followed by a post-processing step to distinguish the different instances. On the other hand, top-down methods typically rely on techniques such as Mask R-CNN [7] for segmentation. However, Mask R-CNN based approaches struggle due to undefined bounding-box scale in EM data volume.

When designing a attention-based framework for 3D mitochondria instance segmentation, a straightforward way is to compute joint spatio-temporal self-attention where all pairwise interactions are modelled between all spatio-temporal tokens. However, such a joint spatio-temporal attention computation is computation and memory intensive as the number of tokens increases linearly with the number of input slices in the volume. In this work, we look into an alternative way to compute spatio-temporal attention that captures long-range global contextual relationships without significantly increasing the computational complexity. Our contributions are as follows:

- We propose a hybrid CNN-transformers based encoder-decoder framework, named STT-UNET. The focus of our design is the introduction of a split spatio-temporal attention (SST) module that captures long-range dependencies within the cubic volume of human and rat mitochondria samples. The SST module independently computes spatial and temporal self-attentions in parallel, which are then later fused through a deformable convolution.
- To accurately delineate the region of mitochondria instances from the cluttered background, we further introduce a semantic foreground-background (FG-BG) adversarial loss during the training that aids in learning improved instance-level features.
- We conduct experiments on three commonly used benchmarks: Lucchi [20], MitoEM-R [36] and MitoEM-H [36]. Our STT-UNET achieves state-of-the-art

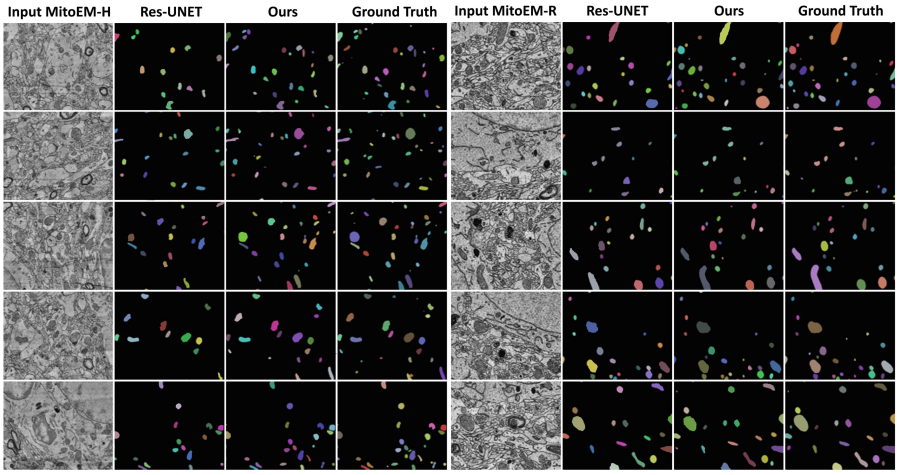

Fig. 1. Qualitative 3D instance segmentation comparison between the recent Res-UNET [16] and our proposed STT-UNET approach on the example input regions from MitoEM-H and MitoEM-R validation sets. Here, we present the corresponding segmentation predictions of the baseline and our approach along with the ground truth. Our STT-UNET approach achieves superior segmentation performance by accurately segmenting 16% more cell instances in these examples, compared to Res-UNET-R.

segmentation performance on all three datasets. On Lucchi test set, our STT-UNET outperforms the recent [4] with an absolute gain of 3.0% in terms of Jaccard-index coefficient. On MitoEM-H val. set, STT-UNET achieves AP-75 score of 0.842 and outperforms the recent 3D Res-UNET [16] by 3.0%. Figure 1 shows a qualitative comparison between our STT-UNET and 3D Res-UNET [16] on examples from MitoEM-R and MitoEM-H datasets.

2 Related Work

Most recent approaches for 3D mitochondria instance segmentation utilize convolution based designs within the "U-shaped" 3D encoder-decoder architecture. In such an architecture, the encoder aims to generate a low-dimensional representation of the 3D data by gradually performing the downsampling of the extracted features. On the other hand, the decoder performs upsampling of these extracted feature representations to the input resolution for segmentation prediction. Although such a CNN-based designs [11,16,34] has achieved promising segmentation results compared to traditional methods, they struggle to effectively capture long-range dependencies due to their limited local receptive field. Inspired from success in natural language processing [32], recently vision transformers (ViTs) [6,13,19,30,31] have been successfully utilized in different computer vision problems due to their capabilities at modelling long-range dependencies and enabling the model to attend to all the elements in the input

sequence. The core component in ViTs is the self-attention mechanism that that learns the relationships between sequence elements by performing relevance estimation of one item to other items. The other attention such as [1,8,10,29,35] have demonstrated remarkable efficacy in effectively managing volumetric data. Inspired by ViTs [10,19] and based on the observation that attention-based vision transformers architectures are an intuitive design choice for modelling long-range global contextual relationships in volume data, we investigate designing a CNN-transformers based framework for the task of 3D mitochondria instance segmentation.

3 Method

3.1 Baseline Framework

We base our approach on the recent Res-UNET [16], which utilizes encoder-decoder structure of 3D UNET [34] with skip-connections between encoder and decoder. Here, 3D input patch of mitochondria volume ($32 \times 320 \times 320$) is taken from the entire volume of ($400 \times 4096 \times 4096$). The input volume is denoised using an interpolation network adapted for medical images [9]. The denoised volume is then processed utilizing an encoder-decoder structure containing residual anisotropic convolution blocks (ACB). The ACB contains three layers of 3D convolutions with kernels ($1 \times 3 \times 3$), ($3 \times 3 \times 3$), ($3 \times 3 \times 3$) having skip connections between first and third layers. The decoder outputs semantic mask and instance boundary, which are then post-processed using connected component labelling to generate final instance masks. We refer to [16] for more details.

Limitations: As discussed above, the recent Res-UNET approach utilizes 3D convolutions to handle the volumetric input data. However, 3D convolutions are designed to encode short-range spatio-temporal feature information and struggle to model global contextual dependencies that extend beyond the designated receptive field. In contrast, the self-attention mechanism within the vision transformers possesses the capabilities to effectively encode both local and global long-range dependencies by directly performing a comparison of feature activations at all the space-time locations. In this way, self-attention mechanism goes much beyond the receptive field of the conventional convolutional filters. While self-attention has been shown to be beneficial when combined with convolutional layers for different medical imaging tasks, to the best of our knowledge, no previous attempt to design spatio-temporal self-attention as an exclusive building block for the problem of 3D mitochondria instance segmentation exists in literature. Next, we present our approach that effectively utilizes an efficient spatio-temporal attention mechanism for 3D mitochondria instance segmentation.

3.2 Spatio-Temporal Transformer Res-UNET (STT-UNET)

Figure 2(a) presents the overall architecture of the proposed hybrid transformers-CNN based 3D mitochondria instance segmentation approach, named STT-UNET. It comprises a denoising module, transformer based encoder-decoder

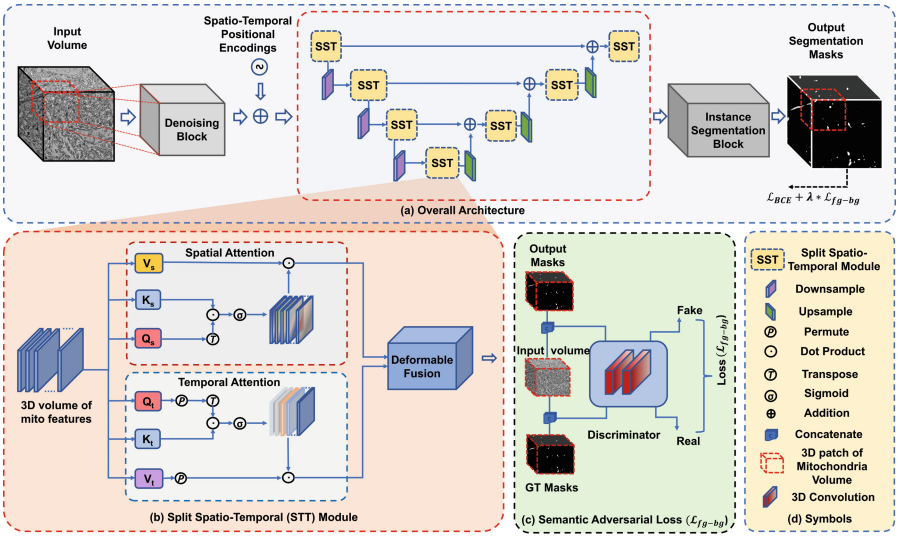

Fig. 2. (a) Overall architecture of our STT-UNET framework for 3D mitochondria instance segmentation. A 3D volume patch of mitochondria is first pre-processed using the interpolation network. The resulting reconstructed volume is then fed to our split spatio-temporal attention based encoder-decoder to generate the semantic-level mitochondria segmentation masks. The focus of our design is the introduction of split spatio-temporal attention (SST) module within the encoder-decoder. **(b)** The SST module first computes spatial and temporal attentions independently, which are later combined through a deformable convolution. Consequently, the semantic masks from the decoder are then input to the instance segmentation module to generate the final instance masks. The entire framework is trained using the standard BCE loss (L_{BCE}) and our semantic foreground-background (FG-BG) adversarial loss (L_{fg-bg}). **(c)** The L_{fg-bg} loss improves the instance-level features, thereby aiding in the better separability of the region of mitochondria instances from the cluttered background.

with split spatio-temporal attention and an instance segmentation block. The denoising module alleviates the segmentation faults caused by anomalies in the EM images, as in the baseline. The denoising is performed by convolving the current frame with two adjacent frames using predicted kernels, thereby generating the resultant frame by adding the convolution outputs. The resulting denoised output is then processed by our transformer based encoder-decoder with split spatio-temporal attention to generate the semantic masks. Consequently, these semantic masks are post-processed by an instance segmentation module using a connected component labelling scheme, thereby generating the final instance-level segmentation output prediction. To further enhance the semantic segmentation quality with cluttered background we introduced semantic adversarial loss which leads to improved semantic segmentation in noisy background.

Split Spatio-Temporal Attention based Encoder-Decoder: Our STT-UNET framework comprises four encoder and three decoder layers. Within

each layer, we introduce a split spatio-temporal attention-based (SST) module, Fig. 2(b), that strives to capture long-range dependencies within the cubic volume of human and rat samples. Instead of the memory expensive joint spatio-temporal representation, our SST module splits the attention computation into a spatial and a temporal parallel stream. The spatial attention refines the instance level features from input features along the spatial dimensions, whereas the temporal attention effectively learns the inter-dependencies between the input volume. The resulting spatial and temporal attention representations are combined through a deformable convolution, thereby generating spatio-temporal features. As shown in Fig 2(b), the normalized 3D input volume of denoised features X of size $(T \times H \times W \times C)$ where T is volume size, $(H \times W)$ is spatial dimension of volume and C is number of channels. The spatial and temporal attention blocks project X through linear layer to generate Q_s, K_s, V_s and Q_t, K_t, V_t. In temporal attention Q_t, K_t, V_t is permuted to generate Q_{tp}, K_{tp}, V_{tp} for temporal dot product. The spatial and temporal attention is defined as,

$$X_s = softmax(\frac{Q_s K_s^T}{\sqrt{d_k}})V_s \tag{1}$$

$$X_t = softmax(\frac{Q_{tp} K_{tp}^T}{\sqrt{d_k}})V_{tp} \tag{2}$$

where, X_s is spatial attention map, X_t is temporal attention map and d_k is dimension of Q_s and K_s. To fuse spatial and temporal attention maps, X_s and X_t, we employ deformable convolution. The deformable convolution generates offsets according to temporal attention map X_t and by using these offsets the spatial attention map X_s is aligned. The deformable fusion is given as,

$$X = \int_{c=1}^{C} \sum_{k_n \in R} W(k_n) \cdot X_s(k_0 + k_n + \Delta K_n) \tag{3}$$

where, C is no of channels, X is spatially aligned attention map with respect to X_t. W is the weight matrix of kernels, X_s is spatial attention map, k_0 is starting position of kernel, k_n is enumerating along all the positions in kernel size of R and ΔK_n is the offset sampled from temporal attention map X_t. We empirically observe that fusing spatial and temporal features through a deformable convolution, instead of concatenation through a conv. layer or addition, leads to better performance. The resulting spatio-temporal features of decoder are then input to instance segmentation block to generate final instance masks, as in baseline.

Semantic FG-BG Adversarial Loss: As discussed earlier, a common challenge in mitochondria instance segmentation is to accurately delineate the region of mitochondria instances from the cluttered background. To address this, we introduce a semantic foreground-background (FG-BG) adversarial loss during the training to enhance the FG-BG separability. Here, we introduce the auxiliary discriminator network D with two layers of 3D convolutions with stride 2 during the training as shown in Fig. 2(c). The discriminator takes the input volume I

along with the corresponding mask as an input. Here, the mask M is obtained either from the ground truth or predictions, such that all mitochondria instances within a frame are marked as foreground. While the discriminator D attempts to distinguish between ground truth and predicted masks (M_{gt} and M_{pred}, respectively), the model Ψ learns to output semantic mask such that the predicted masks M_{pred} are close to ground truth M_{gt}. Let $\mathbf{F}_{gt} = \text{CONCAT}(\mathbf{I}, \mathbf{M}_{gt})$ and $\mathbf{F}_{pr} = \text{CONCAT}(\mathbf{I}, \mathbf{M}_{pred})$ denote the real and fake input, respectively, to the discriminator D. Similar to [11], the adversarial loss is then given by,

$$L_{fg-bg} = \min_{\Psi} \max_{D} \Psi[\log D(F_{gt})] + \Psi[\log(1 - D(F_{pr}))] + \lambda_1 \Psi[D(F_{gt}) - D(F_{pr})] \tag{4}$$

Consequently, the overall loss for training is: $L = L_{BCE} + \lambda \cdot L_{fg-bg}$, Where, L_{BCE} is BCE loss, $\lambda = 0.5$ and L_{fg-bg} is semantic adversarial loss.

Table 1. State-of-the-art comparison in terms of AP on Mit-EM-R and MitoEM-H validation sets. Best results are in bold.

Methods	MitoEM-R	MitoEM-H
Wei [36]	0.521	0.605
Nightingale [24]	0.715	0.625
Li [17]	0.890	0.787
Chen [16]	0.917	0.82
STT-UNET (Ours)	**0.958**	**0.849**

Table 2. State-of-the-art comparison in terms of Jaccard and DSC on Lucchi test set. Best results are in bold.

Methods	Jaccard	DSC
Yuan [37]	0.865	0.927
Casser [2]	0.890	0.942
Res-UNET-R [16]	0.895	0.945
Res-UNET-R + MRDA [4]	0.897	0.946
STT-UNET (Ours)	**0.913**	**0.962**

4 Experiments

Dataset: We evaluate our approach on three datasets: MitoEM-R [36], MitoEM-H [36] and Lucchi [22]. The MitoEM [36] is a dense mitochondria instance segmentation dataset from ISBI 2021 challenge. The dataset consists of 2 EM image volumes (30 μm^3) of resolution of $8 \times 8 \times 30$ nm, from rat tissues (MitoEM-R) and human tissue (MitoEM-H) samples, respectively. Each volume has 1000 grayscale images of resolution (4096×4096) of mitochondria, out of which train set has 400, validation set contains 100 and test set has 500 images. Lucchi [22] is a sparse mitochondria semantic segmentation dataset with training and test volume size of $165 \times 1024 \times 768$.

Implementation Details: We implement our approach using Pytorch1.9 [27] (rcom env) and models are trained using 2 AMD MI250X GPUs. During training of MitoEM, for the fair comparison, we adopt same data augmentation technique from [36]. The 3D patch of size ($32 \times 320 \times 320$) is input to the model and trained using batch size of 2. The model is optimized by Adam optimizer with learning rate of $1e^{-4}$. Unlike baseline [16], we do not follow multi-scale training and

perform single stage training for 200k iterations. For Lucchi, we follow training details of [16,36] for semantic segmentation. For fair comparison with previous works, we use the same evaluation metrics as in the literature for both datasets. We use 3D AP-75 metric [36] for MitoEM-R and MitoEM-H datasets. For Lucchi, we use jaccard-index coefficient (Jaccard) and dice similarity coefficient (DSC).

4.1 Results

State-of-the-Art Comparison: Table 1 shows the comparison on MitoEm-R and MitoEM-H validation sets. Our STT-UNET achieves state-of-the-art performance on both sets. Compared to the recent [16], our STT-UNET achieves an absolute gains of 4.1% and 2.9% on MitoEM-R and MitoEM-H validation sets, respectively. Note that [16] employs two decoders for MitoEM-H. In contrast, we utilize only a single decoder for both MitoEM-H and MitoEM-R sets, while still achieving improved segmentation performance. Fig 3 presents the segmentation predictions of our approach on example input regions from the validation set. Our approach achieves promising segmentation results despite the noise in the input samples. Table 2 presents the comparison on Lucchi test set. Our method sets a new state-of-the-art on this dataset in terms of both Jaccard and DSC.

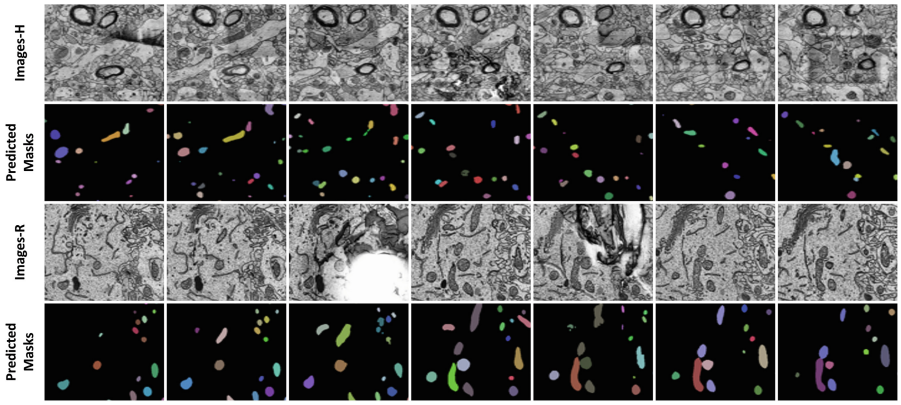

Fig. 3. Qualitative 3D instance segmentation results of our STT-UNET on the example input regions from MitoEM-H and MitoEM-R val sets. Our STT-UNET achieves promising results on these input examples containing noise.

Table 3. Baseline performance comparison.

Methods	MitoEM-R	MitoEM-H
Baseline	0.921	0.823
+ SST	0.948	0.839
+ L_{fg-bg}	**0.958**	**0.849**

Table 4. Ablation study on the impact of feature fusion.

Feature Fusion	MitoEM-R	MitoEM-H
addition	0.950	0.841
concat	0.952	0.842
def-conv	**0.958**	**0.849**

Table 5. Ablation study on the impact of design choice.

Deisgn choice	MitoEM-R	MitoEM-H
spatial	0.914	0.812
spatial-temporal	0.922	0.817
temporal-spatial	0.937	0.832
spatial|temporal	**0.958**	**0.849**

Ablation Study: Table 3 shows a baseline comparison when progressively integrating our contributions: SST module and semantic foreground-background adversarial loss. The introduction of SST module improves performance from 0.921 to 0.941 with a gain of 2.7%. The performance is further improved by 1%, when introducing our semantic foreground-background adversarial loss. Our final approach achieves absolute gains of 3.7% and 2.6% over the baseline on MitoEM-R and MitoEM-H, respectively. We also compare our approach with other attention mechanism in literature such as divided space-time attention [1] and axial attention [35] with our method achieving favorable results with gain of 0.9% and 1.1%, respectively likely due to computing spatial and temporal in parallel and later fusing them through a deformable convolution. Further, we compare our approach with [16] on MitoEM-v2 test set achieving a gain of 4% on MitoEM-R, where the postprocessing from [18] is used to differentiate the mitochondria instances for both methods. Table 4 shows ablation study with feature fusion strategies in our SST module: addition, concat and deformable-conv. The best results are obtained with deformable-conv on both datasets. For encoding spatial and temporal information, we analyze two design choices with SST module: cascaded and split, as shown in Table 5. The best results are obtained using our split design choice (row 3) with spatial and temporal information encoded in parallel and later combined. We also evaluate with different input volumes: 4,8,16,32. We observe best results are obtained when using 32 input volume.

5 Conclusion

We propose a hybrid CNN-transformers based encoder-decoder approach for 3D mitochorndia instance segmentation. We introduce a split spatio-temporal attention (SST) module to capture long-range dependencies within the cubic volume of human and rat mitochondria samples. The SST module computes spatial and temporal attention in parallel, which are later fused. Further, we introduce a semantic adversarial loss for better delineation of mitochondria instances from background. Experiments on three datasets demonstrate the effectiveness of our approach, leading to state-of-the-art segmentation performance.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)
2. Casser, V., Kang, K., Pfister, H., Haehn, D.: Fast mitochondria detection for connectomics. In: Medical Imaging with Deep Learning, pp. 111–120. PMLR (2020)
3. Chen, H., Qi, X., Yu, L., Heng, P.: DCAN: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2016)
4. Chen, Q., Li, M., Li, J., Hu, B., Xiong, Z.: Mask rearranging data augmentation for 3D mitochondria segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 36–46. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_4

5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
8. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
9. Huang, W., Chen, C., Xiong, Z., Zhang, Y., Liu, D., Wu, F.: Learning to restore ssTEM images from deformation and corruption. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12535, pp. 394–410. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66415-2_26
10. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
12. Januszewski, M., et al.: High-precision automated reconstruction of neurons with flood-filling networks. Nat. Meth. **15**(8), 605–610 (2018)
13. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) **54**(10s), 1–41 (2022)
14. Lee, K., Zung, J., Li, P., Jain, V., Seung, H.S.: Superhuman accuracy on the SNEMI3D connectomics challenge. arXiv preprint arXiv:1706.00120 (2017)
15. Li, M., Chen, C., Liu, X., Huang, W., Zhang, Y., Xiong, Z.: Advanced deep networks for 3D mitochondria instance segmentation. arXiv preprint arXiv:2104.07961 (2021)
16. Li, M., Chen, C., Liu, X., Huang, W., Zhang, Y., Xiong, Z.: Advanced deep networks for 3D mitochondria instance segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
17. Li, Z., Chen, X., Zhao, J., Xiong, Z.: Contrastive learning for mitochondria segmentation. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3496–3500. IEEE (2021)
18. Lin, Z., Wei, D., Lichtman, J., Pfister, H.: PyTorch connectomics: a scalable and flexible segmentation framework for EM connectomics. arXiv preprint arXiv:2112.05754 (2021)
19. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: CVPR (2021)
20. Lucchi, A., Li, Y., Smith, K., Fua, P.: Structured image segmentation using kernelized features. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7573, pp. 400–413. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_29
21. Lucchi, A., et al.: Learning structured models for segmentation of 2-D and 3-D imagery. IEEE Trans. Med. Imaging **34**(5), 1096–1110 (2014)
22. Lucchi, A., Smith, K., Achanta, R., Knott, G., Fua, P.: Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features. IEEE Trans. Med. Imaging **31**(2), 474–486 (2011)
23. McBride, H.M., Neuspiel, M., Wasiak, S.: Mitochondria: more than just a powerhouse. Curr. Biol. **16**(14), R551–R560 (2006)

24. Nightingale, L., de Folter, J., Spiers, H., Strange, A., Collinson, L.M., Jones, M.L.: Automatic instance segmentation of mitochondria in electron microscopy data. BioRxiv, pp. 2021–05 (2021)
25. Nunnari, J., Suomalainen, A.: Mitochondria: in sickness and in health. Cell **148**(6), 1145–1159 (2012)
26. Oztel, I., Yolcu, G., Ersoy, I., White, T., Bunyak, F.: Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1195–1200. IEEE (2017)
27. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: NeurIPS, vol. 32 (2019)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: delving into efficient and accurate 3d medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
30. Shamshad, F., et al.: Transformers in medical imaging: a survey. arXiv preprint arXiv:2201.09873 (2022)
31. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
32. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
33. Vazquez-Reina, A., Gelbart, M., Huang, D., Lichtman, J., Miller, E., Pfister, H.: Segmentation fusion for connectomics. In: 2011 International Conference on Computer Vision, pp. 177–184. IEEE (2011)
34. Wang, C., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3D UNet framework for multi-class segmentation on full resolution image. arXiv preprint arXiv:1804.04341 (2018)
35. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C.: Axial-DeepLab: stand-alone axial-attention for panoptic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 108–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_7
36. Wei, D., et al.: MitoEM dataset: large-scale 3D mitochondria instance segmentation from EM images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 66–76. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_7
37. Yuan, Z., Yi, J., Luo, Z., Jia, Z., Peng, J.: EM-Net: centerline-aware mitochondria segmentation in EM images via hierarchical view-ensemble convolutional network. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1219–1222. IEEE (2020)