

# Self-Supervised Neuron Segmentation with Multi-Agent Reinforcement Learning

Yinda Chen<sup>1,2</sup>, Wei Huang<sup>1</sup>, Shenglong Zhou<sup>1</sup>, Qi Chen<sup>1</sup>, Zhiwei Xiong<sup>1,2,\*</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{cyd0806, weih527, slzhou96, qic}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn

## Abstract

The performance of existing supervised neuron segmentation methods is highly dependent on the number of accurate annotations, especially when applied to large scale electron microscopy (EM) data. By extracting semantic information from unlabeled data, self-supervised methods can improve the performance of downstream tasks, among which the mask image model (MIM) has been widely used due to its simplicity and effectiveness in recovering original information from masked images. However, due to the high degree of structural locality in EM images, as well as the existence of considerable noise, many voxels contain little discriminative information, making MIM pretraining inefficient on the neuron segmentation task. To overcome this challenge, we propose a decision-based MIM that utilizes reinforcement learning (RL) to automatically search for optimal image masking ratio and masking strategy. Due to the vast exploration space, using single-agent RL for voxel prediction is impractical. Therefore, we treat each input patch as an agent with a shared behavior policy, allowing for multi-agent collaboration. Furthermore, this multi-agent model can capture dependencies between voxels, which is beneficial for the downstream segmentation task. Experiments conducted on representative EM datasets demonstrate that our approach has a significant advantage over alternative self-supervised methods on the task of neuron segmentation. Code is available at <https://github.com/ydchen0806/dbMiM>.

## 1 Introduction

Neuron segmentation is a crucial task for neuroscientists that allows for the analysis of the distribution and morphology of neurons, providing valuable insights into the connectomics research [Sheridan *et al.*, 2022; Krasowski *et al.*, 2017]. Electron microscopy (EM) is the mainstream method for accurately identifying neural structures, but the dense nature of neurons and the presence of artifacts and deformations in

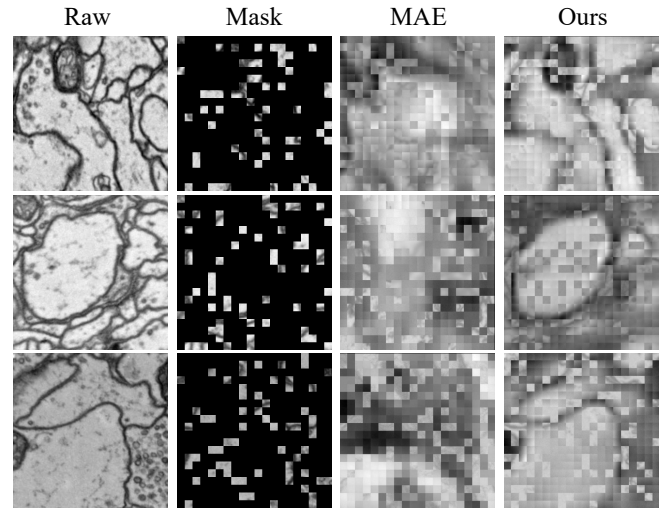


Figure 1: A comparison of the reconstruction effectiveness of our proposed method with MAE. The first column shows the original EM image, the second column shows the image after masking 85% of the voxels, the third column shows the reconstruction using the MAE method, and the fourth column shows the reconstruction using our method.

EM images make the labeling process costly and decrease the credibility of existing annotation data [Deng *et al.*, 2022; Zhou *et al.*, 2019]. Therefore, fully supervised neuron segmentation methods meet great challenges, especially when applied to large scale EM data.

Self-supervised methods have emerged as a solution to the limitations of fully supervised methods, which can be roughly divided into two categories: contrastive learning-based approach and mask image model (MIM)-based approach. The former requires a large number of positive and negative samples [Chen *et al.*, 2020; Grill *et al.*, 2020; You *et al.*, 2022; Chen *et al.*, 2023] and relies heavily on data augmentation [Caron *et al.*, 2021], making it a high-cost option for 3D biomedical images. The latter aims to learn useful structural information in images by masking and recovering certain voxels, which has been recently applied to pretraining biomedical images, showing improvements in downstream tasks [Zhou *et al.*, 2022; Tang *et al.*, 2022; Huang *et al.*, 2022a]. However, the highly localized and

\*Corresponding Author

structured nature of EM data, as well as the existence of considerable noise, make it inefficient to directly employ the existing MIM in extracting useful information for neuron segmentation. It has also been observed that the masking ratio and masking strategy of MIM are highly sensitive and the optimal ones vary greatly across different datasets. Adjusting these configurations to train large models requires significant manual efforts and resources.

In this paper, targeting the neuron segmentation task, we propose a novel decision-based MIM relying on multi-agent reinforcement learning (MARL) [Littman, 1994] for automatically selecting the appropriate masking ratio and masking strategy, which consists of a target network and a policy network. Our approach partitions the input EM volume into patches and treats each patch as a basic control unit. The overall multi-agent task is modeled as a search process for patch masking strategies, where the action space for each patch is to either keep the original voxels or mask them. The feedback of the target network, in the form of the reconstruction loss, serves as the team reward signal for guiding the policy network to adaptively learn masking strategies that are beneficial for the pretraining task [Foerster *et al.*, 2016]. To improve the stability of training and achieve optimal joint decision-making for the entire volume, all agent networks share parameters and are trained in parallel. Furthermore, we introduce the HOG feature as an additional reconstruction loss to enable the target network to learn more structure information. Finally, following a UNETR decoder design [Hatamizadeh *et al.*, 2022], we add a segmentation head to the pretrained target network in the finetuning stage. Experimental results in Figure 1 show that our decision-based MIM achieves clearer reconstruction results than the original MAE [He *et al.*, 2022] in the pretraining phase.

Overall, our main contribution lies in the following aspects:

- 1) We propose an efficient self-supervised method, named decision-based MIM, for EM neuron segmentation using unlabeled EM data. To the best of our knowledge, it is the first effort that large-scale transformer pretraining is conducted on this task.
- 2) We propose a MARL-based approach for searching the optimal masking ratio and masking strategy by treating each patch as an agent with a shared policy, effectively exploring the search space and capturing dependencies between voxels.
- 3) We introduce the HOG feature as an additional reconstruction loss of our decision-based MIM, improving the convergence speed of network training and the performance of the downstream segmentation task.
- 4) We comprehensively demonstrate the effectiveness of our proposed method on two representative EM datasets, especially against alternative self-supervised methods on the task of neuron segmentation.

## 2 Related Work

### 2.1 Neuron Instance Segmentation

In the field of EM image processing, neuron instance segmentation is an important task. [Turaga *et al.*, 2010] first proposed a convolutional neural network based on affinity generation, followed by clustering voxels in the affinity graph into

instances based on post-processing methods such as watershed and LMC [Beier *et al.*, 2017]. In recent years, there have been more advanced networks for the affinity-based approach. Funke *et al.* [Funke *et al.*, 2018] introduced the MALIS loss [Briggman *et al.*, 2009] during the training process to encourage the network to generate correct topological segmentation. [Huang *et al.*, 2022b] introduced an embedding pyramid module to simulate affinity at different scales. [Liu *et al.*, 2022] incorporated both embedding and affinity information and combined it with the graph neural network to further improve the distinguishability of adjacent objects in the feature space. However, due to the anisotropic resolution of 3D EM images in lateral and axial directions, the usage of transformer structures remains unexplored in the field of neuron segmentation. In this paper, we use an affinity-based setup and upgrade the decoder of UNETR [Hatamizadeh *et al.*, 2022] to adapt to the anisotropic EM features.

### 2.2 Mask Image Model (MIM)

The MIM is an important branch of self-supervised learning. Masked autoencoders (MAE) [He *et al.*, 2022] used an asymmetric encoding-decoding structure, encoding only the unmasked patches and using a lightweight decoder to recover the masked patches, which greatly reduced the resources required for computation and quickly became a mainstream structure for MIM. [Feichtenhofer *et al.*, 2022] and [Tong *et al.*, 2022] separately validated the effectiveness of MAE on video datasets and proved that higher image masking ratios can be used in 3D datasets. [Zhou *et al.*, 2022] was the first to introduce MAE to the medical image field. [Bachmann *et al.*, 2022] introduced a multi-modal multi-task adaptation to MAE, resulting in better performance than the original MAE. [Gao *et al.*, 2022] proposed a multi-scale mixed convolution to encode images and achieved improved results on fine-grained downstream tasks such as image segmentation. [Wei *et al.*, 2022] focused on the reconstruction target of the decoder and found that reconstructing artificial features such as HOG and SIFT can facilitate the network to classify and localize objects.

In summary, existing MIM-based methods mainly focus on the design of the encoder architecture, prediction head, and prediction target. Although many works have demonstrated the impact of masking strategies on downstream tasks, there has been little research along this line, and the latest method also requires a search-based masking strategy with a fixed masking ratio [Bandara *et al.*, 2022]. This paper proposes a novel MARL-based approach for adaptively learning the optimal masking ratio and masking strategy, making the pretrained model more robust and achieving better performance on downstream tasks.

### 2.3 Multi-Agent Reinforcement Learning (MARL)

In the field of deep RL, multiple agents working together can improve the efficiency and robustness of the model due to the limited observation and action space of a single agent. Given the complexity of computer vision tasks, MARL is often used to interact with the common environment to make decisions. [Liao *et al.*, 2020] proposed a method for iteratively refining medical image segmentation using interactive MARL, where

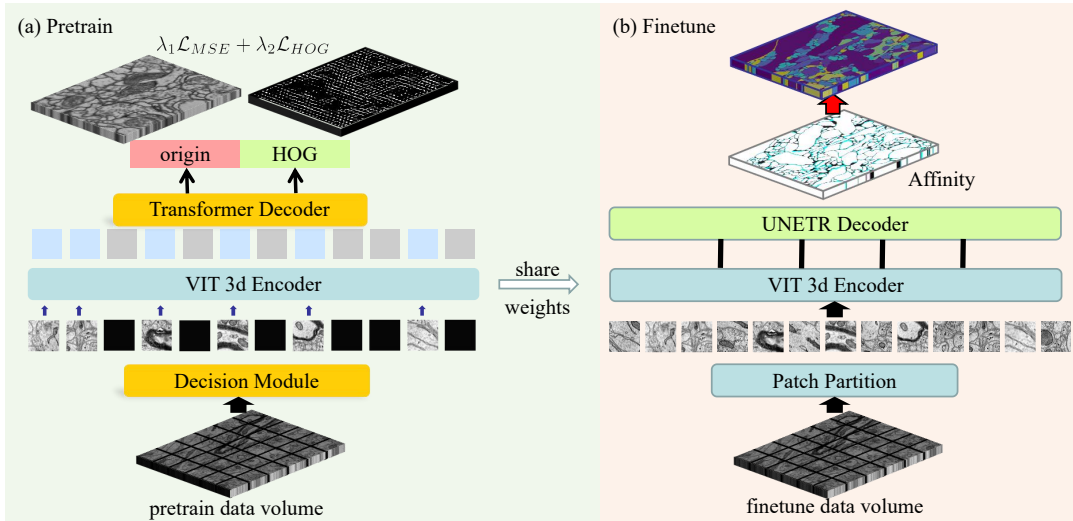


Figure 2: Our proposed network architecture is divided into two main components: (a) the decision-based Mask Image Model (MIM) pretraining process, which employs our proposed decision module to select appropriate patches for masking and then utilizes a 3D Vision Transformer (ViT) encoder to encode the visible patches. The resulting tokens are then passed through a lightweight decoder to reconstruct the original voxels and histograms of oriented gradient (HOG) features. (b) The fine-tuning process for the downstream segmentation task utilizes the encoder weights from the pretraining process and adds a UNETR segmentation head to output the affinity map. The final segmentation results are obtained through post-processing methods such as watershed.

rough image segmentation is initially provided and the network is iteratively refined through user feedback in the form of rewards until the segmentation is sufficient. [Lin *et al.*, 2021a] proposed a method for augmenting images through the use of blocks, with each block acting as an agent and working together to produce optimal data augmentation. Still, the application of MARL in computer vision suffers from the large decision space and difficulty in obtaining rewards. In this paper, to reduce the complexity of the state and the difficulty of searching for RL policies, we first segment the image into patches through a transformer-encoder and treat each patch as an agent. The action space for each patch is limited to only two options, masking or keeping the original voxels, and the rewards are obtained through the reconstruction loss of MAE. Therefore, the masking decision can be naturally modeled as a Markov process and optimized through MARL.

### 3 Proposed Method

Our decision-based MIM consists of two stages of training: pretraining and finetuning. Figure 2 illustrates the overall flow of the network training. In this section, we will first introduce some basic theories of Vision Transformer (ViT) and MARL, and then explain in detail the encoders, decoders, and loss functions of the two stages, as well as specific MARL modeling methods.

#### 3.1 Encoder-Decoder Design

We use ViT [Dosovitskiy *et al.*, 2021] as the backbone architecture for decision-based MIM pretraining and downstream segmentation tasks. To represent high-dimensional data in a ViT, we must transform it into a sequence of patches. Given an input 3D volume  $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ , where  $C$  is the number of channels and  $(H, W, D)$  is the resolution, we reshape

it into a sequence of flattened 3D patches  $\mathbf{x}_p \in \mathbb{R}^{N \times (P^3 \cdot C)}$ . The patch resolution is given by  $(P/4, P, P)$ , and the number of patches is calculated as  $N = \frac{4HW D}{P^3}$ . These patches are then transformed into patch embeddings via a trainable linear projection.

Consistent with the MAE setup, we divide the image patches into visible and masked groups. The encoder in the MAE and ViT architectures processes only the visible blocks. To enhance the performance of our decoder, we incorporated a histogram-of-oriented-gradients (HOG) feature [Dalal and Triggs, 2005], which has been shown to improve pretraining. This is achieved by providing the decoder with various markers, including a patch representation from the encoder and learnable position embeddings. By incorporating positional embedding in all input markers, we enable the decoder to simultaneously recover both the HOG feature and the original voxels, resulting in superior performance compared to the MAE method.

In our pretraining process, we utilize a loss function that combines both the mean squared error (MSE) loss for reconstructing the original voxels and the HOG loss for recovering the HOG features. The HOG feature can be calculated using the following equation

$$\text{HOG}_{i,j} = \frac{\sum_{x \in S_{i,j}} w(x)g(x)}{\sum_{x \in S_{i,j}} w(x)}, \quad (1)$$

where  $\text{HOG}_{i,j}$  is the histogram of oriented gradients for the cell located at position  $(i, j)$ ,  $S_{i,j}$  is the set of voxels in the cell  $(i, j)$ ,  $w(x)$  is a weighting function that assigns a weight to each voxel  $x$  in the cell, and  $g(x)$  is the gradient orientation of the voxel  $x$ .

Our overall loss function can be expressed as

$$\mathcal{L}_{pretrain} = \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{HOG}, \quad (2)$$

where  $\lambda_1$  and  $\lambda_2$  denote the weights assigned to the MSE and HOG losses, respectively. We set  $\lambda_1$  to 0.1 and  $\lambda_2$  to 1.

### 3.2 Decision Module

In our proposed decision module, we model the masking strategy of image patches in MAE as a multi-agent cooperative decision-making problem and adopt a multi-agent reinforcement learning method to solve it. Here, we will introduce in detail the modeling methods of model states, observations, actions, the design of multi-agent team rewards, and the learning method used to update the policy network.

As shown in Figure 3, given the original input batch of image  $\mathbf{x}$ , according to the ViT setup, we divide the image into equal-sized and non-overlapping patches. Our MARL policy aims to determine the overall joint masking policy based on the current input state and the observation of each agent. Our basic setup is as follows.

**State.** The enhancement policy for each patch is closely related to the information of the context, so the decision-making process of MARL requires perceiving the semantic information of the image rather than directly inputting the patches of the raw image. In order to ensure the consistency and convergence of training, we use the target network, i.e. ViT, as the backbone to extract deep semantic features of the image. The global state at time step  $t$  is denoted as  $S^t$  and is visible to all agents.

**Observation.** In addition to capturing global information  $S^t$ , each agent needs to make a masking policy decision based on its own local observation. In general, the observation in MARL tasks is often a part of the global state. Considering these factors, we take the deep feature  $\text{ViT}(P_i)$  of patch  $P_i$  as the current observation value  $O_i^t$  for the  $i$ -th agent. The feature extractor for local features is the same as the one for global features, both using the ViT backbone.

**Action.** The action of the  $i$ -th agent aims to output whether patch  $P_i$  needs to be masked. We define the action of the  $i$ -th agent as a vector  $A_i$ . The joint action space can be represented as  $A = \{A_1, A_2, \dots, A_N\}$ , where  $N$  represents the total number of patches. Since the action space only has two possibilities, masking or keeping the original voxels, the dimension of  $A_i$  is 2. Given the current state  $S^t$  and observation  $O_i^t$ , each agent  $i$  will determine an action  $a_i(S^t, O_i^t) \in A_i$  based on the current policy. The final output is the global joint action  $\mathbf{a}^t = \{a_1^t \cup a_2^t, \dots, \cup a_N^t\}$ . After all patches have taken their corresponding actions through the decision policy, the time step is updated to  $t = t + 1$  and we obtain the enhanced volume through the decision module.

**Rewards.** Rewards are introduced in our MARL decision-making process in order to guide the agents to learn expected behaviors that will improve the main task’s performance through more reasonable masking ratios and masking strategies, allowing the target net to better learn semantic information in the volume. Previous works [Lin *et al.*, 2021b; Zhang *et al.*, 2019] attempted to increase the training loss of the target network based on rewards in order to generate deeper, more difficult-to-learn features. Inspired by these works, we refine the reward design based on the MAE pre-training paradigm. By comparing the loss difference between

the data obtained from the masked data  $\mathbf{x} \cdot \mathbf{a}^{t-1}$  generated by the previous time step’s decision module and the data obtained from the current decision module  $\mathbf{x} \cdot \mathbf{a}^t$ , we compute the reward for the MARL policy. This encourages higher training loss during the MARL decision-making process, as shown by the equation

$$r^t = \mathcal{L}_{\text{pretrain}}(\phi(\mathbf{x} \cdot \mathbf{a}^t)) - \mathcal{L}_{\text{pretrain}}(\phi(\mathbf{x} \cdot \mathbf{a}^{t-1})). \quad (3)$$

In the above equation,  $\mathcal{L}_{\text{pretrain}}$  represents the reconstruction loss generated by the target network, and  $\phi$  denotes the target network. The accumulated reward of one sequence is

$$R^t = \sum_{i=t-T+1}^t \gamma^{i-1} \bar{r}^i, \quad (4)$$

where  $T$  represents the desired time step length to be calculated, the discount factor  $\gamma$  takes a value in  $(0, 1]$ , and  $\bar{r}^t$  is the mean rewards at time  $t$ .

**Policy Learning.** Considering that the action space for MARL decisions is discrete, we utilize the widely used Asynchronous Advantage Actor-Critic (A2C) algorithm [Mnih *et al.*, 2016] to perform MARL policy learning. Since the search space for actions is not large, we use simple convolution and pooling modules to adjust the Actor and Critic networks to the structure shown in Figure 3. The policy network is divided into an Actor and a Critic, which are adapted to the RL training algorithm. The Actor network learns a discrete control policy  $\pi(a_i^t | S^t, O_i^t)$ , while the Critic network aims to estimate the value of the state  $V_\pi(S^t)$ . We model the centralized action value function  $Q$ , which takes in the state information  $S$  and the actions of all agents, and outputs a  $Q$  value for the team, given by

$$Q^\pi(S^t, \mathbf{a}^t) = E_\pi [R_t | S^t, a_1^t, \dots, a_N^t], \quad (5)$$

where  $\mathbf{a}^t$  represents the joint action of all agents, defined as  $\mathbf{a} = \{a_i, \dots, a_N\}$ , and  $R_t$  is the long-term discounted reward, given by equation 4. The advantage function on the policy is then given by

$$A^\pi(S^t, \mathbf{a}^t) = Q^\pi(S^t, \mathbf{a}^t) - V^\pi(S^t), \quad (6)$$

where  $A^\pi(S^t, \mathbf{a}^t)$  is the advantage of taking action  $\mathbf{a}^t$  given state  $S^t$  at time step  $t$ ,  $V^\pi(S^t)$  is the current state estimate output by Critic. It indicates that the actual accumulated reward is independent of the state and reduces the variance of the gradient. We use  $\theta_p$  and  $\theta_v$  to denote the parameters of the Actor and Critic, respectively. The squared value of the dominance function  $A^\pi$  is taken as the loss function to update  $\theta_v$  as

$$\mathcal{L}(\theta_v) = A^\pi(S^t, \mathbf{a}^t)^2. \quad (7)$$

To further achieve cooperative capability, the loss function of the updated Actor  $\theta_p$  is defined as

$$\mathcal{L}(\theta_p) = -\log \pi_\theta(\mathbf{a}^t | S^t) A^\pi(S^t, \mathbf{a}^t), \quad (8)$$

where  $\pi_\theta(\mathbf{a}^t | S^t)$  is the Actor output, that is, the probability of taking each action  $a_i^t$ . The Actor and Critic are jointly trained in an end-to-end manner.

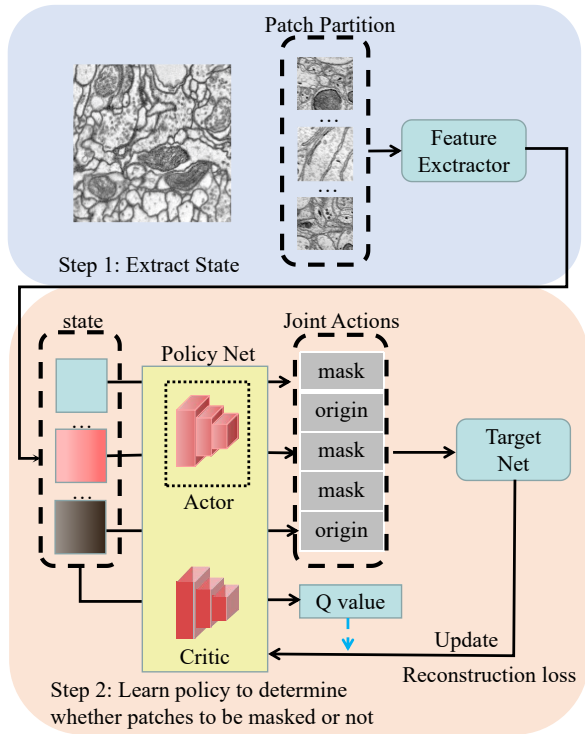


Figure 3: The framework of our proposed decision module. The first step in the process is to extract features from the encoder of the target net. In the second step, the policy network is used to decide whether or not the current patch needs to be masked. The output of the joint action constitutes the final decision.

### 3.3 Neuron Instance Segmentation Method

The UNETR model is specifically designed for 3D segmentation tasks, such as organ segmentation. It is built upon the pretrained ViT encoder from decision-based MIM and a randomly-initialized convolutional decoder. The UNETR architecture is inspired by the U-Net model [Ronneberger *et al.*, 2015], with the inclusion of skip connections between the encoder and decoder at multiple resolutions. The input to the UNETR decoder is a sequence of representations, denoted as  $z_3, z_6, z_9, z_{12}$ , which are reshaped to restore the spatial dimension of  $\frac{H}{P} \times \frac{W}{P} \times \frac{D}{P} \times K$ , where  $K$  is the feature channel.

Starting from the deepest feature,  $z_{12}$ , each representation is processed through a varying number of deconvolutional blocks to increase its resolution by a specific factor. For example,  $z_{12}$  and  $z_9$  are upsampled by a factor of  $\times 2$ ,  $z_6$  is upsampled by a factor of  $\times 4$ , and  $z_3$  is upsampled by a factor of  $\times 8$ . Then, representations at the same spatial resolution, such as  $z_{12}$  and  $z_9$ , are concatenated and further upsampled to match the shape of a shallower feature. This process of concatenation and upsampling is repeated until the full resolution of the input is restored. Finally, the output layer combines the upsampled feature and the original full-resolution input to predict the segmentation map.

In this paper, for the segmentation task of EM neurons, we use a combination of affinity map-based and post-processing approaches. We first create an affinity graph based on the voxel affinity. This graph serves as the foundation for our

post-processing techniques, which utilize both waterz [Funke *et al.*, 2018] and LMC [Beier *et al.*, 2017] to cluster the affinity map and produce the final neuron segmentation results. Affinity-based methods have proven to be highly effective in accurately segmenting and analyzing these complex structures.

## 4 Experiments

### 4.1 Training Strategy

Our strategy consists of two phases: pretraining and fine-tuning. In the pretraining phase, to improve the training efficiency of the framework, we first pretrain decision-based MIM for 100k iterations, synchronously updating the parameters of MAE and the policy network in the decision module, and then fix the parameters of policy network and only update the parameters of MAE for another 100k iterations. In the fine-tuning phase, we load the pretrained ViT weights into the model for the downstream task and train for 200k iterations.

We use the Adam optimizer in both the pretraining and fine-tuning phases, with  $\beta_1 = 0.9, \beta_2 = 0.999$ . The only difference lies in the pretraining process, where we set the learning rate to 0.0001 and perform batch size 16 pretraining on 8 RTX 3090s. In the fine-tuning phase, we adopt a Layer-wise Learning Rate Decay (LLRD) training method, which adjusts the learning rate layer by layer during training. We set the learning rate of the last layer’s parameters to 0.001 and the learning rate of the previous layer’s parameters to 0.95 times the learning rate of the next layer’s parameters. We conduct batch size 8 fine-tuning on 2 RTX 3090s.

### 4.2 Datasets and Evaluation Metrics

**FAFB.** The Full Adult Fly Brain (FAFB) dataset [Zheng *et al.*, 2018] is a highly valuable resource for neuroinformatics research, offering a comprehensive and detailed view of the neural architecture of the *Drosophila melanogaster* (fruit fly) brain. With a size of approximately 40 terabytes, this dataset features high-resolution 3D images with a resolution of approximately 4 nanometers per pixel, as well as manually annotated segmentation data identifying various brain structures such as neurons, glial cells, blood vessels, synapses, and other neuropil regions.

In our work, we utilize the FAFB dataset as a key component in our pretraining process. To optimize the efficiency of our model, we first downsample the original dataset by a factor of 4, carefully curating a selection of 60G images that exhibit exceptional imaging quality from the entire dataset. This strategic selection ensures that our model is trained on the most accurate and reliable data possible, setting the foundation for its future performance.

**CREMI.** The CREMI dataset is derived from the FAFB dataset, which has three manually labeled subvolumes from *drosophila* brain, of which CREMI A has more regular forms, CREMI C has higher size disparities, and CREMI B has in-between segmentation difficulty. Each sub-volume has 125 slices of  $1250 \times 1250$  images, and we choose the first 60 slices for training, 15 slices for validation, and the remaining 50

Dataset	Method	waterz [Funke <i>et al.</i> , 2018]				LMC [Beier <i>et al.</i> , 2017]			
		VOI <sub>S</sub> ↓	VOI <sub>M</sub> ↓	VOI ↓	ARAND ↓	VOI <sub>S</sub> ↓	VOI <sub>M</sub> ↓	VOI ↓	ARAND ↓
CREMI A	superhuman [Lee <i>et al.</i> , 2017]	0.443	<b>0.320</b>	0.763	0.132	0.578	0.272	0.850	<b>0.131</b>
	MALA [Funke <i>et al.</i> , 2018]	0.478	0.627	1.105	0.459	0.574	0.303	0.878	0.146
	UNETR [Hatamizadeh <i>et al.</i> , 2022]	0.495	0.359	0.854	0.153	0.630	0.278	0.908	0.141
	MAE [He <i>et al.</i> , 2022]+UNETR	0.463	0.324	0.787	0.144	0.563	0.268	0.831	0.139
	Dino [Caron <i>et al.</i> , 2021]+UNETR	0.482	0.351	0.833	0.150	0.580	0.287	0.867	0.147
	ours+UNETR	<b>0.411</b>	0.331	<b>0.743</b>	<b>0.131</b>	<b>0.537</b>	<b>0.260</b>	<b>0.797</b>	0.134
CREMI B	superhuman [Lee <i>et al.</i> , 2017]	0.668	0.409	1.076	<b>0.082</b>	0.959	0.232	1.191	<b>0.060</b>
	MALA [Funke <i>et al.</i> , 2018]	0.797	0.512	1.309	0.147	1.060	0.264	1.324	0.084
	UNETR [Hatamizadeh <i>et al.</i> , 2022]	0.937	0.397	1.333	0.103	1.194	0.230	1.424	0.116
	MAE [He <i>et al.</i> , 2022]+UNETR	0.776	0.391	1.167	0.099	0.994	0.224	1.218	0.102
	Dino [Caron <i>et al.</i> , 2021]+UNETR	0.916	0.396	1.312	0.104	1.167	0.229	1.396	0.111
	ours+UNETR	<b>0.642</b>	<b>0.381</b>	<b>1.023</b>	0.092	<b>0.893</b>	<b>0.220</b>	<b>1.113</b>	0.097
CREMI C	superhuman [Lee <i>et al.</i> , 2017]	0.943	0.385	1.328	0.134	1.176	0.260	1.436	0.125
	MALA [Funke <i>et al.</i> , 2018]	<b>0.901</b>	0.621	1.522	0.169	<b>1.137</b>	0.289	1.426	0.127
	UNETR [Hatamizadeh <i>et al.</i> , 2022]	0.996	0.423	1.419	0.158	1.417	0.236	1.653	0.148
	MAE [He <i>et al.</i> , 2022]+UNETR	1.001	0.298	1.299	0.120	1.272	0.214	1.486	0.113
	Dino [Caron <i>et al.</i> , 2021]+UNETR	1.011	0.412	1.423	0.156	1.364	0.234	1.598	0.146
	ours+UNETR	0.925	<b>0.276</b>	<b>1.201</b>	<b>0.107</b>	1.194	<b>0.204</b>	<b>1.398</b>	<b>0.112</b>
AC4	superhuman [Lee <i>et al.</i> , 2017]	0.721	0.295	1.016	<b>0.187</b>	<b>0.770</b>	0.343	1.113	0.110
	MALA [Funke <i>et al.</i> , 2018]	0.734	0.385	1.119	0.305	0.832	0.357	1.189	<b>0.108</b>
	UNETR [Hatamizadeh <i>et al.</i> , 2022]	0.908	0.337	1.245	0.316	1.007	0.340	1.347	0.134
	MAE [He <i>et al.</i> , 2022]+UNETR	0.791	0.306	1.097	0.254	0.888	0.298	1.186	0.120
	Dino [Caron <i>et al.</i> , 2021]+UNETR	0.889	0.329	1.218	0.298	1.001	0.314	1.315	0.129
	ours+UNETR	<b>0.647</b>	<b>0.285</b>	<b>0.931</b>	0.243	0.795	<b>0.284</b>	<b>1.079</b>	0.113

Table 1: Results on the CREMI dataset, VOI<sub>S</sub> represents split error, VOI<sub>M</sub> represents merge error, and VOI is the sum of the two. The final segmentation results are generated by using two classic post-processing methods, waterz and LMC.

slices for testing, which are utilized to validate our method’s performance on segmentation tasks of varying difficulty.

**AC3/AC4.** AC3/AC4 [Kasthuri *et al.*, 2015] are mouse somatosensory cortex datasets with 256 and 100 successive EM images (1024×1024), respectively. The first 80 slices of AC3 are used as the training set, the following 20 slices as the validation set, and the first 50 slices of AC4 are used as the testing set.

**Evaluation Metrics.** We are more interested in the model’s performance in the downstream task in the self-supervised training. To assess the influence of segmentation on EM neurons, we primarily use Variation of Information (VOI) [Nunez-Iglesias *et al.*, 2013] and Adapted Rand Error (Arand) [Arganda-Carreras *et al.*, 2015] metrics. Smaller VOI and ARAND values represent better segmentation results.

### 4.3 Experimental Results

**Decision Making Process.** Decision-based MIM is pre-trained using the FAFB dataset. During pretraining, the policy network in the decision module is also updated, and the real-time change in masking ratio is recorded as in Figure 4. The decision-making process starts with random decisions, resulting in an overall masking ratio of around 0.5. Then, using the reconstruction loss as a reward, the decision-making principles of the agents are updated. After 50k iterations, the decision module converges and produces better reconstruction results with less lost information. During pretraining, it is found that starting with a lower masking ratio is more beneficial for the EM dataset, and gradually increasing the mask-

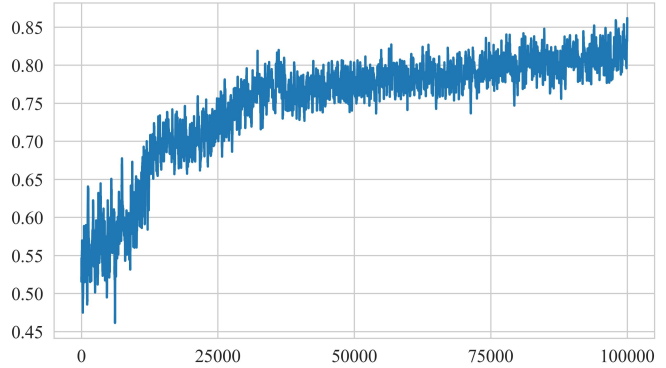


Figure 4: The variation of the masking ratio in the MARL decision process, where the final convergence result shows that the optimal masking ratio fluctuates around 0.83.

ing ratio helps the network’s learning process progress from easy to difficult datasets, which is effective for both upstream reconstruction and downstream segmentation tasks.

**Results on CREMI.** We compare our method to the original MAE pretraining method [He *et al.*, 2022] and the Dino pretraining method [Caron *et al.*, 2021] based on contrastive learning, as well as two affinity-based Unet structures for instance segmentation, superhuman [Lee *et al.*, 2017] and MALA [Funke *et al.*, 2018]. The outcomes of the trials are provided in Table 1 after applying both waterz and LMC post-processing procedures. Figure 5 shows the visualization results. Compared to existing self-supervised methods,

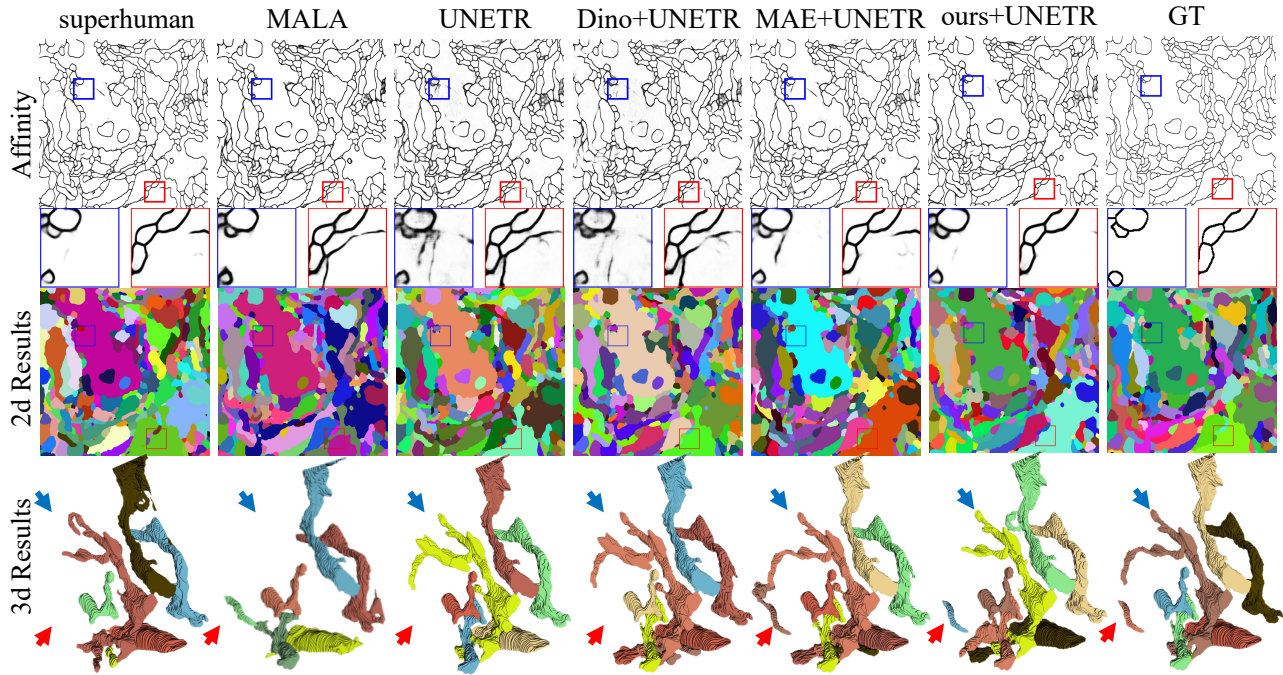


Figure 5: The visualization results in the CREMI C dataset reveal the areas of over-segmentation and under-segmentation produced by methods such as Superhuman, MALA, and UNETR.

MSE	HOG	waterz		LMC	
		VOI ↓	ARAND ↓	VOI ↓	ARAND ↓
✓		1.097	0.254	1.186	0.120
	✓	0.987	0.251	1.103	0.127
✓	✓	<b>0.957</b>	<b>0.245</b>	<b>1.089</b>	<b>0.117</b>

Table 2: Ablation results of the multi-task reconstruction.

our decision-based MIM approach shows a significant improvement for the downstream task and it also performs better than commonly used CNN-based methods in neuron segmentation, as can be seen in Table 1.

**Results on AC3/AC4.** In order to verify the robustness and generalizability of our pretraining method, we further conduct experiments on mouse cortical neural cells. The results in Table 1 indicate that our approach leads to a clear improvement in the downstream segmentation task, even when applied across different species.

#### 4.4 Ablation Study

We use UNETR to perform a comprehensive ablation study on the AC4 dataset.

**Effectiveness of the multi-task reconstruction.** We conduct ablation experiments on the model’s reconstruction targets. As shown in Table 2, it demonstrates that utilizing HOG and MSE losses together for reconstruction in the MIM outperforms using MSE or HOG alone.

**Effectiveness of the decision module.** We compare our proposed method to a straightforward solution that manually

Rate	Decision	waterz		LMC	
		VOI ↓	ARAND ↓	VOI ↓	ARAND ↓
0.65		1.065	0.264	1.196	0.131
0.75		0.997	0.258	1.132	0.129
0.85		0.957	0.245	1.097	0.121
0.95		1.005	0.260	1.121	0.127
/	✓	<b>0.931</b>	<b>0.243</b>	<b>1.079</b>	<b>0.113</b>

Table 3: Ablation results of the decision module.

adjusts the masking ratio. The ablation results, as shown in Table 3, demonstrate that our approach not only eliminates the need for manual adjustment of masking ratios but also outperforms the best results achieved through manual adjustment.

## 5 Conclusion

In this paper, we propose a decision-based MIM approach for neuron segmentation. Our method eliminates the need for manual adjustment of masking ratios and masking strategies, using multi-agent cooperation to search for the optimal solution. Additionally, during the pretraining process, we incorporate a multi-task reconstruction and utilize HOG features to enhance the model’s learning ability. Our method is validated on a variety of EM neuron datasets to demonstrate its generalizability.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62021001.

## References

- [Arganda-Carreras *et al.*, 2015] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, page 142, 2015.
- [Bachmann *et al.*, 2022] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022.
- [Bandara *et al.*, 2022] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhan, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. *arXiv preprint arXiv:2211.09120*, 2022.
- [Beier *et al.*, 2017] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017.
- [Briggman *et al.*, 2009] Kevin Briggman, Winfried Denk, Sebastian Seung, Moritz Helmstaedter, and Srinivas C Turaga. Maximin affinity learning of image segmentation. *Advances in Neural Information Processing Systems*, 22, 2009.
- [Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.
- [Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [Chen *et al.*, 2023] Yinda Chen, Che Liu, Wei Huang, Sibao Cheng, Rossella Arcucci, and Zhiwei Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023.
- [Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.
- [Deng *et al.*, 2022] Shiyu Deng, Wei Huang, Chang Chen, Xueyang Fu, and Zhiwei Xiong. A unified deep learning framework for stem image restoration. *IEEE Transactions on Medical Imaging*, 41(12):3734–3746, 2022.
- [Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [Feichtenhofer *et al.*, 2022] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.
- [Foerster *et al.*, 2016] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- [Funke *et al.*, 2018] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018.
- [Gao *et al.*, 2022] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. Mcmae: Masked convolution meets masked autoencoders. In *Advances in Neural Information Processing Systems*, 2022.
- [Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- [Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.
- [He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [Huang *et al.*, 2022a] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 2022.
- [Huang *et al.*, 2022b] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *Proc. AAAI Conf. Artif. Intell.*, 2022.
- [Kasthuri *et al.*, 2015] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil



- Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015.
- [Krasowski et al., 2017] NE Krasowski, Thorsten Beier, GW Knott, Ullrich Köthe, Fred A Hamprecht, and Anna Kreshuk. Neuron segmentation with high-level biological priors. *IEEE transactions on medical imaging*, 37(4):829–839, 2017.
- [Lee et al., 2017] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.
- [Liao et al., 2020] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9394–9402, 2020.
- [Lin et al., 2021a] Shiqi Lin, Tao Yu, Ruoyu Feng, Xin Li, Xin Jin, and Zhibo Chen. Local patch autoaugmentation with multi-agent collaboration. *arXiv preprint arXiv:2103.11099*, 2021.
- [Lin et al., 2021b] Shiqi Lin, Zhizheng Zhang, Xin Li, Wenjun Zeng, and Zhibo Chen. Selectaugment: Hierarchical deterministic sample selection for data augmentation. *arXiv preprint arXiv:2112.02862*, 2021.
- [Littman, 1994] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.
- [Liu et al., 2022] Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Biological instance segmentation with a superpixel-guided graph. In *IJCAI*, 2022.
- [Mnih et al., 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.
- [Nunez-Iglesias et al., 2013] Juan Nunez-Iglesias, Ryan Kennedy, Toufiq Parag, Jianbo Shi, and Dmitri B Chklovskii. Machine learning of hierarchical clustering to segment 2d and 3d images. *PLoS one*, 8(8):e71715, 2013.
- [Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [Sheridan et al., 2022] Arlo Sheridan, Tri M Nguyen, Dipdip Deb, Wei-Chung Allen Lee, Stephan Saalfeld, Srinivas C Turaga, Uri Manor, and Jan Funke. Local shape descriptors for neuron segmentation. *Nature Methods*, pages 1–9, 2022.
- [Tang et al., 2022] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.
- [Tong et al., 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.
- [Turaga et al., 2010] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H Sebastian Seung. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural computation*, 22(2):511–538, 2010.
- [Wei et al., 2022] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.
- [You et al., 2022] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237, 2022.
- [Zhang et al., 2019] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugmentation. *arXiv preprint arXiv:1912.11188*, 2019.
- [Zheng et al., 2018] Zhihao Zheng, J Scott Lauritzen, Eric Perlman, Camenzind G Robinson, Matthew Nichols, Daniel Milkie, Omar Torrens, John Price, Corey B Fisher, Nadiya Sharifi, et al. A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell*, 174(3):730–743, 2018.
- [Zhou et al., 2019] Shenglong Zhou, Zhiwei Xiong, Chang Chen, Xuejin Chen, Dong Liu, Yueyi Zhang, Zheng-Jun Zha, and Feng Wu. Fast and accurate electron microscopy image registration with 3d convolution. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 478–486. Springer, 2019.
- [Zhou et al., 2022] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*, 2022.