SCCS: Deep Neural Spectral Clustering for Self-Supervised Subcellular Structure Segmentation

Jimao Jiang[1], **Diya Sun**[2], **Tianbing Wang**[2], **and Yuru Pei**[1*]

[1] School of Intelligence Science and Technology, Key Laboratory of Machine Perception (MOE), State Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China
[2] Institute of Artificial Intelligence, Peking University People's Hospital, Peking University, Beijing 100871, China
peiyuru@cis.pku.edu.cn

Abstract

Subcellular structure segmentation is a fundamental task in biological imaging. Existing self-supervised representation learning combined with classical k-means clustering achieved unsupervised image segmentation, but it was constrained by time-consuming test-time pixel-wise feature extraction and clustering synchronization. This study introduces SCCS, a lightweight graph neural network-based **s**pectral **c**lustering framework for end-to-end sub**c**ellular structure **s**egmentation upon superpixel graphs, greatly relieving the computational complexity in test-time numerical spectral clustering and inter-graph label inconsistency. Specifically, SCCS exploits the self-supervised masked autoencoder for representation learning and the construction of superpixel graphs (spG). Unlike per-graph scalar affinity-based spectral clustering, the proposed SCCS parameterizes the mapping from learned deep spG representations to coordinates in the spectral embedding space and the clustering assignments. The SCCS is optimized under unsupervised eigendecomposition and incremental clustering criteria, which synchronize the intra- and inter-graph spectral clustering. The proposed approach is evaluated on a publicly available volumetric electron microscopy dataset. Experiments demonstrate the effectiveness and performance gains of the proposed SCCS over the state-of-the-art in discovering a variety of subcellular structures.

Introduction

Cellular and subcellular structure segmentation plays a critical role in a variety of biological downstream tasks, including biological organization, morphology measurement, and function analysis (Stephens and Allan 2003; Witvliet et al. 2021). With the advent of high-throughput imaging technology, larger-scale microscopic image dataset collections have become efficient and economic (Peddie et al. 2022; Xu et al. 2017). Volumetric electron microscopy (VEM) enables efficient volumetric imaging of the underlying ultrastructures of cells with superior resolutions and scales (Peddie et al. 2022). Manual segmentation involves a tedious annotation burden and reliance on expert knowledge. The deep neural network (DNN)-based segmentation framework has shown superior performances in computer vision and image processing scenarios, where neural networks parameterize the mapping from images to segmentation maps. However, labor-intensive annotations cause scanty annotated data, impeding the generalization capacity of the supervised segmentation models (Heinrich et al. 2021; Mekuc et al. 2020). Additional domain adaptations are required when confronted with novel datasets and cases.

To relieve labor-intensive annotations, self-supervised representation learning and clustering have been applied to discovering pixel-level labels (Wang et al. 2022b, 2023; Melas-Kyriazi et al. 2022). Considering that adjacent pixels with similar feature embedding are likely to belong to the same category of subcellular structures, the local pixel-wise correlations need to be addressed in self-supervised segmentation. Moreover, the global correlation handles the co-occurring subcellular structures for feature enhancement and multi-class segmentation. Transformer-based self-supervised feature learning has shown promising outcomes in spatial correlation modeling of contextual patches in a variety of image processing tasks (Dosovitskiy et al. 2020; Liu et al. 2021; Vaswani et al. 2017). Deep representation learning and pre-trained models, such as DINO and MAE, have been used in graph construction and clustering (Melas-Kyriazi et al. 2022; Xie et al. 2023). However, when confronted with subcellular segmentation of biological VEM images with a large number of fine-grained structures, we need to address two challenging issues.

First, subcellular structures in the high-resolution VEM slices take on a large variety of morphologies, where pixel features have critical influences on clustering and structure segmentation. Existing work used the self-supervised feature extractor to acquire patch representations (Xie et al. 2023). Considering the low granularity of subcellular structures, the patches need to be small enough to bear homogeneous appearances, resulting in a large number of tokens and computationally expensive test-time feature extraction. The situation is even worse when it comes to extracting pixel-level features from VEM images. Economic feature extraction is desirable for efficient subcellular segmentation. Second, considering the random initialization of the clustering centroid, there is no guarantee for consistent inter-image clustering. Additional synchronization, such as Hungarian matching, is required for consistent clustering across images (Melas-Kyriazi et al. 2022). Furthermore, as to spectral clus-

*Corresponding author.

tering, the additional affine transformation has been used to discover consistent spectral embedding across images (Streicher, Cohen, and Gilboa 2022). It is desirable to construct an intra- and inter-image consistent spectral embedding and clustering.

To tackle these challenges, we propose a novel subcellular segmentation framework, SCCS, consisting of self-supervised representation learning and an end-to-end deep neural spectral clustering model for consistent label assignments. To address time-economic feature extraction and label assignments, we construct a superpixel graph (spG) via masked autoencoder (MAE)-based image representation learning and superpixel decomposition. We formulate the subcellular segmentation as a cut over the spG and relieve the test-time cost of pixel-wise feature extraction. Moreover, we build a lightweight GNN for spectral embedding and clustering to ameliorate the numerical eigendecomposition and iterative label assignments. To address the inconsistency labeling across images, we learn the neural spectral clustering under the unsupervised spectral decomposition of graph Laplacian and incremental k-means clustering criteria, avoiding additional clustering synchronization.

We have evaluated the proposed SCCS on dominant subcellular structure segmentation from primary mouse pancreatic islets β cells of the BetaSeg dataset (Heinrich et al. 2021; Müller et al. 2021). Experiments demonstrate the efficiency and performance gains over existing benchmarks. The contributions of this work are as follows:

- We propose SCCS, using learnable spectral clustering to acquire a more consistent understanding of subcellular structures.
- We design an incremental clustering learning framework that capitalizes on the anchor clustering centroid and constructs inter-graph consistent cluster assignments, avoiding post-processing synchronization.
- We demonstrate that our SCCS achieves state-of-the-art performance on self-supervised subcellular structure segmentation from VEM images in comprehensive experiments.

Related Work

Unsupervised Segmentation. Clustering-based methods, such as STEGO (Hamilton et al. 2022) and PiCIE (Cho et al. 2021), achieved self-supervised semantic segmentation. Open-vocabulary semantic segmentation methods of the LSeg (Li et al. 2022) and GroupViT (Xu et al. 2022) adapted CLIP (Radford et al. 2021), which realized zero-shot open-vocabulary semantic segmentation. The pre-trained models, such as DINO (Caron et al. 2021) and Stable Diffusion (Rombach et al. 2021), have been used for unsupervised segmentation and correspondence (Li, Shakhnarovich, and Yeh 2022; Lis et al. 2022; Zhou, Loy, and Dai 2021; Dombrowski et al. 2022). The heuristically designed decoder of the DINO features has been used for instance segmentation (Wang et al. 2022a,b). Anchors defined by text embedding or manual selection enabled dense correspondence via the stable diffusion features (Tang et al. 2023; Rombach et al. 2021; Hedlin et al. 2023). The learned neural features were used to make affinity graphs and spectral clustering (Wang et al. 2022b, 2023; Melas-Kyriazi et al. 2022). Considering the multi-channel features of a neural network, the graph construction relied on feature channel selection. Zhang et al. (Zhang, Yunis, and Maire 2023) addressed the correlation of affinity matrix across neural network layers, which used a gradient descent-based method to solve the eigendecomposition of the joint Laplacian matrix.

When it comes to biomedical image segmentation, self-supervised segmentation of biological images takes advantage of image registration (Liu, Avilés-Rivero, and Schonlieb 2020) and patch-level image clustering to assign pixel labels, greatly relieving the manual annotation burden. Momentum Contrast (MoCo) employed the unsupervised contrastive learning, which used a dynamic dictionary with a queue and a moving-averaged encoder (He et al. 2019). CLMorph utilized unsupervised feature contrastive registration learning for medical image segmentation (Liu, Avilés-Rivero, and Schonlieb 2020). Joint unsupervised learning (JULE) benefited from deep representations and image clusters (Yang, Parikh, and Batra 2016), where deep representation learning and clustering were reciprocal. Moriya et al. (Moriya et al. 2018) used JULE for unsupervised image segmentation, which alternately conducted clustering and refined CNN-based feature extraction using cluster labels. The k-means was applied to the learned representation for medical image segmentation. Han et al. (Han et al. 2022) used variational auto-encoder and metric learning for voxel-level representation and unsupervised cellular segmentation. MAESTER (Xie et al. 2023) utilized the MAE for pixel-level feature extraction, where k-means clustering is used to assign pixel labels. However, when it comes to subcellular images with large amounts of fine-grained structures, there is a computationally expensive test-time feature extraction and a large-scale clustering problem. Patch partitioning is feasible to reduce the clustering complexity, but it requires additional synchronization to ensure consistent cluster assignments across images.

Representation Learning. Transformer has been a mainstream neural network architecture, where a stacking of attention blocks and multi-layer perceptions (MLP) (Vaswani et al. 2017) are used to model long-range attention. Transformer was initiated in the field of natural language processing. The masked language modeling (MLM) used random masking and self-supervised learning for the word prediction (Devlin et al. 2019). Vision transformer (ViT) (Dosovitskiy et al. 2020; Liu et al. 2021) embeds image patch sequences similar to words in a sentence. The MAE has demonstrated potential in self-supervised representation learning (He et al. 2021). The MAE combined with the token embedding has been used in biological image processing, where the semantically relevant tokens of image patches have been used for image clustering and classification (Xie et al. 2023). As to image segmentation, the MAE-based representation enables feature learning of small patches with context information from large view fields. In this work, we utilize the MAE-based feature extractor. Unlike existing work that relied on test-time pixel-level feature extraction with a large computational complexity, we introduce a com-

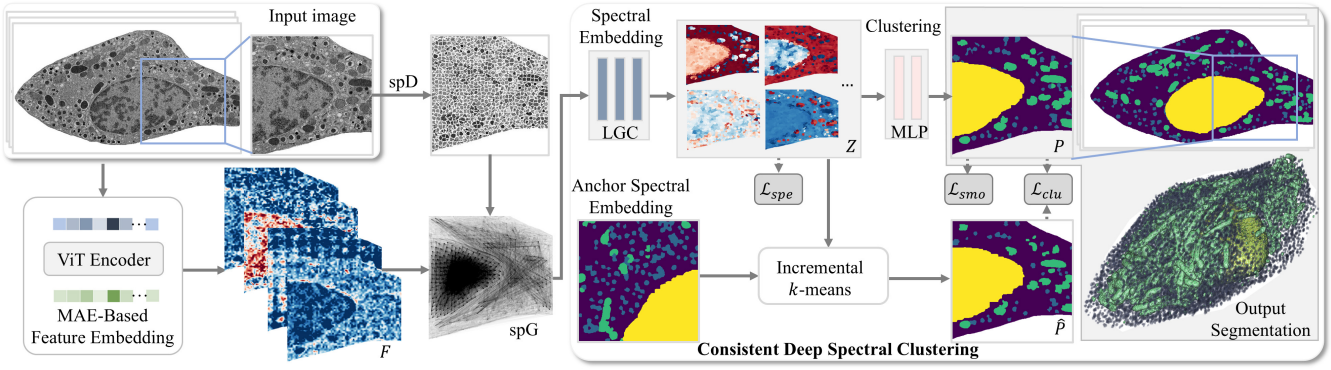

Figure 1: Pipeline overview of our SCCS. We consider subcellular segmentation of VEM images as a deep cut of the superpixel graph (spG) via SLIC-based superpixel decomposition (spD). We compute superpixel-wise features and construct spG with pre-trained transformer-based features F. We construct our inter-image consistent deep neural spectral clustering model by minimizing loss functions $\mathcal{L}_{spe}$ and $\mathcal{L}_{clu}$ to ensure consistent spectral embedding and clustering when given incremental k-means clustering $\hat{P}$. We impose the smoothness regularization $\mathcal{L}_{smo}$ on clustering assignments. The linear graph convolutional network (LGC) is used to approximate spectral bases Z, which are fed to an MLP to produce the final cluster assignment P for the end-to-end semantic segmentation inferences about subcellular structures.

pact image representation of the spG and a neural spectral clustering model for efficient subcellular segmentation.

Method

We aim to find semantic segmentation of multi-class subcellular structures from VEM images $I \in \mathbb{R}^{m\times n}$. Our main idea is to formulate the subcellular segmentation as cuts of spG via a learnable deep neural spectral clustering model, as shown in Figure 1. With that, the spG constructed via transformer-based representation learning is fed to a shallow GNN for end-to-end inference of clustering assignment $P \in \mathbb{R}^{n_s\times k}$ and subcellular segmentation while ensuring that 1) intra-image label consistency for spatially co-occurring subcellular structures is maintained, and 2) inter-image label consistency is retained with the congruent cluster label for the same category of subcellular structures across images. n_s and k denote the spG node number and the cluster number, respectively. In a self-supervised manner, we optimize SCCS using combinational criteria for graph Laplacian eigendecomposition and incremental k-means clustering.

In the following subsections, we first discuss transformer-based representation learning for the construction of the spG specific to VEM images, followed by the presentation of our neural spectral clustering-based subcellular segmentation.

Feature Extraction and spG Construction

We employ the MAE (He et al. 2021) for self-supervised deep representation learning. Given a VEM image with a resolution of $m \times n$, the MAE divides the image into mn/p^2 patches and uses a predefined mask ratio for patch sampling. The ViT encoder receives the unmasked patches and uses embedding to reconstruct the entire image. We up-sample the outcome of the optimized ViT-based encoder as the q-channel image features $F \in \mathbb{R}^{m\times n\times q}$ for downstream segmentation tasks.

spG Construction. Considering the small size of subcellular structures, existing patch-based image clustering methods require patches to be small enough to be inside a specific structure. Moreover, in order to conduct the pixel-level clustering and classification, the MAE-based feature extraction needs to be conducted on patches around each pixel, with great test-time computational complexities (Xie et al. 2023). Unlike time-consuming pixel feature extraction, we conduct SLIC (Achanta et al. 2012)-based superpixel decomposition and construct a weighted spG $\mathcal{G}(\mathcal{S}, \mathcal{E})$ using the pre-trained feature extractor. $\mathcal{S}$ denotes the superpixel node set and $\mathcal{E}$ superpixel-wise connections.

We construct the weighted spG by computing the affinity matrix using q channel pre-trained MAE features. We use the $2q$-channel supervoxel feature F_s as a concatenation of the mean and standard deviation of pixels belonging to a superpixel s. The affinity matrix A is an $n_s \times n_s$ matrix with entries $a_{ij} = \exp \frac{-\|F_{s,i}-F_{s,j}\|_2^2}{\kappa}$ via the RBF kernel, $1 \le i, j \le n_s$. κ denotes the variance of the kernel. Considering the downstream graph embedding and clustering, we modify the affinity matrix with a reduced band width as follows:

$$A \leftarrow \max\left(A - \frac{\max A}{\alpha}, 0\right). \tag{1}$$

The hyperparameter $\alpha > 1$ is related to the repulsion forces to suppress node-wise weak connections (Aflalo et al. 2022). The large value of α tends to produce a low number of clusters. We further remove the negative correlation entries for a sparsified affinity matrix.

Deep Neural Spectral Clustering

We formulate the subcellular segmentation as graph node partition on spG, where spG nodes are divided into k disjoint sets $S_i|_{1\le i\le k}$. $\bigcup_i S_i = \mathcal{S}$, and $S_i \cap S_j = \emptyset$. The node partition can be represented by a $n_s \times k$ dimensional clustering assignment matrix P, where the entry $P_{ij} = 1$ when

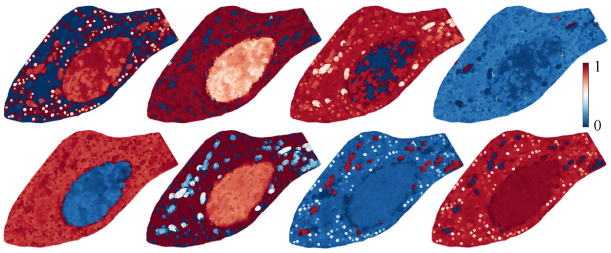

Figure 2: Sampled approximated spectral bases of the graph Laplacian matrix regarding the spG.

node i belongs to the j-th cluster and 0 otherwise. Classical spectral clustering performs the graph partition via eigendecomposition of the graph Laplacian matrix and k-means clustering on the selected spectral bases. Specifically, the graph Laplacian matrix $L = D - A$, where D is a diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$. The u-dimensional embedding $Z \in \mathbb{R}^{n_s \times u}$ is the solution of the eigendecomposition of the graph Laplacian matrix, and $LZ = Z\Lambda$. Λ denotes the diagonal matrix with eigenvalues. The k-means clustering is applied to Z for graph partition.

Existing methods used self-supervised deep feature representation learning to construct the affinity matrix for image segmentation (Wang et al. 2023, 2022b; Melas-Kyriazi et al. 2022). The affinity matrix construction relies on node feature embedding and affects test-time per-graph eigendecomposition. The naive numerical eigendecomposition is known to have a computational complexity of $O(n_s^3)$, which is computationally burdensome for a large graph. Moreover, considering the repetitive occurrence of various categories of subcellular structures across images, we need to address inter-image consistent spectral embedding and clustering. In this work, we present a neural spectral clustering model for consistent cluster assignments.

Neural Spectral Embedding Instead of per-graph eigendecomposition, we leverage a lightweight GNN to parameterize the eigendecomposition of the normalized graph Laplacian matrix. The neural spectral embedding model takes both the node features and the normalized affinity matrix as input, where the l-th linear graph convolution (LGC) (He et al. 2020) updates $q^{(l)}$-channel node embedding $Z^{(l)} \in \mathbb{R}^{n_s \times q^{(l)}}$ with learnable weights $W^{(l)}$. $Z^{(l+1)} = D^{-1/2}AD^{-1/2}Z^{(l)}W^{(l)}$, where Z is initialized as the spG node feature F_s, and $Z^{(0)} = F_s$. Since the graph convolution with the Chebyshev approximation polynomial is a power multiplication of the normalized affinity matrix by removing the nonlinear activation function and the weight, the GCN-based model has been used to discover a set of eigenfunctions (Sun et al. 2022). We define the spectral embedding loss $\mathcal{L}_{spe}$ based on the eigendecomposition criteria of graph Laplacian matrices.

$$\mathcal{L}_{spe} = \sum_{i \neq j} \left[(z_i^t z_j) + \mu (z_i^t L z_j)^2 \right]. \tag{2}$$

z_i and z_j denotes the i-th and j-th column vectors of Z. The first term is used to guarantee the orthogonality of the approximated spectral bases Z. The second term is used to minimize the off-diagonal entry of matrix $Z^T LZ$, considering the diagonalization transform of the graph Laplacian matrix satisfies $\Lambda = Z^T LZ$. Unlike per-graph numerical eigendecomposition, the learned neural spectral embedding module parameterizes the mapping from the MAE-based node feature to its spectral coordinates in the embedding space with knowledge of node-wise correlations. Figure 2 shows sampled spectral bases of a spG. It is interesting to note that spectral bases of different frequencies have activations with respect to various subcellular structures.

Neural Clustering Existing clustering GNN is optimized using N-Cut or the correlation clustering loss for training (Aflalo et al. 2022). Considering the large scale and resolution of biological images, we subdivide the images into smaller FOVs and conduct a graph cut on the spG. We expect to assign the nodes belonging to the same category of subcellular structures across the FOVs to the same cluster. However, the per-graph evaluation of the clustering loss does not handle the cross-graph consistent cluster assignments. We propose to optimize the neural clustering module under the supervision of incremental k-means clustering, which enables us to correct cross-graph inconsistent clustering assignments. The main steps are as follows:

We first define anchor clustering assignments. Considering the subcellular categories of interest, we select one reference spG and conduct k-means clustering to initialize the anchor clustering assignments with cluster centroid $C = \{c_1, \ldots, c_k\}$. Each cluster is associated with a specific category of subcellular structures, and such associations are retained in the following batches for model training.

We secondly incrementally update clustering in the new batch. We draw nodes from spGs in the new batch and assign node s to the nearest center c^* when $c^* = \arg\min_{c \in C} \|z(s) - c\|_2^2$. $z(s)$ denotes the spectral coordinate. We update the node assignment matrix with entry $\hat{P}_{ij} = 1$ when superpixel s_i is assigned to the j-th cluster. The cluster centroids are updated with newly added graph nodes, and $c_i = 1/n_i \sum_{s \in S_i} z(s)$, where n_i denotes the node number in the i-th cluster. We simplify the incremental clustering by fixing the cluster number and batch-updating the cluster centers. Unlike incremental k-means to update the centroid with each newly added node, we conduct centroid updates once per eight spGs.

We thirdly introduce supervision to neural clustering learning. We use the MLP-based clustering module to assign clustering labels. Unlike the per-graph evaluation of clustering criteria, we employ the incremental clustering assignment $\hat{P}$ as the supervision. The binary cross-entropy-based clustering loss $\mathcal{L}_{clu}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{clu} = -\frac{1}{n_s \cdot k} \sum_{1 \leq i \leq n_s, 1 \leq j \leq k} & [\hat{P}_{i,j} \log P_{i,j} \\ & + (1 - \hat{P}_{i,j}) \log(1 - P_{i,j})], \end{aligned} \tag{3}$$

where $\hat{P}_{i,j}$ and $P_{i,j}$ denote the ij-the entry of cluster assignment matrices obtained by the incremental k-means and the neural clustering module. In order to enhance consistent

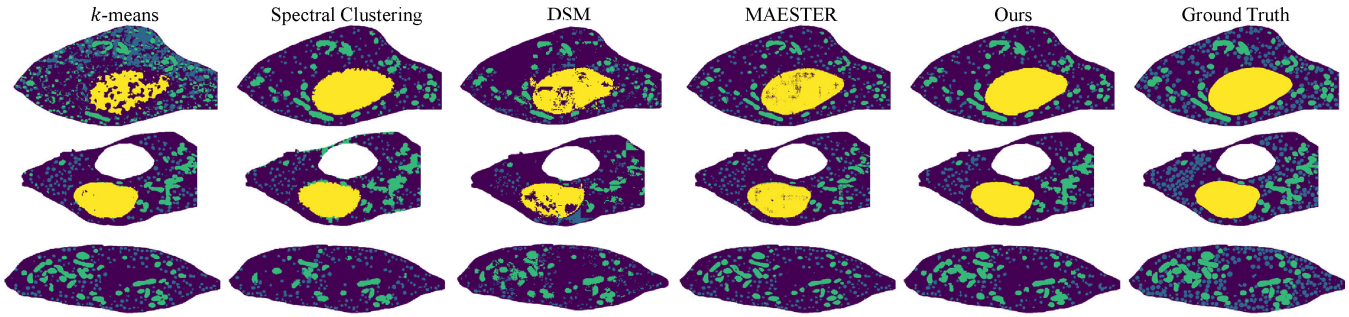

Figure 3: Subcellular structure segmentation by the proposed approach and the compared state-of-the-art, including classical k-means clustering, spectral clustering (Shi and Malik 1997) on pre-trained MAE features and the spG graphs, DSM (Melas-Kyriazi et al. 2022), and MAESTER (Xie et al. 2023). (yellow: nucleus, green: mitochondria, blue: granules)

label assignments, we impose smoothness regularization on clustering labels.

$$\mathcal{L}_{smo} = \frac{1}{n_s^2} \sum_{1 \le i,j \le n_s} (a_{ij} - \eta)\delta(P_i, P_j) + \nu \left| \frac{\sum_i \|P_i\|_1}{n_s} - 1 \right|, \tag{4}$$

where $\delta(p_i, p_j) = 1 - \frac{P_i \cdot P_j}{\|P_i\| \|P_j\|}$, and P_i and P_j denote the i-th and j-th row vectors of matrix P. The scalar value η is used as a similarity threshold. Superpixel pairs with similarity above η are penalized when they are assigned different clustering labels. Conversely, we encourage the superpixel pair with similarity below η to have different labels. The second term is to avoid a trivial solution by requiring the sum of cluster assignment probabilities equal to the node number n_s of the spG. By minimizing $\mathcal{L}_{smo}$, similar superpixels are encouraged to bear the same labels and to promote consistent labeling of subcellular structures. Note that, by employing the proposed neural clustering, the cluster assignment with respect to various subcellular structures is consistent across non-overlapped FOVs, which relieves test-time clustering corrections and synchronization.

Loss The final loss function is defined as a linear combination of the neural spectral embedding and regularized clustering losses.

$$\mathcal{L} = \mathcal{L}_{spe} + \gamma_1 \mathcal{L}_{clu} + \gamma_2 \mathcal{L}_{smo}. \tag{5}$$

Hyperparameters γ_1 and γ_2 are used to trade off the spectral embedding and the regularized clustering in the spectral embedding space.

Experiments

Experimental Setup

Dataset and Metric. To evaluate the efficacy of the proposed method for self-supervised subcellular segmentation, we conduct experiments on standard benchmarks. The approach is trained and evaluated on the primary mouse pancreatic islet β cell dataset named BetaSeg proposed in OpenOrganelle (Heinrich et al. 2021; Müller et al. 2021). The VEM dataset of two pancreatic tissue samples are captured by focused ion beam scanning electron microscopy, consisting of four cell volumes. We compare SOTA methods with the high-dosage group as (Xie et al. 2023). In the preprocessing, we perform cell cropping from the tissue stack into separated volumes. The cell volumes are associated with segmentation maps of subcellular structures by manual annotation or the deep learning models (Müller et al. 2021). In this work, we consider four dominant categories, i.e., nucleus, granules, mitochondria, and the unrecognized, as (Xie et al. 2023). We use the first three cell volumes for training and the remaining volume for testing.

We evaluate all subcellular segmentation using the Dice similarity coefficient (DSC), Intersection over Union (IoU), Sensitivity, and Accuracy, which allow us to measure the consistency between the estimated structural segmentation and the ground truth.

Implemental Details. To reduce the computational complexity and spGs with similar numbers of nodes, we divide the VEM slice into non-overlapping regions with a resolution of 560×560 and a physical size of 5.02 micrometers squared area. We conduct superpixel decomposition using the SLIC algorithm, with the compactness parameter set to 0.2. Each superpixel has an average of 100 pixels. There are approximately 3000 superpixels in each image. As to the MAE feature extractor, the input image region has a resolution of 80×80, which is divided into 5×5 patches. As (Xie et al. 2023), we use a weight of 0.08 to reduce the effectiveness of positional encoding for feature extraction. We set the masking ratio to 0.5. The ViT encoder is composed of 14 transformer layers and one attention head. The MAE-based features channel number q is set to 192. κ in the RBF kernel-based affinity computation is set to 2. We retain $u = 12$ approximated spectral bases. We set the cluster number k to 8. Hyperparameter α in affinity matrix computation is set to 4. The hyperparameter μ in the spectral embedding loss $\mathcal{L}_{spe}$ is set to 1. The scalar threshold η and coefficient ν in $\mathcal{L}_{smo}$ are both set to 0.1. The hyperparameters γ_1 and γ_2 in the loss function $\mathcal{L}$ are set to 1 and 0.1 to balance the criteria of spectral embedding and the regularized clustering.

The LGC-based spectral embedding module consists of three linear graph convolutional layers with 384×96, 96×48, and 48×12 weight matrices. The MLP-based clustering modules consist of two fully connected layers with 12×12

	k-means	DSM	DeepCut	Han et al.	MAESTER	w/o NSC	w/o NC$_{hm}$	w/o NC$_{inc}$	SCCS
nucleus	0.803	0.848	0.861	-	0.943	0.903	0.740	0.945	**0.954**
granules	0.571	0.345	0.333	-	0.556	0.365	0.485	0.500	**0.573**
mitochondria	0.679	0.742	0.694	-	0.778	0.769	0.664	0.785	**0.861**
unrecognized	0.858	0.817	0.778	-	0.868	0.848	0.802	0.858	**0.875**
Average	0.728	0.688	0.666	0.659	0.786	0.721	0.673	0.772	**0.816**

Table 1: Subcellular segmentation accuracy regarding the DSC by compared methods, including k-means, DSM (Melas-Kyriazi et al. 2022), DeepCut (Aflalo et al. 2022), Han et al. (Han et al. 2022), MAESTER (Xie et al. 2023), and variants of the proposed SCCS without using neural spectral clustering (NSC) or neural clustering (NC).

and 12×8 weight matrices. We use the Adam optimizer with a momentum of 0.9 and 0.999. For training the MAE-based feature extractor, we use a learning rate of 1e-5 and a batch size of 32. We set the learning rate to 0.01 with a batch size of 1 for training the neural spectral clustering model. The training is performed on a PC with an NVIDIA RTX 2080Ti GPU, consuming 6 hours with 6,000 iterations. In the online testing process, the feature extraction, spG construction, and neural spectral clustering of a 560×560 image take 1.997 seconds, 1.129 seconds, and 0.003 seconds, respectively.

Experimental Results

Subcellular Segmentation. We summarize our main results on four categories of subcellular structures in Figure 3, Table 1, and Appendix Table 1. SCCS outperforms the compared state-of-the-art, MAESTER (Xie et al. 2023), on all reported categories of structures. In particular, SCCS improves by 0.011 on the nucleus, 0.017 on the granules, 0.083 on the mitochondria, and 0.007 on the unrecognized category regarding the DSC. Note that we only need to extract features for the spG instead of time-consuming patch feature extraction for each pixel (Xie et al. 2023), where our approach is more efficient in test-time feature extraction.

We compare with classical k-means clustering on pre-trained MAE features, where our approach shows a performance gain of 0.088 regarding the DSC. We compare patch clustering on VAE features (Han et al. 2022), where the Mini-Batch k-means was used to conduct large scale clustering. Our approach shows superior clustering performance, with a DSC of 0.816 vs. 0.659 (Han et al. 2022). Unlike the per-graph solution of clustering on the patch graph using pre-trained feature extractors (Melas-Kyriazi et al. 2022; Aflalo et al. 2022), our approach learns neural spectral clustering with consistent inter-graph cluster assignments. We observe performance gains of 0.128 over DSM (Melas-Kyriazi et al. 2022) and 0.150 over DeepCut (Aflalo et al. 2022). Moreover, our SCCS achieves competitive performances with supervised segmentation models (Strudel et al. 2021; Dosovitskiy et al. 2020) (Appendix Table 2).

The proposed SCCS enables self-supervised multi-class subcellular structure segmentation, as shown in Figure 4. We note that SCCS is feasible for spectral embedding and clustering on the spG. The approximated spectral bases of the graph Laplacian matrices have been found to be activated similarly across images regarding subcellular structures. This is because the neural spectral embedding makes it possible to assign similar spectral coordinates to nodes of same category of structures. Moreover, the learned clustering model generates consistent label assignments, where the same category of subcellular structures across images bear congruent labels. We note that SCCS is feasible for spectral clustering on the spG even when confronted noisy images as shown in Appendix Table 3.

Ablation Study

To investigate the necessity of each module in SCCS, we conducted a comparison of SCCS and its variants: (1) SCCS (w/o NC$_{hm}$) that removes the neural clustering module and uses per-graph k-means clustering and Hungarian matching with the ground truth, (2) SCCS (w/o NC$_{inc}$) that removes the neural clustering module and uses incremental k-means, (3) SCCS (w/o NSC) that removes neural spectral clustering modules and conducts numerical spectral clustering as (Melas-Kyriazi et al. 2022). We compare with all variant models using the same MAE-based features and spG graph as ours (Table 1). The experiments demonstrate that even when we use the same refined features and spG graphs, the proposed SCCS outperforms the classical clustering, w/o NC$_{hm}$ and w/o NC$_{inc}$, by 0.143 and 0.044. Moreover, our approach outperforms the numerical spectral clustering (w/o NSC) by a DSC of 0.095. This demonstrates the benefits of learning a neural spectral embedding and clustering module to distill consistent inter-image spectral clustering assignments. Moreover, we provide ablation study of the smoothness regularization (Appendix Figure 1), where the smoothness regularization is feasible to remove inconsistent labeling, especially around structural boundaries.

Parameter Analysis

Cluster Number. Table 2 reports the effectiveness of the cluster number in subcellular segmentation. We note that the performance reaches a local maximum when the cluster number is set to 8. In unsupervised clustering, large or small cluster numbers show limitations. For instance, nucleus segmentation deteriorates when k is set to 6, where the morphology of the nucleus cannot be captured. On the other hand, granule segmentation accuracy decreases with k of 9, where a large number of repetitively occurring fine-grained granules cannot be assigned to the same cluster.

Superpixel Granularity. Unlike the patch-based methods (Xie et al. 2023; Melas-Kyriazi et al. 2022), the spG is more

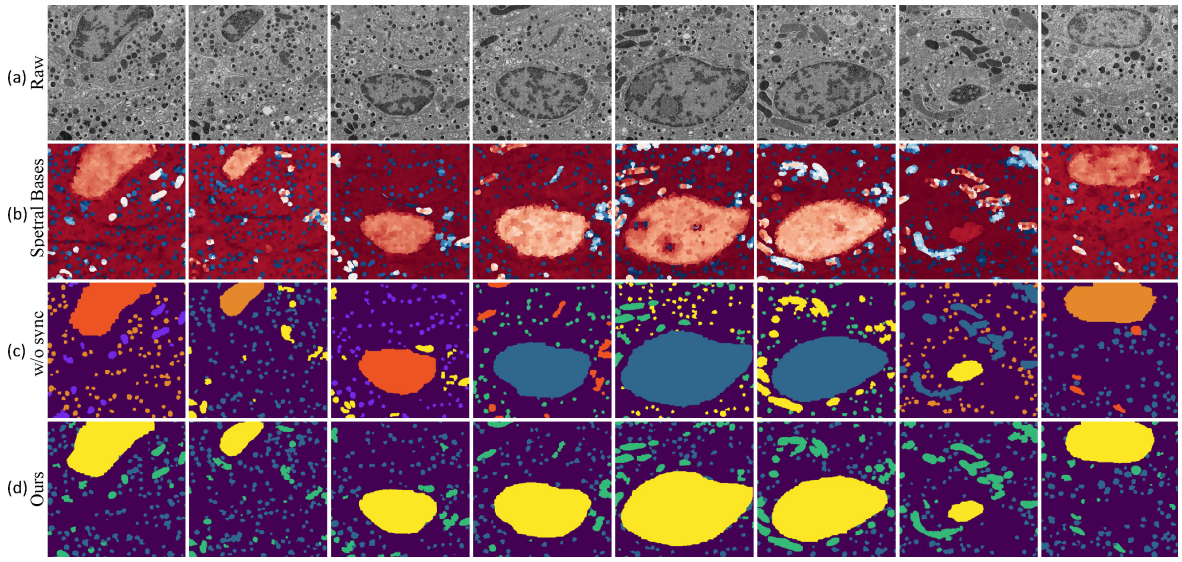

Figure 4: Subcellular segmentation by the proposed SCCS. (a) Input raw images. (b) The approximated spectral bases produced by the proposed neural spectral embedding module. (c) k-means clustering without synchronization. (d) Our approach (yellow: nucleus, green: mitochondria, blue: granules).

k	6	7	8	9
nucleus	0.047	0.934	**0.954**	0.926
granules	0.521	0.521	**0.573**	0.477
mitochondria	**0.876**	0.866	0.861	0.781
unrecognized	0.794	0.871	**0.875**	0.859
Average	0.559	0.798	**0.816**	0.761

Table 2: The DSC when using different cluster number k.

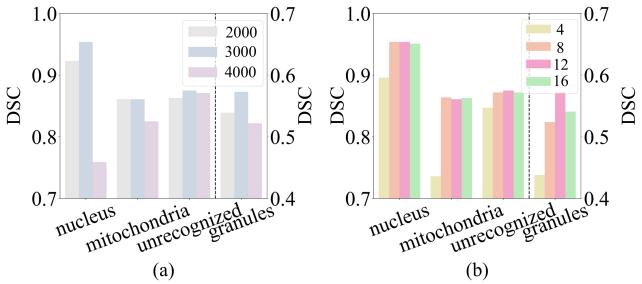

Figure 5: The DSC when using different (a) superpixel number n_s and (b) spectral basis number u.

flexible to represent variable morphologies of the subcellular structures than regular patches. Figure 5 (a) reports segmentation accuracy with superpixel numbers ranging from 2000 to 4000. We note that a local maximum DSC occurs at 3000 superpixels. The spG is feasible to control the graph size and provide a compact representation of the VEM image. In contrast, the patch graph requires small enough patches to fit in the subcellular structures. The graph node number of spG is 6.53 times smaller than the patch graph of DSM (Melas-Kyriazi et al. 2022), allowing for efficient test-time feature extraction and clustering. Moreover, our method is 137.61 times faster than MAESTRER (Xie et al. 2023) which requires time-consuming pixel-level feature extraction.

Spectral Basis Number. We evaluate the effectiveness of the spectral basis number, as shown in Figure 5 (b). Considering the immune nature of the spectral embedding to high-frequency perturbations, we conduct clustering in the low-dimensional spectral embedding space. We observe slight performance variations on most structures when the spectral basis number u varies from 8 to 16. However, the performance decreases greatly when u is set to 4. Moreover, the granule segmentation accuracy is significantly higher when u is set to 12 than others. We think enough spectral bases are required to cover the variations of fine-grained structures, while a large number of spectral bases entails the risk of high-frequency perturbations.

Conclusion

We have presented a novel deep neural spectral clustering model for self-supervised and consistent subcellular structure segmentation. Our key insight is to learn a neural spectral embedding and clustering model on superpixel graphs. This allows for the use of the pre-trained model for efficient test-time feature extraction and label assignments with inter-graph consistency. Specifically, our work is inspired by recent advancements in deep clustering-based image segmentation (Xie et al. 2023; Melas-Kyriazi et al. 2022). We augment these methods with learnable modules that discover spectral clustering, allowing for efficient and consistent labeling of subcellular structures across images and alleviating the need for additional cluster synchronization. Experiments demonstrate that our approach achieves state-of-the-art subcellular structure segmentation accuracy.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62272011, 61876008, Beijing Natural Science Foundation 7232337.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34: 2274–2282.

Aflalo, A.; Bagon, S.; Kashti, T.; and Eldar, Y. C. 2022. DeepCut: Unsupervised Segmentation using Graph Neural Networks Clustering. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 32–41.

Caron, M.; Touvron, H.; Misra, I.; J'egou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640.

Cho, J. H.; Mall, U.; Bala, K.; and Hariharan, B. 2021. PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16789–16799.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dombrowski, M.; Reynaud, H.; Baugh, M.; and Kainz, B. 2022. Foreground-Background Separation through Concept Distillation from Generative Image Foundation Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 988–998.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. *ArXiv*, abs/2203.08414.

Han, H.; Dmitrieva, M.; Sauer, A.; Tam, K. H.; and Rittscher, J. 2022. Self-Supervised Voxel-Level Representation Rediscovers Subcellular Structures in Volume Electron Microscopy. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1873–1882.

He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; and Girshick, R. B. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hedlin, E.; Sharma, G.; Mahajan, S.; Isack, H. N.; Kar, A.; Tagliasacchi, A.; and Yi, K. M. 2023. Unsupervised Semantic Correspondence Using Stable Diffusion. *ArXiv*, abs/2305.15581.

Heinrich, L.; Bennett, D.; Ackerman, D.; Park, W.; Bogovic, J. A.; Eckstein, N.; Petruncio, A.; Clements, J.; Pang, S.; Xu, S.; Funke, J.; Korff, W. L.; Hess, H. F.; Lippincott-Schwartz, J.; Saalfeld, S.; Weigel, A. V.; Team, P.; Ali, R.; Arruda, R.; Bahtra, R.; and Nguyen, D. 2021. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599: 141 – 146.

Li, B.; Weinberger, K. Q.; Belongie, S. J.; Koltun, V.; and Ranftl, R. 2022. Language-driven Semantic Segmentation. *ArXiv*, abs/2201.03546.

Li, J.; Shakhnarovich, G.; and Yeh, R. A. 2022. Adapting CLIP For Phrase Localization Without Further Training. *ArXiv*, abs/2204.03647.

Lis, K.; Rottmann, M.; Honari, S.; Fua, P.; and Salzmann, M. 2022. AttEntropy: Segmenting Unknown Objects in Complex Scenes using the Spatial Attention Entropy of Semantic Segmentation Transformers. *ArXiv*, abs/2212.14397.

Liu, L.; Avilés-Rivero, A. I.; and Schonlieb, C.-B. 2020. Contrastive Registration for Unsupervised Medical Image Segmentation. *IEEE transactions on neural networks and learning systems*, PP.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Mekuc, M. Z.; Bohak, C.; Hudoklin, S.; Kim, B. H.; Romih, R.; Kim, M. Y.; and Marolt, M. 2020. Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data. *Computers in biology and medicine*, 119: 103693.

Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; and Vedaldi, A. 2022. Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8354–8365.

Moriya, T.; Roth, H. R.; Nakamura, S.; Oda, H.; Nagara, K.; Oda, M.; and Mori, K. 2018. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. In *Medical Imaging*.

Müller, A.; Schmidt, D.; Xu, C. S.; Pang, S.; DCosta, J. V.; Kretschmar, S.; Münster, C.; Kurth, T.; Jug, F.; Weigert, M.; et al. 2021. 3D FIB-SEM reconstruction of microtubule–organelle interaction in whole primary mouse β cells. *Journal of Cell Biology*, 220(2).

Peddie, C. J.; Genoud, C.; Kreshuk, A.; Meechan, K. I.; Micheva, K. D.; Narayan, K.; Pape, C.; Parton, R. G.;

Schieber, N. L.; Schwab, Y.; Titze, B.; Verkade, P.; Weigel, A. V.; and Collinson, L. M. 2022. Volume electron microscopy. *Nature reviews. Methods primers*, 2.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Shi, J.; and Malik, J. 1997. Normalized cuts and image segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 731–737.

Stephens, D. J.; and Allan, V. J. 2003. Light Microscopy Techniques for Live Cell Imaging. *Science*, 300: 82 – 86.

Streicher, O. C.; Cohen, I.; and Gilboa, G. 2022. BASiS: Batch Aligned Spectral Embedding Space. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10396–10405.

Strudel, R.; Pinel, R. G.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for Semantic Segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7242–7252.

Sun, D.; Pei, Y.; Zhang, Y.; Xu, T.; Wang, T.; and yan Zha, H. 2022. Dense correspondence of deformable volumetric images via deep spectral embedding and descriptor learning. *Medical image analysis*, 82: 102604.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent Correspondence from Image Diffusion. *ArXiv*, abs/2306.03881.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Neural Information Processing Systems*.

Wang, X.; Girdhar, R.; Yu, S. X.; and Misra, I. 2023. Cut and Learn for Unsupervised Object Detection and Instance Segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3124–3134.

Wang, X.; Yu, Z.; Mello, S. D.; Kautz, J.; Anandkumar, A.; Shen, C.; and Álvarez, J. M. 2022a. FreeSOLO: Learning to Segment Objects without Annotations. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14156–14166.

Wang, Y.; Shen, X.; Yuan, Y.; Du, Y.; Li, M.; Hu, S. X.; Crowley, J. L.; and Vaufreydaz, D. 2022b. TokenCut: Segmenting Objects in Images and Videos With Self-Supervised Transformer and Normalized Cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 15790–15801.

Witvliet, D. K.; Mulcahy, B.; Mitchell, J. K.; Meirovitch, Y.; Berger, D. R.; Wu, Y.; Liu, Y.; Koh, W. X.; Parvathala, R.; Holmyard, D. P.; Schalek, R. L.; Shavit, N.; Chisholm, A. D.; Lichtman, J. W.; Samuel, A. D. T.; and Zhen, M. 2021. Connectomes across development reveal principles of brain maturation. *Nature*, 596: 257 – 261.

Xie, R.; Pang, K.; Bader, G. D.; and Wang, B. 2023. MAESTER: Masked Autoencoder Guided Segmentation at Pixel Resolution for Accurate, Self-Supervised Subcellular Structure Recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3292–3301.

Xu, C. S.; Hayworth, K. J.; Lu, Z.; Grob, P.; Hassan, A. M.; García-Cerdán, J. G.; Niyogi, K. K.; Nogales, E.; Weinberg, R. J.; and Hess, H. F. 2017. Enhanced FIB-SEM systems for large-volume 3D imaging. *eLife*, 6.

Xu, J.; Mello, S. D.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. GroupViT: Semantic Segmentation Emerges from Text Supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18113–18123.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint Unsupervised Learning of Deep Representations and Image Clusters. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5147–5156.

Zhang, X.; Yunis, D.; and Maire, M. 2023. Deciphering 'What' and 'Where' Visual Pathways from Spectral Clustering of Layer-Distributed Neural Representations. *ArXiv*, abs/2312.06716.

Zhou, C.; Loy, C. C.; and Dai, B. 2021. Extract Free Dense Labels from CLIP. In *European Conference on Computer Vision*.