Unsupervised Multi-Scale Segmentation of 3D Subcellular World with Stable Diffusion Foundation Model

Supplementary Material

S1. Cryo-Electron Tomography (cryo-ET)

Cryo-electron tomography is an emerging 3D electron microscopy imaging technique to visualize subcellular objects across various scales inside the cell in their native context. In cryo-ET, a cellular sample or portion of a cellular sample is imaged with an electron microscope. The sample is tilted up to a certain range at both directions (typically $-60°$ to $+60°$) and an image is captured at each titled position [20]. The tilt series images are then backprojected and reconstructed into a 3D voxel image, which is called a tomogram (S1).

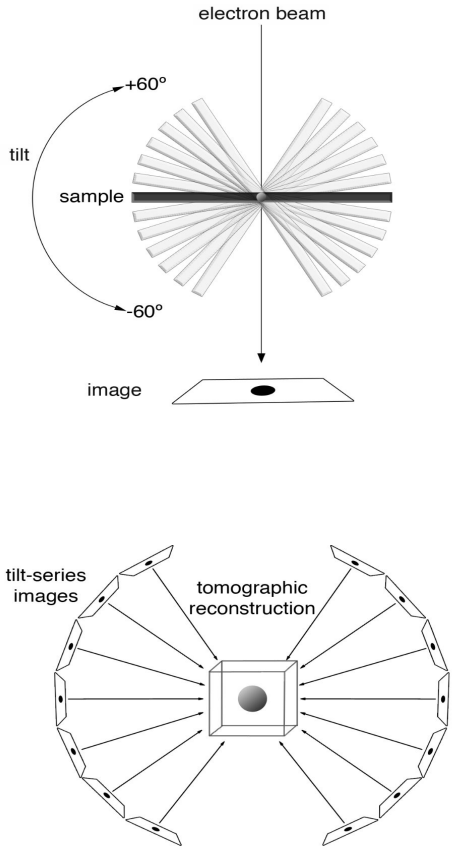

Figure S1. The cryo-ET image acquisition and tomogram reconstruction process. [25]

These tomograms contain *in situ* visualizations of nanometer-scale small macromolecules and micrometer-scale organelles inside a cell and their native spatial organization. However, this unique aspect of tomograms comes with several costs. The tomograms are usually very large (e.g., $4000 \times 6000 \times 1000$ voxels) grayscale volumes and can not be processed as a whole. Even after binning 4 times across each axis, a tomogram is still large (e.g., $1000 \times 1500 \times 250$ voxels). Furthermore, to maintain the native context of the sample specimen, the electron dosage needs to be kept very low. Due to this low electron dose and also because of the complex cytoplasmic environment, the tomograms become very noisy. Due to incomplete tilting, tomograms reconstructed from the tilt series also contain reconstruction-related artifacts [20]- known as missing wedge effect.

Given all these challenges, extracting information of multiscale subcellular objects from the cellular cryogenic tomograms is an extremely difficult process and often requires significant manual efforts. Through automation, this bottleneck can be largely solved, revealing useful insights into the subcellular world that were previously out of reach.

S2. Vision Foundation Models

Currently, vision foundation models (VFM) are emerging in computer vision. VFMs are trained on massive, diverse image datasets. They learn generalized visual representations that can be adapted or fine-tuned for downstream tasks such as classification, segmentation, object detection, and image captioning. The segment anything model (SAM) [7] is a very relevant example that can segment objects directly inside an image with or without a prompt. However, SAM is a supervised segmentation model trained on an extremely large set of annotated images, which does not include cryo-ET images. Hence, we observed that SAM does not perform reasonable segmentation for cryo-ET image slabs. In our work, we instead used self-supervised vision foundation models for feature extraction. Popular examples of such methods include CLIP [16], DINO [2], PLRC [1], and Stable Diffusion [17].

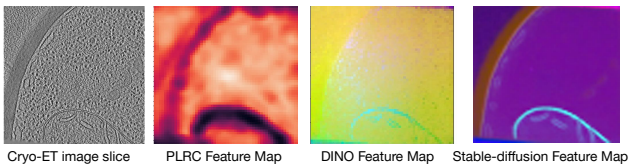

Figure S2. Feature extraction on a single cryo-ET image slab of a *S. Pombe* Tomogram with multiple VFMs

We leveraged several self-supervised vision foundation models (VFMs) to extract features for unsupervised seg-

mentation. In Figure S2, we show extracted features for a sample cryo-ET image slab for multiple VFMs.

S3. Stable Diffusion Foundation Model

The Stable Diffusion model [17] is a state-of-the-art latent text-to-image generative model that excels at generating photorealistic images conditioned on textual descriptions. Developed as a diffusion-based generative framework, Stable Diffusion integrates powerful deep architectures, including a Variational Autoencoder (VAE), a UNet denoising backbone incorporating a Vision Transformer (ViT), and a Transformer-based text encoder. This sophisticated architecture enables robust multimodal representation learning.

Unlike conventional generative models, Stable Diffusion operates within a learned latent space, rather than directly at pixel resolution, which significantly reduces computational complexity while preserving semantic fidelity. It employs a **Denoising Diffusion Probabilistic Model (DDPM)**, progressively transforming Gaussian noise into meaningful latent representations conditioned upon textual input prompts.

The key components of Stable Diffusion include:

- **Variational Autoencoder (VAE):** Encodes high-dimensional images into compact latent representations and subsequently decodes generated latents back into pixel-level images.
- **Text Encoder:** Utilizes a Transformer-based encoder to convert textual inputs into dense embeddings, facilitating conditioning in the diffusion process.
- **UNet Denoiser:** A hierarchical convolutional network combined with Vision Transformer layers, predicting and removing noise in the latent representation space throughout the diffusion steps.
- **Noise Scheduler:** Controls forward (noise addition) and reverse (denoising) processes in latent spaces during training and inference.

In our segmentation pipeline for Cryo-Electron Tomography (Cryo-ET), we leverage the Stable Diffusion model strictly as a frozen feature extractor, without performing any fine-tuning or domain-specific parameter updates. Specifically, we utilize the Vision Transformer backbone within the UNet denoiser to extract dense, high-dimensional features from Cryo-ET images.

Rather than employing Stable Diffusion's generative capacity, we focus on the internal representations obtained from the multi-layer self-attention mechanisms of the ViT backbone. We extract both *query* and *key* embeddings from multiple attention layers, as these embeddings capture rich spatial relationships and structural information essential for accurate unsupervised segmentation.

These embeddings are used to form affinity matrices, characterizing spatial similarity among image patches. Subsequently, spectral clustering is applied to these affinity matrices, allowing eigenvectors to be optimized unsupervisedly to obtain meaningful segmentation masks. By directly using the pre-trained model without modification, our approach efficiently exploits the extensive visual priors embedded in Stable Diffusion, ensuring robust segmentation of complex biological structures in Cryo-ET images.

S4. Self-attention mechanism

The self-attention mechanism is a key component of vision transformers that enables the model to capture long-range dependencies in an image. Each input image is divided into non-overlapping patches of size $p \times p$, forming a set of tokens $\{x_i\}_{i=1}^{N}$, where $N = \frac{HW}{p^2}$ represents the number of patches, where p = 16. These patches are embedded into a higher-dimensional space and passed through the transformer layers, where self-attention operates as follows:

Each patch embedding x_i is linearly transformed into three vectors:

- **Query** (Q_i): Represents the patch's current state in relation to others.
- **Key** (K_i): Represents the identifiable characteristics of a patch.
- **Value** (V_i): Contains the actual feature representation that will be propagated through layers.

S5. Structural Similarity Index Measure (SSIM)

The Structural Similarity Index Measure (SSIM) is an established perceptual metric that quantitatively assesses the visual similarity between two images by modeling human visual perception more effectively than traditional pixel-based metrics, such as Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR). Rather than solely evaluating pixel intensity differences, SSIM quantifies structural information, capturing similarities in luminance, contrast, and structure between image pairs.

Formally, the SSIM between two image patches, denoted as x and y, is defined by:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where:

- μ_x and μ_y represent the mean intensity values of patches x and y, respectively.
- σ_x^2 and σ_y^2 are the corresponding variances, capturing local contrast information within each patch.
- σ_{xy} denotes the covariance between the patches, capturing the joint variations in their structures.
- C_1 and C_2 are small stabilization constants to prevent instability when denominators approach zero, typically set as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where L is the dy-

namic range of pixel intensities, and $K_1, K_2 \ll 1$ (commonly $K_1 = 0.01$, $K_2 = 0.03$).

Within our unsupervised segmentation pipeline tailored specifically for Cryo-Electron Tomography (Cryo-ET) images, the SSIM metric is leveraged during the feature refinement phase, particularly in selecting and aggregating structurally coherent eigenvector-based images (eigenimages).

SSIM is strategically employed to evaluate the structural correspondence between candidate eigenimages and the current aggregated feature (base image). Only eigenimages that exhibit a high SSIM with the base image are considered for integration, alongside additional conditions such as improved diversity score and reduced structural noise. This filtering ensures that the merged eigenimages preserve meaningful structures while suppressing noise and irrelevant features.

S6. UNet training

The training of UNet for membrane segmentation is performed over 30,000 iterations using a batch size of 8. Optimization is performed using stochastic gradient descent (SGD) with an initial learning rate of 0.1, momentum of 0.9, and a weight decay of 1×10^{-5} to prevent overfitting. A StepLR learning rate scheduler is applied to decay the learning rate by a factor of 0.1 every 10,000 iterations. To ensure reproducibility, all experiments are conducted with a fixed random seed.

The UNet architecture implemented here follows the classic encoder–decoder structure, tailored for semantic segmentation tasks. It is composed of two main components: the Encoder, which progressively downsamples the input to capture high-level features, and the Decoder, which upsamples the feature maps to produce a full-resolution segmentation map.

The encoder consists of an initial ConvBlock followed by four DownBlock modules. Each DownBlock applies max pooling to reduce spatial resolution by a factor of two, then passes the result through a ConvBlock consisting of two convolutional layers, batch normalization, LeakyReLU activations, and dropout. The number of feature channels increases at each level, from 16 up to 256, allowing the model to capture increasingly abstract representations. Corresponding dropout rates also increase from 0.05 to 0.5 to provide stronger regularization at deeper layers.

The decoder mirrors this structure, with four UpBlock modules that upsample feature maps using either bilinear interpolation or transposed convolution (in this implementation, transposed convolutions are used). Each UpBlock also performs feature fusion by concatenating the upsampled feature map with the corresponding encoder feature map (skip connection), followed by a ConvBlock. This design enables the decoder to recover spatial detail lost during downsampling. The final segmentation map is produced by a 3×3 convolution that maps the decoder output to the desired number of classes.

S7. DeepETPicker training and inference

DeepETPicker is a deep learning-based method that utilizes a 3D-ResUNet segmentation model with coordinated convolution and multiscale image pyramid inputs to distinguish biological macromolecules from the background in cryo-electron tomograms. The methodology adheres to an established two-stage workflow—training and inference—with identical hyperparameters consistently maintained to ensure methodological rigor. During preprocessing, tomograms undergo coordinate normalization, intensity standardization, and the generation of weak supervision labels—simplified Gaussian-type masks—significantly reducing manual annotation efforts compared to traditional full voxel-level annotations. The training configuration specifies critical parameters: subtomogram size of 72^3 voxels (chosen to balance memory constraints and spatial context), padding size of 12 voxels for the spatial overlap-based strategy, batch size of 8, learning rate of 0.001, and a segmentation threshold of 0.5. These same coordinates of the same tomogram are also used for validation in this pipeline. During the training phase, subtomograms centered on individual particles are extracted to ensure balanced representation. The 3D-ResUNet model, employing feature maps [24, 48, 72, 108], is trained with these weak labels to maintain accuracy while minimizing annotation costs.In the inference stage, identical spatial parameters are maintained. Tomograms are scanned with subtomogram size N and stride S, following the overlap-tile strategy ($S = N - 2 \times padding\_size = 48$ voxels) to eliminate segmentation artifacts at tile edges. A GPU-accelerated pooling-based postprocessing step employing mean-pool non-maximum suppression (NMS) quickly identifies particle centroids from generated segmentation masks.

S8. Dice Coefficient

To measure the accuracy of membrane segmentation, we used a commonly employed metric to evaluate segmentation methods, known as the Dice Score or Dice coefficient. This score measures the overlap between a predicted segmentation and the ground truth.

Given two sets A and B, Dice score is defined as:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

For a 3D voxelized binary segmentation task like ours, A can be regarded as the set of voxels in the ground truth mask and B the set of voxels in the predicted mask.

Then:

$$|A \cap B| = TP, \quad |A| = TP + FN, \quad |B| = TP + FP$$

Substituting into the original Dice score formula:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN}$$

We use the final equation to calculate the Dice scores in the main manuscript.

S9. F1 Score

To measure the accuracy of macromolecule localization, we used the commonly employed metric, the F1 score. We calculate the F1 score using the list of predicted macromolecule locations and the list of ground truth macromolecule locations. The locations are expressed using the macromolecule's center (x, y, z) coordinate. If any predicted macromolecule location has any ground truth location within 10 voxels of euclidean distance, then it is counted a True Positive (TP). If there are multiple predicted macromolecule location within 10 voxels of euclidean distance to a particular ground truth location, then only 1 of the predicted location is counted as True Positive (TP), while the remaining are counted as False Positive (FP). Any predicted macromolecule location not having any ground truth location within 10 voxels of euclidean distance is counted as False Positive (FP). The ground truth locations that do not have any predicted macromolecule location within 10 voxels of euclidean distance is counted as False Negative (FN). Using the count of TP, FP, and FN, we calculate precision and recall as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

.

Finally, the F1 score was calculated as:

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

.

S10. Cellular Cryo-ET datasets used in experiments

We primarily used cellular cryo-ET tomograms of *S. Pombe* for validating our method. These tomograms have a pixel spacing of 13.48 Åand a zyx dimension of either $500 \times 928 \times 960$ or $1000 \times 928 \times 960$. They contain large membranes, organelles, and small macromolecular complexes such as ribosomes and fatty acid synthase (FAS). The tomograms are publicly available on CZI cryo-ET data portal with dataset ID: DS-10001 and also as EMPIAR-10988. We used the 10 tomograms that were imaged using Volta Phase Plate (VPP) and thus have higher contrast for segmentation.

Additionally, we applied our method to *C. Elegans* cellular cryo-ET tomograms and Human RPE1 cellular cryo-ET tomograms to qualitatively validate our method. Since these datasets lacked well-defined ground truth, quantitative evaluation was not possible. The *C. Elegans* tomograms (100 available) have a uniform voxel spacing of 7.56 Å. These tomograms contain multiple subcellular objects, including membranes, ribosomes, and other cellular components. The tomograms are of uniform size with a zyx dimension of $500 \times 1024 \times 1024$. These tomograms exhibited a high noise level and low contrast. To mitigate this and enhance feature extraction, we applied the CCP-EM denoiser [14] prior to segmentation.

The tomograms (3 available) from Human RPE1 (Retinal Pigment Epithelium) cell lines are publicly available at EMPIAR-10989. The tomograms are of size $500 \times 928 \times 928$ with a voxel spacing of 1.348nm. The tomograms visualize regions of cells that contain subcellular objects like actin filaments and microtubules.

S11. Ablation Experiments

We performed ablation experiments to assess the importance of several individual components (feature aggregation, slab correction, and the supervised pseudo-label training) of our pipeline. We conducted experiments for 100 slabs in 2 cryo-ET tomograms in the *S. Pombe* data set. We provide the results in Table S1. The results highlight the need for individual components of our method. Furthermore, generating pseudo-labels for a few slabs in an unsupervised way and then using the pseudo-labels to train a lightweight supervised model to generate the final segmentation mask for all the slabs is a rational strategy, as this improves the final segmentation mask (as evidenced by *w/o PL train*) in Table S1) and also saves time and compute.

	TS_0004 (z=450–550)		TS_0008 (z=200–300)	
Method	Dice_{Mem} (↑)	F1_{Macro} (↑)	Dice_{Mem} (↑)	F1_{Macro} (↑)
w/o Feat. aggr. + PL train	0.431	0.145	0.454	0.318
w/o Slab corr. + PL train	0.427	0.152	0.457	0.355
w/o PL train	0.441	0.171	0.471	0.391
Ours (full)	**0.487**	**0.394**	**0.564**	**0.475**

Table S1. Ablation results. PL train = Pseudo-Label training

References

[1] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16061–16070, 2022. 1

[2] 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. 4, 1

[3] Muyuan Chen, Wei Dai, Stella Y Sun, Darius Jonasch, Cynthia Y He, Michael F Schmid, Wah Chiu, and Steven J Ludtke. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nature methods*, 14(10):983–985, 2017. 1

[4] Muyuan Chen, James M Bell, Xiaodong Shi, Stella Y Sun, Zhao Wang, and Steven J Ludtke. A complete data processing workflow for cryo-et and subtomogram averaging. *Nature methods*, 16(11):1161–1168, 2019. 1

[5] CompVis. Stable diffusion v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022. Accessed: 2025-06-19. 4

[6] Irene de Teresa-Trueba, Sara K Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E Zimmerli, Mauricio Toro-Nahuelpan, Dorothy WC Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, et al. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, 2023. 1, 3

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[8] Lorenz Lamm, Simon Zufferey, Ricardo D Righetto, Wojciech Wietrzynski, Kevin A Yamauchi, Alister Burt, Ye Liu, Hanyi Zhang, Antonio Martinez-Sanchez, Sebastian Ziegler, et al. Membrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography. *biorxiv*, pages 2024–01, 2024. 1, 3

[9] Mart GF Last, Lenard M Voortman, and Thomas H Sharp. Scaling data analyses in cellular cryoet using comprehensive segmentation. *bioRxiv*, pages 2025–01, 2025. 1, 8

[10] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. Deepetpicker: Fast and accurate 3d particle picking for cryo-et using weakly supervised deep learning. https://github.com/cbmi-group/DeepETPicker, 2024. Accessed: 2025-06-13. 6

[11] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. Deepetpicker: Fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15(1):2090, 2024. 1, 2, 4, 6, 7

[12] Julia Mahamid, Stefan Pfeffer, Miroslava Schaffer, Elizabeth Villa, Radostin Danev, Luis Kuhn Cuellar, Friedrich Förster, Anthony A Hyman, Jürgen M Plitzko, and Wolfgang Baumeister. Visualizing the molecular sociology at the hela cell nuclear periphery. *Science*, 351(6276):969–972, 2016. 1

[13] Eva Nogales and Julia Mahamid. Bridging structural and cell biology with cryo-electron microscopy. *Nature*, 628(8006): 47–56, 2024. 1

[14] Ding Sheng Ong and Collaborative Computational Project for Electron cryo-Microscopy (CCP-EM). CCP-EM Denoiser. https://gitlab.com/ccpem/ccpem-denoiser, 2024. GitLab repository. 4

[15] Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, pages 2025–04, 2025. 4, 6

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4, 1, 2

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4, 6

[19] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. 3

[20] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS letters*, 594 (20):3243–3261, 2020. 1

[21] Thorsten Wagner and Stefan Raunser. The evolution of sphire-cryolo particle picking and its application in automated cryo-em processing workflows. *Communications biology*, 3(1):61, 2020. 7

[22] Xudong Wang, Jiale Zhu, Yutong Bai, Ishan Misra, and Stella X Yu. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13288–13298, 2022. 3, 7

[23] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 3, 4, 7

[24] Abigail JI Watson and Alberto Bartesaghi. Advances in cryo-et data processing: meeting the demands of visual proteomics. *Current Opinion in Structural Biology*, 87:102861, 2024. 1

[25] Wikipedia contributors. Cryogenic electron tomography. https://en.wikipedia.org/wiki/Cryogenic_electron_tomography, 2025. Accessed: 2025-11-21. 1

[26] Gong-Her Wu, Charlene Smith-Geater, Jesús G Galaz-Montoya, Yingli Gu, Sanket R Gupte, Ranen Aviner, Patrick G Mitchell, Joy Hsu, Ricardo Miramontes, Keona Q Wang, et al. Cryoet reveals organelle phenotypes in huntington disease patient ipsc-derived and mouse primary neurons. *Nature communications*, 14(1):692, 2023. 1

[27] Xiangrui Zeng, Anson Kahng, Liang Xue, Julia Mahamid, Yi-Wei Chang, and Min Xu. High-throughput cryo-et structural pattern mining by unsupervised deep iterative subtomogram clustering. *Proceedings of the National Academy of Sciences*, 120(15):e2213149120, 2023. 1

[28] Xiao Zhang, David Yunis, and Michael Maire. Deciphering'what'and'where'visual pathways from spectral cluster-

ing of layer-distributed neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4175, 2024. 3, 4, 5, 6

[29] Karel J Zuiderveld et al. Contrast limited adaptive histogram equalization. *Graphics gems*, 4(1):474–485, 1994. 4