This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

CryoSAM: Training-free CryoET Tomogram Segmentation with Foundation Models

Yizhou Zhao[1], Hengwei Bian[1], Michael Mu[1], Mostofa R. Uddin[1], Zhenyang Li[2], Xiang Li[1], Tianyang Wang[2], and Min Xu[1⋆]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] University of Alabama at Birmingham, Birmingham AL 35294, USA

Abstract. Cryogenic Electron Tomography (CryoET) is a useful imaging technology in structural biology that is hindered by its need for manual annotations, especially in particle picking. Recent works have endeavored to remedy this issue with few-shot learning or contrastive learning techniques. However, supervised training is still inevitable for them. We instead choose to leverage the power of existing 2D foundation models and present a novel, training-free framework, CryoSAM. In addition to prompt-based single-particle instance segmentation, our approach can automatically search for similar features, facilitating full tomogram semantic segmentation with only one prompt. CryoSAM is composed of two major parts: 1) a prompt-based 3D segmentation system that uses prompts to complete single-particle instance segmentation recursively with Cross-Plane Self-Prompting, and 2) a Hierarchical Feature Matching mechanism that efficiently matches relevant features with extracted tomogram features. They collaborate to enable the segmentation of all particles of one category with just one particle-specific prompt. Our experiments show that CryoSAM outperforms existing works by a significant margin and requires even fewer annotations in particle picking. Further visualizations demonstrate its ability when dealing with full tomogram segmentation for various subcellular structures. Our code is available at: https://github.com/xulabs/aitom.

Keywords: Cryogenic Electron Tomography (CryoET) · Prompt-based Segmentation · Foundation Models.

1 Introduction

The advancement of Cryogenic Electron Tomography (CryoET) makes it possible to capture macromolecular structures with native conformations at nanometer resolution [3]. In a typical CryoET pipeline, researchers prepare frozen-hydrated samples and expose them to electron beams for imaging. The sample is incrementally tilted, allowing for the collection of multi-view images, i.e., tilt-series. These images can be used for 3D reconstruction, resulting in a 3D density map, the tomogram. Further investigation requires particle picking to

⋆ Corresponding author.

accurately localize and segment sub-cellular structures. To this end, most existing methods [29,23,24,27,6,7,19,17] resort to supervised training or template matching [5], necessitating a large amount of laborious annotation. Some recent works propose to adopt few-shot learning [28] or contrastive learning [8] techniques to ameliorate this issue. However, currently, there is still a need to train on several known categories or at least 20-50 annotations.

Looking out of the CryoET domain, recent years have witnessed a proliferation of general-purpose segmentation models. With the ability to condition on various types of inputs and accomplish different downstream segmentation tasks [15,16,12,14,13], SAM [11] and SEEM [30] have demonstrated a diverse range of capabilities. Furthermore, in the three-dimensional world, SA3D [2] and LERF [10] extend the ability of the implicit 3D representation NeRF [18] with prompt-based segmentation and visual grounding. This progress inspires us to explore segmenting CryoET tomograms with general-domain foundation models. However, there are several obstacles. While we see a tremendous number of 2D foundation models, their counterparts for 3D are relatively scarce, e.g., a general volumetric segmentation model is still absent. Hence, bridging general-domain foundation models to CryoET analysis is not trivial. In addition, general-purpose segmentation models [11,2] are commonly instance-specific while semantic-agnostic. This limits their direct application to semantic-specific particle picking, which requires picking all particles of a category simultaneously.

To overcome these challenges, we present CryoSAM, a training-free approach for prompt-based CryoET tomogram segmentation. Our method introduces a prompt-based 3D segmentation pipeline, bridging the gap between 2D segmentation models and 3D volumetric segmentation. Our intuition is that the silhouettes of a particle are similar in adjacent tomogram slices. Hence, we can segment its 3D structure layer after layer by refining the segmentation mask from the previous plane. Formally, we achieve this by employing a Cross-Plane Self-Prompting mechanism, which recursively propagates and refines segmentation masks along one direction by prompting SAM [11] with segmentation results from preceding planes. This allows us to segment one particle instance with a single prompt. To further segment all particles of a specific category comprehensively, we introduce a Hierarchical Feature Matching strategy for efficient instance-level feature matching. This approach eliminates the need for predefined templates [2,25] and the extraction of subtomograms [26]. Using the mean feature of prompted particles as the query, it filters out regions dissimilar to the query in a coarse-to-fine manner. After filtering, it proposes point prompts in a relatively low resolution and relies on the prompt-based 3D segmentation pipeline to achieve final segmentation results. These designs enable semantic segmentation over a full CryoET tomogram with a single prompt.

Our contributions can be summed up as follows:

- We present a novel, training-free framework, CryoSAM, that takes a full CryoET tomogram and a set of user prompts as input and segments the prompted particle and all particles of the same category. This contrasts with current methods that require supervised training [29,23,8,28].

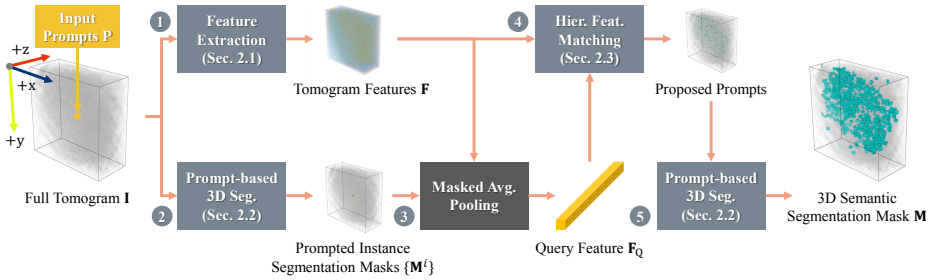

Fig. 1. Framework overview. ❶: We extract per-slice 2D features for three views (z, y, and x) from CryoET tomogram $\mathbf{I}$ and concatenate them as $\mathbf{F}$. ❷: After segmenting the particle(s) prompted by $\mathbf{P}$ with instance segmentation mask(s), ❸: we average pool the masked features to get query feature $\mathbf{F}_Q$. ❹: To efficiently propose prompts for further segmentation, we match $\mathbf{F}_Q$ with $\mathbf{F}$ using Hierarchical Feature Matching. ❺: Finally, we adopt prompt-based 3D segmentation for semantic segmentation results $\mathbf{M}$.

- We introduce Cross-Plane Self-Prompting, which enables 3D volumetric segmentation with 2D foundation models, significantly reducing the labor cost of annotation by leveraging its prompt-based nature.
- We propose a Hierarchical Feature Matching strategy to match instance-level particle features. It cuts down the runtime by 95% compared with naive feature matching, being more efficient and convenient to use.

2 Method

Given a volumetric CryoET tomogram $\mathbf{I} \in \mathbb{R}^{D\times H\times W}$ and N point prompts $\mathbf{P} \in \mathbb{R}^{N\times 3}$ denoting a set of single-category particles, our goal is to segment all particles of the same category as the prompted ones. This process predicts a 3D semantic segmentation mask $\mathbf{M} \in \{0,1\}^{D\times H\times W}$, with the overall pipeline depicted in Fig. 1. D, H and W denote depth, height, and width respectively.

2.1 Feature Extraction

We rely on an off-the-shelf image encoder $\mathcal{E}$ to extract 2D features from tomogram slices $\{\mathbf{I}_z\}_{z=1}^{D}, \{\mathbf{I}_y\}_{y=1}^{H}, \{\mathbf{I}_x\}_{x=1}^{W}$. For each view z, y, and x, we obtain $\mathbf{Z}^{\mathcal{E}} = \{\mathcal{E}(\mathbf{I}_z)\}_{z=1}^{D} \in \mathbb{R}^{D\times h\times w\times C}$, $\mathbf{Y}^{\mathcal{E}} = \{\mathcal{E}(\mathbf{I}_y)\}_{y=1}^{H} \in \mathbb{R}^{d\times H\times w\times C}$, and $\mathbf{X}^{\mathcal{E}} = \{\mathcal{E}(\mathbf{I}_x)\}_{x=1}^{W} \in \mathbb{R}^{d\times h\times W\times C}$, where the lowercase d, h, w are feature resolutions in the latent space. Then we bilinear upsample them to get $\mathbf{Z}, \mathbf{Y}, \mathbf{X} \in \mathbb{R}^{D\times H\times W\times C}$, and aggregate them with a concatenation

$$\mathbf{F} = \{\mathbf{F}_{zyx}\}_{z=1,y=1,x=1}^{D,H,W} = [\mathbf{Z}, \mathbf{Y}, \mathbf{X}] \in \mathbb{R}^{D\times H\times W\times 3C}, \tag{1}$$

where $\mathbf{F}_{zyx}$ is a feature vector in $\mathbf{F}$ with coordinates $[z, y, x]$.

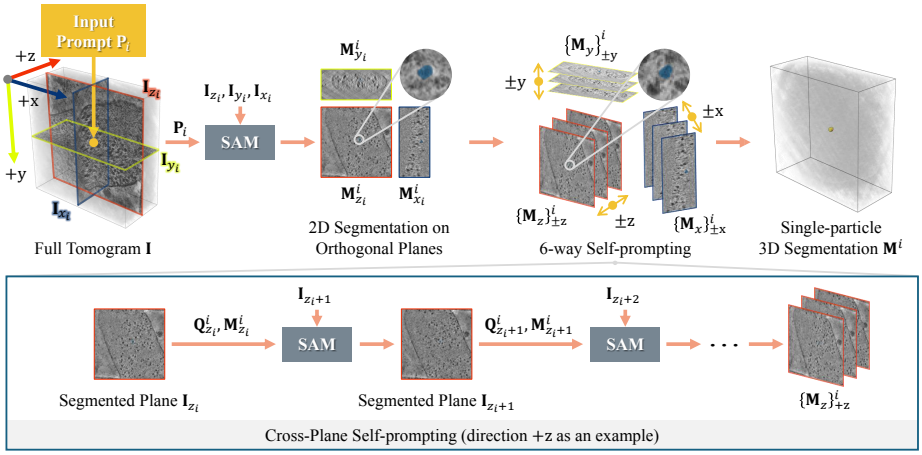

Fig. 2. The pipeline of prompt-based 3D segmentation. After segmenting the orthogonal planes intersect at the point prompt $\mathbf{P}_i$, we iteratively execute Cross-Plane Self-Prompting until we get the complete mask of the particle.

2.2 Prompt-based 3D Segmentation

We propose Cross-Plane Self-Prompting, a mechanism that can propagate segmentation masks along the $\pm$z, $\pm$y, $\pm$x axes, to approach prompt-based 3D segmentation, as illustrated in Fig. 2. The intuition is that the segmentation mask of one particle should be similar for neighboring slices. Hence, we can prompt SAM [11] with the segmentation results from the previous plane to get subsequent results. Formally, we take as input a single point prompt $\mathbf{P}_i = [z_i, y_i, x_i]$ and the three orthogonal planes intersecting at this point, namely, the YX-plane $\mathbf{I}_{z_i}$, the ZX-plane $\mathbf{I}_{y_i}$, and the ZY-plane $\mathbf{I}_{x_i}$. Then, we employ SAM to obtain their 2D segmentation results, with the YX-plane as an example

$$(\mathbf{C}^i_{z_i}, \mathbf{M}^i_{z_i}) = \text{SAM}\left[\mathbf{I}_{z_i}|(x_i, y_i)\right],\ \mathbf{Q}^i_{z_i} = \text{argmax}_{x,y}(\mathbf{C}^i_{z_i}), \tag{2}$$

where $\mathbf{C}^i_*$ are the predicted confidence scores, $\mathbf{M}^i_*$ are the predicted segmentation masks, and $\mathbf{Q}^i_*$ are the coordinates with the highest confidence scores. We use superscript ^{i} to represent the index of the initial point prompt. Then for each direction in $\{\pm z, \pm y, \pm x\}$, we prompt the next tomogram slice with $\mathbf{Q}^i_*$ and $\mathbf{M}^i_*$ from the previous plane, for which we term Cross-Plane Self-Prompting. Taking the +z direction as an example which starts from $z = z_i$, we have

$$(\mathbf{C}^i_{z+1}, \mathbf{M}^i_{z+1}) = \text{SAM}\left[\mathbf{I}_z|\mathbf{Q}^i_z, \mathbf{M}^i_z\right],\ \mathbf{Q}^i_{z+1} = \text{argmax}_{x,y}(\mathbf{C}^i_{z+1}). \tag{3}$$

Here, we benefit from SAM's versatility, which allows it to take both point and mask prompts as inputs. This recursive process continues until the intersection over union (IoU) of the segmentation masks in two adjacent slices drops below a threshold τ_{IoU}, which suggests that prompting the current plane will not get

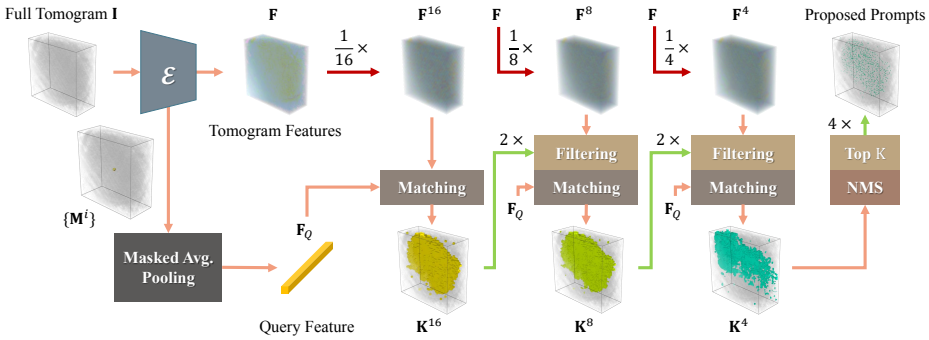

Fig. 3. The pipeline of Hierarchical Feature Matching. We average the tomogram features in the instance segmentation masks to obtain a query feature $\mathbf{F}_Q$. Then we downsample $\mathbf{F}$ into several coarse ones and match them with $\mathbf{F}_Q$ in a coarse-to-fine manner. After the last matching stage, we apply NMS and gather coordinates with top K similarities as prompts to derive final semantic segmentation results.

a result consistent with previous ones. After getting the segmentation masks $\{\mathbf{M}_z\}^i_{\pm \mathrm{z}}, \{\mathbf{M}_y\}^i_{\pm \mathrm{y}}, \{\mathbf{M}_x\}^i_{\pm \mathrm{x}}$ for all 6 directions sequentially, we aggregate a union of all segmentation masks in 3D, i.e., $\mathbf{M}^i = \{\mathbf{M}_z\}^i_{\pm \mathrm{z}} \cup \{\mathbf{M}_y\}^i_{\pm \mathrm{y}} \cup \{\mathbf{M}_x\}^i_{\pm \mathrm{x}}$.

2.3 Hierarchical Feature Matching

Shown in Fig. 3, Hierarchical Feature Matching aims to efficiently search for voxel regions with similar features as the query. For input point prompts $\mathbf{P} = \{\mathbf{P}_i\} \in \mathbb{R}^{N\times 3}$, we obtain an instance segmentation mask for each prompt through prompt-based 3D segmentation, resulting in $\{\mathbf{M}^i\}$. Then, we derive the query feature $\mathbf{F}_Q$ via masked average pooling (MAP)

$$\mathbf{F}_Q = \frac{\sum_i \sum_{zyx} \mathbf{M}^i_{zyx} \odot \mathbf{F}_{zyx}}{\sum_i \|\mathbf{M}^i\|_0}, \tag{4}$$

where $\odot$ is the Hadamard product with broadcasting and $\|\cdot\|_0$ is the 0-norm indicating the number of non-zero voxels. This operation averages features masked by the instance segmentation masks to obtain a mean feature representing the prompted particles. While a brute-force approach can achieve voxel-precise feature matching between $\mathbf{F}_Q$ and $\mathbf{F}$, we empirically show this is neither efficient nor necessary. Instead, we propose to match $\mathbf{F}_Q$ with multi-resolution features in $\mathbf{F}$ in a coarse-to-fine manner, each time keeping only the most similar proportion. We begin with building a feature pyramid

$$\{\mathbf{F}^r\} = \{[\mathbf{Z}^r, \mathbf{Y}^r, \mathbf{X}^r]\}, \tag{5}$$

where $r \in \{16, 8, 4\}$ is the downsampling ratio, and $\mathbf{F}^r \in \mathbb{R}^{\frac{D}{r}\times\frac{H}{r}\times\frac{W}{r}\times 3C}$. $\mathbf{Z}^r \in \mathbb{R}^{\frac{D}{r}\times\frac{H}{r}\times\frac{W}{r}\times C}$ stands for an r times downsampled version of $\mathbf{Z}$, with similar

Table 1. Comparison results for particle picking on EMPIAR-10499 [22].

Method	Annotation Ratio	Precision	Recall	F1 Score	Runtime (min)
EMAN2 [21]	-	26.1	55.3	35.5	2-5
crYOLO [24]	100%	47.8	56.8	52.0	30-40
Huang et al. [8]	5%	49.6	58.1	53.5	5-10
	10%	50.1	58.2	53.8	
	30%	55.9	60.3	58.0	
	50%	53.0	65.1	58.4	
	70%	54.9	66.7	60.2	
CryoSAM (Ours)	< 1% (single prompt)	53.1	55.3	54.2	10-15
	5%	57.8	74.3	65.0	
	10%	58.2	75.1	65.5	
	30%	58.1	75.4	65.6	
	50%	58.0	75.3	65.5	
	70%	58.5	79.4	67.4	

definitions for $\mathbf{Y}^r$ and $\mathbf{X}^r$. Then from the lowest resolution of $\{\mathbf{F}^r\}$, we calculate its point-wise cosine similarity $\mathbf{S}^r = \{\mathbf{S}^r_{zyx}\}_{z=1,y=1,x=1}^{\frac{D}{r},\frac{H}{r},\frac{W}{r}}$ with query $\mathbf{F}_Q$

$$\mathbf{S}^r_{zyx} = \frac{\mathbf{F}_Q \cdot (\mathbf{F}^r_{zyx})^\top}{\|\mathbf{F}_Q\|_2 \cdot \|\mathbf{F}^r_{zyx}\|_2}. \tag{6}$$

For the lowest resolution, we calculate the similarity for all $\frac{D}{r}\frac{H}{r}\frac{W}{r}$ features. Subsequently, we build a mask $\mathbf{K}^r = \mathbf{S}^r \geq \tau_{\text{sim}}$ that filters out regions with low similarity scores and propagates this mask to the next resolution with upsampling. This allows the next round of feature matching to be conducted only on the high-similarity features, thereby greatly reducing the computational complexity. After iterating through the whole downsampling ratio list, we apply non-maximum suppression (NMS) on the coordinates with their similarity scores and keep the top K of them as point prompts. These prompts are then fed into the prompt-based 3D segmentation pipeline for semantic segmentation.

3 Experiment

3.1 Experimental Settings

Datasets and evaluation metrics. Due to the scarcity of CryoET segmentation annotations, we mainly assess the quantitative performance of CryoSAM for particle picking. To this end, we utilize the EMPIAR-10499 dataset [22,9], which comprises 65 tilt-series of native M. pneumoniae cells with annotated ribosomes. We use the prediction from each proposed prompt as an instance segmentation mask to compare with other detection methods [8,21,24] in terms of precision, recall, and F1 score. Results from all 65 tilt-series with 5 random sets of input

Table 2. Ablation study for different feature extractors.

2D Feature Extractor	Annotation Ratio	Precision	Recall	F1 Score
SAM [11]	< 1% (single prompt)	37.4	38.8	38.1
	10%	44.1	60.0	50.8
DINO [1]	< 1% (single prompt)	56.3	52.8	54.5
	10%	63.2	74.4	68.3
DINOv2 [20]	< 1% (single prompt)	55.4	58.8	57.1
	10%	59.8	80.1	68.5

Table 3. Ablation study for different feature matching strategies.

Feature Matching Strategy	Annotation Ratio	Precision	Recall	F1 Score	Runtime (min)
Naive	< 1% (single prompt)	53.5	56.4	54.9	60-65
	10%	60.8	80.7	69.4	
Hierarchical	< 1% (single prompt)	55.4	58.8	57.1	10-15
	10%	59.8	80.1	68.5	

prompts are averaged in our comparison results reported in Tab. 1, while the first 20 tilt-series and a fixed set of input prompts are used in our ablation study. We do not calculate mean average precision (mAP) as our method does not output an explicit score for each segmentation mask.

Implementation details. We use DINOv2 [20] with a ViT-L/14 [4] backbone as the default 2D encoder $\mathcal{E}$ of CryoSAM and SAM [11] with ViT-H as our 2D segmentation model. The IoU threshold τ_{IoU} to determine the end of segmentation mask propagation and the similarity threshold τ_{sim} to filter out dissimilar regions in Hierarchical Feature Matching are both set to 0.5. Top $K = 512$ coordinates in the final stage of Hierarchical Feature Matching are used as proposed prompts for full tomogram semantic segmentation. In all experiments, we do not require any training for CryoSAM. We use a subset of all ground truth coordinates as input prompts. The annotation ratio in tables refers to the proportion of prompted particles to all particles in our scenario.

3.2 Comparison Results

In Tab. 1, CryoSAM demonstrates significant advancements in particle picking compared to three baselines under the same annotation ratio. Using 5 random sets of input prompts, we conducted one-tailed paired t-tests to assess the significance of our improvements over Huang et al. [8], consistently yielding p-values below 0.01. It is also noteworthy that our single-prompt result is better than [8] under 10% annotation, which shows the annotation-efficient property of CryoSAM. Our performance also improves as the number of available prompts increases. This is probably because the averaged features are more robust with the addition of different particle instances in similarity-based matching.

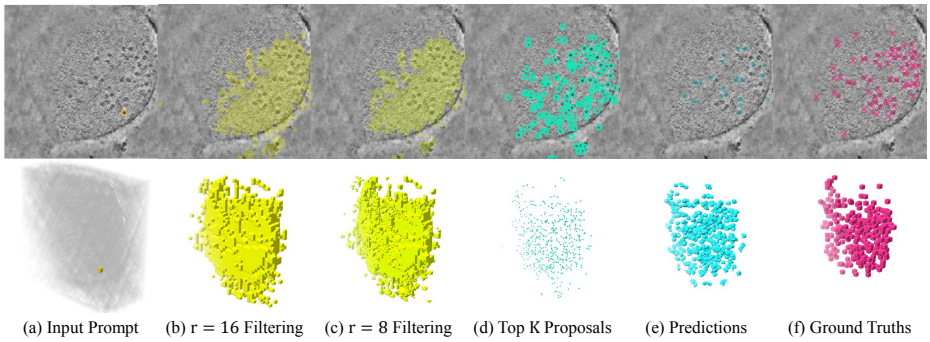

Fig. 4. Intermediate and final results of CryoSAM. In (d) and (f), we show points with coordinates ranging from $z - 20$ to $z + 20$ for demonstration.

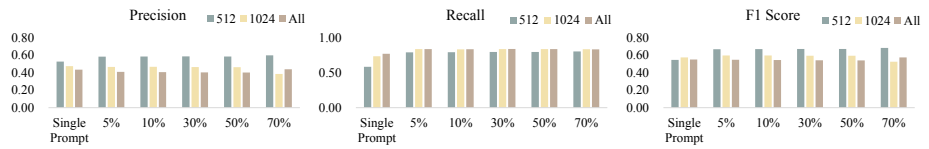

Fig. 5. Ablation study for the number of proposed prompts. 512/1024/All: number of proposed prompts selected for prompt-based semantic segmentation.

3.3 Ablation Study and Analysis

Impact of feature extractors. We ablate the particle picking performance over different 2D feature extractors in Tab. 2. Our results show that using DINO [1] and DINOv2 [20] achieves significantly better results than using the SAM [11] encoder. It follows that DINO and DINOv2 learn more discriminative features with self-supervised training, which is beneficial for accurate feature matching.

Impact of feature matching strategies. We evaluate the effectiveness of Hierarchical Feature Matching in Tab. 3 by replacing it with naive feature matching that only computes voxel-wise similarity in the highest DHW resolution. We see our hierarchical strategy retains a comparable performance while taking a notably shorter time to process. This reflects the robustness of our prompt-based 3D segmentation pipeline, which does not require the input to be voxel-precise.

Impact of the number of proposed prompts. In Fig. 5, we analyze the precision-recall trade-off by varying K. Generally, smaller values of K result in lower recall and higher precision. We make our design choice to set $K = 512$ by selecting the model with the best overall F1 score.

Qualitative analysis. We visualize the whole process of CryoSAM in Fig. 4, which shows it can conduct 3D semantic segmentation with just a single point prompt. See the supplementary for more qualitative results and failure cases.

4 Conclusion

We present CryoSAM, a training-free framework that segments full CryoET tomograms with given prompts. It has two core innovations. First, the proposed Cross-Plane Self-Prompting mechanism bridges the gap between 2D segmentation foundation models and 3D volumetric segmentation. Second, we introduce Hierarchical Feature Matching, which is capable of efficient search for one category of particles. Combining both shows positive synergy in prompt-based full tomogram semantic segmentation, leading to SOTA results in particle picking.

Acknowledgments. This study was partially funded by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148. Additionally, it received support from Oracle Cloud credits and resources provided by Oracle for Research, as well as computational resources from the AMD HPC Fund. MRU was supported by a fellowship from CMU CMLH.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
2. Cen, J., Zhou, Z., Fang, J., Shen, W., Xie, L., Jiang, D., Zhang, X., Tian, Q., et al.: Segment anything in 3d with nerfs. Advances in Neural Information Processing Systems **36** (2024)
3. Doerr, A.: Cryo-electron tomography. Nature Methods **14**(1), 34–34 (2017)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Frangakis, A.S., Böhm, J., Förster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R., Baumeister, W.: Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. Proceedings of the National Academy of Sciences **99**(22), 14153–14158 (2002)
6. Gubins, I., Chaillet, M.L., van Der Schot, G., Veltkamp, R.C., Förster, F., Hao, Y., Wan, X., Cui, X., Zhang, F., Moebel, E., et al.: Shrec 2020: Classification in cryo-electron tomograms. Computers & Graphics **91**, 279–289 (2020)
7. Hao, Y., Wan, X., Yan, R., Liu, Z., Li, J., Zhang, S., Cui, X., Zhang, F.: Vp-detector: A 3d multi-scale dense convolutional neural network for macromolecule localization and classification in cryo-electron tomograms. Computer Methods and Programs in Biomedicine **221**, 106871 (2022)
8. Huang, Q., Zhou, Y., Liu, H.F., Bartesaghi, A.: Accurate detection of proteins in cryo-electron tomograms from sparse labels. In: European Conference on Computer Vision. pp. 644–660. Springer (2022)
9. Iudin, A., Korir, P.K., Salavert-Torres, J., Kleywegt, G.J., Patwardhan, A.: Empiar: a public archive for raw electron microscopy image data. Nature methods **13**(5), 387–388 (2016)

10. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
12. Li, X., Lin, C.C., Chen, Y., Liu, Z., Wang, J., Singh, R., Raj, B.: Paintseg: Painting pixels for training-free segmentation. Advances in Neural Information Processing Systems **36** (2024)
13. Li, X., Wang, J., Li, X., Lu, Y.: Hybrid instance-aware temporal fusion for online video instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1429–1437 (2022)
14. Li, X., Wang, J., Xu, X., Li, X., Raj, B., Lu, Y.: Robust referring video object segmentation with cyclic structural consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22236–22245 (2023)
15. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3402–3413 (2024)
16. Li, X., Wang, J., Xu, X., Yang, M., Yang, F., Zhao, Y., Singh, R., Raj, B.: Towards noise-tolerant speech-referring video object segmentation: Bridging speech and text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2283–2296 (2023)
17. Liu, G., Niu, T., Qiu, M., Zhu, Y., Sun, F., Yang, G.: Deepetpicker: Fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. Nature Communications **15**(1), 2090 (2024)
18. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
19. Moebel, E., Martinez-Sanchez, A., Lamm, L., Righetto, R.D., Wietrzynski, W., Albert, S., Larivière, D., Fourmentin, E., Pfeffer, S., Ortiz, J., et al.: Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. Nature methods (2021)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J.: Eman2: an extensible image processing suite for electron microscopy. Journal of structural biology **157**(1), 38–46 (2007)
22. Tegunov, D., Xue, L., Dienemann, C., Cramer, P., Mahamid, J.: Multi-particle cryo-em refinement with m visualizes ribosome-antibiotic complex at 3.5 å in cells. Nature Methods **18**(2), 186–193 (2021)
23. de Teresa-Trueba, I., Goetz, S.K., Mattausch, A., Stojanovska, F., Zimmerli, C.E., Toro-Nahuelpan, M., Cheng, D.W., Tollervey, F., Pape, C., Beck, M., et al.: Convolutional networks for supervised mining of molecular patterns within cellular context. Nature Methods **20**(2), 284–294 (2023)
24. Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., et al.: Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. Communications biology **2**(1), 218 (2019)

25. Wu, X., Zeng, X., Zhu, Z., Gao, X., Xu, M.: Template-based and template-free approaches in cellular cryo-electron tomography structural pattern mining. Computational Biology (2019)
26. Zeng, X., Kahng, A., Xue, L., Mahamid, J., Chang, Y.W., Xu, M.: High-throughput cryo-et structural pattern mining by unsupervised deep iterative subtomogram clustering. Proceedings of the National Academy of Sciences **120**(15), e2213149120 (2023)
27. Zhang, P.: Advances in cryo-electron tomography and subtomogram averaging and classification. Current opinion in structural biology **58**, 249–258 (2019)
28. Zhou, B., Yu, H., Zeng, X., Yang, X., Zhang, J., Xu, M.: One-shot learning with attention-guided segmentation in cryo-electron tomography. Frontiers in Molecular Biosciences **7**, 613347 (2021)
29. Zhou, L., Yang, C., Gao, W., Perciano, T., Davies, K.M., Sauter, N.K.: A machine learning pipeline for membrane segmentation of cryo-electron tomograms. Journal of Computational Science **66**, 101904 (2023)
30. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)