Adapting Pretrained Vision Transformers from 2D to 3D for Cryo-ET Classification

Yuzhou Wang
Carnegie Mellon University
yuzhouwa@andrew.cmu.edu

Xingjian Li*
Carnegie Mellon University
xingjia2@andrew.cmu.edu

Min Xu*
Carnegie Mellon University
mxu1@cs.cmu.edu

Abstract

Cryogenic electron tomography (Cryo-ET) enables visualization of macromolecular structures in near-native environments, but the resulting subtomograms are noisy and difficult to classify with deep learning models. Although transfer learning has been attempted for Cryo-ET subtomogram tasks, leveraging large-scale image-pretrained Transformers has remained largely unexplored. In this work, we study how such Transformers can be adapted to Cryo-ET subtomogram classification. We propose a simple, effective framework that (i) adapts 2D pretrained Transformer model weights to 3D subtomograms via weight inflation, and (ii) denoises subtomograms with Difference of Gaussian filtering. On both simulated and real subtomogram datasets, our approach enables ViT-B and Swin-B to outperform randomly initialized transformers and strong video-pretrained baselines: the weight-inflated Swin-B achieves **90.70%** (simulated) and **99.29%** (real), while weight-inflated ViT-B reaches **85.40%** and **97.32%**, respectively. These results demonstrate that carefully adapted image-pretrained Transformers provide a strong and practical solution for Cryo-ET subtomogram classification.

1 Introduction

Cryogenic electron tomography (Cryo-ET) reconstructs 3D cellular structures from multiple 2D electron microscopy projections [1, 2]. Cropped volumes containing individual macromolecules, called subtomograms [3, 4], are used for structure identification and classification [5, 6]. However, the data are extremely noisy and limited in quantity, making supervised training difficult and prone to overfitting.

To address this, prior works explored semi-/self-supervised learning [7, 8] or transfer from video-pretrained models [9]. Yet, these strategies underutilize the strong visual priors learned by modern large-scale Transformers such as ViT [10] and Swin [11], which have shown remarkable transferability across natural-image and multimodal domains [12, 13, 14]. Extending such 2D pretrained Transformers to Cryo-ET subtomograms remains an open problem due to their dimensional (2D vs. 3D) and channel (RGB vs. single-volume) mismatch.

We study how image-pretrained Transformers can be effectively adapted to Cryo-ET classification. Our framework combines: (i) *weight inflation*, which converts 2D convolution kernels into 3D ones while preserving pretrained features, and (ii) *Difference-of-Gaussian (DoG)* filtering to denoise subtomograms and construct three input channels aligned with RGB expectations.

On both simulated and real subtomogram datasets, weight-inflated ViT-B and Swin-B substantially outperform randomly initialized and video-pretrained baselines. Our main contributions are:

*Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

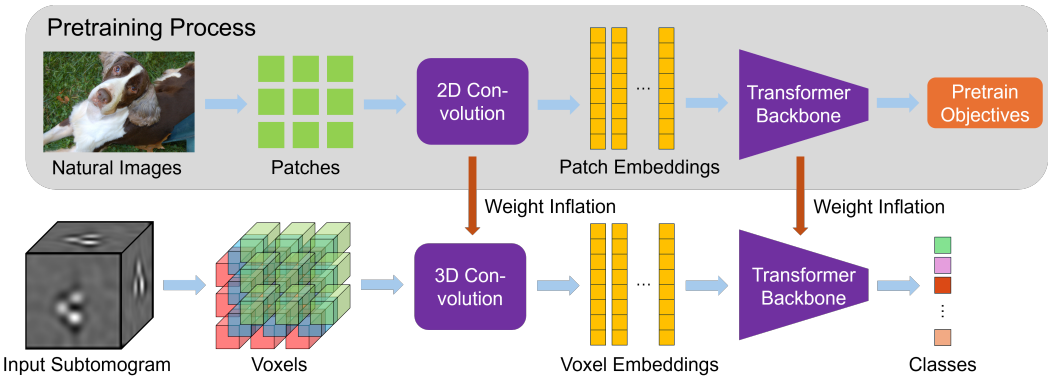

Figure 1: Overall pipeline: starting from a pretrained ViT/Swin transformer on natural images, we adapt the convolution layer and the transformer backbone of the pretrained model to process cryo-et subtomograms by applying weight inflation. Difference-of-Gaussian (DoG) filtering is applied to input subtomograms for denoising before passing to the weight-inflated model.

- A simple, general framework for adapting 2D pretrained Transformers to volumetric Cryo-ET data.
- Empirical evidence that image-pretrained models can surpass strong video-pretrained baselines when properly scaled and denoised.
- New state-of-the-art results on simulated and real Cryo-ET subtomogram classification benchmarks.

2 Related Works

Subtomogram classification. Early work relied on template matching or unsupervised autoencoders, but struggled with extreme noise and conformational heterogeneity [15, 16]. Supervised 3D CNNs improved accuracy yet remained data-hungry [17]. To reduce labeling needs, semi-/self-supervised learning has been explored for Cryo-ET volumes [7]. A parallel direction leverages transfer from video models (e.g., Kinetics-pretrained 3D CNNs and ViViT), showing that pretraining is beneficial but leaving open how to best align features with noisy, small 3D inputs [9, 18].

Transfer learning beyond videos. Large-scale image pretraining transfers strongly in medical imaging, often outperforming domain-specific pretraining when paired with light adaptation [19, 20]. This suggests that high-capacity 2D models can be competitive for Cryo-ET if bridged to volumetric inputs with minimal inductive bias changes.

Transformers for 3D vision and 2D→3D adaptation. Hybrid Transformer–UNet designs (e.g., UNETR/Swin-UNETR) adapt attention to volumetric segmentation [21, 22]. A complementary line directly inflates 2D kernels to 3D to reuse pretrained weights [23, 24]. We follow this inflation path but (i) scale receptive fields to small subtomograms and (ii) replace RGB channels with a simple DoG channelization tailored to low-SNR Cryo-ET. Our results indicate that carefully scaled inflation plus lightweight channel construction can surpass strong video-pretrained baselines on both simulated and real data.

3 Methods

Our pipeline (Figure 1) combines weight inflation of pretrained ViT/Swin backbones with Difference-of-Gaussian (DoG) filtering to adapt to 3D subtomograms and improve robustness to noise.

3.1 Model Adaptation via Weight Inflation

ViT and Swin were originally designed for RGB images, which have shape $C \times H \times W$ with $C = 3$. In contrast, Cryo-ET subtomograms are three-dimensional voxel blocks of shape $H' \times W' \times D'$ without

an explicit channel dimension. This discrepancy introduces two mismatches: (i) a dimensionality mismatch, as pretrained convolution layers operate on 2D kernels, and (ii) a channel mismatch, as subtomograms lack RGB-like channels.

To resolve the dimensionality mismatch between 2D pretrained models and 3D Cryo-ET data, we follow and extend the weight inflation strategy proposed by [24]. Specifically, a 2D convolution kernel of shape $K \times C \times P \times P$ is converted into a 3D kernel of shape $K \times C \times Q \times P' \times P'$, where K is the number of output channels, C the number of input channels, P the original kernel size, and Q the kernel depth. We explore two inflation variants: (i) average inflation, which repeats the 2D kernel Q times along the depth dimension and divides by Q, and (ii) center inflation, which inserts the 2D kernel into the central slice with zeros elsewhere. Unlike prior work that directly transfers 2D kernels into 3D, we account for the smaller spatial scale of subtomograms ($H', W'! \ll! H, W$) by introducing an in-plane resizing step before inflation: either (a) applying 2D average pooling to reduce kernel size from $P!\times!P$ to $P'!\times!P'$ ($P' < P$), or (b) enlarging subtomogram slices to match the pretrained kernel resolution. This adjustment ensures that receptive fields are properly scaled for subtomogram structures, making our approach distinct from standard weight inflation methods.

3.2 Difference-of-Gaussian Filtering

Cryo-ET subtomograms usually have low signal-to-noise ratios (SNR), and the presence of strong background noise can hinder the model's ability to extract structural information. To mitigate this, we apply three-dimensional Difference-of-Gaussian (DoG) filtering [25, 26] as a preprocessing step.

Formally, a 3D Gaussian filter $G \in \mathbb{R}^{(2r+1)\times(2r+1)\times(2r+1)}$ of radius r and variance σ is defined by

$$f(x, y, z) = \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \exp\left(-\frac{x^2 + y^2 + z^2}{2\sigma^2}\right) \tag{1}$$

$$s = \sum_{x=-r}^{r} \sum_{y=-r}^{r} \sum_{z=-r}^{r} f(x, y, z) \tag{2}$$

$$G[x, y, z] = \frac{f(x, y, z)}{s} \tag{3}$$

where $x, y, z \in [-r, r] \cap \mathbb{Z}$.

For each subtomogram, we convolve it with two Gaussian filters of variances (σ_1, σ_2) and subtract the results to obtain a denoised Difference-of-Gaussian (DoG) volume. Repeating this with another pair (σ_3, σ_4) yields a second DoG volume. We then stack the original subtomogram with the two DoG outputs, forming a three-channel input that matches the RGB structure expected by pretrained models while providing multi-scale denoising cues beyond simple channel replication.

4 Experiments

We evaluate our approach on both simulated [27] and real [28] Cryo-ET subtomogram datasets. All experiments use PyTorch with AdamW optimizer and cosine learning rate decay. A detailed ablation study is provided in Appendix 2.

Method	Pretrain Dataset	Simulated Acc.	Real Acc.
ViViT [9]	Kinetics-400	84.8	91.1
3D-ResNet-34 [9]	Kinetics-400	85.9	99.1
ViT-B (random)	-	68.3	81.1
Swin-B (random)	-	80.7	91.6
ViT-B (ours)	ImageNet-21k	85.4	97.3
Swin-B (ours)	ImageNet-22k	**90.7**	**99.3**

Table 1: Classification accuracy (%) on simulated and real Cryo-ET subtomogram datasets.

4.1 Simulated Cryo-ET Subtomogram Dataset

Dataset. We follow [27, 9] to generate simulated subtomograms from ten macromolecular structures (1bxn, 1f1b, 1yg6, 2byu, 2h12, 2ldb, 3gl1, 3hhb, 4d4r, 6t3e) in the Protein Data Bank [29]. Each subtomogram has shape 32^3 voxels with signal-to-noise ratio (SNR) 0.05. We use $3{,}000$ samples for training, $1{,}000$ for validation, and $1{,}000$ for testing.

Baselines. We compare against strong video-pretrained models: ViViT [18] and 3D-ResNet-34 [30], both pretrained on Kinetics-400 [31] following the pipeline in [9]. These represent the standard state-of-the-art in Cryo-ET transfer learning.

Implementation. We evaluate ViT-B/16 [10] pretrained on ImageNet-21k and Swin-B [11] pretrained on ImageNet-22k. For ViT, the 2D patch embedding of size $3 \times 16 \times 16$ is inflated to $3 \times Q \times P' \times P'$ with $Q \in \{1, 3, 5\}$ and $P' \in \{4, 16\}$; for Swin, the stem convolution ($3 \times 4 \times 4$) is similarly inflated. Subtomograms are preprocessed using DoG-based channelization (Sec. 3.2) to form three-channel inputs, and positional encodings are reinitialized for 3D. The final embedding is passed through a fully connected layer for classification

4.2 Real Cryo-ET Subtomogram Dataset

Dataset. We adopt the real Cryo-ET dataset curated in [9, 8], containing seven protein classes extracted from the Noble Single Particle Dataset [28]. Each subtomogram (28^3 voxels) is rescaled to 32^3 and selected via DoG-based particle picking [26]. The dataset includes 400 samples per class (2,800 total), split 3:1:1 for training, validation, and testing.

Baselines and setup. We use the same ViViT and 3D-ResNet-34 baselines as above and apply identical fine-tuning procedures.

4.3 Results and Analysis

As shown in Table 1, our weight-inflated Swin-B achieves 90.7% accuracy on the simulated dataset and 99.3% on the real dataset, outperforming both randomly initialized and video-pretrained models. ViT-B shows similar trends with notable gains over random initialization (68.3→85.4%). These results indicate that image-pretrained Transformers perform better than video-pretrained or randomly initialized Transformers for Cryo-ET subtomogram classification.

The ablation study further (Appendix 2) show that: (1) pretrained inflation consistently improves over random initialization; (2) scaling kernel resolution to subtomogram size provides large boosts, especially for Swin-B; and (3) DoG-based channelization yields stable gains across all settings.

5 Conclusion

We introduced a pipeline to adapt vision Transformers pretrained on RGB images for Cryo-ET subtomogram classification: (i) we adjust kernel sizes and input resolutions before extending 2D convolution kernels into 3D, and (ii) we build a three-channel input by adding two DoG-filtered versions of the subtomogram to the raw volume. These steps addressed both the dimensional and channel mismatches between images and subtomograms. On both simulated and real datasets, the weight-inflated Transformers achieved the best results, outperforming randomly initialized Transformers and video-pretrained baselines.

Limitations and future work. We tested only two Vision Transformer architectures. The selected DoG variances are fixed and not diverse, and the positional encodings for the 3D patches are randomly initialized. Future work could explore the use of pretrained weights from a wider range of vision and biomedical tasks, testing alternative strategies beyond center or average weight inflation, and developing more effective ways to use DoG filtering for channel construction.

Acknowledgments

This work was supported in part by the U.S. National Institutes of Health (NIH) grant R35GM158094.

References

[1] Lu Gan and Grant J. Jensen. Electron tomography of cells. *Quarterly Reviews of Biophysics*, 45(1):27–56, 2012.

[2] Vladan Lučić, Alexander Rigort, and Wolfgang Baumeister. Cryo-electron tomography: the challenge of doing structural biology in situ. *The Journal of Cell Biology*, 202(3):407–419, 2013.

[3] Stephan Nickell, Friedrich Förster, Alexandros Linaroudis, William Del Net, Florian Beck, Reiner Hegerl, Wolfgang Baumeister, and Jürgen M. Plitzko. Tom software toolbox: acquisition and analysis for electron tomography. *Journal of Structural Biology*, 149(3):227–234, 2005.

[4] Friedrich Förster, Ohad Medalia, Nathan Zauberman, Wolfgang Baumeister, and Deborah Fass. Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proceedings of the National Academy of Sciences*, 102(13):4729–4734, 2005.

[5] Friedrich Förster, Sabine Pruggnaller, Anja Seybert, and Achilleas S. Frangakis. Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of Structural Biology*, 161(3):276–286, 2008. The 4th International Conference on Electron Tomography.

[6] Kazuyoshi Murata and Matthias Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(2):324–334, 2018.

[7] Siyuan Liu, Xuefeng Du, Rong Xi, Fuya Xu, Xiangrui Zeng, Bo Zhou, and Min Xu. Semi-supervised macromolecule structural classification in cellular electron cryo-tomograms using 3d autoencoding classifier, 09 2019.

[8] Tarun Gupta, Xuehai He, Mostofa Rafid Uddin, Xiangrui Zeng, Andrew Zhou, Jing Zhang, Zachary Freyberg, and Min Xu. Self-supervised learning for macromolecular structure classification based on cryo-electron tomograms. *Frontiers in Physiology*, Volume 13 - 2022, 2022.

[9] Sabhay Jain, Xingjian Li, and Min Xu. Knowledge transfer from macro-world to micro-world: enhancing 3d cryo-et classification through fine-tuning video-based deep models. *Bioinformatics*, 40(7):btae368, 06 2024.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[15] Xiangrui Zeng, Miguel Ricardo Leung, Tzviya Zeev-Ben-Mordehai, and Min Xu. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. *Journal of Structural Biology*, 202(2):150–160, May 2018.

[16] Chengqian Che, Ruogu Lin, Xiangrui Zeng, Karim Elmaaroufi, John Galeotti, and Min Xu. Improved deep learning-based macromolecules structure classification from electron cryo-tomograms. *Machine Vision and Applications*, 29(8):1227–1236, 2018.

[17] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, Wolfgang Baumeister, Tingying Peng, Benjamin D Engel, and Charles Kervrann. Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. *Nature Methods*, 18(11):1386–1394, 2021.

[18] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[19] Marcia Hon and Naimul Khan. Towards alzheimer's disease classification through transfer learning, 2017.

[20] Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. Epub 2019 Mar 29.

[21] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.

[22] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[23] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[24] Yuhui Zhang, Shih-Cheng Huang, Zhengping Zhou, Matthew P. Lungren, and Serena Yeung. Adapting pre-trained vision transformers from 2d to 3d through weight inflation improves medical image segmentation, 2023.

[25] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 207:187–217, 1980.

[26] Long Pei, Min Xu, Zachary Frazier, and Frank Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC Bioinformatics*, 17(1):405, 2016.

[27] Siyuan Liu, Xuefeng Du, Rong Xi, Fuya Xu, Xiangrui Zeng, Bo Zhou, and Min Xu. Semi-supervised macromolecule structural classification in cellular electron cryo-tomograms using 3d autoencoding classifier. 09 2019.

[28] Alex J Noble, Venkata P Dandey, Hui Wei, Julia Brasch, Jillian Chase, Priyamvada Acharya, Yong Zi Tan, Zhening Zhang, Laura Y Kim, Giovanna Scapin, Micah Rapp, Edward T Eng, William J Rice, Anchi Cheng, Carl J Negro, Lawrence Shapiro, Peter D Kwong, David Jeruzalmi, Amedee des Georges, Clinton S Potter, and Bridget Carragher. Routine single particle cryoem sample and grid characterization by tomography. *eLife*, 7:e34257, may 2018.

[29] H Berman, Talapady Bhat, P Bourne, Z. Feng, G Gilliland, H Weissig, and J Westbrook. The protein data bank and the challenge of structural genomics. (7), 2000-11-01 00:11:00 2000.

[30] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition, 2017.

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

A Ablation Study

Ablation	Model	Pretrain Dataset	Weight Inflation	3D-Convolution Kernal Size	Input Resolution	DoG Variances	Acc on Simulated	Acc on Real
Initialization Strategy	ViT-B	-	Random	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	68.30%	81.07%
	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.60%	87.86%
	ViT-B	ImageNet-21k	Average	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.40%	86.79%
	Swin-B	-	Random	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	80.70%	91.61%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.60%	95.36%
	Swin-B	ImageNet-22k	Average	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	90.00%	94.82%
Kernel Depth	ViT-B	ImageNet-21k	Center	$1 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.20%	86.96%
	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.60%	87.86%
	ViT-B	ImageNet-21k	Center	$5 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	84.90%	87.68%
	Swin-B	ImageNet-22k	Center	$1 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.30%	95.36%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.60%	95.36%
	Swin-B	ImageNet-22k	Center	$5 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.90%	94.29%
Kernel Size & Input Resolution	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.60%	87.86%
	ViT-B	ImageNet-21k	Center	$3 \times 16 \times 16$	$32 \times 128 \times 128$	$(0.5, 1.0), (0.5, 2.0)$	86.10%	97.32%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.60%	95.36%
	Swin-B	ImageNet-22k	Center	$3 \times 4 \times 4$	$32 \times 128 \times 128$	$(0.5, 1.0), (0.5, 2.0)$	89.70%	98.04%
DoG Variances	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	-	79.10%	85.54%
	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	85.60%	87.86%
	ViT-B	ImageNet-21k	Center	$3 \times 4 \times 4$	$32 \times 32 \times 32$	$(1.0, 2.0), (1.0, 4.0)$	84.30%	87.86%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	-	85.30%	92.68%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(0.5, 1.0), (0.5, 2.0)$	89.60%	94.46%
	Swin-B	ImageNet-22k	Center	$3 \times 1 \times 1$	$32 \times 32 \times 32$	$(1.0, 2.0), (1.0, 4.0)$	89.70%	95.36%

Table 2: Ablation results on simulated and real Cryo-ET subtomogram datasets. The table reports accuracy under different choices of initialization strategy, kernel depth, kernel size with input resolution, and DoG variance settings for ViT-B and Swin-B backbones.

Table 2 reports ablations across initialization strategy, kernel depth, kernel size with input resolution, and DoG variance choices. The ablation study suggests several trends. First, pretrained weight inflation substantially outperforms random initialization, confirming the benefit of transferring from large-scale image models. Second, kernel depth has only a modest effect on performance, with depth of 1, 3, and 5 yielding similar results, indicating that depth is not a critical factor for adaptation. Third, adjusting input resolution to better align with pretrained kernel sizes improves accuracy, particularly for Swin-B at higher resolutions. Finally, DoG-based channel augmentation consistently provides gains over raw subtomograms, while the specific variance pairs make only minor differences.

Additional training details. We train single NVIDIA RTX 3090 GPU with mixed precision: forward/backward under `torch.autocast` (FP16) and updates via `torch.amp.GradScaler` (enabled). Optimization uses AdamW with learning rate 5×10^{-5} for randomly initialized models and 2×10^{-5} for pretrained ones, β=$(0.9, 0.999)$, ϵ=10^{-8}, and weight decay 5×10^{-2}. A cosine annealing schedule (`CosineAnnealingLR`) is applied with $T_{\max}$=100 and $\eta_{\min}$=10^{-7}, stepped once per epoch after validation. We set the random seed to 26. The loss is standard cross-entropy.

For data preprocessing, each subtomogram volume is min–max normalized to $[-1, 1]$, replicated to three channels, and optionally converted to a DoG triplet using user-specified sigmas $\sigma_1, \sigma_2, \sigma_3, \sigma_4$: channel 1 is the raw volume, channel 2 is $\mathrm{DoG}(\sigma_1, \sigma_2)$, and channel 3 is $\mathrm{DoG}(\sigma_3, \sigma_4)$. Augmentations (applied only during training when enabled) comprise random 3D crop (`scale` $\in [0.5, 1]$), random flips along all axes, and random affine (isotropic scale ±0.1, rotation $\pm45^\circ$, translation ±3 voxels), each activated independently with probability 0.5.