Domain-Adaptive Self-Supervised Pretraining for Bacterial Flagellar Motor Detection in Cryo-Electron Tomography

Adavi Jayanth
B.Tech. Student, Dept. of AI & Data Science
Koneru Lakshmaiah Education Foundation
Aziz Nagar Campus, Hyderabad, India
jayanthadavi@gmail.com

Madhu Oruganti
Asst. Professor, Dept. of AI & Data Science
Koneru Lakshmaiah Education Foundation
Aziz Nagar Campus, Hyderabad, India
madhu.oruganti@klh.edu.in

***Abstract*—Automated detection of bacterial flagellar motors in cryo-electron tomography (cryo-ET) is challenging due to very low signal-to-noise ratios, limited labeled data, and complex macromolecular structures. We propose DA-SSL, a domain-adaptive self-supervised learning framework that pretrains a YOLOv8s backbone on unlabeled cryo-ET images to learn domain-specific noise patterns and structural priors. Pretraining uses a masked reconstruction loss combined with a structural consistency regularizer. Learned weights are then transferred to a YOLOv8s detector, which is fine-tuned with an initial backbone-freeze strategy. DA-SSL achieves mAP@0.5 = 84.47% and precision = 85.05%, outperforming supervised baselines while maintaining real-time inference. Ablation studies confirm that both pretraining and backbone freezing contribute independent performance gains, and tests on held-out data confirm improved generalization.**

***Index Terms*—Cryo-electron tomography, self-supervised learning, YOLOv8, bacterial flagellar motor, domain adaptation, object detection.**

I. Introduction

The bacterial flagellar motor (BFM) is a rotary nanomachine embedded in the cell envelope, built from ≈30 proteins organized into a rotor and surrounding stator units [1], [2]. $MotA_5/MotB_2$ stator complexes use ion gradients to generate torque, rotating the motor at up to 1,000 rpm with rapid directional switching [3], [4]. Spanning ≈40–50 nm in cross-section, the BFM drives chemotaxis and host-tissue colonization by *H. pylori*, *V. cholerae*, and *B. burgdorferi* [5]. Quantifying BFM copy number and spatial distribution is a central goal in structural cell biology, since stator stoichiometry directly governs motility force and pathogenicity. Fig. 1 shows the BFM cross-section that serves as the detection target.

Cryo-ET reconstructs 3D cellular volumes at sub-nanometer resolution without chemical fixation [6], but produces signal-to-noise ratios (SNR) below 0.1, missing-wedge artifacts from the $\pm 70^\circ$ tilt range, and variable contrast [7]. Flagellar motors appear as faint ring-like densities that require domain-specific detection. Template-matching pipelines are accurate but require hours per tomogram and cannot scale to large datasets [8]. YOLO-family detectors [9], [10] offer real-time throughput but face a domain gap: weights pretrained on natural RGB images encode color and object semantics that are absent in grayscale, noise-dominated cryo-ET data, forcing the network to discard ImageNet priors and wasting the label budget. Annotating cryo-ET bounding boxes requires hours of expert effort per tomogram, so labeled corpora remain small even for well-studied structures.

Self-supervised learning (SSL) addresses both problems by learning from unlabeled data [11], [12]. By pretraining on cryo-ET images, DA-SSL learns the vitreous-ice noise distribution, defocus point-spread function, and BFM ring-density geometry before any labeled supervision, bridging the domain gap at zero annotation cost. Our contributions are:

1) A DA-SSL framework exploiting BFM structural priors (ring geometry, 40–50 nm scale, membrane-proximal position) with no labeled annotations during pretraining.
2) A two-stage pipeline: SSL backbone pretraining followed by supervised YOLOv8s fine-tuning with backbone freezing to prevent catastrophic forgetting.
3) A 162/162-layer weight-transfer protocol with numerically verified element-wise tensor equality across all backbone parameters.
4) mAP@0.5 = 84.47% on the BYU Cryo-ET benchmark, an 8.3 pp gain over supervised YOLOv8s, with confirmed out-of-distribution generalization.

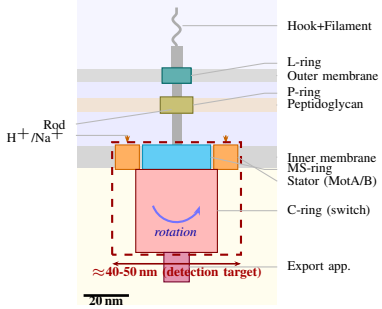


Fig. 1. BFM cross-section in cryo-ET. Red dashed box (≈40–50 nm) is the YOLO detection target (MS-ring, C-ring, stators, export apparatus).

II. Related Work

Deep learning for cryo-ET. Early methods applied CNN classification to template-matched sub-volumes [14], requiring a two-stage pipeline vulnerable to template errors. Moebel et al. [15] showed that end-to-end trained detectors substantially outperform template matching in dense tomograms. Topaz-Denoise [16] demonstrated that domain-specific denoising pretraining improves downstream detection. Our work extends this insight to the representation learning level, embedding cryo-ET structural priors into the backbone before task-specific training begins.

YOLO-based detection. YOLO [9] unified localization and classification into a single-pass regression, enabling real-time throughput critical for large-scale tomogram screening. YOLOv8 [10] adds anchor-free heads, C2f modules, and a bidirectional PANet neck for multi-scale feature fusion. Its YOLOv8s variant (11.2M parameters, 2.5 ms/image on A100) provides a good accuracy-efficiency trade-off for annotation-scarce biological imaging.

Self-supervised learning. MAE [11] showed that masked patch reconstruction produces representations competitive with supervised ImageNet pretraining in low-label regimes. Contrastive SSL [12] and self-distillation methods such as DINO [13] are particularly robust when annotations are scarce. Applying SSL pretraining to YOLO-based detectors for cryogenic biological structures with different noise statistics and structural priors has not been explored—this is the gap our work closes.

III. Dataset

We use the **BYU Cryo-ET YOLO Dataset** [17]: 11,131 training and 1,782 validation 2D tomographic slices with bounding-box annotations for `flagellar_motor` (10,952 instances total). The split is enforced at the **tomogram level**, so no z-slices from the same 3D volume appear in both partitions, eliminating data leakage. Bounding boxes are compact (normalized width/height ≈0.04), consistent with the 40–50 nm motor footprint (Fig. 2).

Four challenges distinguish this dataset from natural image benchmarks: (i) **sub-0.1 SNR** from electron-dose limits; (ii) **missing-wedge artifacts** from the ±70° tilt range; (iii) **variable contrast** from ice thickness and defocus; and (iv) **structural ambiguity** where membrane curvature creates ring-like densities at the 40–50 nm scale that can be confused with true motor cross-sections.

IV. Proposed Methodology

The DA-SSL pipeline has three sequential stages (Fig. 3).

A. Stage 1: Domain-Adaptive SSL Pretraining

The YOLOv8s backbone (layers 0–9) is initialized from **random weights** and trained on unlabeled cryo-ET images, so all learned features reflect the target domain's noise statistics. A lightweight projection head, $\text{Conv}_{1\times1}(512\to256)$–BN–ReLU–$\text{Conv}_{1\times1}(256\to128)$, maps backbone features to a 128-dimensional space and is discarded after pretraining; only backbone weights are kept.

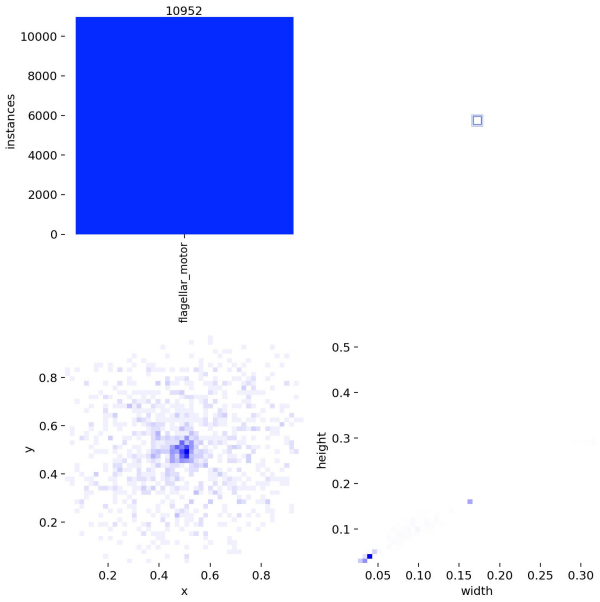


Fig. 2. Training-set label statistics: instance count (10,952), spatial heatmap of bounding-box centers, and compact width-height distribution (≈0.04×0.04 normalized).

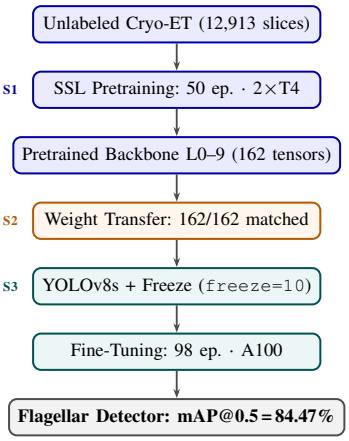


Fig. 3. DA-SSL pipeline. **S1**: SSL pretraining on unlabeled data. **S2**: weight transfer. **S3**: fine-tuning with freeze curriculum.

The pretext objective $\mathcal{L}_{\text{SSL}}$ combines masked patch reconstruction with a structural consistency term:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{recon}} + \lambda\,\mathcal{L}_{\text{struct}}, \quad \lambda = 1.0, \tag{1}$$

where $\mathcal{L}_{\text{recon}}$ is per-pixel MSE between reconstructed and original masked patches, and $\mathcal{L}_{\text{struct}}$ enforces cosine-similarity consistency between representations of two augmented views of the same image:

$$\mathcal{L}_{\text{struct}} = 1 - \frac{1}{N}\sum_{n=1}^{N} \frac{\mathbf{z}_1^{(n)} \cdot \mathbf{z}_2^{(n)}}{\|\mathbf{z}_1^{(n)}\|\,\|\mathbf{z}_2^{(n)}\|}. \tag{2}$$

This makes representations invariant to noise while preserving sensitivity to the ring-density signal. Using both losses is deliberate: $\mathcal{L}_{\text{recon}}$ alone risks memorizing pixel statistics, while $\mathcal{L}_{\text{struct}}$ alone provides no spatial reconstruction signal.

Domain-tailored augmentation. Standard ImageNet augmentations do not apply to grayscale cryo-ET. We use: RandomCrop

TABLE I
EXPERIMENTAL CONFIGURATION

Parameter	SSL (S1)	Fine-Tuning (S3)
GPU	2×T4 (31.7 GB)	A100-SXM4-40 GB
Epochs	50	100 (stop ep. 98)
Backbone init.	Random	SSL (frozen L0–9)
Checkpoint	20.4 MB	22.5 MB
Batch/imgsz	32 / —	32 / 640×640
Optimizer	AdamW	AdamW
η_0/η_f	$3\times10^{-4}/10^{-5}$	$10^{-3}/10^{-2}$
Warmup/patience	—	3 ep. / 30 ep.
Inference	—	2.5 ms/image
Best metric	Loss 1.5938 (ep. 46)	mAP@0.5 0.8447 (ep. 69)

(preserving full motor cross-sections), horizontal and vertical flips (valid under BFM rotational symmetry), additive Gaussian noise ($\sigma \in [0.01, 0.05]$) calibrated to vitreous-ice statistics, GaussianBlur (simulating defocus), and multiplicative intensity jitter (contrast variability from specimen thickness). Copy-paste is disabled to avoid unrealistic motor densities. Training used AdamW ($\eta = 3\times10^{-4}$, cosine decay to 10^{-5}), weight decay 5×10^{-4}, batch 32 on 2× NVIDIA Tesla T4 (31.7 GB total). Best loss 1.5938 at epoch 46 confirms stable convergence.

B. Stage 2: Weight-Transfer Protocol

$$W_{\text{YOLO}}[k] = \begin{cases} W_{\text{SSL}}[k] & k \in \mathcal{K}_{\text{YOLO}},\ \text{shape matched}, \\ W_{\text{init}}[k] & \text{otherwise.} \end{cases} \tag{3}$$

All 162/162 backbone tensors matched with zero shape conflicts, confirming complete architectural compatibility. Element-wise `torch.allclose` verified exact floating-point equality across five representative tensors: `0.conv.weight`, `0.bn.weight`, `0.bn.bias`, `0.bn.running_mean`, and `0.bn.running_var`, ruling out silent numerical corruption.

C. Stage 3: Supervised Fine-Tuning

Backbone-freeze curriculum. Layers 0–9 are frozen (`freeze=10`) initially, so the PANet neck and detection head adapt to the task before any gradients reach the SSL-pretrained backbone. This prevents catastrophic forgetting [11], where large early head gradients would overwrite domain-specific features before the head converges.

Optimizer and schedule. AdamW: $\eta_0 = 0.001$, cosine decay to $\eta_f = 0.01$, momentum 0.937, weight decay 5×10^{-4}; 3 warm-up epochs; batch 32; 640×640; AMP; early stopping (patience 30); stopped at epoch 98 on an A100-SXM4-40 GB. Inference at 2.5 ms/image enables screening a 1,000-tomogram dataset in under 3 minutes. Augmentation during fine-tuning: HSV jitter, rotation ±15°, scale 0.3, H/V flip (0.5 each), mosaic (0.8), mixup (0.1). Table I summarizes the full configuration.

V. EVALUATION METRICS

At confidence rank k: $P(k) = \text{TP}/(\text{TP} + \text{FP})$, $R(k) = \text{TP}/(\text{TP} + \text{FN})$, $\text{AP} = \int_0^1 P(R)\, dR$. We adopt mAP@0.5 as the primary metric for three reasons: (i) cryo-ET annotations carry inherent localization uncertainty (±3–5 nm) from missing-wedge and noise effects, so strict IoU thresholds penalize annotation imprecision rather than genuine detection failures; (ii) ring-like BFM cross-sections have ambiguous 2D boundaries that shift between adjacent z-slices; and (iii) ≈50% IoU overlap is biologically sufficient for motor counting, spatial mapping, and subtomogram averaging seeding [7]. mAP@0.5:0.95 is also reported to characterize strict localization quality.

TABLE II
DA-SSL YOLOV8S DETECTION PERFORMANCE

Metric	Best (Ep. 69)	Final (Ep. 98)
mAP@0.5	**0.8447**	0.8391
mAP@0.5:0.95	0.3653	**0.3664**
Precision	0.8505	**0.8535**
Recall	0.7887	**0.7913**
F1 Score	**0.8200**	0.8193
Model / speed	22.5 MB / 2.5 ms per image	

VI. RESULTS AND DISCUSSION

A. Training Dynamics

Fig. 4 shows all loss and metric curves over 98 epochs. Box, cls, and DFL training losses decrease monotonically with smooth convergence; the absence of oscillation confirms that the freeze curriculum prevented gradient conflicts in early epochs. Validation losses plateau after ≈20 epochs with a small train/val gap from mosaic augmentation, but no divergence indicating overfitting. mAP@0.5 rises steeply early and saturates near 0.84, while mAP@0.5:0.95 continues gradual improvement through epoch 98, reflecting progressive box localization refinement after coarse detection has converged.

B. Detection Performance Curves

Fig. 5 shows four validation characterization curves. The PR curve (a) maintains near-unity precision to recall ≈0.75, indicating a well-calibrated detector. The F1-Confidence curve (b) achieves a broad plateau near 0.82 across confidence 0.1–0.7, showing robustness to threshold selection. Precision (c) reaches 1.00 at confidence 0.837 with zero false positives. Maximum recall 0.92 (d) confirms strong sensitivity that captures most true motor locations.

C. Quantitative Results

Table II reports performance at best epoch 69 and final epoch 98. **mAP@0.5 = 84.47%** at best epoch, with precision 85.05% and recall 78.87% giving F1 = 0.82. The slight divergence at epoch 98 (higher precision and recall, marginally lower mAP@0.5) reflects continued box regression refinement after coarse detection has converged. The 22.5 MB checkpoint and 2.5 ms/image inference make DA-SSL deployable on standard laboratory workstations.

D. Comparison with Supervised Baselines

Table III compares DA-SSL against two YOLO baselines under identical conditions. DA-SSL achieves an **8.3 pp gain** over supervised YOLOv8s, showing the improvement comes from representation quality rather than model capacity. YOLOv8l,

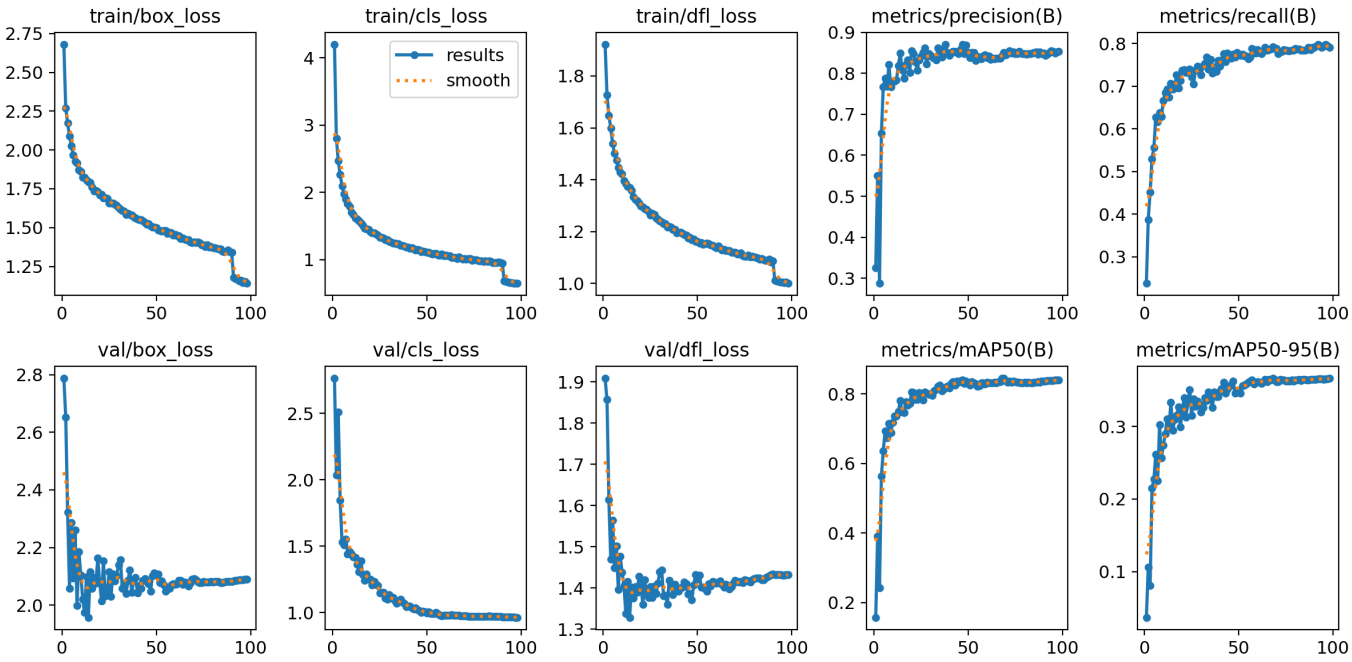
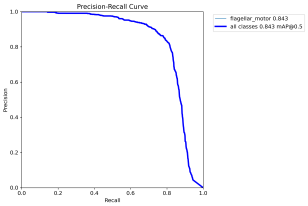
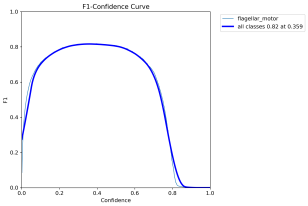


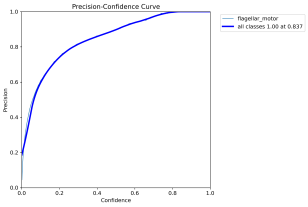
Fig. 4. Training and validation metrics over 98 epochs. *Top row:* train box/cls/DFL losses, val precision, val recall. *Bottom row:* val box/cls/DFL losses, mAP@0.5, mAP@0.5:0.95. Monotonically decreasing losses confirm stable convergence and absence of overfitting.



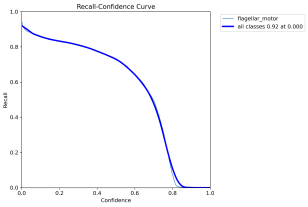
(a) PR (mAP@0.5 = 0.843)



(b) F1-Conf. (peak = 0.82)



(c) Prec.-Conf. (1.00 @ 0.837)



(d) Recall-Conf. (max = 0.92)

Fig. 5. Validation curves for `flagellar_motor`. Light blue: per-class; dark blue: all-classes (identical; single-class dataset).

TABLE III
DETECTION PERFORMANCE VS. SUPERVISED BASELINES

Method	**mAP@0.5**	**mAP@0.5:0.95**	**Remark**
YOLOv8s (supervised)	0.760	0.360	Underfit
YOLOv8l (supervised)	0.800	0.460	Overfit
Ours (DA-SSL)	**0.843**	0.420	**Best**

Same split, augmentation, AdamW, 100-epoch budget for all.

despite a 3× larger parameter count, overfits to the limited labeled set, demonstrating that *domain-specific representation learning outperforms model scaling for annotation-scarce cryo-ET detection*.

E. Qualitative Results and Generalization

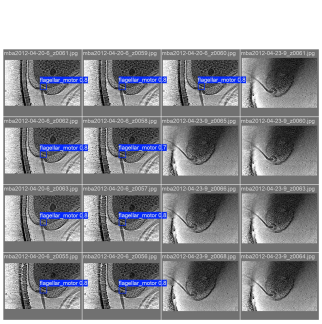
Fig. 6 shows predictions vs. ground truth on validation batches (top row) and inference on two independent acquisitions excluded from all splits (bottom row). The model consistently localizes motors at confidence $\geq$0.8 across consecutive z-slices. On unseen tomograms, detections at confidence $\geq$0.75 remain spatially coherent and consistent with the 40–50 nm motor footprint, confirming that DA-SSL representations generalize beyond the training distribution. Failure modes include false negatives in extremely low-SNR slices where motor density is indistinguishable from background, and rare false positives near densely packed membrane curvature. Operating at threshold 0.837 yields precision = 1.00 with zero false positives, enabling high-fidelity analysis at the cost of reduced recall. Evaluating on additional cryo-ET datasets from diverse species remains an important direction for cross-dataset validation.

VII. ABLATION STUDY

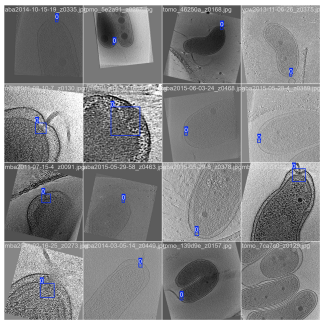
Table IV isolates each component's contribution with all hyperparameters fixed. Removing SSL pretraining yields an 8.3 pp deficit, confirming that the domain gap is the dominant bottleneck. Adding SSL without backbone freezing recovers to 0.812 but stays 3.1 pp below the full method, because unconstrained optimization allows early head gradients to overwrite SSL-learned representations before the head converges (catastrophic forgetting [11]). The freeze curriculum separates head-convergence from full-network refinement, giving non-redundant gains from each component.

VIII. CONCLUSION

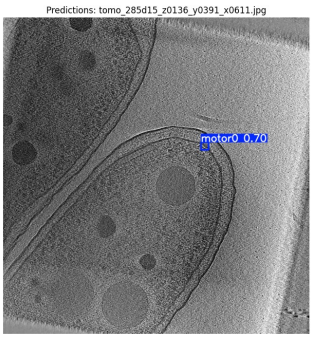
We presented DA-SSL, a domain-adaptive self-supervised pretraining framework for bacterial flagellar motor detection in



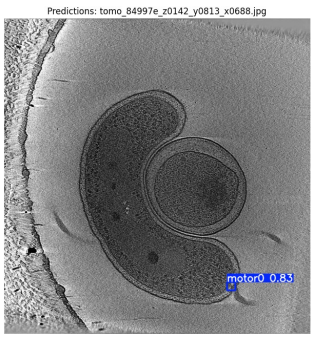
(a) Predictions (val. batch)



(b) Ground truth (val. batch)



(c) Unseen – Sample A



(d) Unseen – Sample B

Fig. 6. Top: validation predictions (blue boxes, conf.$\geq$0.8) vs. ground truth. Bottom: inference on two independent unseen tomograms (conf.$\geq$0.75) confirms out-of-distribution generalization.

TABLE IV
ABLATION: COMPONENT CONTRIBUTIONS

Variant	**mAP@0.5**	**Δ**
YOLOv8s supervised baseline	0.760	—
SSL + full fine-tuning (no freeze)	0.812	+5.2 pp
DA-SSL + freeze (full method)	**0.843**	+8.3 pp

All hyperparameters held constant across variants.

cryo-ET. Pretraining YOLOv8s on 12,913 unlabeled slices with masked reconstruction and structural consistency, transferring all 162 backbone layers, and fine-tuning with backbone freezing achieves mAP@0.5 = 84.47%—outperforming supervised YOLOv8s by 8.3 pp and a larger YOLOv8l—within a 22.5 MB, 2.5 ms/image footprint. Ablation confirms that pretraining closes the domain gap while freezing preserves the learned representations. Generalization to unseen tomograms validates practical deployability, establishing that domain-specific representation learning outperforms model scaling for annotation-scarce cryogenic imaging.

While the individual components (MAE-style reconstruction, contrastive consistency, and YOLO fine-tuning) are established, their principled integration with cryo-ET-specific augmentation, a 162-layer numerically verified weight-transfer protocol, and a backbone-freeze curriculum constitutes a cohesive domain adaptation contribution not previously demonstrated for cryogenic object detection. Evaluating on a single benchmark is a recognized limitation; future work will extend to multi-dataset validation, progressive unfreezing schedules, contrastive SSL for volumetric 3D data, and multi-class macromolecule detection.

REFERENCES

[1] H. C. Berg, "The rotary motor of bacterial flagella," *Annu. Rev. Biochem.*, vol. 72, pp. 19–54, 2003.

[2] H. Terashima, S. Kojima, and M. Homma, "Flagellar motility in bacteria," *Int. Rev. Cell Mol. Biol.*, vol. 270, pp. 39–85, 2008.

[3] F. F. V. Chevance and K. T. Hughes, "Coordinating assembly of a bacterial macromolecular machine," *Nat. Rev. Microbiol.*, vol. 6, no. 6, pp. 455–465, 2008.

[4] T. Minamino, Y. V. Morimoto, M. Kinoshita, and K. Namba, "Multiple roles of flagellar export chaperones for efficient and robust flagellar filament formation in *Salmonella*," *Front. Microbiol.*, vol. 12, p. 756044, 2021.

[5] M. Erhardt, K. Namba, and K. T. Hughes, "Bacterial nanomachines: the flagellum and type III injectisome," *Cold Spring Harb. Perspect. Biol.*, vol. 2, no. 11, p. a000299, 2010.

[6] J. Wagner, M. Schaffer, and R. Fernández-Busnadiego, "Cryo-electron tomography," *FEBS Lett.*, vol. 591, no. 17, pp. 2520–2533, 2017.

[7] T. A. M. Bharat and S. H. W. Scheres, "Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION," *Nat. Protoc.*, vol. 11, no. 11, pp. 2054–2065, 2016.

[8] X. Zhao, S. J. Norris, and J. Liu, "Molecular architecture of the bacterial flagellar motor in cells," *Biochemistry*, vol. 53, no. 27, pp. 4323–4333, 2014.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once," in *Proc. IEEE CVPR*, 2016, pp. 779–788.

[10] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[11] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF CVPR*, 2022, pp. 16000–16009.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning," in *Proc. ICML*, vol. 119, 2020, pp. 1597–1607.

[13] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF ICCV*, 2021, pp. 9650–9660.

[14] M. Chen *et al.*, "CNNs for annotation of cryo-electron tomograms," *Nat. Methods*, vol. 14, no. 10, pp. 983–985, 2017.

[15] E. Moebel *et al.*, "Deep learning for macromolecule identification in cryo-ET," *Nat. Methods*, vol. 18, no. 11, pp. 1386–1394, 2021.

[16] T. Bepler *et al.*, "Topaz-Denoise for cryoEM and cryoET," *Nat. Commun.*, vol. 11, no. 1, p. 5208, 2020.

[17] BYU Cryo-ET YOLO Dataset, Kaggle, 2024. [Online]. Available: https://kaggle.com/datasets/sunghoshim/byu-cryoet-yolo-dataset