SPARSE2GEO: PREDICTION-DERIVED GEOMETRY FOR SPARSELY SUPERVISED SEGMENTATION AND REGIONAL QUALITY ASSESSMENT IN VOLUME ELECTRON MICROSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Volume electron microscopy (vEM) enables nanometer-scale three-dimensional (3D) imaging of cellular ultrastructure, but semantic segmentation remains annotation-intensive. Sparse supervision reduces annotation cost, yet supervision confined to sparsely sampled locations provides limited constraints on object extent, boundaries, and 3D structure. Signed distance fields (SDFs) continuously encode object interior, exterior, and distance to the boundary, but are difficult to construct directly from sparse labels. Regional segmentation quality also remains difficult to assess without ground truth. To address both challenges, we introduce Sparse2Geo, a unified framework that exploits prediction-derived geometry for both sparsely supervised segmentation and regional quality assessment. In Stage I, Sparse2Geo constructs a teacher SDF from 3D consensus geometry aggregated from fixed full-volume predictions of an initial sparsely supervised model, providing geometric and boundary-aware supervision without additional dense semantic targets. In Stage II, local predictions are re-encoded as normalized local SDFs and combined with aligned raw volumetric patches to estimate regional segmentation quality without ground truth at inference. Our experiments use volumes from BetaSeg and UroCell, two public vEM datasets spanning mouse pancreatic β cells and urinary bladder urothelial cells, comprising approximately 2.32 billion and 67 million voxels, respectively. Sparse2Geo achieves the highest Macro Dice among the compared segmentation methods (90.63% and 71.44%), improving over ϵ-Seg by 4.13 and 15.33 percentage points, respectively. For regional quality assessment, Sparse2Geo achieves the lowest Macro MAE on BetaSeg, reducing it by 12.3% relative to the best adapted external baseline, while incorporating local SDF geometry reduces Macro MAE by 35.1% relative to Raw-only on UroCell.

1 INTRODUCTION

Volume electron microscopy (vEM) enables nanometer-scale three-dimensional (3D) imaging of cellular ultrastructure, providing a powerful basis for quantitative studies across biological scales, ranging from synaptic connectivity to intracellular organelle morphology and spatial organization (Takemura et al., 2013; Midorikawa et al., 2022). Extracting biologically meaningful information from these large volumes depends on reliable semantic segmentation, but dense voxel-level annotation remains labor-intensive, limiting the scalability of supervised segmentation.

Sparse supervision reduces annotation burden but weakly constrains object extent, boundary localization, and 3D morphological continuity when labels are restricted to sparsely sampled locations (Machireddy et al., 2023; Rahnamai Kordasiabi et al., 2025). Signed distance fields (SDFs) provide continuous boundary-distance geometry (Xue et al., 2020), yet require sufficiently complete foreground–background structure that sparse labels do not provide. At deployment, ground truth is also unavailable for directly measuring regional segmentation quality, while discrete prediction masks lack explicit continuous distance-to-boundary geometry for characterizing subtle local errors.

Existing methods address individual aspects of these challenges. Pseudo-labeling propagates discrete semantic predictions into unlabeled regions (Takaya et al., 2021), but does not provide contin-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

uous boundary-distance constraints. u-Segment3D aggregates segmentations from orthogonal views to obtain 3D consensus geometry (Zhou et al., 2025), while Spatial-SAM incorporates SDF-based spatial representations for 3D segmentation (Huang et al., 2026). These geometry-aware approaches demonstrate the value of structural information, but do not formulate prediction-derived geometry as supervision for sparsely supervised segmentation. For quality assessment, Robinson et al. (2018) directly regress segmentation quality from image and segmentation information, whereas GANREC–REGNET (Zaman et al., 2023) exploits reconstruction cues for quality estimation. Neither explicitly re-encodes the candidate prediction into continuous boundary-distance geometry. Together, these limitations leave an under-explored need for a common geometric representation that supports both sparse segmentation learning and regional quality assessment.

To address this gap, we introduce **Sparse2Geo**, which uses prediction-derived geometry for both tasks. Stage I derives a teacher SDF from 3D consensus geometry to provide geometric supervision without additional dense semantic targets, while Stage II combines normalized local SDFs with aligned raw volumetric patches for quality assessment without ground truth at inference. Across approximately 2.32 billion BetaSeg voxels and 67 million UroCell voxels, Sparse2Geo improves Macro Dice over ϵ-Seg by 4.13 and 15.33 percentage points, respectively, while reducing Macro MAE by 12.3% relative to the best adapted external baseline on BetaSeg and by 35.1% relative to Raw-only on UroCell. These results highlight prediction-derived geometry as a common geometric basis for reliable vEM analysis. In summary, our main contributions are:

1. We propose **Sparse2Geo**, a unified framework that uses prediction-derived geometry for sparsely supervised segmentation and regional quality assessment in vEM.
2. We derive 3D consensus geometry and a teacher SDF from fixed full-volume predictions to provide geometric supervision through SDF regression and boundary-region consistency without additional dense semantic targets.
3. We re-encode fixed local predictions as normalized local SDFs and combine them with aligned raw volumetric patches for regional quality assessment without ground truth at inference.

2 Related Work

2.1 Sparse Supervision in Microscopy Segmentation

Sparse supervision reduces annotation cost by using limited labels rather than dense masks. Machireddy et al. (2023) show that deep learning models segment 3D FIB-SEM ultrastructures from sparse manual labels, with boundary propagation for cell separation. ϵ-Seg (Rahnamai Kordasiabi et al., 2025) learns semantic representations from extremely sparse labels through hierarchical variational modeling, inpainting, and sparse-label contrastive learning. However, neither approach constructs prediction-derived 3D boundary-distance supervision across the full volume. 4S (Takaya et al., 2021) expands limited annotations through repeated training, inference, and pseudo-labeling, but propagates discrete semantic targets rather than continuous geometric constraints. These differences motivate prediction-derived continuous geometry as supervision for sparse segmentation.

2.2 Geometry-Aware Segmentation and Distance-Field Representations

Geometry-aware segmentation introduces structural information beyond discrete semantic labels. Xue et al. (2020) jointly predict segmentation and signed distance maps for shape-aware predictions, while Luo et al. (2021) enforce consistency between segmentation and a geometry-aware level-set representation in semi-supervised learning. These methods incorporate geometry within segmentation learning rather than deriving supervision from fixed predictions of a sparsely supervised model. u-Segment3D (Zhou et al., 2025) aggregates 2D segmentations from orthogonal views into a consensus 3D segmentation, demonstrating the value of cross-view geometric integration, but focuses on consensus construction rather than geometric supervision for sparse learning. Spatial-SAM (Huang et al., 2026) further introduces 3D SDF memory for spatial guidance in 3D electron microscopy segmentation, but uses geometry as a prior rather than as prediction-derived supervision for sparse segmentation learning. Together, these differences motivate prediction-derived geometric supervision from fixed full-volume predictions.

2.3 Segmentation Quality Assessment and Error Detection

Segmentation quality assessment aims to identify unreliable predictions when ground truth is unavailable at inference. Roy et al. (2019) use Bayesian model uncertainty to derive structure-wise quality measures, providing an indirect proxy for segmentation accuracy rather than explicitly modeling prediction geometry. Robinson et al. (2018) directly predict Dice scores from image and segmentation inputs, while GANREC–REGNET (Zaman et al., 2023) combines reconstruction error with segmentation information for patch-wise quality estimation, but neither explicitly encodes continuous boundary-distance geometry. More recently, SegQC (Specktor-Fadida et al., 2025) predicts voxelwise segmentation errors from image–mask inputs and derives both quality metrics and error regions, but retains a discrete-mask representation of the candidate prediction. Together, these approaches lack explicit boundary-distance encoding of prediction geometry, motivating prediction-derived geometry for regional quality assessment.

3 Method

As shown in Figure 1, **Sparse2Geo** consists of two stages connected through prediction-derived geometry. Given a vEM volume X with predefined sparse semantic supervision, Stage I derives a teacher SDF from fixed full-volume predictions and uses it to supervise a separately initialized final segmenter through SDF regression and boundary-region consistency. Only the predefined sparse locations provide semantic targets, while unexposed dense class values are not used to construct prediction-derived geometric supervision. Stage II then re-encodes the fixed final prediction as normalized local SDFs and combines them with aligned raw volumetric patches for regional quality assessment. Ground truth is used only to construct regional quality targets after regions are fixed and is not required at inference. Thus, prediction-derived geometry serves as training-time supervision in Stage I and as an input representation in Stage II.

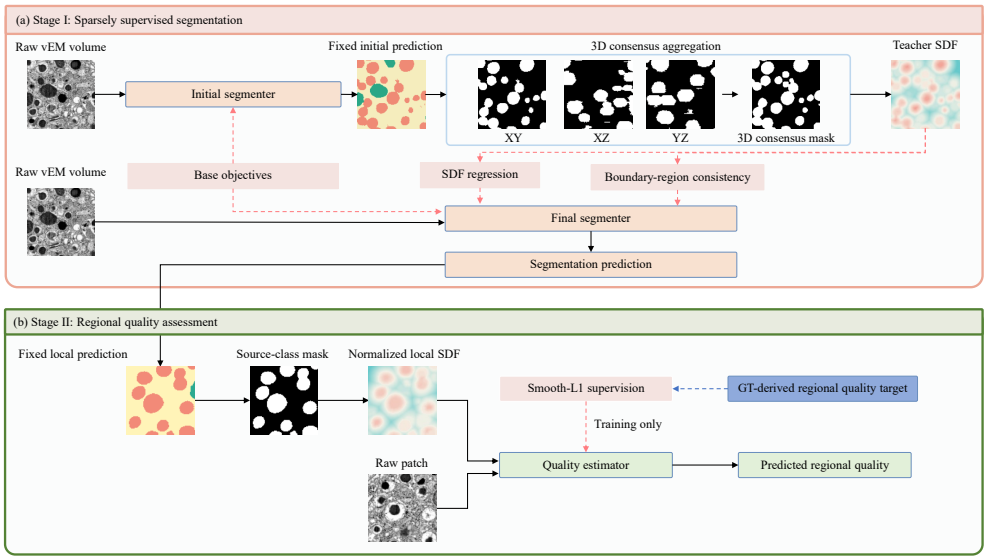

Figure 1: **Overview of Sparse2Geo.** (a) An initial segmenter maps raw vEM input to a fixed full-volume prediction. Its target-class mask undergoes XY/XZ/YZ processing and aggregation to form a 3D consensus mask, from which a teacher SDF is derived. A separately initialized final segmenter takes raw vEM input and produces the final segmentation prediction, using the base objectives and supervision derived from the teacher SDF during training. (b) Local crops of the fixed final prediction are converted into source-class masks and then normalized local SDFs. Each SDF is concatenated with its aligned raw volumetric patch and fed to the quality estimator to produce a regional quality estimate. Ground-truth-derived regional quality targets are used for training and evaluation, but are not estimator inputs or required at inference. Solid black arrows denote data and inference flow, coral dashed arrows training-time supervision, and blue dashed arrows ground-truth-derived target flow.

3.1 STAGE I: PREDICTION-DERIVED GEOMETRIC SUPERVISION

Stage I consists of a sparse semantic segmentation model, prediction-derived consensus geometry, teacher SDF supervision, and boundary-region segmentation–geometry consistency. The sparse semantic segmentation model is first trained with the base objectives and then frozen as the initial segmenter, which produces fixed full-volume predictions. Consensus geometry and a teacher SDF are then derived from these predictions. The teacher SDF provides geometric supervision for the final segmenter through SDF regression and boundary-region consistency.

Sparse semantic segmentation model. We build on the two-dimensional hierarchical variational backbone of ϵ-Seg (Rahnamai Kordasiabi et al., 2025) by attaching a dense semantic prediction head, with semantic cross-entropy applied only at supervised locations. This head forms the semantic output path of Sparse2Geo and is used in the controlled baseline, while Sparse2Geo uses an auxiliary SDF head and the geometric objectives introduced below without additional dense semantic targets.

Let Ω_s denote the set of supervised patch–location pairs in a training batch. Given semantic logits $\mathbf{z}_\theta^{(i)}(\mathbf{u})$ and the corresponding label $y^{(i)}(\mathbf{u})$,

$$\mathcal{L}_{\text{seg}} = \frac{1}{|\Omega_s|} \sum_{(i,\mathbf{u})\in\Omega_s} \ell_{\text{CE}} \left(\mathbf{z}_\theta^{(i)}(\mathbf{u}), y^{(i)}(\mathbf{u}) \right), \tag{1}$$

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{seg}} + \alpha\mathcal{L}_{\text{IP}} + \beta\mathcal{L}_{\text{KL}} + \gamma\mathcal{L}_{\text{CL}}, \tag{2}$$

where ℓ_{CE} denotes semantic cross-entropy, $\mathcal{L}_{\text{IP}}$, $\mathcal{L}_{\text{KL}}$, and $\mathcal{L}_{\text{CL}}$ are the inpainting, KL regularization, and contrastive objectives retained from ϵ-Seg, with weights α, β, and γ. We train the base model with $\mathcal{L}_{\text{base}}$ and then freeze it as the initial segmenter f_{θ_0}.

Prediction-derived consensus geometry. For training volume X, the frozen initial segmenter produces a fixed full-volume prediction $\hat{Y}^{(0)} = f_{\theta_0}(X)$. For target class $k^\star$, we define $B_{k^\star}(\mathbf{v}) = \mathbf{1}_{\hat{Y}^{(0)}(\mathbf{v})=k^\star}$. We re-slice this binary mask in the XY, XZ, and YZ orientations, apply slice-wise connected-component labeling and small-component filtering, and aggregate the stacks into

$$C_{k^\star} = \mathcal{A}\left(\mathcal{R}_{xy}(B_{k^\star}), \mathcal{R}_{xz}(B_{k^\star}), \mathcal{R}_{yz}(B_{k^\star})\right), \tag{3}$$

where $\mathcal{R}_a$ denotes orientation-specific re-slicing, component labeling, and filtering. We implement $\mathcal{A}$ using indirect 2D-to-3D aggregation from u-Segment3D (Zhou et al., 2025). The consensus mask is derived entirely from fixed predictions and used only for training-time geometric supervision.

Signed-distance geometric supervision. Let $\mathcal{C}_{k^\star} = \{\mathbf{v} : C_{k^\star}(\mathbf{v}) = 1\}$. The corresponding clipped teacher SDF is

$$\phi_{k^\star}(\mathbf{v}) = \text{clip}_{[-\tau,\tau]}\left[d(\mathbf{v}, \mathcal{C}_{k^\star}^{\text{c}}) - d(\mathbf{v}, \mathcal{C}_{k^\star})\right], \tag{4}$$

where $d(\mathbf{v}, \mathcal{S})$ denotes voxel-space Euclidean distance and $\tau > 0$ is the Stage I clipping radius. Under this convention, values are positive inside the consensus mask and negative outside.

The separately initialized final segmenter f_θ includes an auxiliary single-channel SDF head. Let $\mathbf{c}_i$ be the volume-space anchor of patch x_i and $\mathcal{P}_{k^\star}^+ = \{i : C_{k^\star}(\mathbf{c}_i) = 1\}$. Here, ℓ_{SL1} denotes Smooth-L1, defined for residual e as $e^2/(2\kappa)$ when $|e| < \kappa$ and $|e| - \kappa/2$ otherwise. For $\mathcal{P}_{k^\star}^+ \neq \emptyset$,

$$\mathcal{L}_{\text{sdf}} = \frac{1}{|\mathcal{P}_{k^\star}^+|} \sum_{i\in\mathcal{P}_{k^\star}^+} \ell_{\text{SL1}} \left(\hat{\phi}_\theta(x_i) - \phi_{k^\star}(x_i) \right), \tag{5}$$

where $\hat{\phi}_\theta(x_i)$ and $\phi_{k^\star}(x_i)$ are the predicted and aligned target SDF patches. The loss is applied voxelwise with mean reduction, and the Stage I transition parameter is reported in Appendix A.2. We set $\mathcal{L}_{\text{sdf}} = 0$ when $\mathcal{P}_{k^\star}^+ = \emptyset$.

Boundary-region segmentation–geometry consistency. Let $\phi_{k^\star}^{(i)}(\mathbf{u})$ denote the aligned teacher SDF value at location $\mathbf{u}$ in patch i. We define

$$s_{k^\star}^{(i)}(\mathbf{u}) = \sigma\left(\frac{\phi_{k^\star}^{(i)}(\mathbf{u})}{T}\right), \tag{6}$$

$$\Omega_B = \left\{(i,\mathbf{u}) : \mathbf{u} \in \Omega_{\text{valid}}^{(i)},\ \left|\phi_{k^\star}^{(i)}(\mathbf{u})\right| \leq \delta\right\}, \tag{7}$$

where σ is the logistic sigmoid, T controls the soft transition, δ is the boundary-band width, and $\Omega^{(i)}_{\text{valid}}$ denotes the set of valid, non-ignored voxels.

For $\Omega_B \neq \emptyset$, the consistency loss is

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\Omega_B|} \sum_{(i,\mathbf{u})\in\Omega_B} \ell_{\text{BCE}} \left(z^{(i)}_{\theta,k^\star}(\mathbf{u}), s^{(i)}_{k^\star}(\mathbf{u}) \right), \tag{8}$$

where $z^{(i)}_{\theta,k^\star}(\mathbf{u})$ is the target-class semantic logit and ℓ_{BCE} is binary cross-entropy with logits. We set $\mathcal{L}_{\text{cons}} = 0$ when $\Omega_B = \emptyset$. Within this consistency loss, manual annotations are used only to determine valid voxels, while the consistency target is derived entirely from the teacher SDF.

Overall objective and inference. The final Stage I objective is

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{base}} + \lambda_{\text{sdf}}\mathcal{L}_{\text{sdf}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}. \tag{9}$$

Here, λ_{sdf} and λ_{cons} weight the two geometric objectives. Consensus masks and teacher SDFs are generated offline from the fixed predictions of the frozen initial segmenter and remain unchanged during final training. At inference, only the dense semantic prediction path is used. Consensus aggregation, teacher SDF construction, and the auxiliary SDF head are not required. Implementation details are provided in Appendix A.2.

3.2 Stage II: Regional Segmentation Quality Assessment

We use the frozen final Stage I segmenter f_{seg} to obtain the fixed prediction $\hat{Y} = f_{\text{seg}}(X)$ for Stage II. For local region i, let $\mathbf{c}_i$ denote its sampling anchor, $k_i = \hat{Y}(\mathbf{c}_i)$ its source class, and X_i and $\hat{Y}_i$ the aligned raw and prediction crops. Region centers are determined from the fixed prediction without ground truth, and each Stage II unit is defined as a fixed local region rather than a connected component or biological object. The source class k_i is used to construct local geometry and define the class-conditional quality target.

Local prediction-derived geometry. We extract the source-class binary mask $M_i(\mathbf{v}) = \mathbf{1}_{\hat{Y}_i(\mathbf{v})=k_i}$ and define $\mathcal{M}_i = \{\mathbf{v} : M_i(\mathbf{v}) = 1\}$. Its clipped local SDF is

$$S_i(\mathbf{v}) = \text{clip}_{[-\tau,\tau]} \left[d(\mathbf{v}, \mathcal{M}_i^c) - d(\mathbf{v}, \mathcal{M}_i) \right], \tag{10}$$

using the same voxel-space distance and sign convention, with Stage II clipping radius $\tau > 0$. Before being used as model input, the SDF is normalized as $\widetilde{S}_i(\mathbf{v}) = S_i(\mathbf{v})/\tau \in [-1, 1]$. Because S_i is a deterministic re-encoding of M_i, it introduces no additional semantic information.

Regional quality target. Let Ω^q_i denote the valid labeled voxels within the central 3D evaluation core and Y^{GT}_i the aligned ground-truth crop. Define $P_i = \{\mathbf{v} \in \Omega^q_i : \hat{Y}_i(\mathbf{v}) = k_i\}$ and $G_i = \{\mathbf{v} \in \Omega^q_i : Y^{\text{GT}}_i(\mathbf{v}) = k_i\}$. For $|P_i| + |G_i| > 0$, the class-conditional local Dice target is

$$q^{\text{GT}}_i = \frac{2|P_i \cap G_i|}{|P_i| + |G_i|}. \tag{11}$$

Ground truth is accessed only after regions are fixed and is not used for region selection or model input. Regions without a valid target are excluded. Thus, q^{GT}_i measures local segmentation quality rather than confidence, uncertainty, or instance-level quality.

Appearance–geometry quality estimator. The Stage II estimator jointly encodes raw appearance and local prediction-derived geometry:

$$\hat{q}_i = g_\psi(X_i, \widetilde{S}_i), \tag{12}$$

where g_ψ denotes the regional quality estimator, which uses a final sigmoid to constrain $\hat{q}_i$ to $[0, 1]$.

Training and inference. Let $\mathcal{D}_q$ denote the set of regions with valid quality targets. We train the quality estimator using Smooth-L1 against the class-conditional local Dice target:

$$\mathcal{L}_{\mathrm{II}} = \frac{1}{|\mathcal{D}_q|} \sum_{i \in \mathcal{D}_q} \ell_{\mathrm{SL1}} \left(\hat{q}_i - q_i^{\mathrm{GT}} \right). \tag{13}$$

At inference, ground truth is not required. The local SDF is derived from the fixed segmentation prediction, and the estimator uses only the aligned raw region and local SDF to produce $\hat{q}_i$. Implementation details are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate Sparse2Geo on two publicly available vEM datasets acquired using focused ion beam scanning electron microscopy (FIB-SEM). **BetaSeg** (Müller et al., 2021) provides volumetric images of primary mouse pancreatic β cells. We segment nuclei, insulin secretory granules, and mitochondria. We use `high_c1`–`high_c3` for training and validation, and hold out `high_c4` for final evaluation. **UroCell** (Žerovnik Mekuč et al., 2022) captures the intracellular ultrastructure of urinary bladder urothelial cells. We segment mitochondria, endolysosomes, and fusiform vesicles. We use `fib1-0-0-0`, `fib1-3-3-0`, and `fib1-4-3-0` for training and validation, and hold out `fib1-3-2-1` for final evaluation. Both datasets are used to evaluate segmentation and regional quality assessment. For BetaSeg Stage I, we use extremely sparse point-level semantic supervision, exposing only approximately 0.05% of the valid semantic labels. No dense semantic masks are used as training targets. A representative visualization of the sparse point-level supervision and detailed protocols are provided in Appendix A.1.

Evaluation metrics. **For segmentation,** we report per-class **Dice** and **Macro Dice** to evaluate segmentation overlap, with Macro Dice averaging equally across all evaluated classes. For the Stage I target class on BetaSeg (Class 2, granules), we report Average Symmetric Surface Distance (**ASSD**), 95th-percentile Hausdorff Distance (**HD95**), and Surface Dice at 1- and 2-voxel tolerances (**SD@1** and **SD@2**). Lower ASSD and HD95 and higher Surface Dice indicate better boundary agreement. **For regional quality assessment, Macro MAE** is the primary metric for regional quality estimation, with lower values indicating accurate estimates. **Macro Spearman** ρ measures rank agreement. **AUPRC** evaluates detection of low-quality regions defined by $q^{\mathrm{GT}} < 0.90$ and $q^{\mathrm{GT}} < 0.80$, with higher values indicating better detection performance. Macro MAE and Macro Spearman ρ average equally across the three source classes, whereas AUPRC is computed over all evaluable regions.

4.2 Compared Methods

Segmentation methods. We benchmark Sparse2Geo against three representative supervision paradigms. **Self-supervised methods** include MAESTER (Xie et al., 2023) and the method proposed by Han et al. (2022). Their clustering-based outputs are mapped to semantic classes using a single global one-to-one assignment. **Sparsely supervised methods** include LABKIT (Arzt et al., 2022), BiaPy U-Net (Ronneberger et al., 2015; Franco-Barranco et al., 2025), and ϵ-Seg (Rahnamai Kordasiabi et al., 2025). On BetaSeg, ϵ-Seg and Sparse2Geo use the same extremely sparse point-level supervision, exposing approximately 0.05% of valid semantic labels, while BiaPy U-Net uses an approximately matched annotation budget. **Fully supervised references** include Vanilla ViT (Dosovitskiy et al., 2021) and Segmenter (Strudel et al., 2021), both trained with dense semantic annotations. Further details are provided in Appendix A.2.

Regional quality assessment methods. We compare Sparse2Geo with statistical and learned quality-assessment baselines. **Statistical baselines** include Class mean, which predicts the mean training quality for each source class, and Occupancy + Class Ridge, which regresses quality from source-class occupancy and class metadata. **Learned quality-assessment baselines** include an approach adapted from Robinson et al. (2018), which directly regresses segmentation quality from image and segmentation inputs, and an adapted GANREC–REGNET (Zaman et al., 2023), which combines reconstruction error with segmentation information for regional quality estimation. Adaptation details are provided in Appendix A.3.

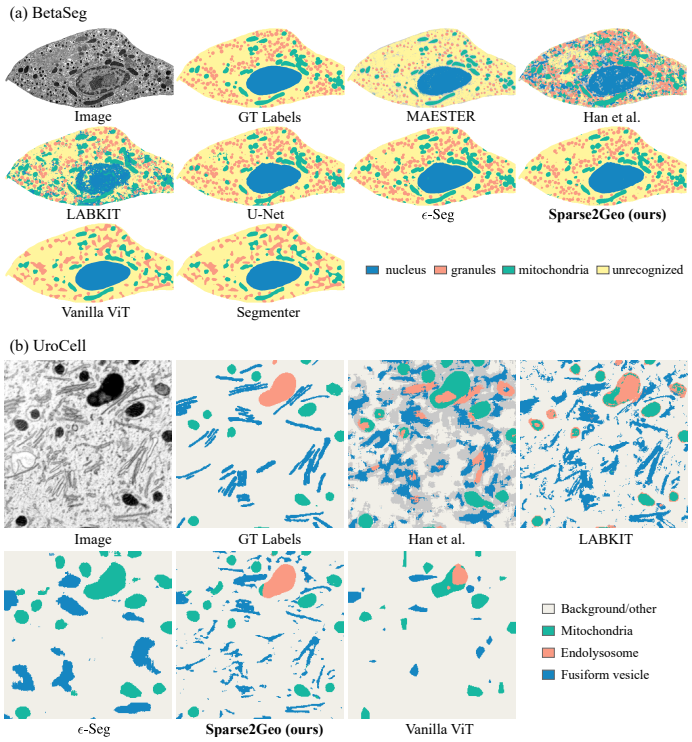

Figure 2: **Qualitative comparison of segmentation on BetaSeg and UroCell.** Representative regions are shown from (a) the BetaSeg evaluation volume `high_c4` and (b) the UroCell evaluation volume `fib1-3-2-1`. Within each dataset, all methods are shown on the same region using a common semantic color mapping.

4.3 Overall Results

Sparsely supervised segmentation. On BetaSeg, Sparse2Geo achieves the highest Macro Dice among the evaluated methods at 90.63%, improving ϵ-Seg by 4.13 percentage points (Table 1). On UroCell, Sparse2Geo again achieves the highest Macro Dice at 71.44%, improving ϵ-Seg by 15.33 percentage points. Together, these results support the effectiveness of prediction-derived geometric supervision for sparsely supervised segmentation across both vEM datasets. Qualitative examples are shown in Figure 2, with full UroCell results in Appendix B.1.

Regional quality assessment. On BetaSeg, Sparse2Geo achieves the best Macro MAE, Macro Spearman ρ, and AUPRC$_{<0.80}$ among the evaluated methods (Table 2). Compared with GANREC–REGNET, Sparse2Geo reduces Macro MAE by 12.3%, increases Macro Spearman ρ from 0.7597 to

Table 1: **Full-volume segmentation results on BetaSeg.** U, N, G, and M denote unrecognized, nucleus, granules, and mitochondria. * denotes clustering-based outputs. Best and second-best results across all evaluated methods are shown in bold and underlined, respectively.

Paradigm	Model	U (%)	N (%)	G (%)	M (%)	Macro Dice (%)
Self-supervised	MAESTER*	82.29	94.45	49.37	78.16	76.07
	Han et al.*	40.31	67.95	57.01	54.21	54.87
Sparsely supervised	LABKIT	77.13	60.89	76.49	59.67	68.55
	U-Net	<u>88.01</u>	95.83	<u>79.88</u>	76.51	85.06
	ϵ-Seg	87.96	97.71	78.47	81.85	<u>86.50</u>
	Sparse2Geo (ours)	**92.27**	97.77	**85.08**	**87.38**	**90.63**
Fully supervised	Vanilla ViT	87.70	<u>98.83</u>	69.05	82.24	84.46
	Segmenter	87.67	**98.85**	68.50	<u>82.76</u>	84.45

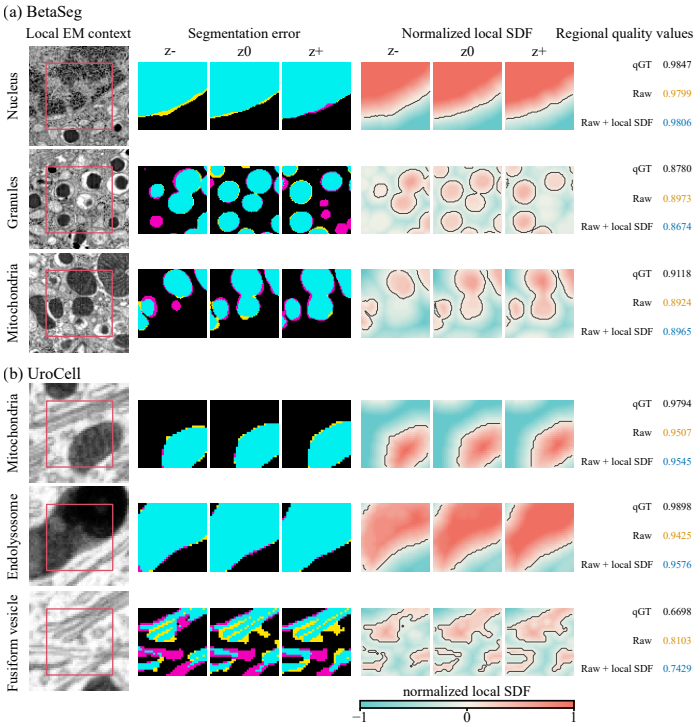

Figure 3: **Qualitative regional quality estimation on BetaSeg and UroCell.** Columns show local EM context, segmentation error, normalized local SDF, and regional quality values. Red boxes mark the central evaluation core, and each row represents one fixed 3D region, with z-, z0, and z+ showing representative relative depths. Error maps show true positives in cyan, false negatives in magenta, false positives in yellow, and background or non-evaluable voxels in black. Numeric annotations report q^{GT}, Raw, and Raw + local SDF regional quality values, which correspond to the full 3D region rather than individual slices.

0.8000, and increases $\mathrm{AUPRC}_{<0.80}$ from 0.3436 to 0.6814. On UroCell, Raw + local SDF reduces Macro MAE by 35.1%, from 0.2276 to 0.1477, and increases Macro Spearman ρ from 0.3125 to 0.7421, with improvements in both AUPRC metrics. These results support the value of combining raw appearance with prediction-derived boundary-distance geometry for regional quality assessment. Qualitative examples are shown in Figure 3, with full UroCell results in Appendix C.1.

Table 2: **Regional quality estimation on 597 evaluable BetaSeg `high_c4` regions.** † marks adapted methods. "–" indicates undefined Spearman correlation for Class mean. Best and second-best results are shown in bold and underlined, respectively.

Method	Macro MAE ↓	Macro ρ ↑	$\mathrm{AUPRC}_{<0.90}$ ↑	$\mathrm{AUPRC}_{<0.80}$ ↑
Class mean	0.0352	–	0.8636	0.1165
Occupancy + Class Ridge	0.0331	0.3817	0.8798	0.2500
Robinson et al.†	0.0356	0.3050	0.8248	0.1569
GANREC–REGNET†	<u>0.0277</u>	<u>0.7597</u>	**0.9721**	<u>0.3436</u>
Sparse2Geo (Raw + local SDF)	**0.0243**	**0.8000**	<u>0.9617</u>	**0.6814**

4.4 Ablation Studies

Progressive segmentation ablation. Table 3 evaluates the dense semantic head, representative-slice validation, and two geometric objectives. Adding the dense head to ϵ-Seg raises Macro Dice from 86.50% to 88.88%, and representative-slice validation further improves it to 90.07%. From this

Table 3: **Progressive segmentation ablation on BetaSeg.** Dense head denotes the added dense semantic prediction head while sparse semantic supervision remains unchanged. Rep.-val denotes the representative-slice validation protocol used for checkpoint selection and early stopping. Best results are shown in bold.

Configuration				Dice (%)				
Dense head	Rep.-val	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{sdf}}$	U	N	G	M	Macro
×	×	×	×	87.96	97.71	78.47	81.85	86.50
✓	×	×	×	90.64	**98.21**	84.22	82.44	88.88
✓	✓	×	×	91.82	97.77	83.90	86.78	90.07
✓	✓	✓	×	92.05	98.06	84.59	87.10	90.45
✓	✓	×	✓	92.17	97.77	85.01	87.18	90.53
✓	✓	✓	✓	**92.27**	97.77	**85.08**	**87.38**	**90.63**

Table 4: **Class 2 (granules) region and boundary quality on BetaSeg.** Results correspond to the segmentation ablation configurations in Table 3. Best results are shown in bold.

Configuration				Class 2 quality				
Dense head	Rep.-val	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{sdf}}$	Dice ↑ (%)	ASSD ↓ (vox.)	HD95 ↓ (vox.)	SD@1 ↑ (%)	SD@2 ↑ (%)
✓	✓	×	×	83.90	1.50	9.22	69.99	86.21
✓	✓	✓	×	84.59	1.48	9.17	71.43	86.83
✓	✓	×	✓	85.01	1.44	8.83	72.28	87.43
✓	✓	✓	✓	**85.08**	**1.42**	**8.77**	**72.37**	**87.51**

baseline, $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{sdf}}$ individually reach 90.45% and 90.53%, while their combination achieves the best Macro Dice of 90.63%. These results show that the two geometric objectives provide complementary gains. **Target-class boundary quality.** Combining the two geometric objectives achieves the best performance across all five Class 2 metrics (Table 4), indicating improvements in both region overlap and boundary quality. **Regional representation ablation.** For Stage II, adding the local SDF to Raw reduces Macro MAE by 21.1% and yields the best Macro MAE and both AUPRC metrics (Table 5), showing that prediction-derived boundary-distance geometry complements raw appearance.

Table 5: **Regional representation ablation on BetaSeg.** Best results are shown in bold.

Input	Macro MAE ↓	Macro ρ ↑	AUPRC$_{<0.90}$ ↑	AUPRC$_{<0.80}$ ↑
Raw	0.0308	0.5917	0.8894	0.3604
Local SDF	0.0260	**0.8205**	0.9563	0.5908
Raw + local SDF	**0.0243**	0.8000	**0.9617**	**0.6814**

5 Conclusion

We presented **Sparse2Geo**, a unified geometric learning framework that exploits prediction-derived continuous geometry as both a geometric supervisory signal for sparsely supervised 3D segmentation and a structural representation for regional quality assessment in volume electron microscopy (vEM). In **Stage I**, prediction-derived 3D consensus geometry and teacher SDF supervision effectively regularize segmentation learning, improving volumetric overlap and boundary fidelity under extreme annotation scarcity without additional dense semantic targets. In **Stage II**, normalized local SDFs provide fine-grained boundary-distance information that complements raw appearance and substantially enhances regional quality assessment without ground truth at inference. Experiments demonstrate benefits across cellular contexts and target structures. Together, these findings highlight prediction-derived geometry as a powerful geometric supervisory signal and an expressive structural descriptor, bridging sparse segmentation learning and prediction reliability in 3D bioimage analysis. Although fixed predictions may propagate systematic artifacts downstream, future iterative co-refinement among segmentation, geometry, and quality estimation may offer further gains.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and organization, and to provide feedback on the presentation of the methodology, experimental design, and interpretation of results. All AI-assisted suggestions were independently reviewed against the implementation, experimental records, and cited literature. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The main methodological definitions, objectives, and training and inference protocols are described in Sections 3 and 4.1. Appendix A provides additional reproducibility details, including data splits and supervision protocols, Stage I and Stage II implementation settings, optimization and architecture details, baseline provenance and adaptations, and ground-truth access. Additional controlled analyses and supporting results are reported in Appendices B and C. Random seeds and evaluation protocols are reported where applicable. Code, data, and other reproduction materials will be publicly released upon acceptance.

REFERENCES

Matthias Arzt, Joran Deschamps, Christopher Schmied, Tobias Pietzsch, Deborah Schmidt, Pavel Tomancak, Robert Haase, and Florian Jug. LABKIT: Labeling and segmentation toolkit for big image data. *Frontiers in Computer Science*, 4:777728, 2022.

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. In *International Conference on Learning Representations*, 2021.

Daniel Franco-Barranco, Jesús A. Andrés-San Román, Ivan Hidalgo-Cenalmor, Lenka Backová, Aitor González-Marfil, Clément Caporal, Anatole Chessel, Pedro Gómez-Gálvez, Luis M. Escudero, Donglai Wei, Arrate Muñoz-Barrutia, and Ignacio Arganda-Carreras. BiaPy: accessible deep learning on bioimages. *Nature Methods*, 22(6):1124–1126, 2025.

Hongqing Han, Mariia Dmitrieva, Alexander Sauer, Ka Ho Tam, and Jens Rittscher. Self-supervised voxel-level representation rediscovers subcellular structures in volume electron microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1874–1883, 2022.

Yikai Huang, Renmin Han, Yuxuan Wang, Youcheng Cai, and Ligang Liu. Spatial-SAM: Spatially consistent 3D electron microscopy segmentation with SDF memory and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22763–22772, 2026.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8801–8809, 2021.

Archana Machireddy, Guillaume Thibault, Kevin G. Loftis, Kevin Stoltz, Cecilia E. Bueno, Hannah R. Smith, Jessica L. Riesterer, Joe W. Gray, and Xubo Song. Segmentation of cellular ultrastructures on sparsely labeled 3D electron microscopy images using deep learning. *Frontiers in Bioinformatics*, 3:1308708, 2023.

Keiko Midorikawa, Ayaka Tateishi, Kiminori Toyooka, Mayuko Sato, Takuto Imai, Yutaka Kodama, and Keiji Numata. Three-dimensional nanoscale analysis of light-dependent organelle changes in Arabidopsis mesophyll cells. *PNAS Nexus*, 1(5):pgac225, 2022.

Andreas Müller, Deborah Schmidt, C. Shan Xu, Song Pang, Joyson Verner D'Costa, Susanne Kretschmar, Carla Münster, Thomas Kurth, Florian Jug, Martin Weigert, Harald F. Hess, and Michele Solimena. 3D FIB-SEM reconstruction of microtubule–organelle interaction in whole primary mouse β cells. *Journal of Cell Biology*, 220(2):e202010039, 2021.

Sheida Rahnamai Kordasiabi, Damian Nogare, and Florian Jug. ϵ-Seg: Sparsely supervised semantic segmentation of microscopy data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30567–30590. Curran Associates, Inc., 2025.

Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V. Valindria, Mihir M. Sanghvi, Nay Aung, José M. Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Bernhard Kainz, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Daniel Rueckert, and Ben Glocker. Real-time prediction of segmentation quality. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 578–585. Springer International Publishing, 2018.

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*, pp. 234–241. Springer International Publishing, 2015.

Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

Bella Specktor-Fadida, Liat Ben-Sira, Dafna Ben-Bashat, and Leo Joskowicz. SegQC: a segmentation network-based framework for multi-metric segmentation quality control and segmentation error detection in volumetric medical images. *Medical Image Analysis*, 103:103638, 2025.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7262–7272, 2021.

Eichi Takaya, Yusuke Takeichi, Mamiko Ozaki, and Satoshi Kurihara. Sequential semi-supervised segmentation for serial electron microscopy image with small number of labels. *Journal of Neuroscience Methods*, 351:109066, 2021.

Shin-ya Takemura, Arjun Bharioke, Zhiyuan Lu, Aljoscha Nern, Shiv Vitaladevuni, Patricia K. Rivlin, William T. Katz, Donald J. Olbris, Stephen M. Plaza, Philip Winston, Ting Zhao, Jane Anne Horne, Richard D. Fetter, Satoko Takemura, Katerina Blazek, Lei-Ann Chang, Omotara Ogundeyi, Mathew A. Saunders, Victor Shapiro, Christopher Sigmund, Gerald M. Rubin, Louis K. Scheffer, Ian A. Meinertzhagen, and Dmitri B. Chklovskii. A visual motion detection circuit suggested by Drosophila connectomics. *Nature*, 500(7461):175–181, 2013.

Ronald Xie, Kuan Pang, Gary D. Bader, and Bo Wang. MAESTER: Masked autoencoder guided segmentation at pixel resolution for accurate, self-supervised subcellular structure recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3292–3301, 2023.

Yuan Xue, Hui Tang, Zhi Qiao, Guanzhong Gong, Yong Yin, Zhen Qian, Chao Huang, Wei Fan, and Xiaolei Huang. Shape-aware organ segmentation by predicting signed distance maps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12565–12572, 2020.

Fahim Ahmed Zaman, Tarun Kanti Roy, Milan Sonka, and Xiaodong Wu. Patch-wise 3D segmentation quality assessment combining reconstruction and regression networks. *Journal of Medical Imaging*, 10(5):054002, 2023.

Manca Žerovnik Mekuč, Ciril Bohak, Eva Boneš, Samo Hudoklin, Rok Romih, and Matija Marolt. Automatic segmentation and reconstruction of intracellular compartments in volumetric electron microscopy data. *Computer Methods and Programs in Biomedicine*, 223:106959, 2022.

Felix Y. Zhou, Zach Marin, Clarence Yapp, Qiongjing Zou, Benjamin A. Nanes, Stephan Daetwyler, Andrew R. Jamieson, Md Torikul Islam, Edward Jenkins, Gabriel M. Gihana, Jinlong Lin, Hazel M. Borges, Bo-Jui Chang, Andrew Weems, Sean J. Morrison, Peter K. Sorger, Reto Fiolka, Kevin M. Dean, and Gaudenz Danuser. Universal consensus 3D segmentation of cells from 2D segmented stacks. *Nature Methods*, 22(11):2386–2399, 2025.

A Implementation and Reproducibility Details

A.1 Data and Evaluation Protocols

For BetaSeg, `high_c1–high_c3` are used for training and validation, and `high_c4` is held out for final evaluation. Stage I uses sparse point-level semantic supervision on eligible 2D slices. Under the SPC33 setting, up to 33 supervised locations per class are exposed on each eligible slice, corresponding to approximately 0.05% of the valid semantic labels in the training and validation volumes. SPC33 is used as the primary setting to maintain a highly sparse annotation budget and a matched supervision level for Sparse2Geo and the ϵ-Seg baseline. Label-budget sensitivity is further examined in Appendix B.3. Only these sampled locations provide semantic training targets, while the surrounding pixels in each input patch are not exposed as dense semantic labels. Dense benchmark annotations are used to define the sampled sparse locations and valid regions and to provide reference labels for validation and evaluation. Unexposed dense class values are not used as semantic training targets or to construct prediction-derived geometric supervision.

A representative visualization of the sparse point-level supervision is shown in Figure 4.

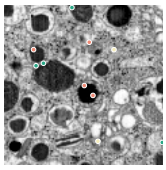

Figure 4: **Visualization of sparse point-level supervision on BetaSeg.** A representative crop from a single XY slice at a fixed z-coordinate is shown. Colored markers indicate sampled supervised locations over the raw EM image. Only a local subset of the sampled locations is shown for clarity.

For UroCell, `fib1-0-0-0`, `fib1-3-3-0`, and `fib1-4-3-0` are used for training and validation, while `fib1-3-2-1` is held out for final evaluation.

Stage II regions and source classes are determined from the fixed segmentation prediction before ground truth is accessed. Ground truth is then used only to construct the class-conditional regional quality target within the predefined evaluation core and is not used for region selection or provided as model input.

A.2 Stage I Settings

Stage I uses the ϵ-Seg backbone with a dense semantic prediction head and 64×64 two-dimensional patches. The final segmenter is initialized separately from the frozen initial segmenter used to construct the consensus geometry and teacher SDF, which are generated offline from fixed full-volume predictions from the training and validation volumes and remain unchanged during final training. For UroCell, endolysosome (Class 2) is used as the target class for prediction-derived geometric supervision.

Figure 5 summarizes the architecture of the final Stage I segmenter.

Primary Stage I settings are summarized in Table 6.

Stage I comparison baselines. MAESTER uses the authors' released checkpoint, whereas ϵ-Seg is retrained from the official code under SPC33. Han et al., LABKIT, Vanilla ViT, and Segmenter are reproduced and evaluated under the same full-volume evaluation protocol. The BiaPy U-Net uses an approximately SPC33-matched localized patch-label budget, and clustering-based outputs are mapped to semantic classes using a single global one-to-one assignment.

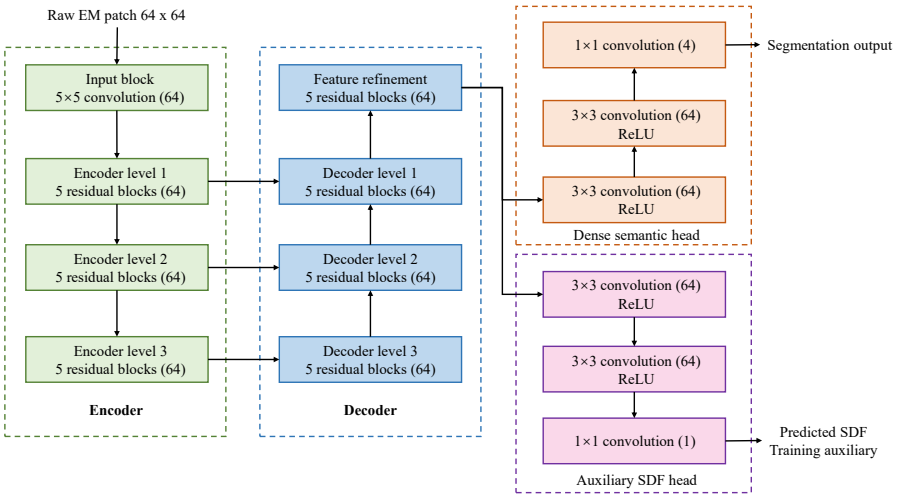

Figure 5: **Simplified architecture of the final Stage I segmenter.** The ϵ-Seg backbone provides shared features to a dense semantic prediction head and an auxiliary SDF head. The semantic head produces the segmentation output, while the auxiliary head predicts the target-class signed distance field (SDF) during training and is not required at inference. Internal backbone operations are omitted for clarity. Numbers in parentheses indicate channel counts.

Table 6: **Primary Stage I implementation settings.**

Item	Setting
Input	64×64 2D patches
Sparse supervision	SPC33 with localized semantic cross-entropy
Inherited objective weights	$\alpha = 1.0$, $\beta = 0.01$, $\gamma = 0.1$
Teacher construction	Class 2 with XY/XZ/YZ re-slicing, component filtering, and u-Segment3D consensus aggregation
Teacher SDF	Voxel-space signed distance clipped to $[-16, 16]$ without normalization
SDF regression	Smooth-L1 loss with transition parameter 1.0 and mean reduction
Geometric weights	$\lambda_{\text{sdf}} = 0.05$, $\lambda_{\text{cons}} = 0.05$
Boundary consistency	Temperature $T = 1$ and boundary width $\delta = 1$ voxel
Optimization	Adamax with learning rate 3×10^{-5} and batch size 128
Validation / model selection	Highest Macro Dice on 45 fixed representative validation slices (15 per training/validation volume, XY stride 16) for checkpoint selection and early stopping
Random seed	42

A.3 Stage II Settings and Adapted Baselines

Regional construction. Stage II regions are sampled from a fixed segmentation prediction and balanced across the three foreground source classes. Training and validation regions are split spatially to avoid randomly partitioning potentially overlapping crops. Regional quality is evaluated within the central core, and a target is valid only when at least 50% of the core voxels have valid ground truth and $|P_i| + |G_i| > 0$.

Spatial settings are summarized in Table 7.

Table 7: **Spatial settings for Stage II regional quality estimation.** Region counts are reported per source class for training, validation, and evaluation.

Dataset	Input region	Evaluation core	SDF clip	Regions per class
BetaSeg	$32 \times 96 \times 96$	$16 \times 64 \times 64$	16	1000 / 200 / 200
UroCell	$16 \times 48 \times 48$	$8 \times 32 \times 32$	8	1000 / 200 / 200

For BetaSeg, the nominal training, validation, and evaluation pools contain 3,000, 600, and 600 regions, of which 2,998, 599, and 597 remain after target-validity filtering. The scale-adapted UroCell evaluation contains 600 regions, with 200 per source class.

Quality estimator. Each raw crop is independently normalized using its first and 99th intensity percentiles and clipped to $[0, 1]$. The default estimator receives the aligned raw EM region and normalized local SDF. The source-class binary mask is used both to construct the local SDF and as the matched Raw + Mask representation control.

Figure 6 summarizes the architecture of the regional quality estimator, with implementation settings provided in Table 8.

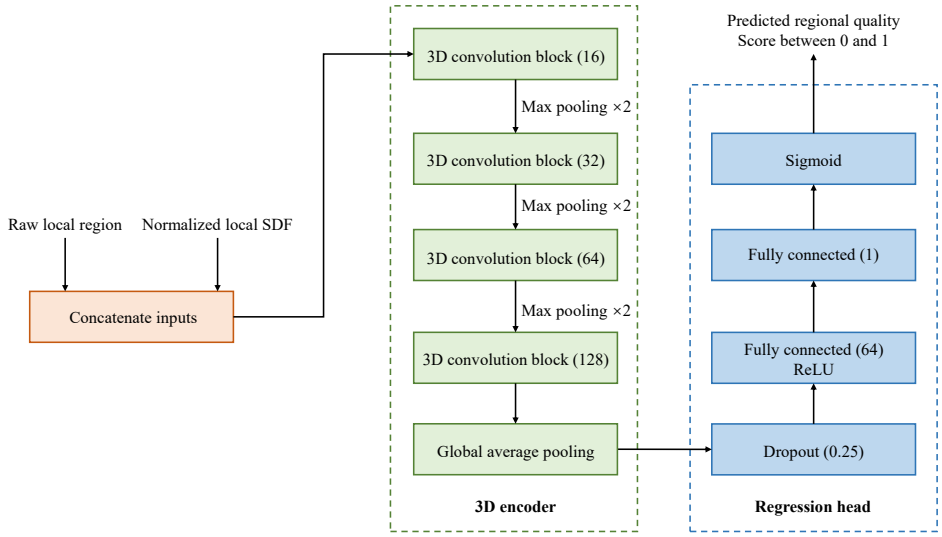

Figure 6: **Architecture of the Stage II regional quality estimator.** The aligned raw local region and normalized local SDF are concatenated along the channel dimension and encoded by four 3D convolutional blocks with channel widths 16, 32, 64, and 128. Each block contains two $3 \times 3 \times 3$ convolutions, each followed by GroupNorm and ReLU, with max pooling after the first three blocks. Global average pooling and the regression head produce a scalar regional quality estimate in $[0, 1]$. The regression head uses dropout with probability 0.25 and a final sigmoid output.

During training, spatial flips are applied jointly to aligned channels, and raw-containing inputs receive mild intensity scale and shift augmentation. The Raw, local SDF, Raw + local SDF, and Raw + Mask variants otherwise use the same estimator and training protocol.

Table 8: **Primary Stage II quality-estimator settings.**

Item	Setting
Encoder	Four 3D convolutional blocks with 16/32/64/128 channels
Normalization	GroupNorm
Global aggregation	Adaptive average pooling
Regression head	Dropout 0.25 with $128 \to 64 \to 1$ and sigmoid output
Objective	Smooth-L1 loss with transition parameter 0.05 and mean reduction
Optimizer	AdamW with learning rate 10^{-4} and weight decay 10^{-4}
Batch size	8
Training	Up to 80 epochs with early-stopping patience 15
Checkpoint selection	Lowest validation Macro MAE
Random seed	42

Simple baselines. The Class mean baseline predicts the mean training quality for each source class. Occupancy + Class Ridge fits a ridge regressor using source-class occupancy in the input region and evaluation core together with source-class metadata.

Adapted external quality assessment baselines. The two adapted external baselines use the same fixed $32 \times 96 \times 96$ BetaSeg regions, `high_c1–high_c3` training and validation data, source-class regional Dice targets over the central evaluation core, and the same evaluation metrics as Sparse2Geo.

Robinson adaptation. Following Robinson et al. (2018), we use an image–segmentation input with a 3D ResNet-50 quality regressor. Each input concatenates the normalized raw region with four one-hot channels from the frozen Stage I prediction, representing the four semantic classes. The network predicts one sigmoid quality estimate per foreground class, and only the estimate corresponding to the region's source class is supervised with mean-squared error. Training uses Adam with learning rate 10^{-5}, batch size 2, up to 80 epochs, and early-stopping patience 15.

GANREC–REGNET adaptation. We retain the two-stage reconstruction and quality-estimation formulation of GANREC–REGNET (Zaman et al., 2023). Because dense semantic targets are not exposed as model inputs under our sparse BetaSeg protocol, the dense reference-mask input is replaced by the source-class mask from the frozen Stage I prediction. The normalized raw region is mapped from $[0, 1]$ to $[-1, 1]$ before network input.

RECNET reconstructs the raw region using an adversarial objective and an L1 reconstruction term weighted by 100. After RECNET is frozen, REGNET receives the absolute reconstruction error together with the source-class prediction mask and predicts a quality map whose spatial mean gives the regional quality estimate. The scalar regional quality target is broadcast over this map and optimized with mean-squared error. RECNET and REGNET use Adam with learning rates of 10^{-4} and 4×10^{-5}, respectively, batch size 2, up to 100 epochs, and early-stopping patience 15.

Model selection uses only the training and validation data. Robinson and REGNET are selected by validation Macro MAE, whereas RECNET is selected by validation reconstruction L1. All external comparisons use the same fixed 597 BetaSeg evaluation regions, with no evaluation-set model selection, retraining, or hyperparameter tuning. Neural-network configurations use random seed 42 unless otherwise stated.

B Additional Stage I Results

B.1 UroCell Evaluation

Results are reported in Table 9.

Sparse2Geo achieves the highest Macro Dice of 71.44%, improving ϵ-Seg by 15.33 percentage points and providing supporting evidence on a second vEM dataset.

Table 9: **Stage I segmentation results on the UroCell evaluation volume.** B/O, M, E, and F denote background/other, mitochondria, endolysosome, and fusiform vesicle, respectively. * denotes a clustering-based output. Best and second-best results across all evaluated methods are shown in bold and underlined, respectively.

Paradigm	Model	B/O (%)	M (%)	E (%)	F (%)	Macro Dice (%)
Self-supervised	Han et al.*	48.25	59.40	11.03	30.26	37.24
Sparsely supervised	LABKIT	88.29	**84.13**	30.17	49.68	63.07
	ϵ-Seg	90.30	62.78	21.78	49.59	56.11
	Sparse2Geo (ours)	<u>92.87</u>	<u>80.36</u>	**58.60**	**53.91**	**71.44**
Fully supervised	Vanilla ViT	**94.22**	75.60	<u>52.33</u>	<u>50.75</u>	<u>68.23</u>

B.2 Teacher Construction

Table 10 evaluates consensus construction and target-class sensitivity under the same SPC33 setting and $\lambda_{\mathrm{sdf}} = \lambda_{\mathrm{cons}} = 0.05$.

Table 10: **Sensitivity of Stage I geometric supervision to consensus construction and target class on the BetaSeg `high_c4` evaluation volume.** All values are fixed-seed point estimates under the same full-volume evaluation protocol.

Setting	Macro Dice (%)
Controlled SPC33 baseline	90.07
Aggregated Class 2 geometry	90.63
Class 2 geometry without aggregation	89.58
Aggregated Class 1 geometry	90.14
Aggregated Class 3 geometry	89.73

Aggregated Class 2 geometry gives the highest Macro Dice among the evaluated settings, whereas the Class 2 configuration without aggregation does not improve over the controlled baseline.

B.3 Sparse-Label Sensitivity

Table 11 reports label-budget sensitivity with the dense semantic prediction head and evaluation protocol kept fixed.

Table 11: **Label-budget sensitivity of the base Stage I segmenter on BetaSeg.** Prediction-derived geometric objectives are disabled in this analysis.

SPC	Macro Dice (%)
1	81.09
10	89.56
33	90.07
66	90.05
132	89.72

Performance improves from SPC1 to SPC10 and changes little between SPC33 and SPC66. SPC33 corresponds to approximately 0.05% of valid labels and is therefore used as the primary sparse-supervision setting.

C Additional Stage II Results

C.1 UroCell Evaluation

Results are reported in Table 12.

Table 12: **Scale-adapted Stage II regional quality estimation on UroCell with 600 regions, 200 per source class.** Best results are shown in bold.

Input	Macro MAE ↓	Macro ρ ↑	AUPRC$_{<0.90}$ ↑	AUPRC$_{<0.80}$ ↑
Raw	0.2276	0.3125	0.8939	0.8342
Raw + local SDF	**0.1477**	**0.7421**	**0.9745**	**0.9487**

Raw + local SDF improves all four reported metrics over Raw, reducing Macro MAE by 35.1% and increasing Macro Spearman ρ from 0.3125 to 0.7421 under the scale-adapted UroCell setting.

C.2 Prediction Representation Controls

Table 13 compares the local SDF with the binary source-class mask and explicit source-class metadata under the same evaluation protocol. For Raw + ClassID, a three-element one-hot source-class vector is concatenated to the globally pooled regional feature before regression.

Table 13: **Stage II representation and metadata controls on the same 597 evaluable BetaSeg `high_c4` regions.**

Input	Macro MAE ↓	Macro ρ ↑	AUPRC$_{<0.90}$ ↑	AUPRC$_{<0.80}$ ↑
Raw	0.0308	0.5917	0.8894	0.3604
Raw + ClassID	0.0319	0.6082	0.8730	0.2586
Raw + Mask	0.0286	0.8279	0.9638	0.5854
Raw + local SDF	0.0243	0.8000	0.9617	0.6814

Compared with Raw + Mask, Raw + local SDF reduces Macro MAE by 15.0% and increases AUPRC$_{<0.80}$ from 0.5854 to 0.6814, while Raw + Mask is slightly higher in Macro Spearman ρ and AUPRC$_{<0.90}$. These results support an advantage of the continuous SDF representation for regional quality estimation and identification of low-quality segmentation regions beyond the binary mask. Raw + ClassID does not improve Macro MAE over Raw, suggesting that explicit source-class metadata alone does not account for the gain from the local SDF.

C.3 Context Size Sensitivity

Table 14 compares the primary $32 \times 96 \times 96$ input (P32) with a $64 \times 64 \times 64$ input (P64) using identical region centers, source classes, evaluation cores, and quality targets. Thus, only the input context changes between the two settings.

Table 14: **Context-size sensitivity on matched BetaSeg `high_c4` regions with a fixed $16\times64\times64$ evaluation core.**

Input context	Macro MAE ↓	Macro ρ ↑	AUPRC$_{<0.90}$ ↑	AUPRC$_{<0.80}$ ↑
P32: $32 \times 96 \times 96$	0.0243	0.8000	0.9617	0.6814
P64: $64 \times 64 \times 64$	0.0260	0.7741	0.9017	0.5639

P32 performs better than P64 on all four reported metrics.