HASP-NET: MANIFOLD-GUIDED POINT-CLOUD SEGMENTATION OF HIGH-ASPECT-RATIO SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Point cloud segmentation has advanced substantially in recent years, yet scenes containing high-aspect-ratio (HAR) structures and small attached components remain challenging. In these scenes, elongated structures often extend far beyond local neighborhoods while lying close to geometrically similar objects, making it difficult to simultaneously preserve long-range continuity and avoid feature interference from adjacent structures. Sparse sampling and scene cropping further weaken structural cues, while conventional point-wise metrics may underrepresent errors such as local breaks and missed attachments. To address this gap, we propose HASP-Net, a manifold-guided segmentation framework built on PTv3 for modeling long-range structural continuity and local semantic variation. The Long-range Track Module represents elongated structures as low-dimensional manifolds in 3D space, constructs geometry-consistent support chains, and propagates context bidirectionally along them to capture long-range dependencies while reducing interference from nearby structures. The Geometry-guided Interface Module focuses on regions with complex local manifold geometry and refines their features through geometry-aware point-set reasoning, reliability-gated residual updates, and semantic contrastive supervision. Together, the two modules jointly model long-range manifold structure and local geometric variation for improved segmentation of elongated structures and their attached components. However, existing benchmarks typically focus on specific domains and lack a unified setting for evaluating these challenges. We introduce HAR3D-Seg, a unified benchmark covering transmission, railway, and vascular scenes, with the Semantic Continuity Score (SCS) and Cross-class Adjacency Score (CAS) for structural evaluation. Averaged across the three scene types, HASP achieves 74.17% mIoU, 80.92% mAcc, 96.16% OA, 2.95 SCS, and 1.13 CAS, demonstrating strong segmentation performance across diverse scenes.

1 INTRODUCTION

Point-cloud semantic segmentation is essential for understanding complex 3D scenes and has been widely used in applications such as infrastructure asset management, digital twin construction, automated inspection, railway surveying, industrial monitoring, and biomedical imaging. Many of these scenes contain both high-aspect-ratio (HAR) structures and the attached components, such as conductors with insulators and spacers, rails with overhead contact lines, industrial cables and pipelines, and branching vascular structures. Accurate segmentation of these scenes requires maintaining consistent semantic predictions over long spatial ranges while also distinguishing small attached components from nearby structures with similar geometric features.

Although point cloud segmentation has made substantial progress in recent years, most methods still aggregate features within Euclidean neighborhoods or locally serialized groups, making it difficult to explicitly model dependencies over long spatial ranges or distinguish support along an elongated structure from proximity across a semantic boundary. For example, PointNet++ (Qi et al., 2017b) and KPConv (Thomas et al., 2019) construct hierarchical or deformable local neighborhoods, Point Transformer (Zhao et al., 2021) uses neighborhood attention, and RandLA-Net (Hu et al., 2020), PointNeXt (Qian et al., 2022), and Point Transformer V3 (Wu et al., 2024) emphasize efficient local or serialized aggregation.

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.

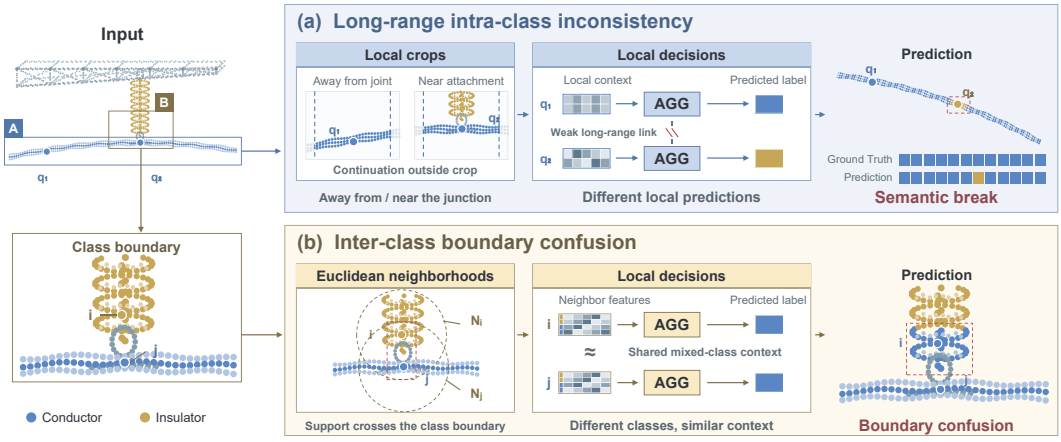


Figure 1: Two challenges in high-aspect-ratio (HAR) segmentation. (a) Long-range intra-class inconsistency along the same structure. (b) Inter-class boundary confusion near attached components. AGG denotes local feature aggregation.

However, existing methods exhibit two coupled failure modes in HAR scenes, as shown in Figure 1. First, long-range intra-class inconsistency arises when point q_1 away from an attachment and q_2 near an insulator receive different local contexts without sufficient along-wire propagation. Second, inter-class boundary confusion occurs when overlapping Euclidean neighborhoods mix features of points i and j from different classes near attachments. Enlarging the neighborhood can recover some along-structure context but may also increase cross-class interference, motivating evaluation of both long-range semantic continuity and local cross-class discrimination.

To address these challenges, we propose HASP-Net, which models elongated structures as low-dimensional manifolds and exploits their long-range continuity and local geometric variation. The Long-range Track Module approximates manifold connectivity with ordered geometric support chains and aggregates context bidirectionally according to distance and geometric reliability. Near junctions, the Geometry-guided Interface Module adapts feature refinement to local manifold variations through reliability-gated residual updates conditioned on neighborhood features and geometric embeddings. Semantic contrastive supervision further promotes inter-class discrimination where geometrically connected regions have different semantic labels.

Existing benchmarks primarily focus on individual domains and provide limited support for jointly evaluating elongated structures and their fine-grained attached components (Qiu et al., 2024; Lavado et al., 2025; Yang et al., 2020). Besides, standard point-wise metrics also do not explicitly assess structural continuity or cross-class adjacency. To address these limitations, we introduce HAR3D-Seg, a unified benchmark for segmenting HAR structures and their attached components across transmission, railway, and vascular scenes. Beyond standard segmentation metrics, we introduce the Semantic Continuity Score (SCS) and Cross-class Adjacency Score (CAS) to assess long-range semantic continuity and local cross-class discrimination, respectively. In summary, this work makes three contributions:

- **HASP-Net**: A manifold-guided framework combining a Long-range Track Module for bidirectional propagation along discrete manifold supports and a Geometry-guided Interface Module for feature refinement guided by local manifold geometry.
- **HAR3D-Seg**: A unified benchmark covering transmission, railway, and vascular scenes, with SCS and CAS complementing standard segmentation metrics to evaluate long-range semantic continuity and local cross-class discrimination.
- **Experimental results**: HASP achieves 74.17% mIoU, 80.92% mAcc, 96.16% OA, 2.95 SCS, and 1.13 CAS averaged across the three scene types, demonstrating strong segmentation performance on HAR structures and their attached components.

2 Related Work

2.1 Point-cloud segmentation and scalable representations

PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) proposed permutation-invariant and hierarchical representations for point clouds, while DGCNN (Wang et al., 2019) and KPConv (Thomas et al., 2019) improved local geometric modeling through dynamic graph neighborhoods and deformable kernels. Sparse convolution (Graham et al., 2018; Choy et al., 2019) and point-voxel methods (Tang et al., 2020) further improve scalability by exploiting spatial sparsity, enabling efficient processing of large-scale scenes. More recent methods, including RandLA-Net (Hu et al., 2020), PointNeXt (Qian et al., 2022), Point Transformer (Zhao et al., 2021), and Point Transformer V3 (Wu et al., 2024), further improve the balance between segmentation accuracy and efficiency for large-scale point clouds. Despite substantial progress in point cloud segmentation, existing methods still mainly rely on local feature aggregation, making it difficult to capture long-range structural continuity and fine-grained interactions between elongated structures and the attached components.

2.2 Datasets and benchmarks for large-scale high-aspect-ratio structures

Existing benchmarks cover a wide range of large-scale and fine-grained 3D understanding tasks. Semantic3D (Hackel et al., 2017), S3DIS (Armeni et al., 2016), ScanNet (Dai et al., 2017), SemanticKITTI (Behley et al., 2021), and DALES (Varney et al., 2020) provide diverse indoor and outdoor scenes, while WHU-Railway3D (Qiu et al., 2024) and TS40K (Lavado et al., 2025) focus on large-scale infrastructure, and IntrA (Yang et al., 2020) provides fine-grained annotations of branching vascular structures. Although these datasets address scene scale, sensing diversity, and annotation detail, they provide limited support for jointly evaluating long-range semantic continuity and local cross-class discrimination in HAR scenes. HAR3D-Seg addresses this gap through a unified benchmark covering transmission, railway, and vascular scenes, with SCS and CAS complementing standard segmentation metrics to evaluate these shared structural challenges.

2.3 Structural, topological, and graph-based reasoning

Superpoint-based methods (Landrieu & Simonovsky, 2018; Robert et al., 2023) group points into geometrically consistent regions and propagate features over region-level graphs, while dynamic graph methods (Wang et al., 2019) capture local geometric relationships through adaptive neighborhoods. Scene graph approaches (Feng et al., 2025; Zemskova & Yudin, 2025) model object-level context and semantic relations. Open-vocabulary approaches explore Gaussian-based panoptic segmentation (Zhai et al., 2025) and point–entity contrastive learning (Wang et al., 2025). Structure-oriented methods further introduce geometry-aware objectives (Lin et al., 2025) or semantic encodings (Luo et al., 2025) for structural prediction. These approaches model structural relationships at different spatial and semantic levels. HASP-Net focuses on long-range continuity along elongated structures by approximating longitudinal manifold connectivity with geometry-consistent support chains. Directional context propagation along these chains is followed by local refinement to distinguish neighboring semantic regions.

3 Method

3.1 Problem formulation and method overview

Problem formulation. The following method description is based on the transmission scenes of HAR3D-Seg, with XYZ coordinates as input. For each sampled scene, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^3$ denote the XYZ point set and $\mathbf{Y} = \{y_i\}_{i=1}^{N}$ the corresponding labels. The task is to learn a point-wise predictor $f_\theta : \mathbb{R}^{N\times 3} \to \mathbb{R}^{N\times K}$ that maps the input coordinates to class logits, where K is the number of semantic classes.

Method overview. We view HAR segmentation as learning a semantic field over low-dimensional geometric supports, where long-range consistency within a component must coexist with local class transitions between nearby or connected components. This motivates coupling structural context with local semantic discrimination. As shown in Figure 2, HASP consists of a PTv3 backbone

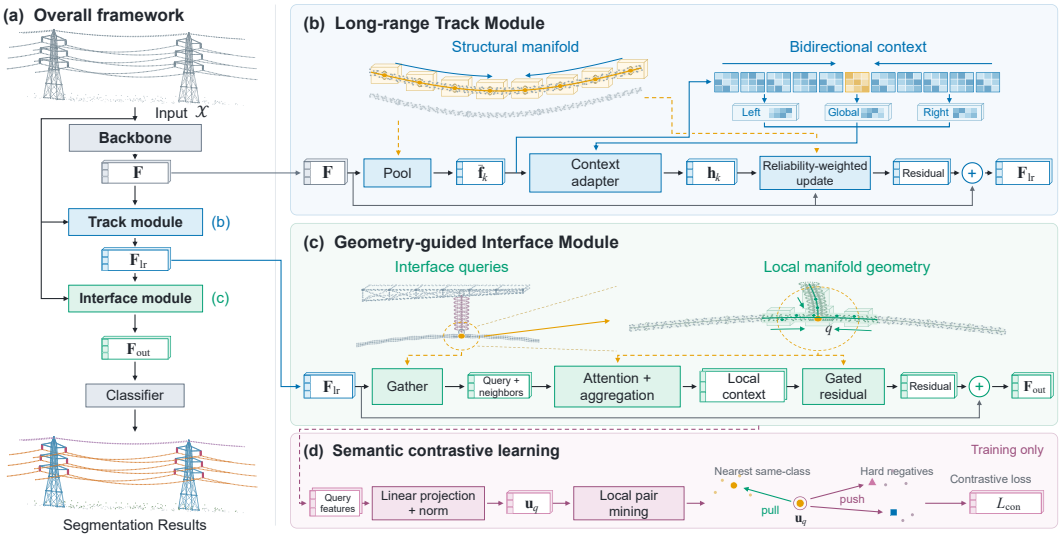


Figure 2: Overview of HASP. (a) Overall framework with sequential track and interface refinement. (b) Bidirectional context aggregation along geometric tracks. (c) Geometry-guided local refinement with gated residuals. (d) Semantic contrastive learning.

E_ϕ, two sequential refinement modules, and a semantic head H_ψ. The Long-range Track Module aggregates context along ordered geometric supports that approximate longitudinal manifold structures, allowing distant observations to share structural evidence. Since geometric continuity does not imply semantic uniformity, the Geometry-guided Interface Module subsequently refines query features using local geometry and query–context discrepancies. This sequential design enables local discrimination to draw on long-range structural context, while semantic contrastive supervision encourages class separation where geometry alone is ambiguous.

3.2 Long-range Geometric Track Context

The Long-range Track Module aggregates context along ordered geometric supports that approximate longitudinal manifold structure. Given point coordinates $\mathcal{X}$ and features $\mathbf{F}$, it outputs updated point features $\mathbf{F}_{\mathrm{lr}}$ for the subsequent Geometry-guided Interface Module. Specifically, the module first partitions each scene into $4\,\mathrm{m}$ cells, estimates their local orientation, width, support, and reliability q_k, and connects geometrically consistent cells into tracks. Each cell serves as an anchor, and tracks with fewer than four anchors or shorter than $16\,\mathrm{m}$ are discarded. Each anchor k has mean feature $\bar{\mathbf{f}}_k$, descriptor $\mathbf{d}_k$, reliability q_k, and arc-length coordinate s_k. Let $\mathcal{T}_d(k)$ contain the preceding, following, or all anchors on its track, indexed by $d \in \{\mathrm{left}, \mathrm{right}, \mathrm{global}\}$. Distance- and reliability-weighted contexts are computed as:

$$\mathbf{c}_k^d = \frac{\sum_{r\in\mathcal{T}_d(k)} q_r e^{-|s_r - s_k|/\tau} \varphi(\bar{\mathbf{f}}_r)}{\sum_{r\in\mathcal{T}_d(k)} q_r e^{-|s_r - s_k|/\tau}},$$
$$\mathbf{h}_k = \mathcal{A}\left(\bar{\mathbf{f}}_k, \mathbf{c}_k^{\mathrm{left}}, \mathbf{c}_k^{\mathrm{right}}, \mathbf{c}_k^{\mathrm{global}}, \mathbf{d}_k\right). \tag{1}$$

Here, $\mathbf{c}_k^d$ denotes the context for direction d, and $\mathbf{h}_k$ is the fused anchor context. The functions φ and $\mathcal{A}$ denote a learned feature mapping and context adapter, respectively. Arc-length weighting follows the track, while q_r downweights uncertain supports. We set $\tau = 64\,\mathrm{m}$. The fused anchor context is then propagated to assigned points to obtain $\mathbf{F}_{\mathrm{lr}}$ through the following residual update:

$$\mathbf{f}_i^{\mathrm{lr}} = \begin{cases} \mathbf{f}_i + q_{\pi(i)}\,\mathrm{MLP}_{\mathrm{pt}}\left([\mathbf{f}_i, \mathbf{h}_{\pi(i)}, \mathbf{d}_{\pi(i)}]\right), & \pi(i) \geq 0, \\ \mathbf{f}_i, & \pi(i) = -1, \end{cases} \tag{2}$$

where $\pi(i)$ is the anchor assignment, -1 indicates an invalid assignment, and $[\cdot]$ denotes concatenation. The residual adds shared context while retaining point-specific features.

3.3 GEOMETRY-GUIDED INTERFACE REFINEMENT

The Long-range Track Module aggregates context along ordered supports approximating longitudinal manifold structure, while Euclidean neighborhoods near attachments may still span adjacent supports or distinct semantic regions within a continuous support, causing cross-class feature mixing. Therefore, the geometry-guided interface module leverages the discrepancy between the query and the context as a cue for local variation and conditions the residual correction based on geometric descriptors and support reliability, thereby refining geometry-selected query features. This allows local semantic distinctions to be learned within the shared structural context.

For each geometry-selected query $q \in \mathcal{Q}$, we construct a neighborhood $\mathcal{N}(q)$ from XYZ coordinates using a $1\,\mathrm{m} \times 1\,\mathrm{m}$ grid. Each neighborhood contains up to 16 cells within $3\,\mathrm{m}$. Query features from $\mathbf{F}_{\mathrm{lr}}$ are projected and normalized to obtain $\mathbf{h}_q$, with $\mathbf{h}_j$ computed from mean cell features in the same way. Their scaled dot-product similarities yield softmax weights a_{qj}, which define the context $\mathbf{m}_q = \mathcal{P}(\sum_{j\in\mathcal{N}(q)} a_{qj}\bar{\mathbf{f}}_j^{\mathrm{lr}})$, where $\bar{\mathbf{f}}_j^{\mathrm{lr}}$ is the mean cell feature, and $\mathcal{P}$ applies projection and normalization. To adapt the correction to local support, we use the same weights to aggregate geometric embeddings into $\mathbf{g}_q$ and cell reliability scores into ρ_q. These quantities guide the following gated residual update:

$$\begin{aligned} \gamma_q &= \sigma\left(G_{\mathrm{gate}}\left([\mathbf{h}_q, \mathbf{m}_q, \mathbf{g}_q, \rho_q]\right)\right), \\ \mathbf{f}_q^{\mathrm{out}} &= \mathbf{f}_q^{\mathrm{lr}} + \gamma_q R\left([\mathbf{h}_q, \mathbf{m}_q, \mathbf{h}_q - \mathbf{m}_q, \mathbf{g}_q]\right), \end{aligned} \tag{3}$$

where σ is sigmoid. The networks G_{gate} and R predict the correction strength and residual, respectively. We obtain $\mathbf{F}_{\mathrm{out}}$ by updating query features while leaving the others unchanged.

3.4 SEMANTIC CONTRASTIVE LEARNING AND TRAINING OBJECTIVE

To distinguish locally similar shapes, we apply contrastive learning to valid interface queries. A shared linear head maps features to normalized embeddings $\mathbf{u}_q$. The contrastive loss for anchor q is

$$\ell_{\mathrm{con}}(q) = -\log \frac{\exp(\mathbf{u}_q^\top \mathbf{u}_{p(q)}/\tau_{\mathrm{con}})}{\sum_{r\in\{p(q)\}\cup\mathcal{N}_q} \exp(\mathbf{u}_q^\top \mathbf{u}_r/\tau_{\mathrm{con}})}, \tag{4}$$

where $p(q)$ is the nearest distinct same-class candidate, $\mathcal{N}_q$ contains different-class candidates selected for their high embedding similarity to the anchor, and $\tau_{\mathrm{con}} = 0.1$ is the temperature. We average the losses over anchors to obtain L_{con}. The overall training objective is

$$L = L_{\mathrm{CE}} + L_{\mathrm{Lovasz}} + \lambda_{\mathrm{con}} L_{\mathrm{con}}, \tag{5}$$

where L_{CE} and L_{Lovasz} denote the point-wise cross-entropy and multiclass Lovász–Softmax losses(Berman et al., 2018), respectively, and λ_{con} controls the contrastive loss weight.

4 HAR3D-SEG BENCHMARK

4.1 BENCHMARK DESIGN AND COVERAGE

HAR3D-Seg is a unified benchmark for semantic segmentation of high-aspect-ratio structures in transmission, railway, and vascular scenes. It evaluates long-range semantic consistency and local discrimination under common metrics and predefined splits. Newly collected UAV-LiDAR transmission scenes provide long-span context and fine-grained annotations, while railway and vascular scenes are incorporated from public collections (Qiu et al., 2024; Yang et al., 2020).

As summarized in Table 1, the transmission scenes contain 7,241 million points and 22 classes, offering finer semantic granularity than the compared infrastructure datasets. Figure 3 illustrates elongated structures and nearby semantic boundaries across the three scene types. Aspect-ratio measurements are detailed in Appendix B.

4.2 EVALUATION PROTOCOL

HAR3D-Seg retains scene-specific labels and inputs, with fixed evaluation references shared across methods. Transmission scenes use 1,043 training and 100 test scenes with XYZ inputs. High vegetation is merged into low vegetation, while mIoU and mAcc retain a fixed denominator of 22 class entries.

Table 1: Reference datasets and HAR3D-Seg. Counts describe source collections, with evaluated splits given in Section 4.2. "n.r." denotes a count not reported on a comparable basis.

Dataset / benchmark	Acquisition	Scene type	Scenes	Points(M)	Classes
ECLAIR (Melekhov et al., 2024)	Airborne LiDAR	Utility	1,364	594.022	11
TS40K (Lavado et al., 2025)	Airborne LiDAR	Transmission	24,355	n.r.	5
EHVTL (Yu et al., 2023)	Airborne LiDAR	Transmission	65	318	6
500KV3D (Zhang et al., 2024)	Airborne LiDAR	Transmission	42	29.543	6
HAR3D-Seg (ours)	UAV LiDAR	Transmission	1,143	7,241	22
	Mobile LiDAR	Railway	120	4,614	11
	MRA-derived surfaces	Vascular	116	4.524	2

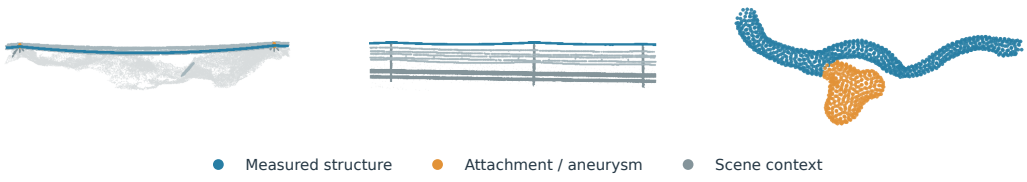


Figure 3: Three scene types in HAR3D-Seg, with measured structures highlighted in blue.

Railway scenes use Urban, Rural, and Plateau train/validation splits of 24/8, 36/12, and 12/4, with XYZ, intensity, and scene-relative height as inputs. Results are reported on the validation sets because test labels are unavailable. Vascular scenes contain 116 annotated objects with vessel/aneurysm labels and use five-fold cross-validation with normalized XYZ and surface normals. Sampling, annotation, and evaluation details are provided in Appendices A and C.

4.3 Evaluation Metrics

HAR3D-Seg reports mIoU, mAcc, and overall accuracy (OA), with mIoU as the primary benchmark metric. Semantic Continuity Score (SCS) and Cross-class Adjacency Score (CAS) further assess long-range continuity and local semantic discrimination. As illustrated in Figure 4, predictions with identical mIoU can differ in structural quality and SCS and CAS are defined as

$$\mathrm{SCS} = \ln \frac{1+\epsilon}{1-\overline{\ell^{\max}/L}+\epsilon}, \qquad \mathrm{CAS} = \ln \frac{1+\epsilon}{1-\overline{N_{\mathrm{correct}}/N_{\mathrm{pair}}}+\epsilon}, \tag{6}$$

where L is the observable path length, and $\ell^{\max}$ sums the longest correct segment in each observable interval. N_{pair} and N_{correct} count all cross-class reference pairs and those with both endpoints correctly classified, respectively. Overbars denote reference averages, and $\epsilon = 10^{-6}$. Evaluation details are provided in Appendix C.

5 Experiments

5.1 Implementation Details

HASP-Net is trained for 80 epochs on transmission scenes and 100 epochs on railway and vascular scenes of HAR3D-Seg. The transmission configuration uses AdamW with batch size 1, weight decay 0.05, learning rates of 3×10^{-4} and 3×10^{-5} for default parameters and backbone blocks, respectively, and a maximum contrastive weight of 0.03. Further settings are provided in Appendix C. HASP-Net is compared with RandLA-Net (Hu et al., 2020), KPConv (Thomas et al., 2019), PointNeXt (Qian et al., 2022), ECKConv (Kim et al., 2026), and PTv3 (Wu et al., 2024) under the HAR3D-Seg evaluation protocol. Metrics follow Section 4.3.

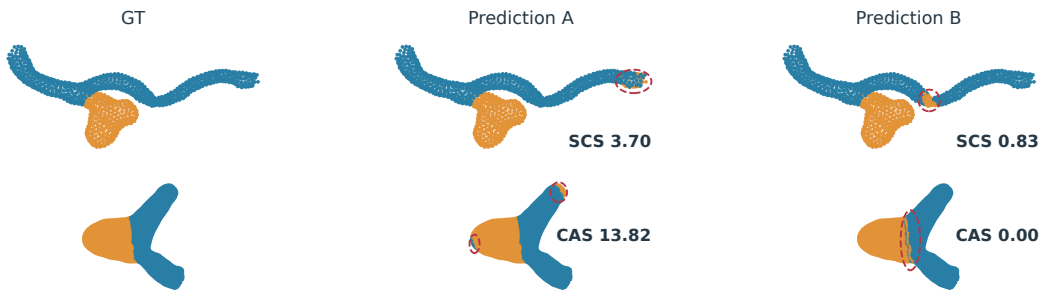


Figure 4: Controlled predictions on vascular scenes with identical mIoU within each row. SCS distinguishes continuity errors, while CAS distinguishes interface errors.

Table 2: Overall HAR3D-Seg results. mIoU, mAcc, and OA are percentages.

Method	mIoU	mAcc	OA	SCS	CAS
RandLA-Net (Hu et al., 2020)	57.28	68.19	88.57	**3.58**	0.62
KPConv (Thomas et al., 2019)	61.37	73.35	88.28	2.86	0.87
PointNeXt (Qian et al., 2022)	47.36	51.45	85.70	1.87	0.55
ECKConv (Kim et al., 2026)	56.81	70.18	87.13	2.57	0.75
PTv3 (Wu et al., 2024)	72.59	79.80	89.86	2.74	1.07
HASP	**74.17**	**80.92**	**96.16**	2.95	**1.13**

5.2 Main Results

Overall benchmark results. Table 2 reports scores averaged across the three scene types. HASP achieves 74.17% mIoU, 80.92% mAcc, 96.16% OA, 2.95 SCS, and 1.13 CAS. The semantic and structural scores jointly characterize class-wise recognition, point-wise accuracy, long-range continuity, and local discrimination. These results are consistent with long-range aggregation providing structural context and interface refinement with contrastive supervision improving separation between neighboring classes.

Transmission scenes. As shown in Table 3, HASP(Full) achieves 62.27% mIoU, exceeding PTv3 by 3.29 points, with the highest SCS and CAS. Both HASP variants preserve long-range context through scene-wide sampling without spatial cropping (Appendix C.2). Figure 5 illustrates how shared long-range context supports consistent predictions along elongated structures, while local refinement helps distinguish small attached components.

Railway scenes. As shown in Table 4, HASP achieves the highest mIoU and CAS on all three railway subsets, with mIoU gains of 0.34–0.50 points over PTv3 and higher SCS throughout. These improvements are consistent with long-range aggregation supporting elongated structures and interface refinement reducing confusion between nearby classes.

Vascular scenes. As shown in Table 5, HASP achieves $92.15 \pm 2.85\%$ mIoU, 96.02% mAcc, and 96.98% OA, with mAcc and OA exceeding PTv3 by 0.84 and 0.42 percentage points, respectively, consistent with improved discrimination between connected vessel and aneurysm regions.

5.3 Ablation Study

As shown in Table 6, Long-range and Interface improve Base by 1.37 and 0.93 mIoU points, respectively. Their combination achieves 61.17% mIoU, further increasing to 62.27% with CL. These gains suggest that structural context complements local refinement, while contrastive supervision helps distinguish geometrically similar classes. Sequential fusion also outperforms parallel fusion in mIoU, likely because local refinement can adapt to the propagated context and correct cross-class mixing.

Table 3: Segmentation results on the transmission scenes of HAR3D-Seg.

Method	mIoU	mAcc	OA	SCS	CAS
RandLA-Net (Hu et al., 2020)	26.32	38.24	79.82	0.90	0.14
KPConv (Thomas et al., 2019)	40.04	62.58	76.15	1.63	0.79
PointNeXt (Qian et al., 2022)	14.85	17.62	81.66	0.41	0.07
ECKConv (Kim et al., 2026)	27.37	46.71	74.63	1.00	0.29
PTv3 (Wu et al., 2024)	58.98	70.40	78.06	1.83	1.02
HASP(Base)	58.49	67.12	95.50	1.83	1.00
HASP(Full)	**62.27**	**72.33**	**96.35**	**1.92**	**1.03**

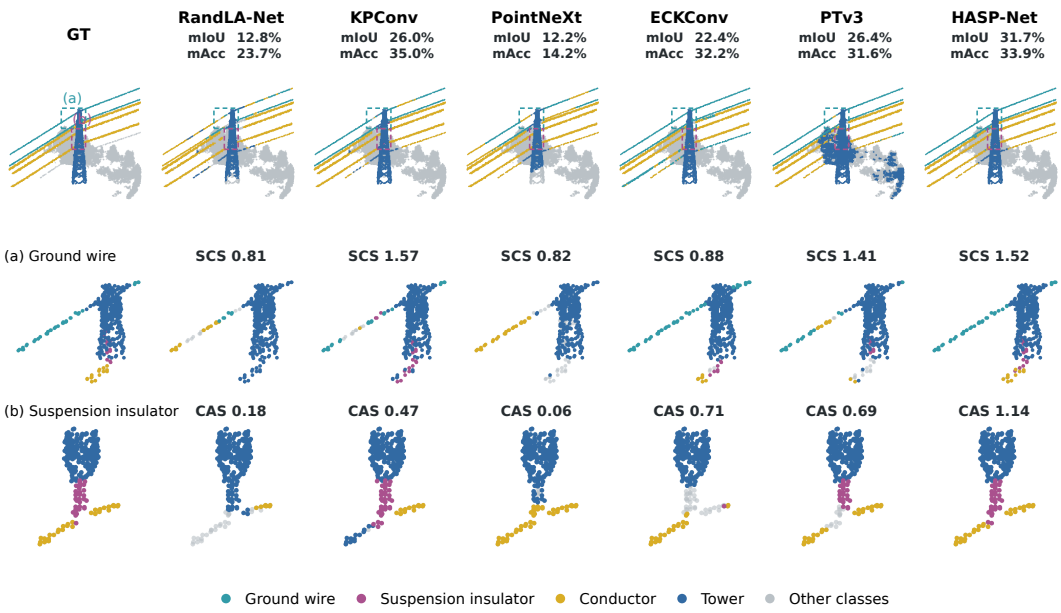


Figure 5: Qualitative comparison on a transmission scene from HAR3D-Seg.

5.4 Inference Efficiency

Figure 6 compares full evaluation pipelines on 100 transmission test scenes. HASP-Net requires 2.673 s per scene, compared with 22.467–653.224 s for the crop-based baselines. The baselines require repeated inference and voting across local crops, whereas HASP-Net processes a scene-wide sample without spatial cropping, reducing redundant computation while preserving long-range context.

6 Conclusions and Limitations

In this study, HAR3D-Seg is established as a unified benchmark for high-aspect-ratio point cloud segmentation across transmission, railway, and vascular scenes. SCS and CAS complement semantic accuracy by assessing long-range continuity and local cross-class discrimination. HASP addresses these challenges through manifold-guided propagation, geometry-guided interface refinement, and contrastive supervision, achieving 74.17% benchmark mIoU with improved SCS and CAS over other methods. Despite these improvements, geometry-consistent support chains lack explicit semantic constraints and may propagate cross-class interference. Unit-level pooling may also mix class features before interface attention, while query-level gating does not explicitly distinguish their contributions. These limitations motivate further investigation into semantic-aware routing and boundary-preserving aggregation.

Table 4: Segmentation results on the Urban, Rural, and Plateau validation subsets of HAR3D-Seg railway scenes.

	Urban			Rural			Plateau		
Method	mIoU	SCS	CAS	mIoU	SCS	CAS	mIoU	SCS	CAS
RandLA-Net (Hu et al., 2020)	70.30	**2.29**	0.62	65.24	2.10	0.69	47.88	**13.82**	2.15
KPConv (Thomas et al., 2019)	71.34	2.02	0.65	66.86	2.11	0.90	38.75	4.15	2.24
PointNeXt (Qian et al., 2022)	66.39	1.64	0.55	57.74	1.78	0.69	38.77	3.82	2.30
ECKConv (Kim et al., 2026)	70.42	1.87	0.93	61.81	**2.23**	0.95	36.75	4.17	2.21
PTv3 (Wu et al., 2024)	80.21	1.96	1.23	71.09	1.70	1.04	51.73	3.59	2.19
HASP	**80.65**	2.08	**1.26**	**71.43**	1.74	**1.12**	**52.23**	4.53	**2.33**

Table 5: Five-fold cross-validation results on HAR3D-Seg vascular scenes. mIoU is reported as mean $\pm$ SD.

Method	mIoU $\pm$ SD	mAcc	OA	SCS	CAS
RandLA-Net (Hu et al., 2020)	84.38 ± 1.96	91.83	93.57	3.76	0.58
KPConv (Thomas et al., 2019)	85.08 ± 2.20	90.42	94.13	**4.19**	0.55
PointNeXt (Qian et al., 2022)	72.92 ± 3.85	84.04	87.96	2.78	0.41
ECKConv (Kim et al., 2026)	86.74 ± 2.89	92.08	94.72	3.96	0.61
PTv3 (Wu et al., 2024)	91.10 ± 1.65	95.18	96.56	3.98	0.69
HASP	$\mathbf{92.15} \pm 2.85$	**96.02**	**96.98**	4.15	**0.78**

Table 6: Ablation results on the transmission scenes of HAR3D-Seg. LR, IR, and CL denote long-range aggregation, interface refinement, and contrastive supervision, respectively.

Variant	LR	IR	CL	mIoU	mAcc	OA	Att. mIoU	Att. mAcc	V-ins. IoU	SCS	CAS
Base	–	–	–	58.49	67.12	95.50	87.45	93.83	74.76	1.83	1.00
Long-range	yes	–	–	59.86	67.18	95.56	89.43	93.70	**85.55**	1.82	0.94
Interface	–	yes	–	59.42	69.31	96.24	85.20	94.79	61.44	1.82	0.95
Interface+CL	–	yes	yes	61.22	71.10	95.90	88.14	93.98	83.55	1.84	**1.08**
Joint	yes	yes	–	61.17	72.14	95.20	86.31	94.59	73.69	1.83	0.95
Joint+CL	yes	yes	yes	**62.27**	**72.33**	**96.35**	88.57	93.72	81.34	**1.92**	1.03
Joint (parallel)	yes	yes	–	60.73	68.12	95.64	**89.51**	**95.12**	84.22	1.81	0.95
Parallel+CL	yes	yes	full	61.00	68.61	95.77	89.36	94.02	84.98	1.81	0.96

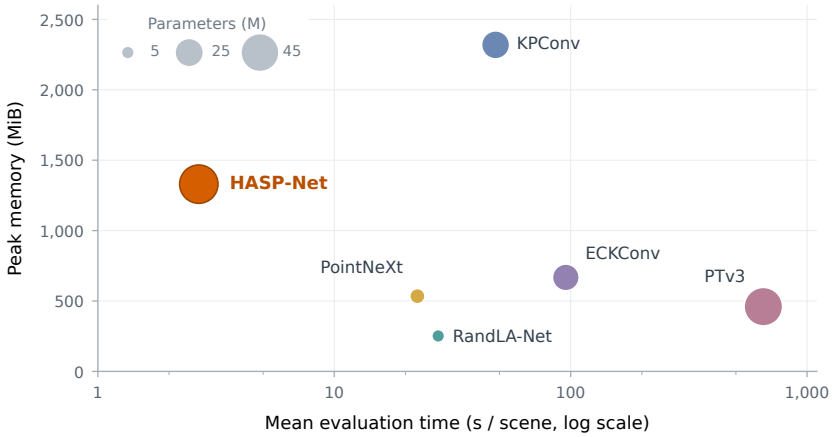


Figure 6: Computational cost on transmission test scenes in HAR3D-Seg using an NVIDIA A100.

AI USE STATEMENT

Generative AI was used for language editing and manuscript organization. The authors verified all technical content and take responsibility for the final submission.

ETHICS STATEMENT

The dataset contains infrastructure scenes, so its release will follow applicable data permissions and protect sensitive location information. Model predictions are intended to support inspection and require human review.

REPRODUCIBILITY STATEMENT

A partial release of the transmission scenes in HAR3D-Seg is available at `https://huggingface.co/datasets/tccx18/Towerdataset`. This release does not yet cover the full benchmark used in this study. The complete benchmark and HASP source code will be released upon acceptance, together with predefined splits, label mappings, structural references, evaluation scripts, and training configurations.

REFERENCES

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1534–1543, 2016.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The SemanticKITTI dataset. *The International Journal of Robotics Research*, 40(8–9):959–967, 2021.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421, 2018.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3070–3079, 2019.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2432–2443, 2017.

Mingtao Feng, Chenbo Yan, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Hyperrectangle embedding for debiased 3d scene graph prediction from rgb sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6410–6426, 2025.

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, 2018.

Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1/W1:91–98, 2017.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11105–11114, 2020.

Jaein Kim, Hee Bin Yoo, Dong-Sig Han, and Byoung-Tak Zhang. Learning coordinate-based convolutional kernels for continuous SE(3) equivariant and efficient point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9986–9995, 2026.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567, 2018.

Diogo Lavado, Ricardo Santos, André Coelho, João Santos, Alessandra Micheletti, and Claudia Soares. Learning under noisy labels, spurious points, and diverse structures: TS40K, a 3d point cloud dataset of rural terrain and electrical transmission systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7326–7336, 2025.

Chao Lin, Shuhei Abe, Shitao Zheng, Xianfeng Li, and Pang jo Chun. A structure-oriented loss function for automated semantic segmentation of bridge point clouds. *Computer-Aided Civil and Infrastructure Engineering*, 40(6):801–816, 2025.

Chuanyu Luo, Nuo Cheng, Wenyu Chen, and Pu Li. Semantic encoding supervises LiDAR semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(10):16973–16985, 2025.

Iaroslav Melekhov, Anand Umashankar, Hyeong-Jin Kim, Vlad Serkov, and Dusty Argyle. ECLAIR: A high-fidelity aerial LiDAR dataset for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision Workshops*, pp. 7627–7637, 2024.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5099–5108, 2017b.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting pointnet++ with improved training and scaling strategies. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23192–23204, 2022.

Bo Qiu, Yuzhou Zhou, Lei Dai, Bing Wang, Jianping Li, Zhen Dong, Chenglu Wen, Zhiliang Ma, and Bisheng Yang. WHU-Railway3D: A diverse dataset and benchmark for railway point cloud semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20900–20916, 2024.

Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17149–17158, 2023. doi: 10.1109/ICCV51070.2023.01577.

Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Jiwen Lu, Mingkui Tan, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pp. 685–702, 2020.

Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6411–6420, 2019.

Nina Varney, Vijayan K Asari, and Quinn Graehling. Dales: A large-scale aerial LiDAR data set for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 717–726, 2020.

Yan Wang, Baoxiong Jia, Ziyu Zhu, and Siyuan Huang. Masked point-entity contrast for open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14125–14136, 2025.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, 2019.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851, 2024.

Xi Yang, Ding Xia, Taichi Kin, and Takeo Igarashi. IntrA: 3d intracranial aneurysm dataset for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2653–2663, 2020.

Hao Yu, Zhengyang Wang, Qingjie Zhou, Yuxuan Ma, Zhuo Wang, Huan Liu, Chunqing Ran, Shengli Wang, Xinghua Zhou, and Xiaobo Zhang. Deep-learning-based semantic segmentation approach for point clouds of extra-high-voltage transmission lines. *Remote Sensing*, 15(9):2371, 2023.

Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8885–8895, 2025.

Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, and Guofeng Zhang. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14114–14124, 2025.

Keying Zhang, Ruirui Cai, Xinqiao Wu, Jiguang Zhao, and Ping Qin. iBALR3D: imbalanced-aware long-range 3d semantic segmentation. *Computer Sciences & Mathematics Forum*, 9(1):6, 2024.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

A Dataset and annotation

This section describes the acquisition, structural diversity, scene geometry, and semantic annotations of the transmission scenes in HAR3D-Seg. The examples and statistics below characterize HAR structures and their small attached components, together with the class imbalance that makes fine-grained segmentation challenging.

Corridor coverage and acquisition. The transmission scenes comprise 1,043 training scenes and 100 test scenes acquired by UAV–LiDAR along 110, 220, and 500 kV transmission corridors, with expert annotations for 22 point-wise classes. They cover diverse support configurations, attachment layouts, and terrain conditions. Training and test scenes are partitioned by geographic region and come from different transmission lines, with no spatial overlap.

Diversity of towers and poles. Figure 7 shows nine selected support configurations from the transmission scenes. Panels (a–f) include suspension and tension supports, while panels (g–i) show a four-circuit steel-tube pole, a wineglass-shaped lattice tower, and a compact distribution pole. Additional configurations vary in shaft geometry, crossarm arrangements, bracing patterns, and attachment layouts. These variations occur within the tower and distribution-tower labels.

Scene scale and terrain complexity. The transmission training set contains 6.237 billion points, with per-scene counts ranging from 0.308M to 61.566M. Figure 9 and Table 7 summarize point counts and spatial extents. Horizontal extent is defined as $\max(x_{\max} - x_{\min}, y_{\max} - y_{\min})$, and vertical extent as $z_{\max} - z_{\min}$, using source LAS headers. Vertical extent includes terrain relief and above-ground objects across the entire scene. It does not measure individual tower height or ground clearance. Training scenes have a median vertical extent of 62.2 m, with 215 scenes exceeding 100 m and 33 exceeding 200 m. Figure 8 shows a steep descent, a steep ascent, and a deep-valley crossing. These examples illustrate substantial changes in support elevation and separation between wire trajectories and the underlying terrain.

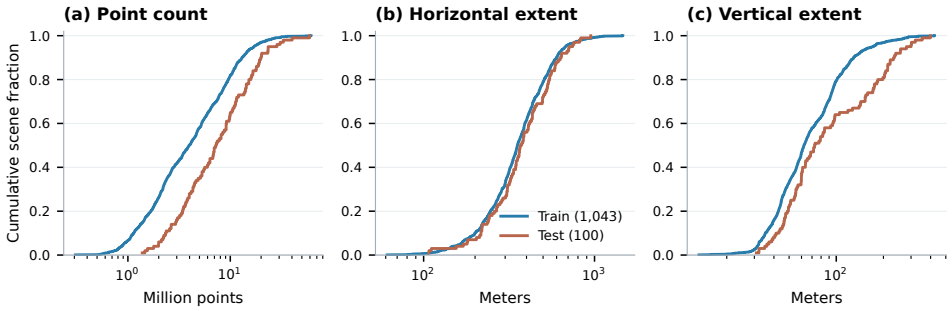


Figure 9: Cumulative distributions of per-scene point count, horizontal extent, and vertical extent for the transmission training and test sets. Horizontal axes use logarithmic scales.

Table 7: Scene-scale statistics for the transmission scenes of HAR3D-Seg, computed from source LAS headers. P10–P90 spans the central 80% of scenes.

Statistic	Training (1,043)	Test (100)
Total points	6.237B	1.004B
Mean points per scene	5.980M	10.045M
Median points per scene	3.964M	7.247M
Point-count range	0.308–61.566M	1.381–59.473M
Median horizontal extent	354.0 m	370.8 m
Horizontal extent range	61.3–1,465.0 m	107.4–950.8 m
Median vertical extent	62.2 m	72.9 m
Vertical extent P10–P90	35.3–129.0 m	42.9–232.1 m
Vertical extent range	13.3–424.4 m	30.9–399.1 m
Scenes with vertical extent > 100 m	215 (20.6%)	36 (36.0%)

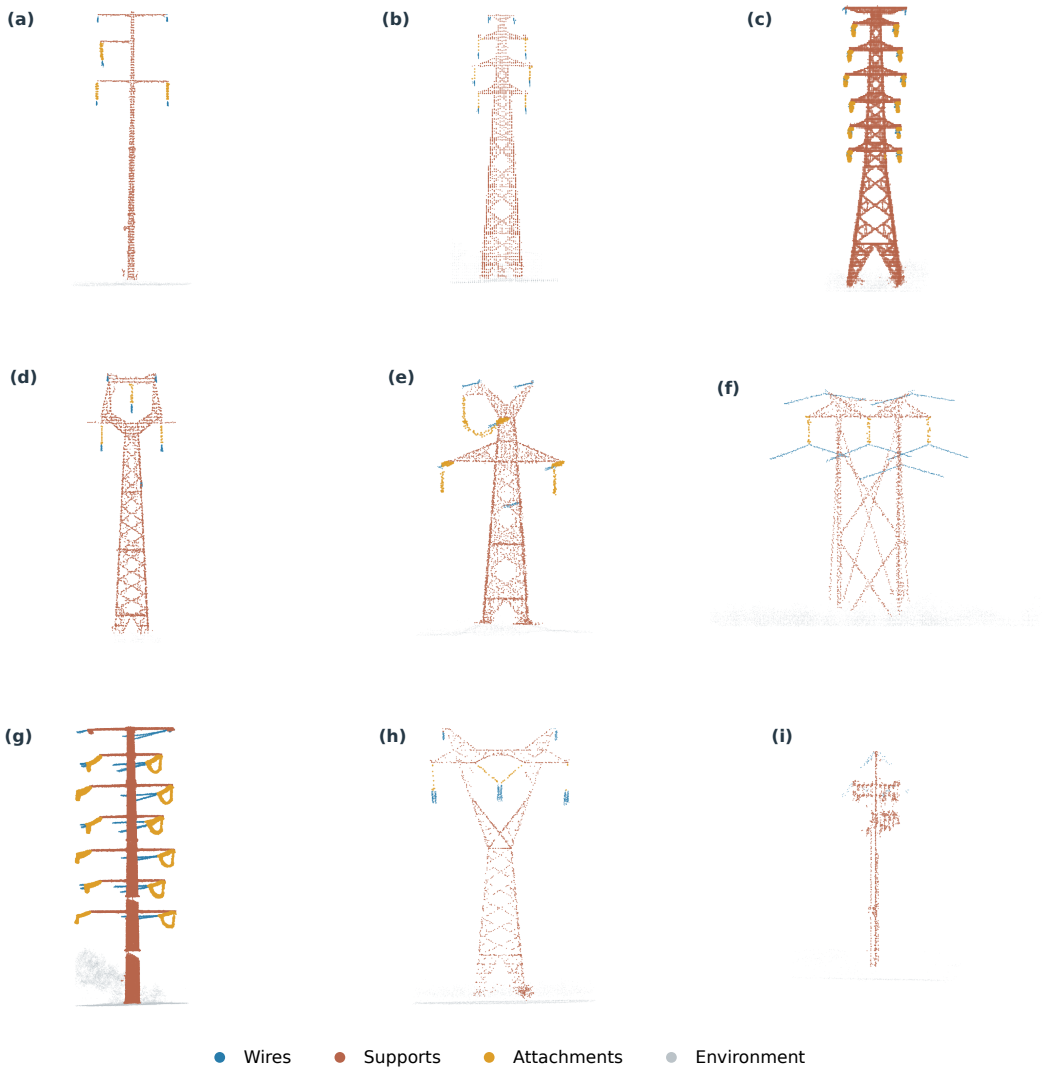


Figure 7: Selected support configurations in the transmission scenes of HAR3D-Seg. In reading order, steel-tube, Gan-shaped, drum-shaped, cat-head, ram-horn, paired concrete poles, four-circuit steel-tube, wineglass lattice, and distribution pole.

High-aspect-ratio structures and semantic granularity. Conductors, ground wires, and optical cables extend over long distances while retaining narrow cross-sections. Slender support members introduce additional elongated geometry, while compact attachments connect or separate nearby structures at wire–support interfaces. Their coexistence creates substantial geometric scale differences and closely adjacent semantic boundaries, requiring both long-range prediction consistency and fine-grained local discrimination.

The taxonomy comprises five long-asset, five attachment, two support, and ten context classes. Table 8 lists the 22 source labels. During loading, high vegetation (ID 9) is merged into low vegetation (ID 8), while the model retains 22 outputs and evaluation uses a fixed 22-class denominator.

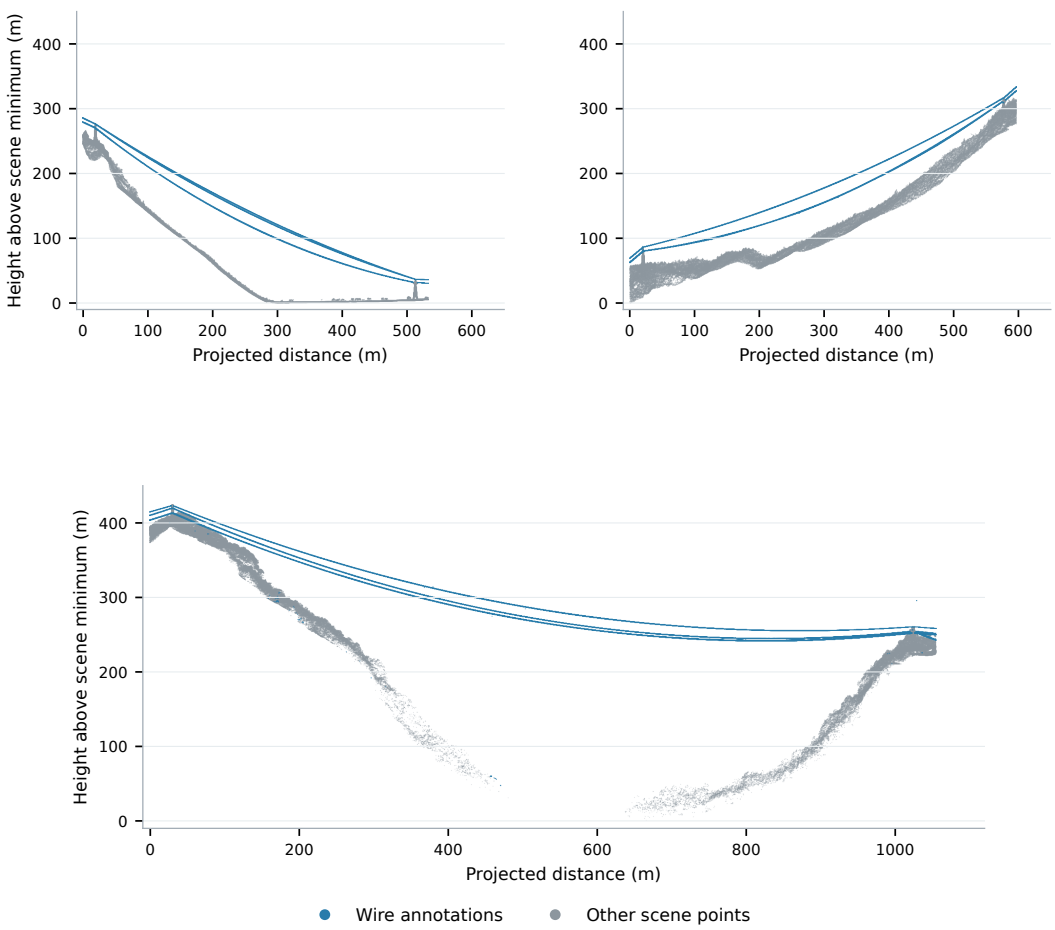


Figure 8: Selected high-relief transmission training scenes. In reading order, steep descent, steep ascent, and deep-valley crossing.

Table 8: Source-label distribution in the 1,043 transmission training scenes.

ID	Class	Points	Share (%)	Role
0	jumper	9,713,387	0.1557	attachment
1	street furniture	330,219	0.0053	context
2	ground	857,184,873	13.7427	context
3	building	47,486,538	0.7613	context
4	road	42,430,740	0.6803	context
5	greenhouse	13,504,377	0.2165	context
6	railway	651,242	0.0104	context
7	vehicle	676,025	0.0108	context
8	low vegetation	4,969,643,300	79.6748	context
9	high vegetation	11,794,269	0.1891	context
10	conductor	98,338,245	1.5766	long asset
11	tower	153,485,571	2.4607	support
12	strain insulator	8,432,149	0.1352	attachment
13	V-insulator	245,754	0.0039	attachment
14	spacer	38,605	0.0006	attachment
15	distribution tower	684,779	0.0110	support
16	distribution conductor	2,039,017	0.0327	long asset
17	shield wire	106,483	0.0017	long asset
18	suspension insulator	1,540,406	0.0247	attachment
19	ground wire	17,571,825	0.2817	long asset
20	optical cable	497,287	0.0080	long asset
21	other	1,010,607	0.0162	context

Class imbalance and small components. Vegetation and ground account for 79.86% and 13.74% of the transmission training points, whereas the five attachment classes together contribute only 0.32%. V-insulators and spacers contain just 245,754 and 38,605 points, respectively. Figures 10 and 11 illustrate this imbalance and the small components that define important semantic distinctions at structural interfaces. Their low point counts limit their influence on overall accuracy, motivating the group-level and class-wise evaluations in Tables 10 and 11, respectively.

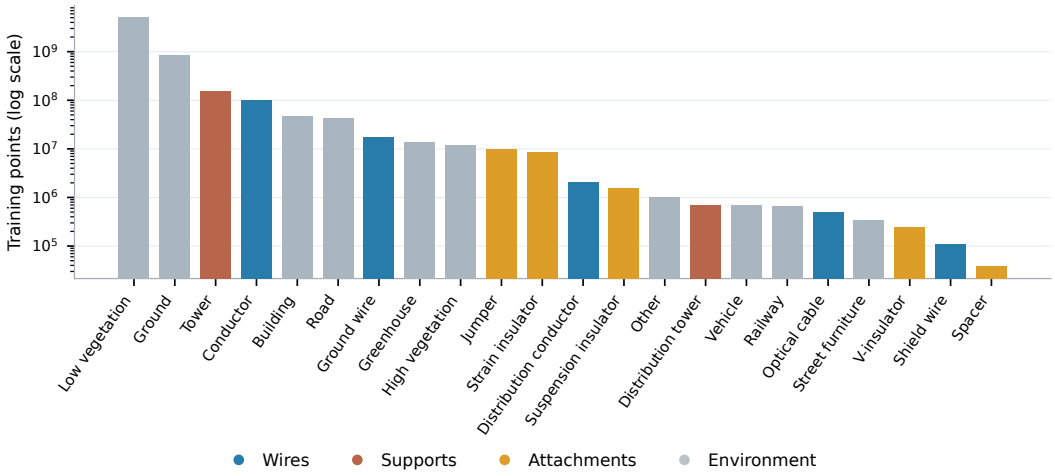


Figure 10: Class distribution and semantic-role imbalance in the transmission training set.

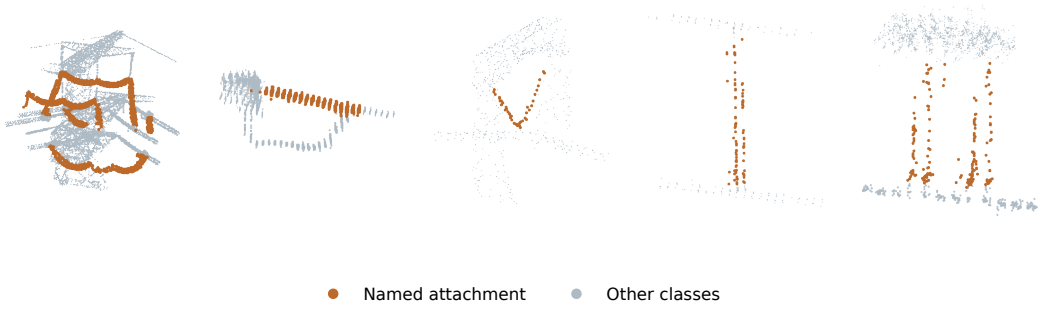


Figure 11: Attached components in the transmission scenes of HAR3D-Seg. From left to right, jumper, strain insulator, V-insulator, spacer, and suspension insulator.

B STRUCTURE-LEVEL ASPECT-RATIO MEASUREMENTS

We measure structure-level aspect ratios across the transmission, railway, and vascular scenes of HAR3D-Seg. Aspect ratio is computed as the supported center-path length divided by an effective cross-sectional diameter. Let $\gamma(s)$ denote the center path with arc length L, and let $A(s) > 0$ denote the cross-sectional area in the plane normal to the local tangent. We define

$$d_A(s) = 2\sqrt{A(s)/\pi}, \qquad d_{\text{eff}} = \text{P90}_{s\sim ds/L}\, d_A(s), \qquad \rho_A = \frac{L}{d_{\text{eff}}}. \tag{7}$$

The effective diameter is the arc-length-weighted 90th percentile of the local equivalent diameters. Measurements are estimated from observed point clouds. Table 9 summarizes the ranges of the measured structures, and Figure 12 shows illustrative examples.

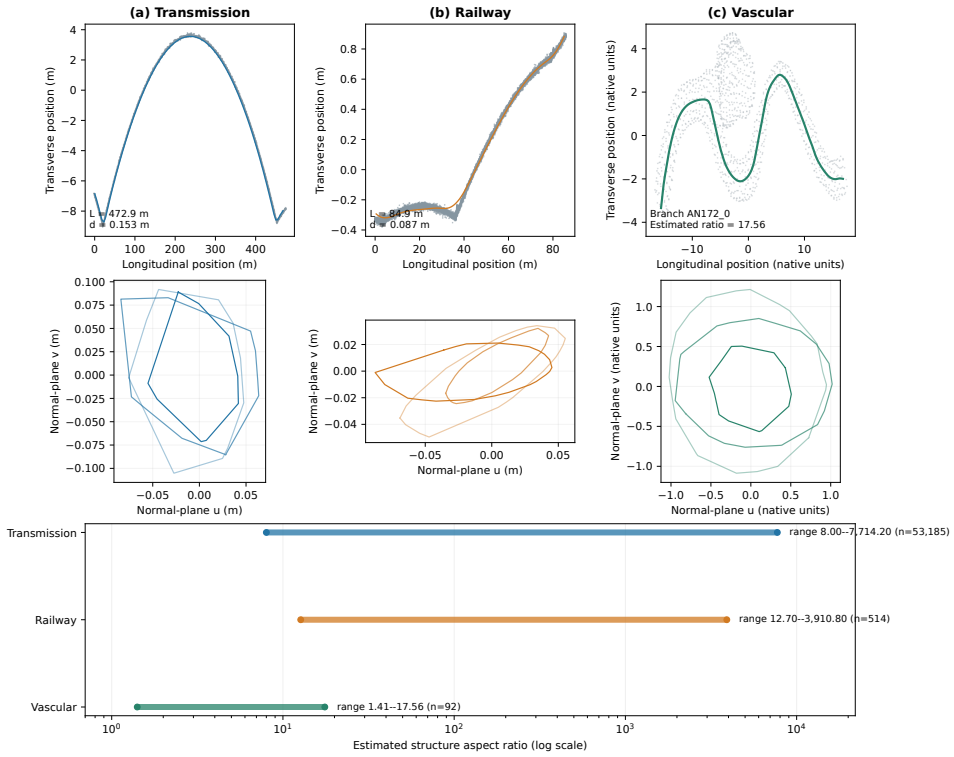


Figure 12: Illustrative structures and estimated aspect ratios. Gray points show the observed geometry, and colored curves indicate measured center paths. The lower panel shows the ranges reported in Table 9.

Table 9: Estimated aspect-ratio ranges for measured structures in HAR3D-Seg.

Scene type	Structure	Aspect-ratio range
Transmission	Wire candidates	8.0–7,714.2
Railway	Overhead-wire candidates	12.7–3,910.8
Vascular	Vessel branches	1.41–17.56

C Evaluation protocol and reproducibility

C.1 Structural references and metric aggregation

SCS and CAS use fixed references constructed offline from coordinates, ground-truth labels, and source point IDs. All methods share these references, which are used only for evaluation. The reference construction details below refer to the transmission scenes of HAR3D-Seg. Benchmark-level scores average the final scores of the three scene types, after averaging the three railway subsets and the five vascular folds.

For SCS, reference voxels within each class form spatial graphs, with additional PCA-direction and lateral-distance constraints for elongated classes. Paths are split at branches or direction changes and divided into $2\,\mathrm{m}$ units. Units containing reference voxels are observable, and valid paths require at least three observable units. A unit is correct when more than half of its points receive the path label. Within each observable interval, only the longest consecutive correct run contributes to $\ell^{\max}$, without bridging gaps. SCS is reported as $\mathrm{N/A}$ when no valid path exists.

For CAS, nearest cross-class point pairs are formed between adjacent regions according to a fixed class-pair list, using distance thresholds specified for the corresponding scene type. Duplicate pairs are removed. A pair is correct only when both endpoints are correctly classified. Region-level

correct-pair ratios are averaged over the class-pair list. CAS is reported as $\mathrm{N/A}$ when no valid pair exists.

C.2 Experimental protocols

For the transmission scenes of HAR3D-Seg, we use XYZ-only inputs on a $0.25\,\mathrm{m}$ grid. PTv3 is trained on spherical crops capped at 120,000 points, whereas both HASP variants use scene-wide samples capped at two million points. HASP(Base) removes both refinement modules and contrastive supervision while retaining the same training and inference pipeline. Source label 9 (high vegetation) is merged into label 8 (low vegetation), and mIoU and mAcc use a fixed 22-class denominator, including zero-valued entries.

For railway scenes, evaluation follows the predefined Urban, Rural, and Plateau splits. All methods are evaluated on the same original source points, with mIoU computed over a fixed set of 11 classes. HASP/PTv3 and RandLA-Net/KPConv use a $0.06\,\mathrm{m}$ grid with inverse projection. PointNeXt/ECKConv use $0.06\,\mathrm{m}$ centroids with nearest-neighbor projection and EMA voting. These processing differences preserve a common evaluation point set and shared structural references.

For vascular scenes, evaluation follows five-fold cross-validation. Each model is trained on four folds for 100 epochs and evaluated on the remaining fold. Within each fold, all methods are evaluated on the same original vertices, and mIoU is computed from the pooled confusion matrix. SCS and CAS use identical structural references across methods.

C.3 Implementation details

This subsection describes track construction, interface query selection, and reliability estimation for HASP on the transmission scenes of HAR3D-Seg. Track construction forms ordered geometric supports for long-range feature aggregation. Each scene is partitioned into $4\,\mathrm{m}$ cells. Cells containing at least four points are described by their PCA direction, width, linearity, and point count. Candidate anchors are connected according to distance, direction, and width consistency, retaining only mutually selected forward and backward links. Tracks with fewer than four anchors or shorter than $16\,\mathrm{m}$ are discarded. Track context updates assigned point features through residual refinement, while other points remain unchanged.

Interface query selection identifies locations for local feature refinement. Eligible points are ranked by local changes in direction, width, and density, then selected across cells in round-robin order. The query budget is $B_q = \min(8192, \lceil 0.05N \rceil)$, where N is the number of input points. Each query gathers at most 16 neighboring cells within $3\,\mathrm{m}$ using a $1\,\mathrm{m} \times 1\,\mathrm{m}$ grid. Refinement is skipped when no query is eligible.

Reliability estimation provides geometric confidence for context aggregation and local refinement. Cell reliability q_k combines linearity, alignment with the selected track, and normalized point support. Query reliability $\rho_q = \sum_{j \in \mathcal{N}(q)} a_{qj} q_j$ uses the same attention weights as feature aggregation and serves as an input to the refinement gate. Empty or zero-weight contexts are represented by zero vectors.

D Additional class analysis

This section analyzes semantic-group and class-level results on the transmission scenes of HAR3D-Seg. Base and Joint+CL correspond to HASP(Base) and HASP(Full), respectively. The first four rows of Table 10 form disjoint groups. Infrastructure combines attachments, wires, and supports, while insulators are a subset of attachments. Compared with Base, Joint+CL improves mIoU in all four groups, with the largest gains on supports ($+8.29$ points) and context ($+5.16$ points). Wire mAcc increases by 2.10 points, consistent with long-range context helping recover points along elongated structures. Context mAcc increases by 9.76 points, indicating improved mean class recall. Attachments gain 1.12 mIoU points despite a 0.11-point decrease in mAcc, showing that improved segmentation does not necessarily increase recall for every group.

Table 11 further illustrates the class-specific gains. Compared with Base, Joint+CL increases tower IoU by 13.33 points with only a 0.67-point rise in recall, indicating that the improvement mainly

Table 10: Semantic-group mIoU and mAcc on the transmission scenes of HAR3D-Seg.

Group	Base		Interface+CL		Joint		Joint+CL	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
attachments	87.45	93.83	88.14	93.98	86.31	**94.59**	**88.57**	93.72
wires	64.28	67.25	65.46	68.53	65.87	68.74	**66.16**	**69.35**
supports	75.88	88.67	82.39	91.12	79.61	91.63	**84.17**	**92.23**
context	37.63	49.39	41.42	56.95	42.56	58.71	**42.79**	**59.15**
infrastructure	75.87	81.89	77.73	82.90	76.68	**83.33**	**78.50**	83.32
insulators	83.01	92.78	85.32	92.82	81.64	**93.27**	**85.59**	92.15

comes from fewer false positives. Interface+CL already achieves 95.33% tower IoU, consistent with local refinement and contrastive supervision helping distinguish neighboring structures. Optical-cable IoU and recall increase by 7.41 and 7.37 points, respectively, suggesting better recovery of missed points along elongated structures. However, Interface+CL achieves slightly higher optical-cable IoU, indicating that long-range aggregation does not improve every class equally.

Table 11: Selected class IoU and recall on the transmission scenes of HAR3D-Seg.

	Base		Interface+CL		Joint		Joint+CL	
Class	IoU	Recall	IoU	Recall	IoU	Recall	IoU	Recall
Jumper	93.56	**95.70**	92.53	94.25	**93.75**	95.59	93.18	95.56
Conductor	95.54	97.11	95.49	97.87	**96.78**	**98.30**	96.72	98.07
Tower	82.73	97.51	95.33	97.49	88.82	97.68	**96.06**	**98.18**
Strain insulator	86.45	93.06	85.07	95.06	84.64	**95.82**	**87.28**	95.09
V-insulator	74.76	94.92	**83.55**	93.88	73.69	**94.97**	81.34	89.74
Spacer	**94.69**	95.11	92.21	97.19	92.88	**97.56**	92.89	96.60
Distribution tower	69.02	79.83	69.44	84.75	70.40	85.58	**72.28**	**86.27**
Suspension insulator	87.81	90.35	87.33	89.53	86.60	89.03	**88.16**	**91.63**
Ground wire	89.89	96.34	88.64	92.66	**92.05**	96.47	91.83	**97.10**
Optical cable	50.21	51.76	**57.88**	**59.45**	56.52	58.06	57.62	59.13

E ADDITIONAL QUALITATIVE RESULTS

Figures 13–15 compare ground truth with HASP and baseline predictions on the transmission and railway scenes of HAR3D-Seg.

In the transmission line examples, several baselines confuse ground wires with conductors or assign sparse distribution-conductor points to vegetation and other classes. HASP produces more consistent ground-wire labels and recovers distribution-conductor points missed by PTv3. These results are consistent with long-range context supporting predictions where local evidence is weak. Near structural interfaces, HASP better distinguishes strain and suspension insulators from neighboring structures and recovers distribution-tower points confused with vegetation. These examples support the role of interface refinement and contrastive supervision in reducing local cross-class confusion.

The railway Plateau example shows a sparse overhead-line fragment. HASP recovers more overhead-line points than PTv3, although confusion with poles and other classes remains. This example illustrates both the benefit of structural context and the remaining difficulty of distinguishing nearby classes under sparse observations.

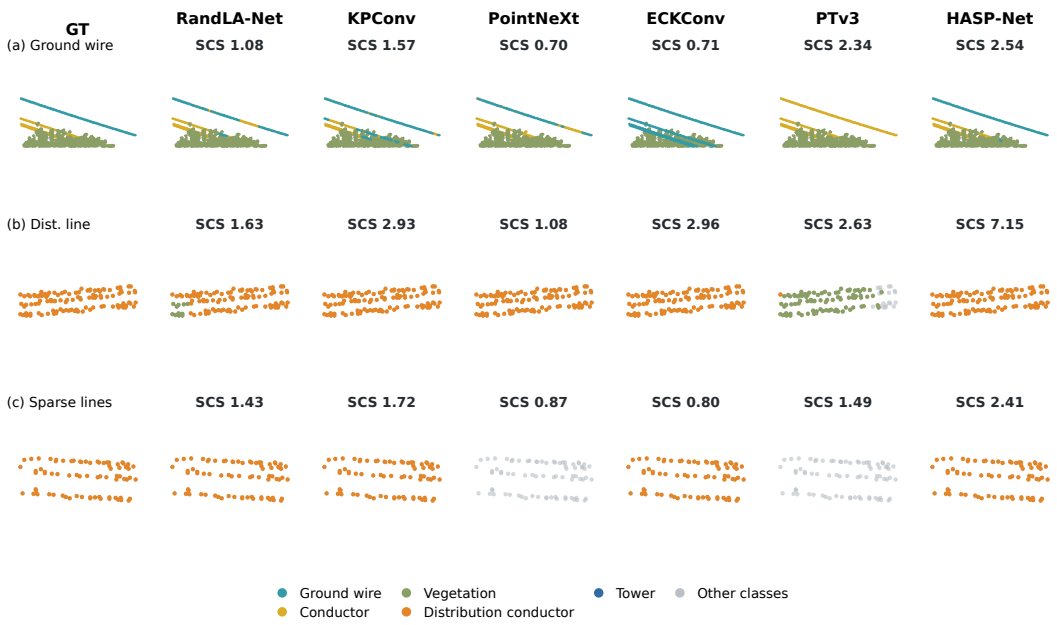


Figure 13: Long-range continuity comparisons in the transmission scenes of HAR3D-Seg. Rows show ground-wire, distribution-conductor, and sparse distribution-conductor examples. Columns show ground truth and predictions from six methods, with SCS reported.

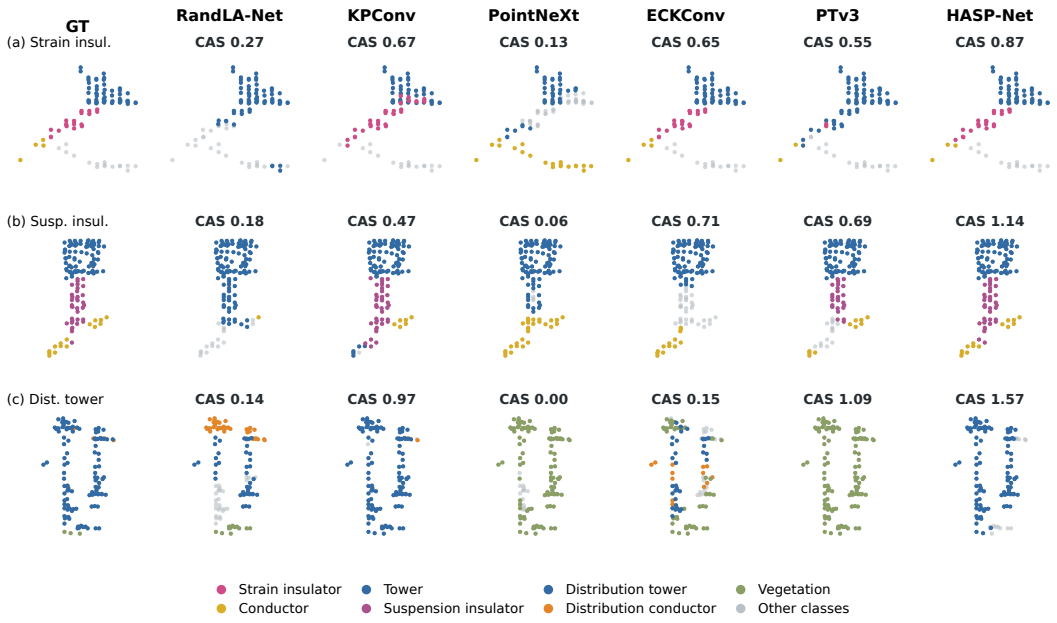


Figure 14: Local semantic discrimination in the transmission scenes of HAR3D-Seg. Rows show strain-insulator, suspension-insulator, and distribution-tower examples. Columns show ground truth and predictions from six methods, with CAS reported.

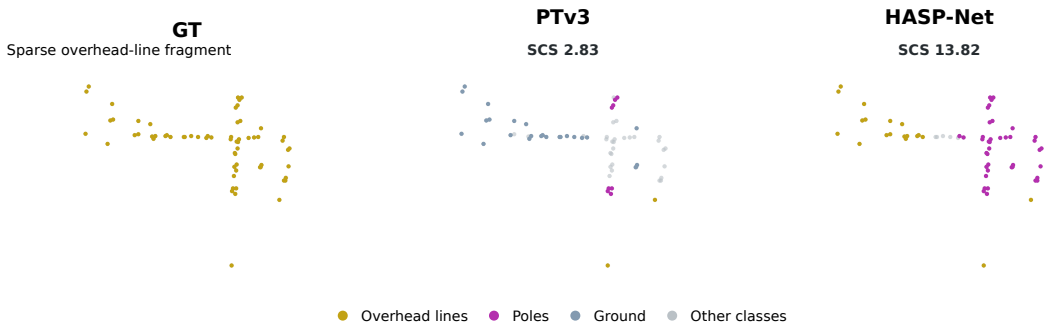


Figure 15: Sparse overhead-line example from the Plateau subset of the railway scenes in HAR3D-Seg. Columns show ground truth, PTv3, and HASP, with SCS reported.