SERVE: Semantic Evidence from Return Verification for LiDAR Segmentation

Anonymous authors
Paper under double-blind review

Abstract

An unfamiliar object can receive a confident semantic label when a LiDAR model only compares known classes. We propose Semantic Evidence from Return Verification (SERVE), which combines appearance support with class-conditioned range compatibility. A full-scan encoder supplies appearance support, while a separate predictor excludes the target angular cell and produces a distribution over log range for each class. Its density at the recorded range enters the supervised class decision alongside appearance support. The strongest joint class support supplies both the semantic label and the unknown score. Training uses normal semantic labels and recorded ranges. On STU, SERVE reaches 73.94% validation and 61.74% test average precision, compared with 21.36% and 15.09% for a matched appearance model, while improving normal semantic mIoU by 1.24 points. At a 1% normal false-positive rate, it recovers 81.6% of unknown points confidently assigned to known classes by the appearance model. Matched ablations assess the contributions of target exclusion, class-conditioned measurement support, and joint semantic training. These results support integrating learned range compatibility into semantic recognition.

1 Introduction

LiDAR segmentation must recognize familiar traffic participants while identifying objects outside its training vocabulary. A classifier can confidently label an unfamiliar road obstacle as a car or road because confidence expresses a preference among the available classes. A recorded range provides an additional quantity whose compatibility with each candidate class can be learned from normal scans. Real road-anomaly benchmarks such as STU provide a setting for testing whether this compatibility improves unknown detection (Nekrasov et al., 2025).

Existing methods estimate unknownness from classifier outputs, normal feature structure, or auxiliary anomalous examples (Liu et al., 2020; Vojíř et al., 2024; Mosco et al., 2026; Li et al., 2026). In a LiDAR scan, neighboring ranges also contain information about a target return. Under a locally continuous road hypothesis, surrounding ground returns can support a concentrated range prediction; a car hypothesis can support a different continuation. Class-conditioned prediction expresses these alternatives. Our hypothesis is that their compatibility with the measured range complements appearance when recognizing unknown objects. We test this selectivity by comparing unknown recall at fixed normal false-positive rates.

We introduce Semantic Evidence from Return Verification (SERVE). For every point, a full-scan encoder supplies appearance support for each known class. A separate context encoder predicts a range distribution for each class from neighboring angular cells. We use return verification to mean evaluating this learned conditional density at the recorded range. The resulting measurement support quantifies class-conditioned range compatibility. SERVE combines it with appearance support within each class and uses the resulting energies for semantic labeling and unknown detection. Normal semantic labels supervise this combined decision, training the range predictor through semantic competition.

The prediction target is a recorded physical measurement that stays fixed while the semantic representation learns. The predictor excludes the complete target cell before any cross-cell interaction, so its prediction depends on surrounding returns and the candidate class. A normalized predictive distribution accounts for both residual error and predictive spread: broad predictions accommodate

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.

variation at the cost of lower peak density. Normal likelihood and appearance compactness objectives train absolute support, complementing the relative class preferences learned by cross-entropy.

SERVE brings class-conditioned prediction of recorded measurements into dense semantic recognition. Its decision rule adds an energy term when appearance and measurement favor different classes. We derive this relationship and evaluate its effect on unknown ranking through matched training and fixed-weight inference comparisons. Under matched normal training data, SERVE improves STU test AP by 46.65 points over its appearance control. Removing measurement evidence from class supervision reduces validation AP by 25.67 points, supporting its direct participation in semantic learning.

2 Related Work

Open-world LiDAR perception. Classifier confidence and energy provide inexpensive unknown scores (Hendrycks & Gimpel, 2017; Liu et al., 2020). RbA and PixOOD develop class rejection and normal-feature modeling for dense perception (Nayal et al., 2023; Vojíř et al., 2024); LIDO uses prototype geometry for normal-only LiDAR anomaly segmentation (Mosco et al., 2026). Other methods train with auxiliary anomalous observations: REAL scales objects, LiON introduces diverse synthetic objects, DOODLE generates examples with diffusion, and NDP combines synthetic geometry with a learned prior over network predictions (Cen et al., 2022; Xu et al., 2025; Oh et al., 2026; Li et al., 2026). COVAL varies the visibility of generated anomalies to learn consistent representations (Yang et al., 2026). SERVE learns from normal semantic labels and recorded ranges, using neighboring returns to test each class's prediction of a held-out measurement. Target exclusion defines the information available to this predictor.

Reconstruction as class support. C2AE conditions reconstruction on class hypotheses, and CoReSeg extends class-conditioned reconstruction to open-set segmentation (Oza & Patel, 2019; Nunes et al., 2022). CSSR uses class-specific autoencoders to reconstruct learned semantic features and trains classification through reconstruction-error logits (Huang et al., 2023). SERVE uses a fixed physical target: it predicts the recorded range from target-excluded context and combines the resulting normalized conditional density with appearance support. This choice ties class evidence to an observable quantity throughout representation learning. Joint energy models share class scores between recognition and generative learning over complete inputs (Grathwohl et al., 2020). SERVE combines an appearance support function and a conditional measurement density in a discriminative class energy.

Prediction from restricted context. Blind-spot self-supervision excludes an observation before predicting it (Batson & Royer, 2019; Laine et al., 2019; Krull et al., 2020). Conditional neural processes parameterize distributions at queries from observed context (Garnelo et al., 2018; Kim et al., 2019), and masked point-cloud learning exploits spatial prediction objectives (Pang et al., 2022). SERVE applies restricted-context prediction to semantic recognition: each class predicts the target return, and its density at the recorded range contributes to the class decision. Proper scoring rules provide the basis for normal likelihood fitting (Gneiting & Raftery, 2007). Studies of generative likelihood show that anomaly rankings depend on the modeled observation and its coordinates (Nalisnick et al., 2019; Le Lan & Dinh, 2021); SERVE specifies log range as its target and evaluates measurement support jointly with appearance.

3 Semantic Evidence from Return Verification

3.1 One Support Function for Labeling and Rejection

A scan is $X = \{(\mathbf{x}_i, I_i)\}_{i=1}^{N}$, with position $\mathbf{x}_i$ and intensity I_i. Write $r_i = \|\mathbf{x}_i\|_2$, $\boldsymbol{\omega}_i = \mathbf{x}_i / r_i$, and $z_i = \log(r_i / r_0)$ with $r_0 = 1\,\mathrm{m}$. The known vocabulary is $\mathcal{C} = \{1, \ldots, C\}$. A normal training label specifies either a class or an allowed set $Y_i \subseteq \mathcal{C}$ when source labels are coarser than the target vocabulary.

For each class c, SERVE computes appearance support $a_{ic} \in (0, 1]$ and a log-range density $p_{ic} = p_\theta(z_i \mid C_i, \boldsymbol{\omega}_i, c)$ from target-excluded context C_i. Let M bound the density and κ_i indicate whether

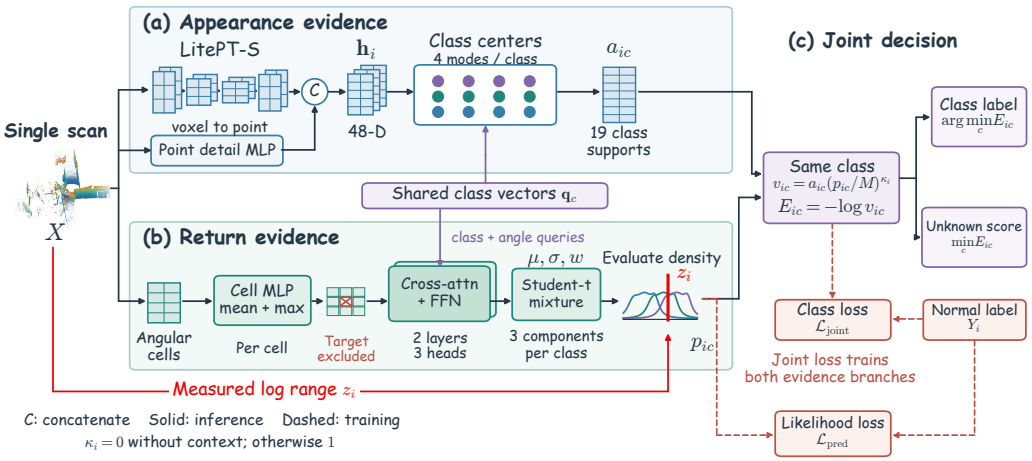


Figure 1: SERVE. The full-scan branch produces appearance support. Independently encoded neighboring cells supply a separate predictor. Each class predicts a distribution over the held-out target's log range. Combining its density at the measured range with appearance support yields the semantic label, unknown score, and supervised class probabilities. The indicator κ_i selects appearance-only inference for empty context. A recorded input scan and schematic density curves illustrate the computation.

any prescribed context cell is available. Define

$$s_{ic} = -\log a_{ic}, \qquad g_{ic} = \kappa_i(\log M - \log p_{ic}), \qquad E_{ic} = s_{ic} + g_{ic}. \tag{1}$$

The class support is $v_{ic} = \exp(-E_{ic})$. The semantic prediction and unknown score are

$$\hat{y}_i = \arg\min_c E_{ic}, \qquad u_i = \min_c E_{ic}. \tag{2}$$

A point is rejected when u_i exceeds a threshold. With no neighboring cell, $g_{ic} = 0$ for all classes and inference uses appearance support alone. At fixed appearance energies, this fallback gives $u_i = \min_c s_{ic}$, whereas supported prediction gives $u_i \geq \min_c s_{ic}$. Context availability therefore introduces a structural difference in the score scale; Appendix B.5 derives its dependence on M.

Same-class evidence combination. Let $s_i^* = \min_c s_{ic}$ and $g_i^* = \min_c g_{ic}$. Then

$$u_i = s_i^* + g_i^* + \Delta_i, \qquad \Delta_i = \min_c[(s_{ic} - s_i^*) + (g_{ic} - g_i^*)] \geq 0. \tag{3}$$

The additional term vanishes exactly when at least one class minimizes both energies. Independently taking the best appearance and the best measurement explanation drops Δ_i, even when they belong to different classes. For example, appearance supports $(0.9, 0.1)$ and measurement supports $(0.1, 0.9)$ produce a best same-class product of 0.09, versus 0.81 after independent maximization. This is a pointwise score increase. Its effect on detection depends on the relative increments for normal and unknown points, which we evaluate using fixed-weight readouts.

3.2 Appearance and Conditional Return Prediction

Multimodal appearance. A LitePT-S backbone (Yue et al., 2026) encodes the scan with 5 cm voxels. Point features derived from coordinates, intensity, and within-voxel offsets preserve detail lost by voxel aggregation. Concatenating both representations gives an embedding $\mathbf{h}_i \in \mathbb{R}^D$, $D = 48$. A shared class vector $\mathbf{q}_c$ and $A = 4$ offsets represent normal variation:

$$\mathbf{t}_{cm} = \sqrt{D}\,\frac{\mathbf{q}_c + \boldsymbol{\delta}_{cm}}{\|\mathbf{q}_c + \boldsymbol{\delta}_{cm}\|_2}, \qquad a_{ic} = \frac{1}{A}\sum_{m=1}^{A} \exp\left(-\frac{\|\mathbf{h}_i - \mathbf{t}_{cm}\|_2^2}{D\tau}\right), \tag{4}$$

with $\tau = 0.2$. Common norms, temperature, and mode count set a shared scale. The same class vectors query the measurement predictor and receive gradients from both branches under the joint objective.

Excluding the target before prediction. Azimuth and elevation partition returns into 0.5° cells. A separate pointwise encoder processes raw returns, and independent within-cell mean/max pooling produces one token per occupied cell. For a target cell, two cross-attention layers query eight neighboring offsets at each of radii one and three cells, giving 16 context positions. Queries contain the candidate class and target angular position. The predictor receives only these neighboring tokens and queries; all target-cell returns are excluded before cross-cell interaction. After predicting the distribution parameters, the model evaluates the density at the target's recorded range. This order holds for every return in the target cell, including multiple returns along nearby directions.

Predicting a distribution of log ranges. Each class predicts a $K = 3$ component Student-t mixture with three degrees of freedom. For target cell j, ray offsets (d_a, d_e) normalized by the half-cell width define $\boldsymbol{\phi}_i = (d_a, d_e, d_a^2, d_a d_e, d_e^2)$. The component means and scales are

$$\mu_{ick} = \bar{z}_{C_j} + \alpha_{jck} + \boldsymbol{\beta}_{jck}^{\mathsf{T}} \boldsymbol{\phi}_i, \qquad \sigma_{jck} = \sigma_{\min} + \mathrm{softplus}(\rho_{jck}), \tag{5}$$

where $\bar{z}_{C_j}$ averages neighboring cells' mean log ranges and $\sigma_{\min} = 10^{-3}$. Softmax weights w_{jck} define

$$p_{ic} = \sum_{k=1}^{K} w_{jck} \frac{2}{\pi\sqrt{3}\,\sigma_{jck}} \left(1 + \frac{(z_i - \mu_{ick})^2}{3\sigma_{jck}^2}\right)^{-2}. \tag{6}$$

Each mixture integrates to one in log-range space and is bounded by $M = 2/(\pi\sqrt{3}\,\sigma_{\min})$. Hence all energies in Equation (1) are nonnegative. Increasing a component's scale tenfold reduces its peak density tenfold, so measurement support accounts for predictive spread as well as residual error. Multiple components represent alternative surface continuations.

3.3 Training Measurement Evidence through Semantic Competition

Semantic supervision uses $P_{ic} = \mathrm{softmax}_c(-E_{ic})$ and an allowed-set loss:

$$\ell_{\text{joint},i} = -\log \sum_{c \in Y_i} P_{ic}, \qquad \frac{\partial \ell_{\text{joint},i}}{\partial \log p_{ic}} = \kappa_i (P_{ic} - P_{ic}^Y), \tag{7}$$

where $P_{ic}^Y = \mathbf{1}[c \in Y_i] P_{ic} / \sum_{d \in Y_i} P_{id}$ and the partial derivative holds the other intermediate inputs fixed. For a fine label, the true class receives a likelihood-increasing gradient and competing classes receive the opposite gradient. Measurement prediction thereby participates directly in the semantic decision.

Cross-entropy depends on relative energies: adding the same scalar to every class leaves P_{ic} unchanged but shifts u_i. We therefore also train absolute support. Context logits η_{ic} define weights $\pi_{ic}^Y = \exp \eta_{ic} / \sum_{d \in Y_i} \exp \eta_{id}$ over allowed classes. The predictive and compactness losses are

$$\ell_{\text{pred},i} = -\log \sum_{c \in Y_i} \pi_{ic}^Y p_{ic}, \qquad \ell_{\text{compact},i} = -\log \left(\frac{1}{|Y_i|} \sum_{c \in Y_i} a_{ic} \right). \tag{8}$$

For fine labels, predictive training is true-class negative log likelihood (NLL). For coarse labels, it fits the mixture over the allowed classes. An appearance classification loss uses $\mathrm{softmax}(-s)$, and a context classification loss uses $\mathrm{softmax}(\eta)$, both with the same allowed-set formulation. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{joint}} + 0.5\mathcal{L}_{\text{appearance}} + 0.05\mathcal{L}_{\text{compact}} + 0.2\mathcal{L}_{\text{pred}} + 0.2\mathcal{L}_{\text{context}}. \tag{9}$$

Classification losses balance distinct allowed label sets within each frame. Joint supervision uses up to $4{,}096$ normal queries; the appearance loss covers all reliably labeled points. Predictive and context losses exclude queries with $\kappa_i = 0$. Appendix C gives exact reductions.

The objective trains normalized conditional densities for a discriminative decision: predictive NLL rewards fit to normal ranges, while semantic cross-entropy rewards class separation. The fitted density is a learned compatibility model shaped by both terms. Appendix B derives how classification can shift its optimum away from the true conditional distribution. Section 4 evaluates predictive coverage and anomaly ranking as separate properties. The combined quantity v_{ic} represents class support.

Table 1: Normal development mIoU and point-level anomaly detection (%). The upper block reports three-seed means under matched normal-only training; bold marks its best results. The lower block reports published results with each method's original backbone, training data, and supervision: LIDO (Mosco et al., 2026), NDP (Li et al., 2026), and COVAL (Yang et al., 2026). COVAL's ML and AL scores occupy separate rows; n/r denotes an unreported value.

	201	STU validation			STU test		
Model	mIoU↑	AP↑	AUROC↑	FPR95↓	AP↑	AUROC↑	FPR95↓
Matched normal-only training							
Softmax	61.84	12.60	96.40	20.84	7.31	95.05	25.47
Energy	61.84	18.72	97.12	16.35	12.64	96.51	20.08
Appearance	61.92	21.36	97.86	12.48	15.09	97.05	16.22
CSSR-style	62.48	33.81	98.12	10.77	25.09	97.45	12.98
Separate objectives	62.10	48.27	98.59	6.71	37.42	98.03	8.91
SERVE	**63.16**	**73.94**	**99.41**	**1.74**	**61.74**	**99.08**	**3.18**
Published external references							
LIDO	n/r	27.53	95.05	34.86	14.99	93.67	34.29
NDP-EE	n/r	74.24	99.53	1.43	61.31	99.26	3.30
COVAL (ML)	n/r	64.35	97.34	17.97	n/r	n/r	n/r
COVAL (AL)	n/r	62.16	97.70	7.33	n/r	n/r	n/r

4 Experiments

4.1 Evaluation Design

Normal training and selection. All matched models start from official nuScenes-pretrained LitePT-S weights, followed by source semantic training and normal STU adaptation with source replay (Caesar et al., 2020). They share initialization of common modules, supervision, and training budgets. Checkpoints are selected using normal development mIoU, with the normal objective breaking ties; anomaly labels are reserved for evaluation. Appendix C specifies the data, label mapping, optimization schedule, and three paired random seeds.

Metrics and populations. We report normal development mIoU over the 14 classes present in STU sequence 201, and anomaly AP, AUROC, and FPR95 on STU validation and hidden test. Official point evaluation retains distances from 2.5 to 50 m, applies the benchmark label mask, and includes a scan when it contains at least five valid anomaly points (Nekrasov et al., 2025). Objects with one to four points remain included in eligible scans. Metrics pool valid point predictions before reduction. Fixed-FPR comparisons set each model's threshold on the same evaluation normal population, giving a shared false-alarm budget.

Matched controls. The appearance control is independently trained using appearance classification and compactness. The separate-objectives control retains SERVE's architecture, shared class vectors, predictive losses, and joint inference, and uses appearance support alone in the classification loss. This comparison tests direct supervision of measurement evidence through class competition. Our dense adaptation of CSSR reconstructs backbone features with class-specific autoencoders and uses reconstruction errors as class logits. Standard softmax and energy use a linear class head. Matched runs share backbone initialization, scan order, normal supervision, and update budget. Full definitions appear in Appendix D.

4.2 Unknown Detection and Normal Perception

SERVE raises validation AP from 21.36 to 73.94 and test AP from 15.09 to 61.74 relative to the matched appearance model (Table 1). Test FPR95 falls from 16.22% to 3.18%, alongside a 1.24-point increase in normal development mIoU. The separate-objectives model combines predictive training with joint inference and reaches 48.27% validation AP. Adding direct measurement-based

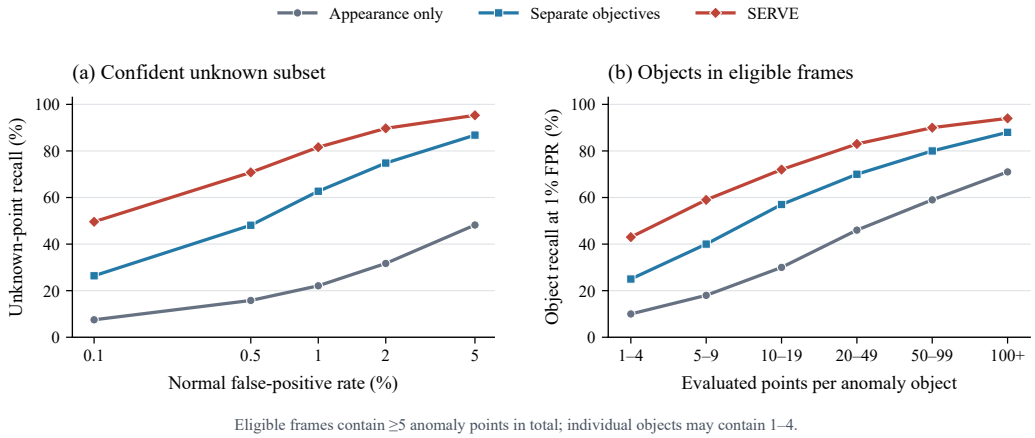


Figure 2: Verification under fixed false-alarm budgets. Left: point recall on unknown points that the independently trained appearance model assigns a known class with confidence at least 0.9. The subset is fixed across methods within each seed. Right: recall of anomaly instances by valid-return count at 1% pooled normal FPR; an instance is detected when at least half its valid points are detected. Instances with 1–4 points are retained in eligible scans. Values are three-seed means.

classification gradients in SERVE brings a further 25.67-point gain. Fixed-weight readouts below test measurement support's contribution during inference.

The published references place these results alongside existing STU methods under their respective training settings. SERVE's test AP and FPR95 are 61.74% and 3.18%, close to NDP-EE's 61.31% and 3.30%. NDP-EE has higher AUROC on both splits and stronger validation results, including 74.24% AP versus 73.94%. SERVE achieves this performance using normal-only training and leads all three anomaly metrics within the matched comparison. Its three-seed standard deviations are 1.08 points for validation AP and 1.26 for test AP (Appendix F).

4.3 Unknown Detection at High Semantic Confidence

For each seed, the independently trained appearance model fixes an unknown-point subset with maximum known-class softmax probability at least 0.9. We evaluate every method's unknown score on these same point identities. This subset tests rejection of points for which the appearance model strongly favors one known class.

At 1% normal FPR, SERVE detects 81.6% of this subset, versus 22.1% for appearance and 62.7% for separate objectives (Figure 2). Relative to appearance, 62.1% of the subset is newly recovered and 2.6% is lost, giving the 59.5-point net gain. At 5% FPR, SERVE reaches 95.3% recall. Normal decisions also improve: SERVE corrects appearance errors on 1.74% of normal points and introduces errors on 0.96%, a net point-accuracy gain of 0.78 points. These paired changes indicate that the additional evidence can revise a known-class decision as well as raise an unknown score.

Recall improves across object sizes. In the per-scan instance-coverage diagnostic, SERVE recovers 43% of one-to-four-point instances and 94% of instances with at least 100 valid points at the same 1% normal FPR. The appearance model recovers 10% and 71%, respectively. The gains extend measurement-based detection to sparse objects, which remain the more difficult group for both models.

4.4 Contributions of the Model Components

Target exclusion improves rejection evidence. Allowing the target token into attention reduces AP to 39.62%, while normal predictive NLL improves from -2.08 to -4.72. In this comparison, access to the target makes range prediction easier and weakens the resulting rejection score. Target-cell exclusion removes that measurement from its own predictor; the density is evaluated at the recorded range after prediction.

Table 2: Validation ablations (%). R retrains a model; I changes inference with fixed SERVE weights. The common-density readout adds identical measurement energy to every class, preserving appearance-only semantic labels. Independent minima changes only the unknown score.

Variant	Type	201 mIoU↑	AP↑	FPR95↓
SERVE	R	63.16	73.94	1.74
Separate objectives	R	62.10	48.27	6.71
Target cell available	R	63.03	39.62	10.34
One predictive component	R	62.79	67.46	3.24
Without predictive NLL	R	62.31	61.85	4.61
Without appearance compactness	R	62.94	68.17	2.98
Appearance-only readout	I	62.73	29.42	10.88
Common-density readout	I	62.73	51.93	5.38
Independent minima score	I	63.16	59.18	4.36

Class-specific support matters beyond geometric surprise. Replacing each class density by their uniform mixture gives a common-density readout. It reaches 51.93% AP, below SERVE's 73.94%. Because the added energy is identical across classes, its labels exactly match the appearance-only readout, at 62.73 mIoU. Separately minimizing appearance and measurement energies also reduces AP, to 59.18%, while leaving joint labels unchanged by construction. These fixed-weight comparisons measure the effects of class conditioning and the disagreement term in Equation (3) on the evaluation ranking.

Absolute-support objectives and multimodality complement classification. Removing predictive NLL or appearance compactness reduces AP to 61.85% or 68.17%. Replacing the three-component predictor with one component reduces it to 67.46%. Together, these comparisons show how normal predictive fitting, concentrated appearance support, and multiple plausible return patterns contribute to the complete model.

4.5 Predictive Quality and Computation

On supported, fine-labeled normal development queries, SERVE's central 90% predictive interval has 90.7% coverage and mean width 2.84 m; the predictive median has 0.58 m mean absolute error. The separate-objectives model attains slightly lower NLL (-2.16 versus -2.08), while SERVE achieves higher anomaly AP. Joint training therefore improves rejection alongside a small change in normal predictive fit. Appendix F reports how coverage and interval width vary with distance.

Inference encodes each cell once and shares context across point queries. On an RTX 5080 Laptop GPU, processing takes 58.6 ms per scan versus 43.8 ms for appearance, with peak allocated GPU memory of 4.46 versus 3.62 GB. Synchronized timing covers scan preparation, host-to-device transfer, encoding, prediction, and scoring, adding 14.8 ms for measurement verification. Appendix E specifies the measurement procedure.

5 Discussion and Conclusion

The STU comparisons connect improved rejection to how measurement evidence is learned and combined. Predictive training combined with joint inference improves over appearance; training through the joint class decision adds a further gain. Fixed-weight readouts show the value of matching appearance and range explanations within the same class. These improvements accompany higher normal development mIoU.

Local predictability defines concrete failure modes. A smooth unfamiliar surface can receive strong support when surrounding returns belong to the same object. A normal boundary can receive weak support when its observed continuation has low predicted density. Target exclusion governs access to the target cell, while neighboring cells retain their original object content. Empty-context points additionally omit the nonnegative measurement energy. Reporting rejection separately by context

availability would quantify the effect of that scoring choice; object-size groups characterize a different property.

SERVE integrates learned range compatibility into semantic recognition. Target-cell exclusion withholds the measured range from its own predictor, class-conditioned distributions express alternative local predictions, and joint supervision trains their contribution to the class decision. The modeling target is the distribution of a recorded return given neighboring cells. The STU comparisons evaluate when that statistical support improves labeling and rejection.

Reproducibility Statement

Equations (1)–(9) define the model and losses. Appendices A and C describe the architecture, normal supervision, and loss reductions. Appendix D defines the matched comparisons, and Appendix E specifies the evaluation populations and metrics. Appendix B derives the score properties, and Appendix F reports seed variation and predictive diagnostics.

References

Joshua Batson and Loic Royer. Noise2Self: Blind Denoising by Self-Supervision. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Mingqian Tang, Ming Liu, and Michael Yu Wang. Open-world Semantic Segmentation for LIDAR Point Clouds. In *Proceedings of the European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136980315.pdf.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional Neural Processes. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.03263.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-Specific Semantic Reconstruction for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4214–4228, 2023. doi: 10.1109/TPAMI.2022.3200384. URL https://doi.org/10.1109/TPAMI.2022.3200384.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive Neural Processes. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.05761.

Alexander Krull, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug. Probabilistic Noise2Void: Unsupervised Content-Aware Denoising. *Frontiers in Computer Science*, 2:5, 2020. doi: 10.3389/fcomp.2020.00005. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2020.00005/full.

Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-Quality Self-Supervised Deep Image Denoising. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper/8920-high-quality-self-supervised-deep-image-denoising.html.

Charline Le Lan and Laurent Dinh. Perfect Density Models Cannot Guarantee Anomaly Detection. *Entropy*, 23(12):1690, 2021. doi: 10.3390/e23121690. URL https://www.mdpi.com/1099-4300/23/12/1690.

Zizhao Li, Zhengkang Xiang, Jiayang Ao, Feng Liu, Joseph West, and Kourosh Khoshelham. Neural Distribution Prior for LiDAR Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3035–3045, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Neural_Distribution_Prior_for_LiDAR_Out-of-Distribution_Detection_CVPR_2026_paper.html.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020. URL https://arxiv.org/abs/2010.03759.

Simone Mosco, Daniel Fusaro, and Alberto Pretto. Learning to Identify Out-of-Distribution Objects for 3D LiDAR Anomaly Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17204–17214, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Mosco_Learning_to_Identify_Out-of-Distribution_Objects_for_3D_LiDAR_Anomaly_Segmentation_CVPR_2026_paper.html.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.09136.

Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. RbA: Segmenting Unknown Regions Rejected by All. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2211.14293.

Alexey Nekrasov, Malcolm Burdorf, Stewart Worrall, Bastian Leibe, and Julie Stephany Berrio Perez. Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11875–11885, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Nekrasov_Spotting_the_Unexpected_STU_A_3D_LiDAR_Dataset_for_Anomaly_CVPR_2025_paper.html.

Ian Nunes, Matheus B. Pereira, Hugo Oliveira, Jefersson A. dos Santos, and Marcus Poggi. Conditional Reconstruction for Open-Set Semantic Segmentation. In *IEEE International Conference on Image Processing*, pp. 946–950, 2022. doi: 10.1109/ICIP46576.2022.9897407. URL https://arxiv.org/abs/2203.01368.

Changgyoon Oh, Hyeonseong Kim, Daehyun We, Jongoh Jeong, Yujeong Chae, and Kuk-Jin Yoon. DOODLE: Diffusion-based Out-of-Distribution Learning for Open-set LiDAR Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2273–2283, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Oh_DOODLE_Diffusion-based_Out-of-Distribution_Learning_for_Open-set_LiDAR_Semantic_Segmentation_WACV_2026_paper.html.

Poojan Oza and Vishal M. Patel. C2AE: Class Conditioned Auto-Encoder for Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2307–2316, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Oza_C2AE_Class_Conditioned_Auto-Encoder_for_Open-Set_Recognition_CVPR_2019_paper.html.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked Autoencoders for Point Cloud Self-supervised Learning. In *Proceedings of the European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136620591.pdf.

Tomáš Vojíř, Jan Šochman, and Jiří Matas. PixOOD: Pixel-Level Out-of-Distribution Detection. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/07720.pdf.

Shaocong Xu, Pengfei Li, Qianpu Sun, Xinyu Liu, Yang Li, Shihui Guo, Zhen Wang, Bo Jiang, Rui Wang, Kehua Sheng, Bo Zhang, Li Jiang, Hao Zhao, and Yilun Chen. LiON: Learning Point-Wise Abstaining Penalty for LiDAR Outlier DetectioN Using Diverse Synthetic Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/32968.

Longyu Yang, Jun Liu, Yap-Peng Tan, Fumin Shen, Heng Tao Shen, Xiaofeng Zhu, and Ping Hu. Counterfactual Occlusion-Aware Learning via Visibility Intervention for LiDAR Anomaly Detection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=Hly2QYcuLW.

Yuanwen Yue, Damien Robert, Jianyuan Wang, Sunghwan Hong, Jan Dirk Wegner, Christian Rupprecht, and Konrad Schindler. LitePT: Lighter Yet Stronger Point Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24173–24184, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yue_LitePT_Lighter_Yet_Stronger_Point_Transformer_CVPR_2026_paper.html.

A Model Specification and Inference

The appearance branch combines LitePT-S features, assigned back to original points, with pointwise position, intensity, and within-voxel offsets. Their projection supplies the embedding in Equation (4).

The measurement branch independently encodes raw coordinates, intensity, and unit direction, then pools means and maxima within each 0.5° angular cell. Context comprises the eight surrounding offsets at radii one and three cells. Azimuth wraps; invalid elevation positions and empty cells are masked. The entire target cell is excluded before attention, and each retained token depends only on its own cell. Class vectors and target angular position form queries for two cross-attention layers. Output heads predict Student-t mixture parameters, direction-dependent means, and context logits η_{ic}. An empty-context state keeps attention finite; $\kappa_i = 0$ removes measurement energy and predictive supervision.

Inference evaluates all 19 hypotheses at each original return, using its measured log range and the fully normalized mixture density, computed by log-sum-exp. Semantic labels and unknown scores retain original point order. Ground-truth labels and allowed sets enter supervision and evaluation only.

An optional normal-reference scale uses 129 quantile knots from normal development scores. Duplicate knots are merged, and positive-slope tails extend the strictly increasing piecewise-linear transform. The same transform applies across classes and context states, preserving raw-score ordering, AP, and AUROC in exact arithmetic.

B Mathematical Properties

B.1 Density Normalization and a Finite Objective Bound

Proposition 1. *For any predicted means, nonnegative weights summing to one, and scales at least $\sigma_{\min} > 0$, Equation* (6) *is a density in z bounded by $M = 2/(\pi\sqrt{3}\sigma_{\min})$. Therefore $E_{ic} \geq 0$. For the stated $\sigma_{\min} = 10^{-3}$, so that $M > 1$, the objective in Equation* (9) *is bounded below by $-0.2 \log M$.*

Proof. The location-scale Student-t component integrates to one and reaches its maximum $2/(\pi\sqrt{3}\sigma)$ at its location. A convex mixture preserves normalization and cannot exceed the common component bound M. Appearance support is an average of numbers in $(0, 1]$, so $s \geq 0$; the density bound gives $g \geq 0$ for $\kappa = 1$, and $g = 0$ otherwise. The allowed-class predictive mixture also cannot exceed M, so its NLL is at least $-\log M$. All other losses are nonnegative. Averaging with normalized nonnegative reduction weights preserves these bounds. □

Numerically, $\log M = 5.9068664$ for the stated scale floor. A continuous density can exceed one, giving a negative NLL; subtraction from $\log M$ makes the measurement energy nonnegative.

B.2 Prediction Invariance under Target-Cell Changes

Proposition 2. *Fix model parameters, target directions, the non-target cell contents, and neighbor identities. Altering ranges, intensities, or the number of returns inside the target cell, while retaining the queried target directions and cell membership, leaves that cell's predicted mixture parameters unchanged.*

Proof. Each retained context token is a function only of points in its own non-target cell. The query uses class identity and target angular position. Thus every input to the attention layers and output heads is unchanged. Their output parameters, and the directional means at the fixed queried directions, are unchanged. The altered target measurement can change the subsequent density evaluation, and the unrestricted appearance branch can also change. □

The proposition establishes invariance of the forward computation at fixed parameters. Independent cell encoding and exclusion before attention preserve its premises. Target and neighboring returns can remain statistically dependent and can belong to the same physical object.

B.3 Class Agreement, Ranking, and Semantic Gradients

Subtracting $s_i^* + g_i^*$ inside $\min_c(s_{ic} + g_{ic})$ proves Equation (3). Every summand defining Δ_i is nonnegative. Its minimum is zero if and only if a common minimizing class exists. Thus the unknown score combines weak absolute support, represented by $s_i^* + g_i^*$, with disagreement between the preferred classes, represented by Δ_i.

For an unknown point a and a normal point n, write $B_i = s_i^* + g_i^*$. Their score difference is

$$u_a - u_n = (B_a - B_n) + (\Delta_a - \Delta_n). \tag{10}$$

Ranking depends on the difference between the two disagreement terms. For example, let a normal point have $s_n = (1, 5)$ and $g_n = (5, 1)$, and an unknown point have $s_a = (2, 5)$ and $g_a = (2, 5)$. Then $B_n = 2$, $B_a = 4$, $\Delta_n = 4$, and $\Delta_a = 0$. Adding disagreement reverses their order to $u_n = 6 > u_a = 4$. These positive measurement energies are compatible with Equation (6): overlapping components centered at the observation with scale $\sigma_{\min} e^{g_{ic}}$ give $p_{ic} = Me^{-g_{ic}}$.

For a general allowed set, differentiation of $-\log \sum_{d \in Y_i} P_{id}$ with respect to the class logit gives $P_{ic} - P_{ic}^Y$. The logit is $-s_{ic} + \kappa_i \log p_{ic} - \kappa_i \log M$, yielding Equation (7). Gradients to shared parameters sum the contributions through appearance and measurement prediction. The separate-objectives control retains this parameter sharing while directing its classification gradient through appearance support.

Softmax is invariant to $E_{ic} \mapsto E_{ic} + b_i$ for every class, whereas $u_i \mapsto u_i + b_i$. The classification loss therefore leaves a common energy shift unconstrained. The compactness and predictive terms act on absolute support and provide gradients along this otherwise unconstrained direction.

B.4 Effect of Joint Supervision on Density Fitting

The logarithmic score is proper when optimizing the predictive distribution alone under its specified conditioning variables (Gneiting & Raftery, 2007). The classification term changes that objective. Consider equally likely classes and a binary measurement, with true probabilities $p(z = 1 \mid c = 1) = 0.8$ and $p(z = 1 \mid c = 2) = 0.2$. Suppose fixed appearance support already matches the true posterior: $(0.8, 0.2)$ at $z = 1$ and $(0.2, 0.8)$ at $z = 0$. Parameterize predicted measurement probabilities symmetrically by t and $1 - t$.

The joint posterior at $z = 1$ is $q(t) = 4t/(1 + 3t)$. Up to terms constant in t, the expected joint cross-entropy plus 0.2 predictive NLL is

$$L(t) = 1.2[-0.8 \log t - 0.2 \log(1 - t)] + \log(1 + 3t) + \text{const.} \tag{11}$$

Its derivative at the true $t = 0.8$ is $3/3.4 > 0$, and its minimizer is approximately 0.5763. Classification shifts the predictive optimum because appearance already carries information about the class. The combined objective fits measurement support for a discriminative decision; empirical NLL and interval coverage characterize the resulting predictive distribution.

B.5 Score Coordinates and Empty Context

The density is defined relative to $dz = d \log(r/r_0)$, conditional on an existing return. Converting to a density in meters gives $p_R(r) = p_Z(\log(r/r_0))/r$. The resulting NLL adds $\log r$, which is common across classes at one point but varies across points. This transformation preserves class preferences and can change the unknown ranking. Log range is therefore part of the score definition and represents relative range variation (Le Lan & Dinh, 2021).

For fixed appearance energies, empty context gives $u^{(0)} = s^*$. Supported prediction gives $u^{(1)} = \min_c(s_c + g_c)$. For $c_s \in \arg\min_c s_c$,

$$0 \le \min_c g_c \le u^{(1)} - u^{(0)} \le g_{c_s}. \tag{12}$$

The fallback assigns measurement support one and omits this nonnegative score increment. For example, at the center of a single component with scale 0.1, the measurement energy is $g = \log(0.1/0.001) = \log 100$. This comparison holds appearance fixed; changing the actual neighboring returns can also change the full-scan appearance representation.

Proposition 3. *Fix the model, predictive scales, and context indicators. Replacing the density bound M by a looser bound $M' = \lambda M$, for a constant $\lambda \geq 1$, gives*

$$E'_{ic} = E_{ic} + \kappa_i \log \lambda, \qquad u'_i = u_i + \kappa_i \log \lambda. \tag{13}$$

Semantic labels, class probabilities, and the objective in Equation (9) *remain unchanged. Unknown-score ordering within each context state is preserved; ordering across the two states can change.*

Proof. Substitution in Equation (1) adds the same $\kappa_i \log \lambda$ to every class at point i. The minimum and softmax identities yield the stated outputs. The joint classification loss is invariant to this common shift, and the other losses use appearance support, predictive densities, and context logits directly. Each context state receives a constant score shift, while the difference between states increases by $\log \lambda$. □

The implemented rule fixes the tight bound determined by $\sigma_{\min}$. Its choice also fixes a relative offset between context states. A shared strictly increasing normal-reference transform preserves this ordering. Measuring the frequency and rejection performance of each state determines the empirical effect of the fallback.

C Normal Supervision and Optimization

C.1 Data and Label Construction

Fine labels define singleton allowed sets Y_i within the 19-class vocabulary; coarse source labels include all compatible classes, such as vegetation and trunk. Empty allowed sets retain input context but contribute no semantic supervision.

nuScenes supplies $28{,}130$ training scans from 700 scenes and $6{,}019$ development scans from 150 disjoint scenes. Supplementary labels cover rider-associated bicycle points in $1{,}586$ training frames from 74 scenes, 328 parking points in three training scenes, and $1{,}328$ building points in four training frames. Source checkpoint selection includes 3,366 supplementary building points in one development frame. Training and development scenes, logs, and scan identities are disjoint.

STU sequence 206 supplies 449 normal training scans; sequence 201 supplies 682 normal development scans. Training and selection scans contain no target anomaly label. Fine normal supervision covers 2.5–50 m and excludes STU labels 0, 1, 2, 52, and 99; all nonempty returns remain inputs. Sequence 206 lacks fine parking and bicyclist labels, so half of source replay positions prioritize those classes; the remainder follow the ordinary source sequence. Matched models share labels and replay identities. nuScenes intensity is divided by 255; STU retains its recorded scale.

C.2 Exact Loss Reductions

Let V contain reliably labeled normal points and $Q \subseteq V$ at most 4,096 queries, sampled to cover represented labels before filling the quota. Define $Q^+ = \{i \in Q : \kappa_i = 1\}$. For $S \subseteq V$, let $\mathcal{G}(S)$ contain its distinct allowed sets and $S_G = \{i \in S : Y_i = G\}$. The balanced reduction is

$$\mathcal{B}_S(\ell) = \frac{1}{|\mathcal{G}(S)|} \sum_{G \in \mathcal{G}(S)} \frac{1}{|S_G|} \sum_{i \in S_G} \ell_i. \tag{14}$$

Joint, appearance, and context classification use $\mathcal{B}_Q$, $\mathcal{B}_V$, and $\mathcal{B}_{Q^+}$, respectively. Compactness averages over Q; predictive NLL averages over Q^+. Empty reductions contribute zero. Each coarse point enters once in its allowed-set group; per-frame losses are averaged over the effective batch.

C.3 Optimization and Selection

Matched models receive two source passes and eight target passes, with one source replay per four target scans: 28,130 source and 2,244 target optimizer updates. AdamW uses effective batch size two and rotation augmentation. Selection uses central frames from 50 source development scenes and 68 uniformly spaced target development scans. Each model maximizes normal mIoU under its own semantic readout, breaking ties by its normal objective; the selected source checkpoint initializes

adaptation. Runs use seeds 206, 307, and 409. Matched pairs share common-module initialization, scan order, augmentations, and query identities; additional modules use separate random streams. Every model completes the shared update budget before checkpoint selection.

D Matched Models and Ablations

Standard classifier. The same backbone and point-detail encoder feed a linear class head. Allowed-set class-balanced cross-entropy trains the logits. Softmax uses $1 - \max_c P_c$ for unknownness; energy uses $-\log \sum_c \exp l_c$ at temperature one. Both readouts use the same fitted weights, explaining their identical mIoU.

Independent appearance model. This independently trained model retains Equation (4), scores $u_i = \min_c s_{ic}$, and minimizes $\mathcal{B}_Q(\ell_{\text{appearance}}) + 0.5\mathcal{B}_V(\ell_{\text{appearance}}) + 0.05\mathcal{L}_{\text{compact}}$. The two classification terms use their respective populations, Q and V.

Separate objectives. This model replaces E by s in the coefficient-one classification term and retains the architecture, other losses, and shared class vectors. It uses $E = s + g$ at inference. The change removes measurement density from direct class competition while preserving parameter sharing.

CSSR-style dense adaptation. Class-specific feature autoencoders reconstruct the point embedding with a 12-dimensional bottleneck. CSSR's absolute reconstruction error $d_{ic} = \|\mathbf{h}_i - A_c(\mathbf{h}_i)\|_1$ supplies logits $-0.1d_{ic}$ for joint backbone and autoencoder training (Huang et al., 2023). Allowed-set losses on Q and V have coefficients one and 0.5. The semantic label is $\hat{c}_i = \arg\min_c d_{ic}$.

Unknown scoring retains relative reconstruction error and first- and second-order activation support. Write $\mathbf{b}_i = |\mathbf{h}_i|$ elementwise. On unaugmented normal training points, group observations by predicted class into T_c and compute

$$\boldsymbol{\mu}_c = \frac{1}{|T_c|}\sum_{i\in T_c}\mathbf{b}_i, \quad \widetilde{\boldsymbol{\mu}}_c = \frac{\boldsymbol{\mu}_c}{\sum_d \boldsymbol{\mu}_d + \epsilon}, \quad G_c = \frac{1}{|T_c|}\sum_{i\in T_c}\mathbf{b}_i\mathbf{b}_i^{\mathrm{T}}, \tag{15}$$

where the ratio is elementwise and $\epsilon = 10^{-8}$. The three known-support components are

$$r_{1i} = \frac{-d_{i\hat{c}_i}}{\|\mathbf{h}_i\|_1^2 + \epsilon}, \qquad r_{2i} = \mathbf{b}_i^{\mathrm{T}}\widetilde{\boldsymbol{\mu}}_{\hat{c}_i}, \qquad r_{3i} = \langle G_{\hat{c}_i}, \mathbf{b}_i\mathbf{b}_i^{\mathrm{T}}\rangle_F. \tag{16}$$

The unknown score is $u_i = -\sum_{j=1}^{3}(r_{ji} - m_j)/(s_j + \epsilon)$, with global means m_j and standard deviations s_j estimated on normally augmented training points. Empty predicted-class groups have zero activation statistics. The adaptation treats each LiDAR point as a single-position feature map.

Retrained ablations. The target-available model appends its own cell token to the 16 context positions. The one-component model uses $K = 1$; the no-NLL and no-compactness models set the indicated coefficient to zero. Each retrains under the shared schedule and selection rule.

Fixed-weight readouts. Appearance-only sets $g = 0$ in both outputs. Common-density replaces every p_{ic} by $\bar{p}_i = C^{-1}\sum_c p_{ic}$, preserving a normalized density while making its energy common across classes. Its semantic labels are therefore exactly the appearance labels. Independent minima retains joint semantic labels and scores $s_i^* + g_i^*$, removing only Δ_i. All readouts use the fitted SERVE parameters, the same evaluated point identities, and a scalar threshold shared across classes.

E Evaluation Populations and Reductions

E.1 Semantic and Pointwise Anomaly Metrics

Normal development IoU uses the pooled confusion matrix. The 14 ground-truth-present classes in sequence 201 contribute equally to mIoU; motorcycle, other vehicle, bicyclist, motorcyclist, and

other ground are excluded from this mean. All 19 classes remain prediction candidates, and predictions into absent classes count as false negatives for the true classes.

STU evaluation applies the distance and semantic masks before selecting scans with at least five valid anomaly points. Each valid point contributes once; all valid normal points form the false-positive denominator, and small individual anomaly objects remain included. Validation contains 1,960 eligible scans and 193,879,969 valid points: 193,792,470 normal and 87,499 anomalous points.

AP, AUROC, and FPR95 are computed on pooled scores separately for each seed, then averaged across runs. AP uses the precision–recall step integral with anomaly as positive. FPR95 follows the benchmark threshold convention for 95% anomaly recall. Hidden-test outputs preserve original point identities.

E.2 Paired Decisions and Object Size

For each seed, the appearance model fixes $H = \{i : y_i \text{ is anomalous}, \max_c P_{ic}^{\text{app}} \geq 0.9\}$. For an operating FPR, all models choose a scalar threshold on the same evaluation normal population. On H, "recovered" means reference miss and candidate detection; "lost" means reference detection and candidate miss. Both percentages use $|H|$, so their difference equals the change in recall under the shared false-alarm budget.

Normal corrections and introduced errors compare the same labeled points in sequence 201. Their difference gives the change in pooled point accuracy; class IoU is computed separately from the confusion matrix.

An anomaly instance is detected when at least half its valid points exceed the threshold. Size bins use the post-mask return counts in Figure 2. Each scan-instance pair receives equal weight within its bin; an object observed in several scans contributes once per scan.

E.3 Predictive Diagnostics and Runtime

Predictive diagnostics include all fine-labeled, context-supported normal queries in the diagnostic population. The true-class mixture CDF supplies central quantiles $q_{0.05}$ and $q_{0.95}$ in log range. Coverage uses $z_i \in [q_{0.05}, q_{0.95}]$; interval width in meters is $r_0(e^{q_{0.95}} - e^{q_{0.05}})$. Point prediction uses the median $r_0 e^{q_{0.5}}$, which remains finite for the Student-t distribution's heavy tails in log space. NLL, coverage, interval width, and absolute error are averaged over individual queries.

Runtime uses batch size one, identical scan order, and FP32 on an RTX 5080 Laptop GPU with 16 GB memory. After 50 warmup scans, synchronization brackets preparation, transfer, model computation, and per-point output construction for each validation scan. Disk loading and global metric computation occur outside the timed region. Peak allocated GPU memory is reset after model loading and reported for the full inference pass, including model parameters.

F Additional Results and Interpretation

F.1 Seed Variation

Table 3: Individual matched runs and sample standard deviations across three training seeds (percentage points).

Model	Metric	Seed 206	Seed 307	Seed 409	Mean ± SD
Appearance	Validation AP	20.41	21.27	22.40	21.36 ± 1.00
Separate objectives	Validation AP	47.02	48.37	49.42	48.27 ± 1.20
SERVE	Validation AP	72.80	74.08	74.94	73.94 ± 1.08
SERVE	Test AP	60.58	61.56	63.08	61.74 ± 1.26
SERVE	Normal mIoU	62.80	63.20	63.48	63.16 ± 0.34

SERVE improves validation AP over the separate-objectives control in all three paired runs. Its validation AP ranges from 72.80% to 74.94%, and its test AP ranges from 60.58% to 63.08%.

F.2 Known-Class Performance

Table 4: Normal development IoU on sequence 201 for its 14 ground-truth-present classes (%). Values are three-seed means; mIoU averages the 14 class values.

Class	Appearance	SERVE	Difference
Car	90.10	90.80	+0.70
Bicycle	46.20	48.30	+2.10
Truck	64.30	66.10	+1.80
Person	51.70	53.90	+2.20
Road	92.40	92.80	+0.40
Parking	27.60	30.10	+2.50
Sidewalk	76.80	77.70	+0.90
Building	82.50	83.00	+0.50
Fence	44.90	46.10	+1.20
Vegetation	88.20	88.60	+0.40
Trunk	45.60	46.50	+0.90
Terrain	70.80	71.70	+0.90
Pole	48.10	49.50	+1.40
Traffic sign	37.68	39.14	+1.46
mIoU	61.92	63.16	+1.24

IoU improves across all 14 evaluated surface and object classes. Parking shows the largest gain, from 27.60% to 30.10%, and remains the lowest-IoU class. Person and bicycle improve by 2.20 and 2.10 points, respectively, while road improves from 92.40% to 92.80%.

F.3 Predictive Fit across Models and Distances

Table 5: Normal predictive diagnostics and validation anomaly AP. Diagnostics use supported fine-labeled queries, NLL in log-range coordinates, and central 90% predictive intervals.

Model	NLL↓	Coverage (%)	Width (m)	Median MAE (m)	Val. AP (%)
Separate objectives	−2.16	90.2	2.71	0.56	48.27
Target cell available	−4.72	95.8	0.41	0.06	39.62
SERVE	−2.08	90.7	2.84	0.58	73.94

Table 6: SERVE predictive diagnostics by range. Query shares give each stratum's weight in the overall means for the supported fine-labeled diagnostic population.

Range (m)	Query share (%)	NLL	Coverage (%)	Width (m)	Median MAE (m)
$[2.5, 10)$	30	−2.45	93.0	1.40	0.30
$[10, 20)$	35	−2.20	92.0	2.20	0.45
$[20, 35)$	25	−1.75	88.2	4.00	0.80
$[35, 50]$	10	−1.38	85.5	6.50	1.325

Coverage decreases from 93.0% at 2.5–10 m to 85.5% at 35–50 m, while interval width increases from 1.40 to 6.50 m. The mean absolute error of the predicted median rises from 0.30 to 1.325 m. The overall 90.7% coverage averages over this distance-dependent pattern. At long range, widening predictive intervals still leave a greater fraction of observations outside the nominal 90% interval.

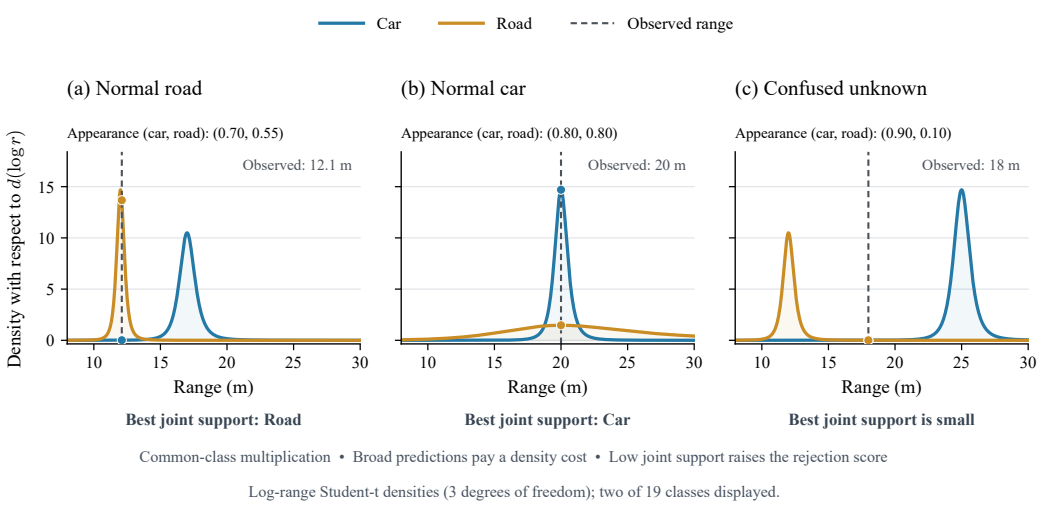


Figure 3: Class decisions illustrated with hand-specified Student-t densities. Left: measurement support corrects an appearance preference for car at a normal road return. Center: a concentrated car prediction provides greater density than a broad road prediction at the observed range. Right: weak measurement support under both classes raises the unknown score despite strong car appearance. The vertical axis gives density per unit log range; the horizontal axis displays range in meters.

F.4 Interpretation of Failure Cases

Equation (3) decomposes the unknown score into best-class appearance energy, best-class measurement energy, and disagreement between the preferred classes. Figure 3 illustrates how these terms affect class decisions. A foreign object can receive a low unknown score when one normal class explains both its appearance and local return pattern. A boundary between normal surfaces can receive a high score when the observed continuation has low density under every class prediction. Sparse context increases ambiguity, and empty context gives the appearance-only rule.