Neural Distribution Prior for LiDAR Out-of-Distribution Detection

Supplementary Material

1. Implementation Details

We initialize the model using a Mask4Former checkpoint pretrained on SemanticKITTI [1] and Panoptic CUDAL [10]. The model is then fine-tuned for up to 10 epochs on the downstream datasets, with Perlin noise–synthesized OOD samples included during training. Optimization uses AdamW with a learning rate of 2×10^{-4} and a batch size of 8 on NVIDIA A100 GPUs. For the NDP matrix ψ, the embedding dimension d is set to 16 unless indicated otherwise. In the SOE loss, the soft OOD target β is fixed to 0.9 for all experiments unless otherwise specified. To compensate for the scarcity of auxiliary OOD points compared to in-distribution points, their loss contribution is weighted 10000 times higher than that of ID points.

For the Perlin Raise algorithm, the patch radius r is sampled from $[0.75, 1.5]$, the noise strength α is set to 0.4, and the target ratio ρ is fixed at 0.3.

2. Explanation of Evaluation Metrics

Point-level Evaluation Metrics Point-level evaluation metrics for LiDAR OOD detection include AUROC, FPR@95, and Average Precision (AP). These metrics are widely used in OOD detection and anomaly segmentation [2, 4, 9, 12].

AUROC assesses how well the OOD score separates OOD points from ID points across all possible thresholds. It is obtained by ranking points by their OOD scores and measuring how consistently OOD points receive higher scores than ID points. Because it is threshold-free, AUROC reflects the overall separability of the score function. However, this metric is not ideal for scenarios with severe ID/OOD imbalance, and AP is therefore often used as the main evaluation metric [2, 9].

FPR@95 measures the reliability of the detector at a high-recall operating point. We first determine the score threshold that correctly identifies 95% of OOD points, and then evaluate the proportion of ID points incorrectly flagged as OOD at this threshold.

Average Precision evaluates the quality of OOD detection under the precision-recall trade-off. By sweeping the score threshold from high to low, AP quantifies how well the detector maintains high precision as it covers more OOD points. The AP score is the integral of the resulting precision–recall curve, typically approximated through monotonic interpolation. AP is especially informative for LiDAR OOD segmentation because it naturally handles the severe imbalance between ID and OOD points.

Object-level Evaluation Metrics The STU benchmark [9] provides fine-grained instance masks for all OOD objects and adopts Panoptic Quality (PQ) [7] as the primary metric for object-level anomaly segmentation. PQ evaluates instance-level performance by combining Segmentation Quality (SQ) and Recognition Quality (RQ). For a class c, it is defined as:

$$\mathrm{PQ}_c = \underbrace{\frac{\sum_{(p,g)\in TP_c} \mathrm{IoU}(p,g)}{|TP_c|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP_c|}{|TP_c| + \frac{1}{2}|FP_c| + \frac{1}{2}|FN_c|}}_{\text{Recognition Quality (RQ)}}. \tag{1}$$

A predicted object is counted as a true positive (TP) if it overlaps with a ground-truth instance with Intersection over Union (IoU) greater than 50%. Unmatched predictions are counted as false positives (FP), and missed ground-truth instances as false negatives (FN). Ignore regions are excluded from evaluation and predictions inside these regions are not penalized.

For in-distribution classes, the final PQ score is obtained by averaging PQ_c over all classes. For anomaly segmentation, all OOD objects are grouped into a single class, and PQ is reported for this aggregated category.

To quantify anomaly recall, STU also reports the Unknown Quality (UQ) metric [11]:

$$\mathrm{UQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + |FN|}}_{\text{Recall Quality (RecallQ)}}. \tag{2}$$

Unlike PQ, UQ does not penalize false positives, allowing the metric to focus purely on the model's ability to retrieve anomaly instances. As with PQ, an IoU threshold of 50% is required to count a prediction as a true positive. However, anomaly segmentation in LiDAR scenes requires both high anomaly recall and careful control of false positives, since excessive false alarms can negatively affect downstream planning [9].

3. Additional Visualization

Fig. 1 and Fig. 2 illustrate the qualitative performance of our method. Across diverse environments, including narrow urban alleys and unstructured rural roads, the model consistently identifies a broad range of OOD objects such as armchairs, fallen branches, packages, and yoga mats. Our method also substantially reduces the false positive rate. In addition, baseline approaches such as MaxLogit [6] and RbA [8] incorrectly label tree trunks as OOD in forest environments, where dense geometry and cluttered backgrounds

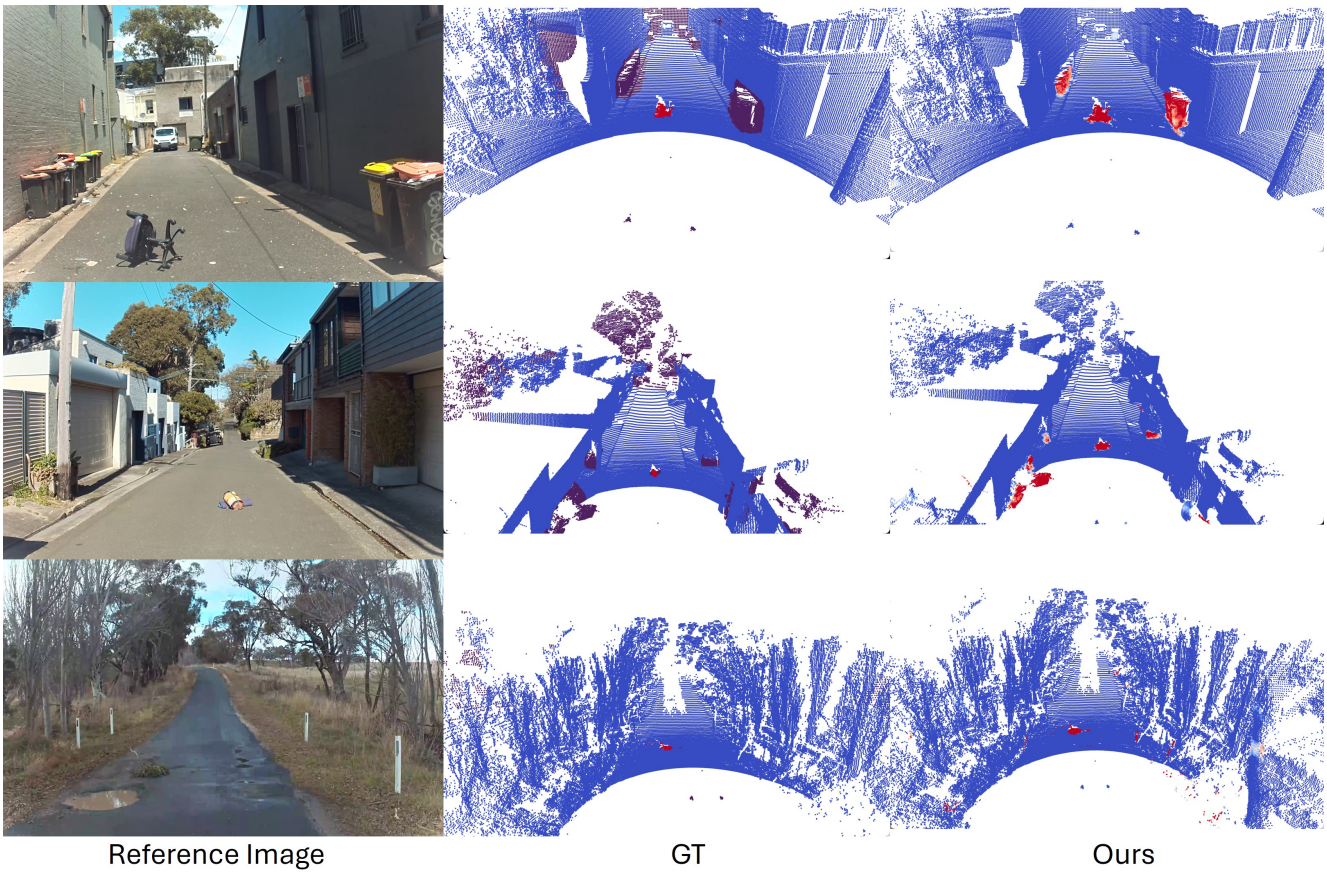


Figure 1. Additional visualization of OOD score map on the STU benchmark with image reference. Points are labeled as inlier, anomaly, and unlabeled. Our approach yields precise and coherent anomaly masks while maintaining a low false-positive rate on inlier regions.

make boundary estimation difficult. Our model maintains reliable predictions in these complex scenes, avoiding such false positives and producing cleaner and more consistent OOD masks under challenging structural variability.

In addition, as showin in Fig. 3, we provide visualizations of the OOD map on SemanticKITTI, where our method still performs well, demonstrating generalization across datasets.

4. Additional Results

Tab. 1 presents an ablation study on the template size d of the NDP matrix ψ, where d determines the dimensionality of the vectors stored in ψ as the learnable prior. A moderate NDP size yields the best performance: $d = 16$ achieves the highest AP (74.24%) and a strong AUROC (99.53%). Overall, NDP is not highly sensitive to this hyperparameter. Smaller values of d store fewer parameters and struggle to capture the dynamics of the logit distribution, whereas larger values introduce additional parameters that are more difficult to optimize and may lead to overfitting.

We validated our method using a lightweight MinkUNet [5] backbone. As shown in Tab. 2, our method consistently improves OOD detection performance.

Table 1. Ablation study of template size d in NDP matrix ψ, where d is the dimensionality of the vectors stored in ψ as the learnable prior.

d	AUROC ↑	FPR@95 ↓	AP ↑
8	99.42	1.21	70.29
16	**99.53**	**1.43**	**74.24**
32	99.20	1.67	70.14

Table 2. OOD Detection Results of NDP-EE using various backbones

Method	AUROC ↑	FPR@95 ↓	AP ↑
Static Extended Energy	98.33	2.94	58.36
NDP-EE	**99.19**	**2.89**	**70.29**

Tab. 3 and Tab. 4 report the in-distribution per-class panoptic segmentation performance on the STU and Se-

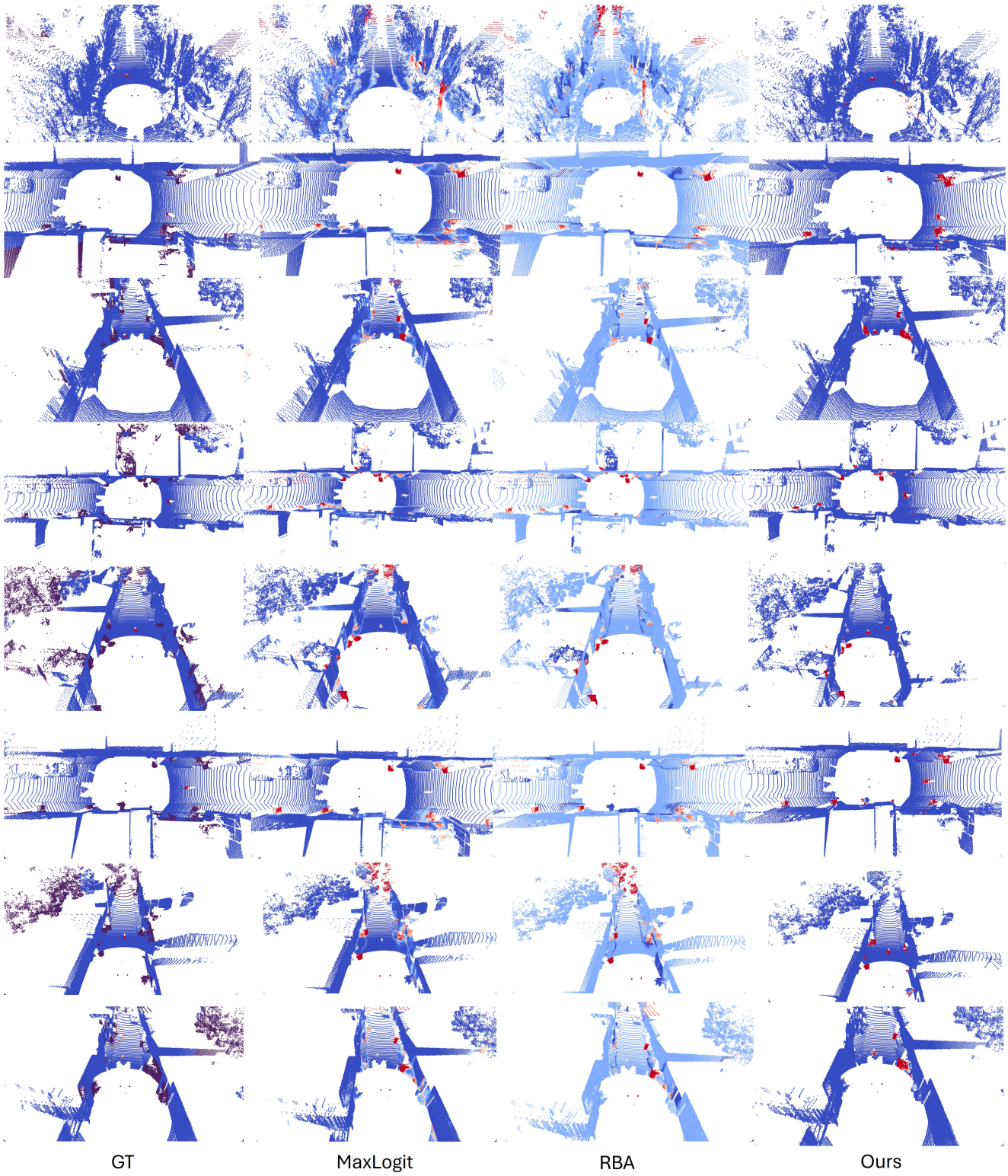


Figure 2. Additional visualization of OOD score map on the STU benchmark. Points are labeled as inlier, anomaly, and unlabeled. Our approach yields precise and coherent anomaly masks while maintaining a low false-positive rate on inlier regions.

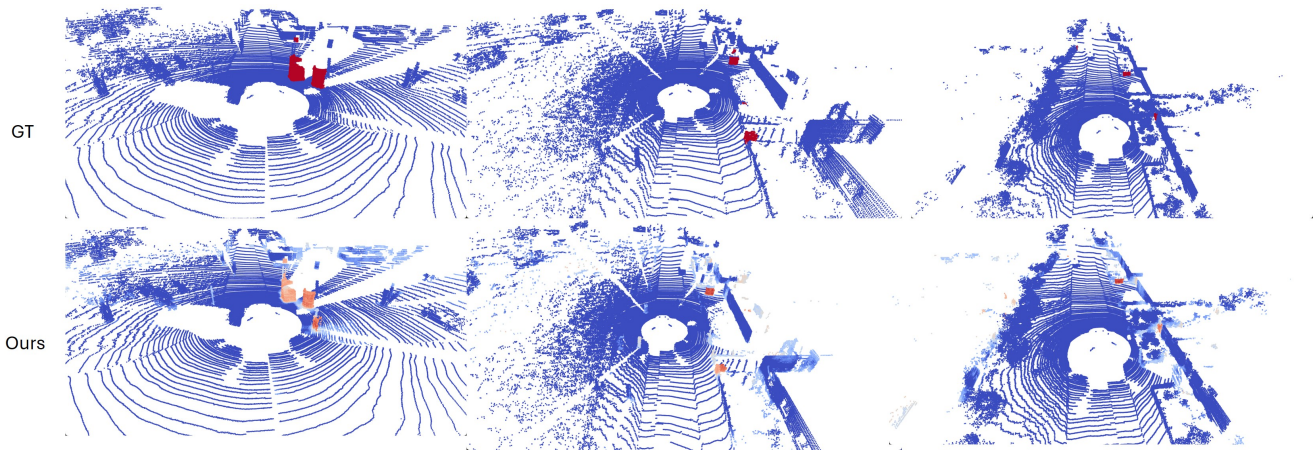


Figure 3. Visualization of OOD score map on the SemanticKITTI. Points are labeled as inlier, anomaly, and unlabeled. Our approach yields precise and coherent anomaly masks while maintaining a low false-positive rate on inlier regions.

Table 3. In-distribution per-class performance of the methods on the validation set of STU [9] dataset. Our model retains comparable panoptic segmentation performance to standard Mask4Former training.

Method	void	car	truck	bicycle	person	road	sidewalk	parking	building	vegetation	trunk	terrain	fence	pole	traffic sign	PQ
Mask4Former [13]	–	80.99	37.28	47.65	80.99	71.46	17.74	0.0	84.08	89.73	29.34	30.79	47.6	59.62	60.96	52.73
Mask4Former-void [3]	0.07	23.88	20.78	1.01	43.30	38.24	20.03	11.11	48.45	43.09	20.20	17.31	30.80	27.26	33.16	26.96
Mask4Former-NDP	–	77.42	48.58	51.47	76.05	40.37	12.37	0.0	90.83	92.84	31.00	65.09	36.14	51.28	59.77	52.37

Table 4. In-distribution per-class performance of the methods on validation sets of SemanticKITTI [1]. Our model retains comparable panoptic segmentation performance to standard Mask4Former training.

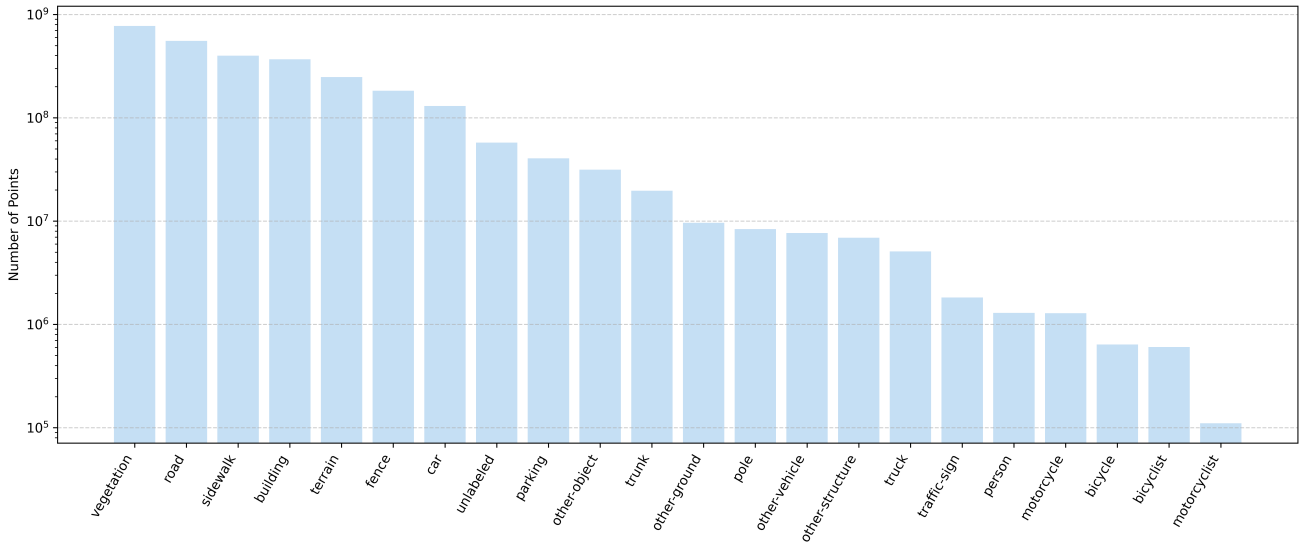
Method	void	car	truck	bicycle	motorcycle	other vehicle	person	bicyclist	motorcyclist	road	sidewalk	parking	other ground	building	vegetation	trunk	terrain	fence	pole	traffic sign	PQ
Mask4Former [13]	–	93.53	59.39	62.55	64.82	54.36	79.61	89.16	25.01	93.24	77.90	28.79	0.0	87.27	87.28	51.08	59.92	24.85	56.76	58.14	60.72
Mask4Former-void [3]	6.08	74.36	47.00	32.19	43.34	33.30	42.90	68.75	00.33	93.35	77.07	19.01	0.0	82.77	81.34	47.56	56.94	19.98	54.48	36.82	47.97
Mask4Former-NDP	–	93.22	60.77	60.17	68.99	56.62	80.25	87.93	0.0	92.90	77.29	25.44	0.0	87.28	86.88	52.01	59.89	23.61	58.31	56.73	59.38

manticKITTI validation sets. Our model (NDP-EE) preserves segmentation accuracy comparable to the standard Mask4Former [13] baseline. On STU [9], Mask4Former-NDP achieves a PQ of 52.37, matching the closed-set performance of Mask4Former and substantially surpassing variants trained with void classification. On SemanticKITTI [1], Mask4Former-NDP maintains strong segmentation quality with a PQ of 59.38, closely tracking the original closed-set Mask4Former and outperforming other OOD-training-based counterparts. These results indicate that the incorporation of the proposed NDP module does not compromise closed-set segmentation performance.

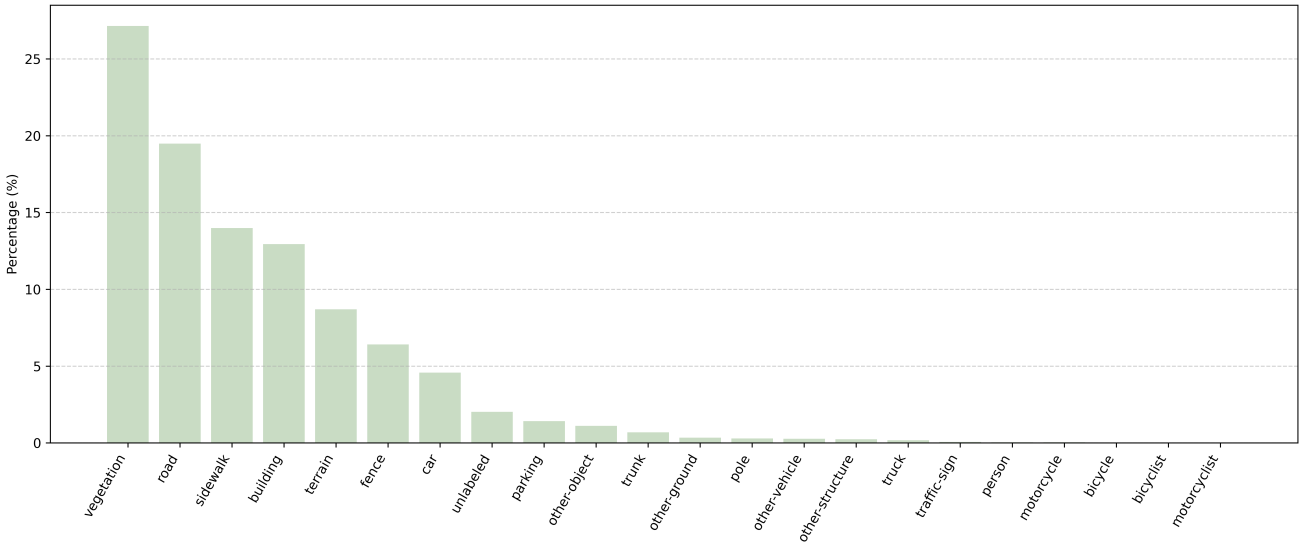
5. Dataset Statistics

For OOD detection, class imbalance is especially severe in LiDAR data and makes anomaly discrimination more difficult. This motivates the use of adaptive mechanisms such as distribution-aware priors or dynamic reweighting.

As shown in Fig. 4, SemanticKITTI [1] exhibits an extremely long-tailed distribution. Vegetation, road, and sidewalk account for the majority of points. Vegetation alone contributes more than one quarter of the dataset, and the top four to five classes collectively comprise more than half of all annotated points. In contrast, classes such as motorcyclist, bicyclist, bicycle, person, and traffic sign appear in very small quantities, often below one percent of the total



(a) Per-class point counts (log scale).



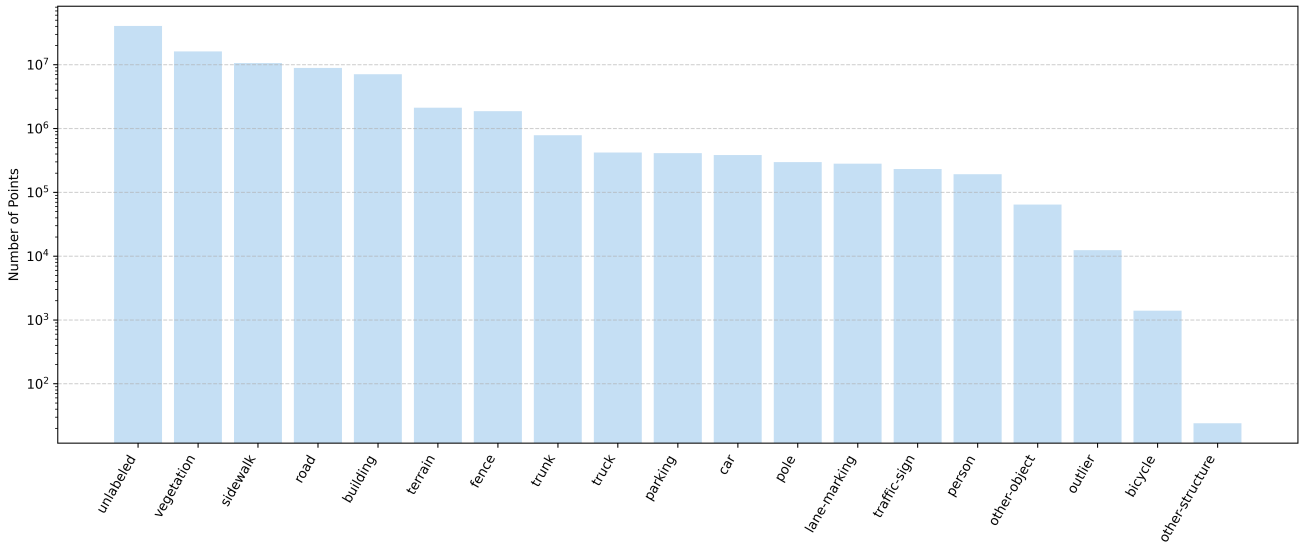
(b) Per-class point percentage.

Figure 4. Class distribution in the SemanticKITTI dataset. In the dataset, vegetation, road, and sidewalk account for most points. Vegetation alone contributes more than a quarter of all points, and the top four to five classes collectively exceed half of the dataset. In contrast, many classes such as motorcyclist, bicyclist, bicycle, person, and traffic sign appear in extremely small proportions. These categories often fall below one percent of the total point count.

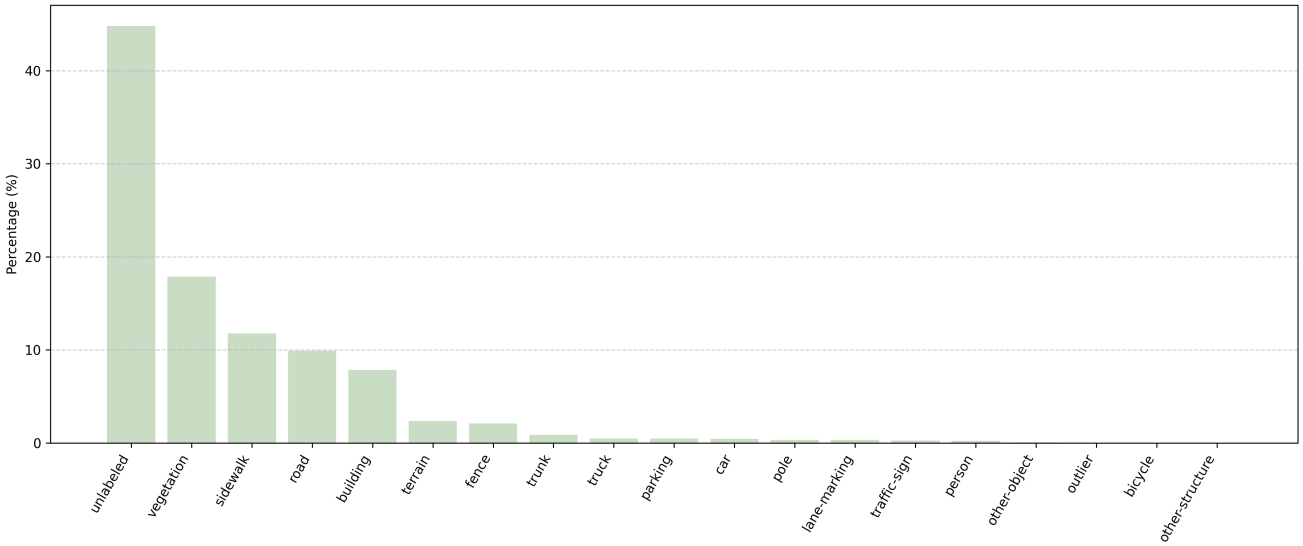
point count.

In STU [9], most evaluation sequences provide only three labels: inlier, anomaly, and unlabeled, without detailed in-distribution class annotations. We therefore use sequence 201, which includes full semantic labels, as a representative example. As shown in Fig. 5, this scene contains over 90 million LiDAR points. Similar to SemanticKITTI, a few head classes, including vegetation, road, and sidewalk, dominate the point cloud, while rare categories such as person, traffic sign, and bicycle account for less than 0.5% of all points.

Our innovation directly targets this issue. By introducing a learnable distribution prior and reweighting logits through a class-dependent attention mechanism, the proposed framework models the characteristic prediction patterns of each class rather than assuming a uniform inlier distribution. This enables more faithful calibration across both head and tail categories and substantially improves OOD



(a) Per-class point counts (log scale).



(b) Per-class point percentage.

Figure 5. Class distribution in the STU dataset. Using the sequence 201 with full annotation as an example, the scene contains over 90 million LiDAR points in total. A few dominant classes, such as vegetation, road, and sidewalk, occupy most of the point cloud. Rare classes like person, traffic-sign, and bicycle comprise less than 0.5% of all points.

scoring in long-tailed LiDAR scenes.

6. Visualization of OOD Samples Generated by Perlin Noise

The Perlin Raise augmentation produces synthetic OOD regions highlighted in blue, which exhibit substantial variation in geometry and scale. As shown in Fig. 6, these OOD insertions span small localized perturbations to larger, irregular structures that integrate coherently with the surrounding scene layout. This diversity yields a wide range of anomaly shapes that are not repetitive and do not correspond to any in-distribution semantic category. The resulting samples provide a rich and varied training signal for OOD detection, enabling the model to learn more generalizable decision boundaries and reducing susceptibility to overfitting on narrowly defined auxiliary OOD data.

Although Perlin noise does not explicitly model occlusion, we observe that it still performs well in practice. Future work may incorporate more realistic geometric constraints, such as occlusion-aware generation.

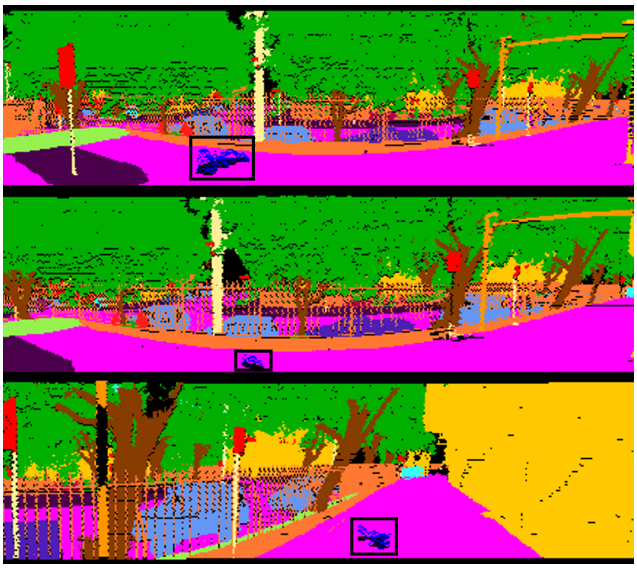


Figure 6. Range-view visualization of Perlin Raise–generated OOD samples. Blue regions denote synthetic anomalies.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 4

[2] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. *International Conference on Computer Vision Workshop (ICCV'W)*, 2019. 1

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *International Journal on Computer Vision (IJCV)*, 2021. 4

[4] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *International Conference on Machine Learning (ICML)*, 2022. 1

[7] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[8] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. RbA: Segmenting Unknown Regions Rejected by All. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[9] 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 *"Conference on Computer Vision and Pattern Recognition (CVPR)"*, 2025. 1, 4, 5

[10] Tzu-Yun Tseng, Alexey Nekrasov, Malcolm Burdorf, Bastian Leibe, Julie Stephany Berrio Perez, Mao Shan, and Stewart Worrall. Panoptic-CUDAL Technical Report: Rural Australia Point Cloud Dataset in Rainy Conditions. *arXiv preprint arXiv:2503.16378*, 2025. 1

[11] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying Unknown Instances for Autonomous Driving. In *Conference on Robot Learning (CoRL)*, 2019. 1

[12] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28, 2024. 1

[13] Kadir Yilmaz, Jonas Schult, Alexey Nekrasov, and Bastian Leibe. Mask4Former: Mask Transformer for 4D Panoptic Segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 4