UniD-Shift: Towards Unified Semantic Segmentation via Interpretable Shared–Private Multimodal Decomposition

Supplementary Material

A. Additional Ablation Studies

To examine the contribution of each architectural choice, we conduct a comprehensive set of ablation experiments summarised in Table 7. The results reveal several consistent patterns. Replacing ResNet with SAM leads to a substantial improvement, with the gain reaching more than five percentage points when paired with SPVCNN. This trend reflects the benefit of dense semantic structure extracted by the foundation model. The transition from SPVCNN to SPTNet also produces a measurable increase, indicating that long-range geometric aggregation enhances the stability of point-wise representation. When both SAM and SPTNet are used simultaneously, the unimodal baseline becomes considerably stronger and forms a reliable basis for evaluating the contribution of fusion.

The second group of experiments focuses on the influence of the shared-private fusion design. When applied to weaker unimodal features, the fusion module alters the accuracy only slightly, yet its effect becomes more pronounced once the backbone features attain a higher level of semantic and geometric completeness. The comparison between experiments that differ only in the fusion choice demonstrates this trend clearly: the shift from KL-based alignment to shared-private fusion consistently produces improvement on top of both SAM-based and SPTNet-based settings. This behaviour indicates that the fusion mechanism benefits from well-structured modality inputs and reorganises them into a more coherent representation.

Table 7. Full ablation studies evaluating the influence of the 2D encoder, 3D backbone, and fusion strategy on segmentation accuracy.

Exp	2D Module	3D Module	Fusion	mIoU (%)
1(Baseline)	ResNet	SPVCNN	KL	68.2
2	ResNet	SPTNet	KL	70.1
3	SAM	SPVCNN	KL	73.0
4	SAM	SPTNet	KL	73.4
5	ResNet	SPVCNN	Shared-Private fusion	Not Converged(64.8)
6	ResNet	SPTNet	Shared-Private fusion	70.3
7	SAM	SPVCNN	Shared-Private fusion	74.1
8 (Ours)	SAM	SPTNet	Shared-Private fusion	**74.8**

The final configuration attains the highest accuracy at 74.8%. The margin over the strongest unimodal setting confirms that the improvement is not solely attributable to the use of SAM or SPTNet. Instead, the interaction created by the shared-private formulation provides additional refinement, enabling the network to maintain stable semantic relations across modalities. The combined results illustrate how encoder strength and fusion structure reinforce one another and highlight the importance of coordinated multimodal design.

B. Network Architecture Details

B1. Overall Architecture Overview

The proposed framework consists of a dual-branch encoder, a shared-private feature decomposition module, a lightweight fusion block, and a 3D segmentation head. The 2D branch processes the RGB image through a frozen vision foundation model, while the 3D branch extracts geometric structure using a sparse convolution-transformer backbone. Both branches are projected into a unified latent space and decomposed into shared and private components. The fusion block aggregates the shared features to form a consistent multimodal representation, which is decoded to produce point-level predictions.

B2. 2D Branch Details

The 2D branch adopts the SAM ViT-L vision backbone, using the official pretrained weights sam_vit_l_0b3195.pth. The encoder is kept frozen during training to retain stable visual priors and reduce optimization complexity. The model extracts hierarchical image features and produces dense representations that serve as the visual input to the multimodal pipeline.

Since SAM outputs high-dimensional visual tokens, a projection head is applied to map the features into the unified latent space. A linear layer followed by layer normalization compresses the SAM output into a 256-dimensional feature vector for each pixel aligned with valid LiDAR projections. This representation ensures that the 2D branch provides semantically rich information while maintaining compatibility with the 3D geometric features. The projected features are then used as the input to the shared-private decomposition module.

B3. 3D branch details

The 3D branch is built upon the SPTNet encoder, which couples sparse convolution with transformer modeling to capture geometric structure at multiple spatial scales. Sparse convolution operates on voxelized point clouds while preserving the irregular distribution of LiDAR points. Only non-empty voxels participate in computation, reducing the cost of 3D processing and maintaining the spatial layout of the scene. The encoder contains six SpConv

blocks, each responsible for aggregating local geometry within a limited neighborhood. These blocks strengthen the description of fine structural patterns, yet their locality restricts the ability to capture long-range relations common in large outdoor environments.

To address this limitation, each SpConv block is followed by a transformer block that extends the receptive field beyond the convolutional neighborhood. The feature map from sparse convolution is linearly projected to form query, key, and value tensors, and multi-head attention is applied to model interactions across distant regions. This mechanism enhances the representation of the global structure and supports the integration of contexts that span multiple object scales. The transformer output is combined with the corresponding convolutional feature map through a skip connection, producing a unified representation that maintains both local detail and global coherence. By stacking six pairs of SpConv and transformer blocks, the encoder generates a hierarchical geometric feature pyramid, which serves as the 3D input to the shared-private fusion module described in the main paper.

C. Implementation Details

C1. Training Configuration

All experiments are conducted on eight NVIDIA RTX 4090 GPUs with 24 GB memory. The model adopts a hidden dimension of 256 and six hierarchical scales $\{2, 4, 8, 16, 16, 16\}$ as specified in the configuration file. The 2D branch employs a frozen SAM encoder without any parameter updates, while the 3D branch uses an SPTNet backbone trained end to end. During training, the loader applies geometric augmentation to point clouds and photometric perturbations to images, while the projection indices between 2D pixels and 3D points are precomputed to ensure stable multimodal alignment.

All trainable parameters belong to the 3D backbone and the fusion module, while the SAM encoder remains frozen. The total training loss follows the formulation in the main paper and includes 3D segmentation, 2D segmentation, cross-modal KL alignment, and the two decomposition terms. The loss weights are set as $\lambda_{\text{seg2D}} = 1$, $\lambda_{\text{xm}} = 1$, $\lambda_{\text{gram}} = 0.05$, and $\lambda_{\text{diff}} = 0.05$. Optimization uses stochastic gradient descent with a learning rate of 0.24, momentum of 0.9, and a weight decay of 10^{-4}, together with cosine annealing over 80 epochs. Mixed-precision training is enabled, and each GPU handles a batch of 8 samples. This configuration ensures stable convergence for all experiments reported in the main paper.

C2. Inference and Runtime Measurement

Inference is performed without test-time augmentation. Sparse convolutions reconstruct point-level features using the stored inverse voxel mapping, and semantic predictions are obtained by applying an $\arg\max$ operation on the class logits. The runtime reported in the main paper is measured using the same hardware environment and identical batch configuration to ensure fair comparison. All measurements reflect end-to-end forward time, including feature extraction, fusion, and final decoding.

D. Detailed Performance Evaluation

Table 8 presents a detailed per-class evaluation on the nuScenes test set. Compared with LiDAR-only methods, multimodal approaches (LC) generally achieve stronger performance, highlighting the benefit of incorporating image cues for semantic understanding. UniD-Shift achieves competitive overall performance, with clear improvements over several multimodal baselines in categories that require cross-modal reasoning, such as bus, motorcycle, and pedestrian. In addition, the method maintains stable predictions across both geometry-dominant classes (e.g., barrier and terrain) and appearance-sensitive categories (e.g., traffic cone and vegetation), demonstrating balanced semantic modeling. These results further validate that the proposed shared-private decomposition effectively integrates complementary information from 2D and 3D modalities, leading to robust performance across diverse semantic classes.

E. Failure Cases

Figure 6 illustrates representative cases in which the proposed framework encounters difficulties with distant and low-resolution targets. In these situations, the projected image features become severely compressed, and the corresponding point cloud samples contain only a few valid returns. Limited geometric evidence and reduced appearance detail create conditions in which the fused representation loses part of the fine structural patterns. Models that operate solely on point clouds occasionally retain these distant instances because their predictions rely more directly on sparse geometric clusters, whereas the multimodal fusion tends to emphasize regions with stronger cross-view correspondence.

The wider scene layout remains stable in the majority of samples, and the predicted masks maintain coherent spatial organization even when these local errors appear. The observed behavior indicates that the remaining challenges concentrate on long-range perception and extremely sparse regions. These conditions highlight potential directions for refinement, including improved handling of distant depth intervals and more adaptive treatment of sparse samples within the fusion process.

Table 8. Quantitative results on the nuScenes test set. Numbers for some baselines are taken from their original papers.

Method	Modality	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	other flat	sidewalk	terrain	manmade	vegetation	mIoU(%)
PolarNet [71]	L	69.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.5	96.6	67.1	77.7	72.1	87.1	84.5
Cylinder3D [76]	L	77.2	82.8	29.8	84.3	89.4	53.0	79.3	77.2	73.4	84.6	69.1	97.7	70.2	80.3	75.5	90.4	87.6
CMDFusion [8]	L	80.8	83.5	45.7	94.5	91.4	76.7	87.0	77.2	73.0	85.6	77.3	97.4	69.2	79.5	75.5	91.0	88.5
SPVCNN++ [48]	L	81.1	86.4	43.1	91.9	92.2	75.9	75.7	83.4	77.3	86.8	77.4	97.7	71.2	81.1	77.2	91.7	89.0
SVQNet [9]	L	81.3	84.5	41.8	93.4	92.5	69.2	85.5	83.7	78.4	84.5	77.5	97.2	70.4	81.7	77.9	91.8	90.2
LidarMultiNet [66]	L	81.4	80.4	48.4	84.3	90.0	71.5	87.2	85.2	80.4	86.9	74.8	97.8	67.3	80.7	76.5	92.1	89.6
LiDARFormer [78]	L	81.5	84.4	40.8	84.7	92.6	72.7	91.0	84.9	81.7	88.6	73.8	97.9	69.3	81.7	77.4	92.4	89.6
SphereFormer [25]	L	81.9	83.3	39.2	94.7	92.5	77.5	84.2	84.4	79.1	88.4	78.3	97.9	69.0	81.5	77.2	93.4	90.2
2DPASS [61]	LC	80.8	81.7	55.3	92.0	91.8	73.3	86.5	78.5	72.5	84.7	75.5	97.6	69.1	79.9	75.5	90.2	88.0
PMF [81]	LC	77.0	82.1	40.3	80.9	86.4	63.7	79.2	79.8	75.9	81.2	67.1	97.3	67.7	78.1	74.5	89.9	88.5
2D3DNet [13]	LC	80.0	83.0	59.4	88.0	85.1	63.7	84.4	82.0	76.0	84.8	71.9	96.9	67.4	79.8	76.0	92.1	89.2
MSeg3D [29]	LC	81.1	83.1	42.5	94.9	92.0	67.1	78.6	85.7	80.5	87.5	77.3	97.7	69.8	81.2	77.8	92.4	90.1
U2MKD [47]	LC	84.2	86.0	67.3	93.0	92.1	79.0	89.3	84.8	80.1	87.8	77.0	97.8	70.6	81.5	78.0	93.1	90.7
TASeg [54]	LC	84.6	87.1	69.4	90.5	92.2	78.7	90.4	86.3	81.9	88.3	75.9	97.8	70.9	81.0	78.2	93.4	91.2
UniD-shift(ours)	LC	81.2	83.4	61.9	92.0	86.7	73.0	85.7	81.4	76.1	87.4	68.4	97.7	67.6	80.6	76.6	92.3	89.3

Notes. L: LiDAR-only. **LC**: LiDAR-Camera fusion.

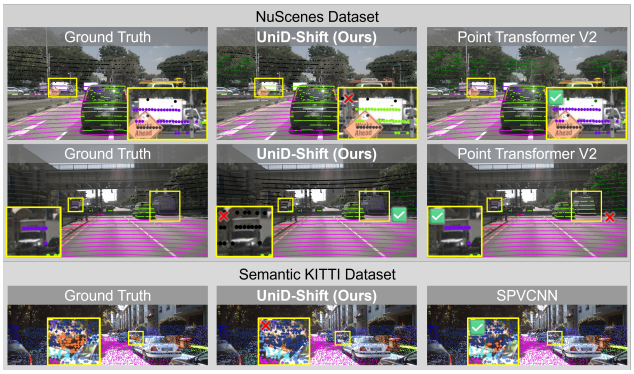

Figure 6. Representative failure cases on the nuScenes and Semantic KITTI validation datasets.

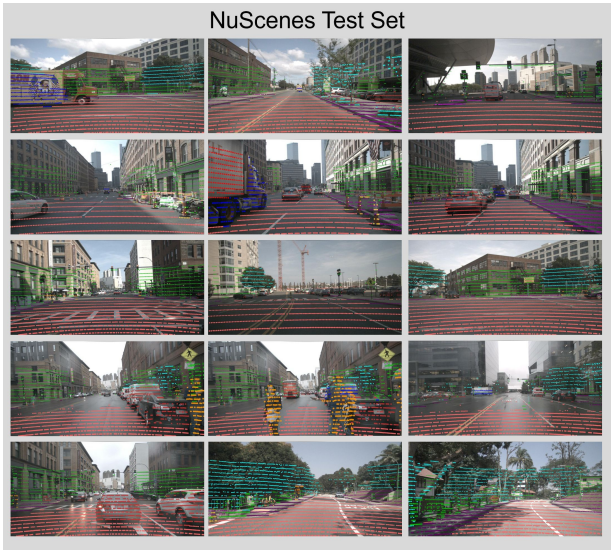

Figure 7. More visualization of object segmentation performance on the Nuscenes Test set.

F. More Visualization

Additional visualizations are provided in Fig. 9 and Fig. 10, offering extended visual evidence of the segmentation of UniD-Shift over a broad range of outdoor environments.

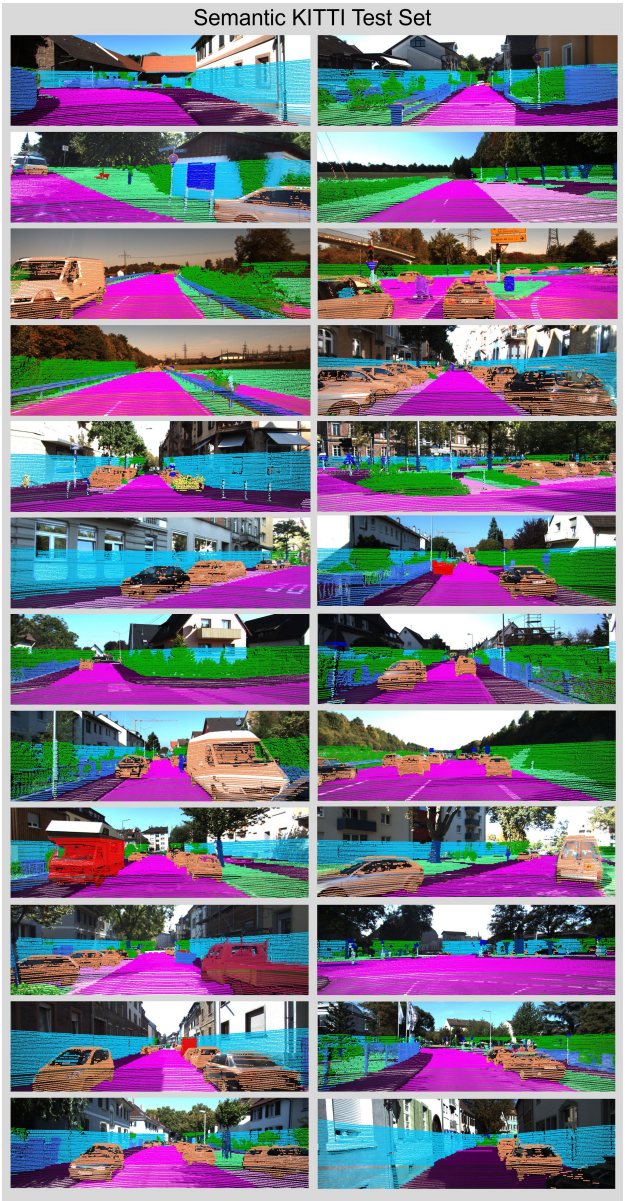

Figure 8. More visualization of object segmentation performance on the SemanticKitti Test set.

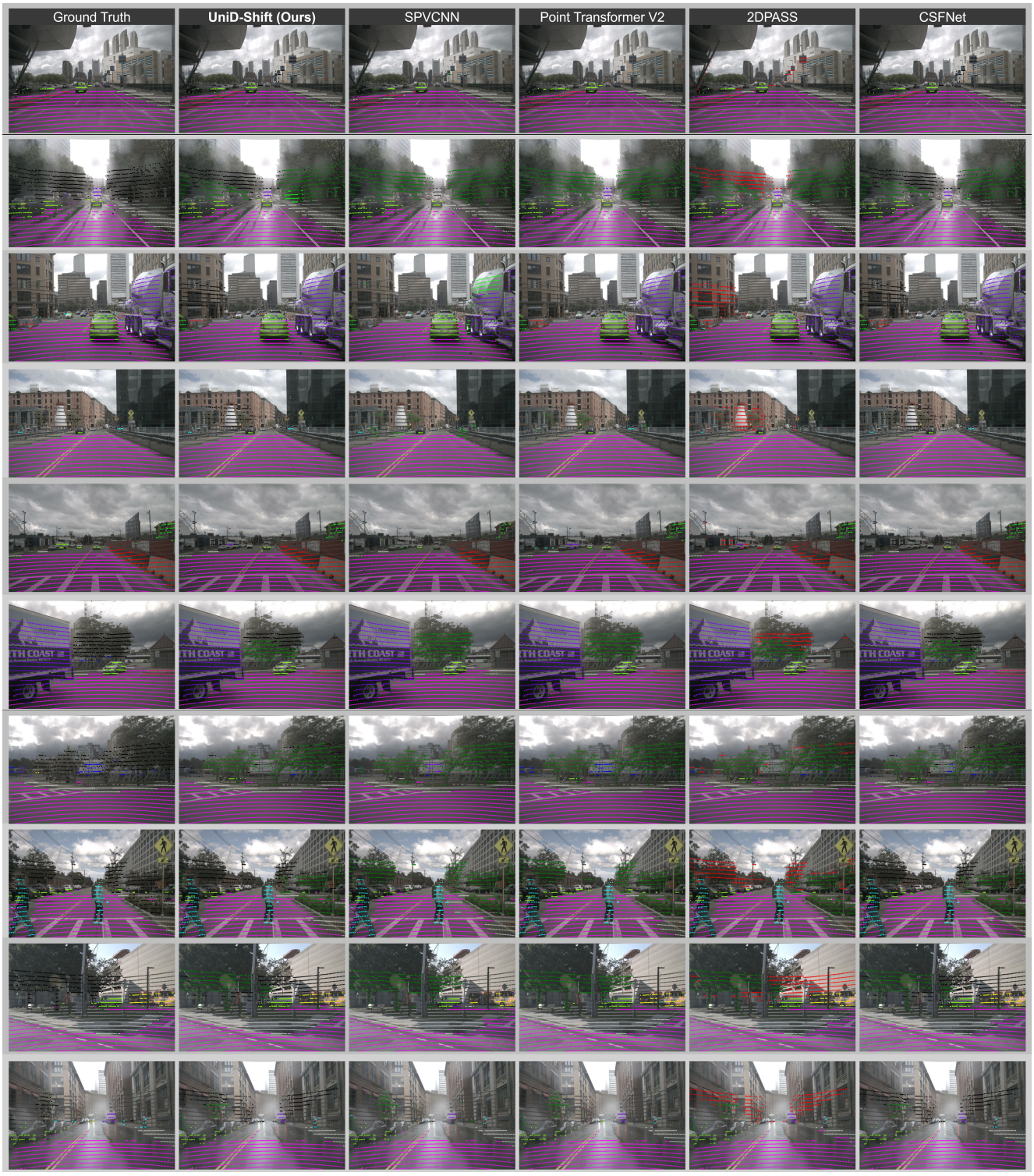

Figure 9. More visualizations comparing object segmentation performance with other methods on the nuScenes validation set.

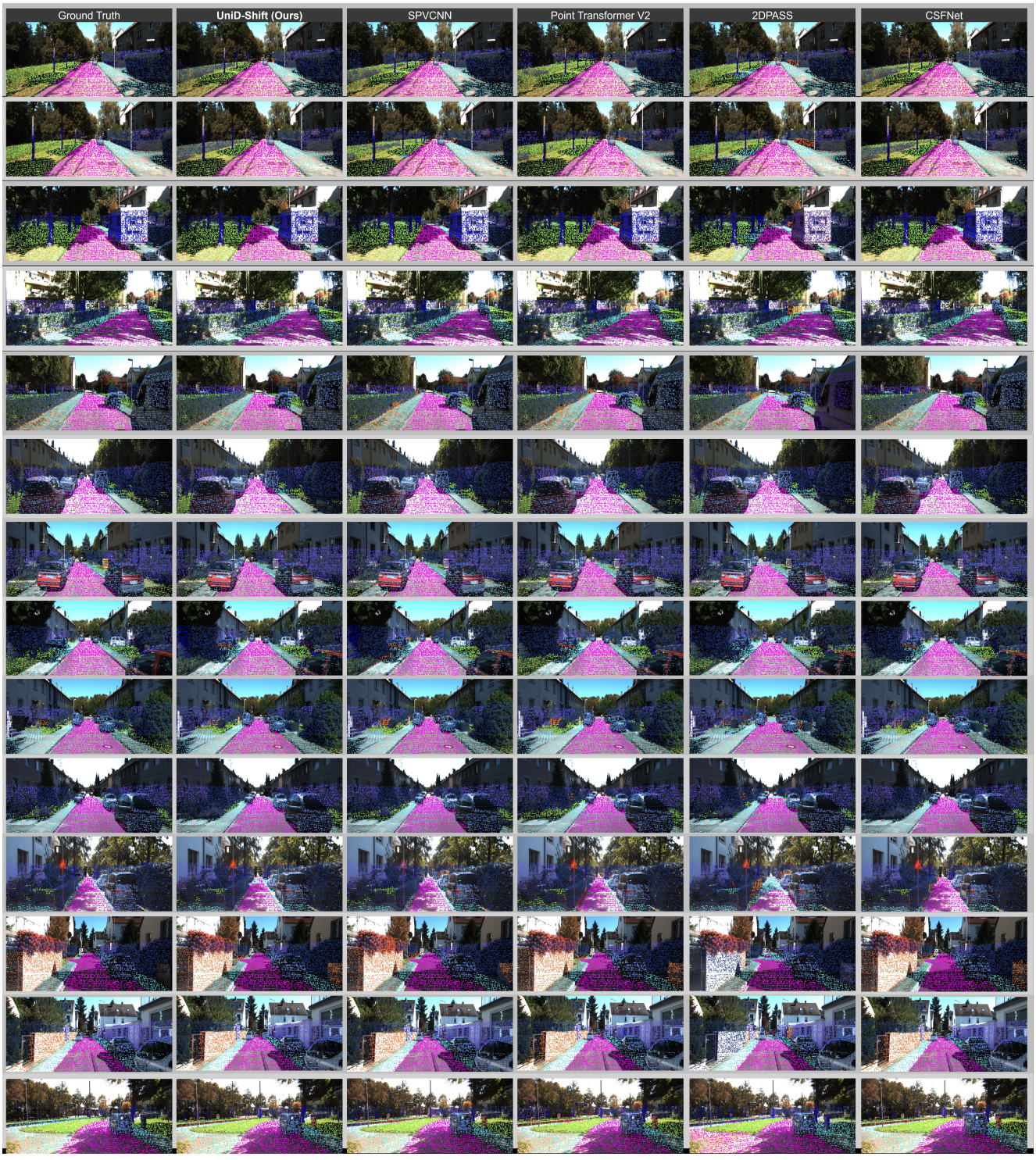

Figure 10. More visualizations comparing object segmentation performance with other methods on the SemanticKitti validation set.