Large Scene Generation with Cube-Absorb Discrete Diffusion

Supplementary Material

This supplementary document provides a comprehensive overview of our implementation details (Section A), visualization for ablation studies (Section F) and additional visualization for main results (Section G).

A. Implementation Details

A.1. Full List of Training Hyperparameters

All training hyperparameters for our CADD models are listed in Table 8.

Hyperparameter	Value
Number of Layers	5
Hidden size	1152 / 576 / 576
FFN inner hidden size	1152 / 576 / 576
Warmup Steps	2500
Batch Size	96 / 16 / 4
Max Steps	200k
Gradient Clipping	1.0
Adam ϵ	1e-8
Attention heads	16
Attention head size	72 / 48 / 48
Dropout	0.1
Peak Learning Rate	3e-4
Weight Decay	0
Learning Rate Decay	Linear
AdamW β_1	0.9
AdamW β_2	0.999

Table 8. Hyperparameters for CADD Models. Hyperparameters separated by '/' indicate variations across the three levels: level 1 / level 2 / level 3.

A.2. Datasets Pre-processing

In this paper, we use two outdoor scene datasets for our experiments: CarlaSC [66] and KITTI-360 [24]. CarlaSC is based on simulated road scenes, while KITTI-360 comes from real-world environments. Due to the differences in their sources and label structures, we apply customized preprocessing steps to ensure compatibility for experimentation. The details are provided below.

A.2.1. CarlaSC

CarlaSC, extensively used in our main experiments and ablation studies, is a synthetic dataset composed of outdoor road point cloud scenes. The dataset originally includes 23 semantic labels, which are grouped into 11 classes based on the official guidelines in [28]. These 11 semantic classes, with 0 representing the unclassified category, are detailed in Table 9. The dataset comprises 18 scenes for training, 3 for validation, and 3 for testing. For our experiments, we use a high-resolution version of CarlaSC, where each scene is represented as $256^2 \times 16$ voxels. This resolution corresponds to a physical area covering 25.6 meters in both forward and backward directions from the radar scanner and a vertical height of up to 3 meters.

Index	Label	Index	Label
1	Building	6	Road
2	Fences	7	Ground
3	Other	8	Sidewalk
4	Pedestrian	9	Vegetation
5	Pole	10	Vehicle

Table 9. Merged semantic labels for the CarlaSC dataset. This table presents the 10 consolidated classes used in our experiments, with 0 representing unclassified elements that are excluded from the list.

A.2.2. KITTI-360

The KITTI-360 dataset features a variety of environments, including inner-city traffic, residential areas, highways, and countryside roads. It consists of 11 distinct sequences, each capturing a continuous driving trajectory, with 9 sequences designated for training and 2 for testing. Since semantic labels for the test set are unavailable, we further divide the training data, using 7 sequences (sequence 0, 2, 3, 4, 5, 6, and 7) for model training and the remaining 2 sequences (sequence 9 and 10) for evaluation. To generate ground-truth voxels, we segment accumulated LiDAR scans into chunks measuring $51.2\ m \times 51.2\ m \times 12.8\ m$ and voxelize the points at a resolution of $256 \times 256 \times 64$. The original dataset includes 46 categories, but these are remapped or removed to match the 11 categories used in CarlaSC, as outlined in Table 10. Certain categories, such as moving objects, were removed due to their limited relevance in semantic segmentation. These excluded categories constitute only a small fraction of the labels, ensuring the remaining labels align the CarlaSC dataset.

B. None-overfitting Verification

We use structural similarity (SSIM) to confirm that a generated scene differs from its nearest neighbor in the CarlaSC dataset. Specifically, we generate 1k scenes and identify their closest matches in the training set using SSIM. Likewise, we sample 1k validation scenes and find their nearest counterparts in the training set. The average SSIM of these scenes is calculated and presented in Table 11. Figure 7

Index	Original Labels	Mapped Index	Mapped Labels	Index	Original Labels	Mapped Index	Mapped Labels
0	Unlabeled	0	Unlabeled	23	Sky	remove	-
1	Ego Vehicle	remove	-	24	Person	4	Pedestrian
2	Rectification Border	remove	-	25	Rider	4	Pedestrian
3	Out of ROI	remove	-	26	Car	10	Vehicle
4	Static	remove	-	27	Truck	10	Vehicle
5	Dynamic	-	-	28	Bus	10	Vehicle
6	Ground	7	Ground	29	Caravan	10	Vehicle
7	Road	6	Road	30	Trailer	10	Vehicle
8	Sidewalk	8	Sidewalk	31	Train	10	Vehicle
9	Parking	7	Ground	32	Motorcycle	3	Other
10	Rail Track	3	Other	33	Bicycle	3	Other
11	Building	1	Building	34	Garage	3	Other
12	Wall	2	Fence	35	Gate	3	Other
13	Fence	2	Fence	36	Stop	3	Other
14	Guard Rail	2	Fence	37	Smallpole	5	Pole
15	Bridge	3	Other	38	Lamp	3	Other
16	Tunnel	3	Other	39	Trash Bin	3	Other
17	Pole	5	Pole	40	Vending Machine	3	Other
18	Polegroup	5	Pole	41	Box	3	Other
19	Traffic Light	5	Pole	42	Unknown Construction	3	Other
20	Traffic Sign	5	Pole	43	Unknown Vehicle	10	Vehicle
21	Vegetation	9	Vegetation	44	Unknown Object	3	Other
22	Terrain	7	Ground	45	License Plate	remove	-

Table 10. Conversion of KITTI-360 labels to match CarlaSC's 11 categories. Labels marked as 'remove' are not present in CarlaSC, while those labeled with '-' are excluded from semantic segmentation based on the original settings.

Data	SSIM
Generated	0.70
Validation Set	0.65

Table 11. Average SSIM between each generated scene and the closest scene in the training set.

shows the SSIM distribution, demonstrating that CADD learns the training set distribution and generates new samples rather than merely memorizing it.

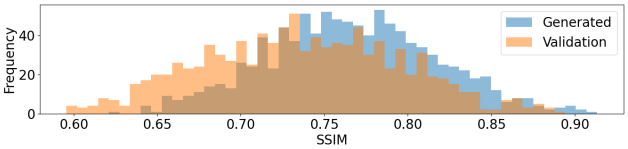

Figure 7. Data retrieval visualization of 1000 samples.

C. Higher Resolution Results

The method can generate 1024-resolution or even higher-resolution scenes by adding another diffusion layers. In the main text, We used 256 resolution for fair comparison with PDD. Fig. 8 shows that the generated $1024 \times 1024 \times 256$ results on the KITTI-360 dataset with a voxel size of $0.1m \times 0.1m \times 0.1m$. Compared to the 256 resolution, the higher-resolution output captures finer structural details.

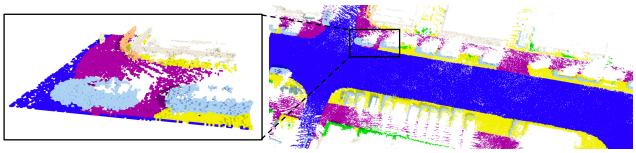

Figure 8. Generation result at $1024 \times 1024 \times 256$ on KITTI-360.

D. Transferability and Infinite Scene Generation

Table 12 shows our model's improved performance when transferred from CarlaSC to KITTI-360. Fine-tuning effectively adapts it to complex object distributions and scene dynamics. We can also generate infinite 3D scenes using sub-scene conditional generation, similar to PDD. Figure 9 shows the results. These results will be included in the supplementary materials.

Finetuned Scales	Conditioned	F3D↓	MMD↓
None	×	0.085	0.022
$l = 1$	×	0.066	0.019
$l = 1, 2, 3$	×	**0.065**	**0.015**
None	✓	0.130	0.018
$l = 2, 3$	✓	**0.093**	**0.017**

Table 12. Generation results on KITTI-360. Finetuned Scales set to None indicates training from scratch and others stand for fine-tuning corresponding pre-trained CarlaSC model.

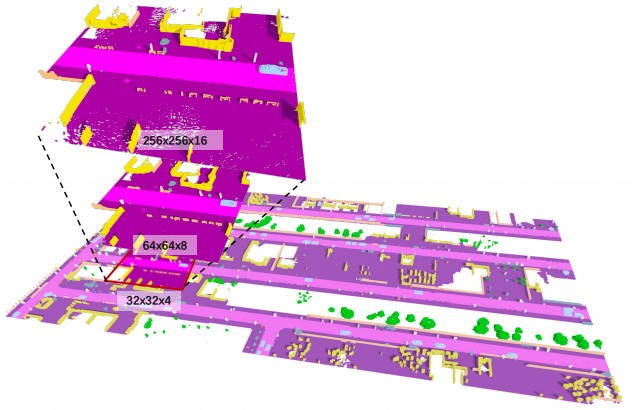

Figure 9. Infinite scene generation results.

E. Additional Ablation Studies

E.1. Cube Size Selection

We performed an ablation study on cube size selection in SCDT for $128^2 \times 8 \rightarrow 256^2 \times 16$ upsampling (Table 13). The default cube size is 8. The results show that a smaller cube size leads to performance degradation, possibly due to a reduced receptive field, while a larger cube size does not improve performance. This study will be added to the supplementary materials.

Cube Size	mIoU↑	MA↑	F3D↓	MMD↓
4	92.43	94.13	0.207	0.093
8	**95.18**	**97.32**	0.175	**0.078**
16	95.13	97.29	**0.174**	0.080

Table 13. Ablation study on different cube sizes in SCDT.

F. Visualization of Ablation Studies

F.1. Diffusion Process Selection

Figure 10 illustrates the generation results of different diffusion processes in the coarse-to-fine generation task ($128^2 \times 8 \rightarrow 256^2 \times 16$), including uniform diffusion, absorb diffusion, and the proposed cube-absorb diffusion. Among these, the cube-absorb diffusion demonstrates the closest alignment with the ground truth, whereas other methods exhibit varying degrees of noise. This difference arises from the cube-absorb diffusion starting from a coarse-grained initial state, in contrast to other approaches that initiate generation from noise or a fully masked state.

F.2. DiT Architecture

Figure 11 presents the generation results of Cube-DiT and the proposed SCDT in the coarse-to-fine generation task ($128^2 \times 8 \rightarrow 256^2 \times 16$). The proposed SCDT delivers comparable generation performance to Cube-DiT while significantly reducing memory consumption (6.38G compared to 17.09G).

F.3. Hierarchy Configuration

Figure 12 shows the generation results of our model at different hierarchy resolutions and depths on the CarlaSC dataset: (a) $256^2 \times 16$; (b) $64^2 \times 4 \rightarrow 256^2 \times 16$; (c) $64^2 \times 4 \rightarrow 128^2 \times 8 \rightarrow 256^2 \times 16$; (d) $32^2 \times 2 \rightarrow 64^2 \times 4 \rightarrow 128^2 \times 8 \rightarrow 256^2 \times 16$. Hierarchical models clearly outperform single-level models, demonstrating their superior ability to represent 3D structures. Besides, the model's performance is robust to both the initial resolution and hierarchy depth.

G. Additional Visualization of Main Results

Figures 13 and 14 present additional visualizations of unconditional generation results on CarlaSC and KITTI-360, respectively. Our model excels in generating realistic scenes, including roads, vehicles, and crossroads, producing results that closely resemble real-world data. Notably, the generated scenes align more closely with the details of the ground truth, particularly for the complex and challenging KITTI-360 dataset.

Figures 15 and 16 present additional visualizations of conditional generation results on CarlaSC and KITTI-360, respectively. Our results are much closer to the ground truth scenes than those of the comparison method, demonstrating CADD's ability to align details between the generated outputs and coarse inputs.

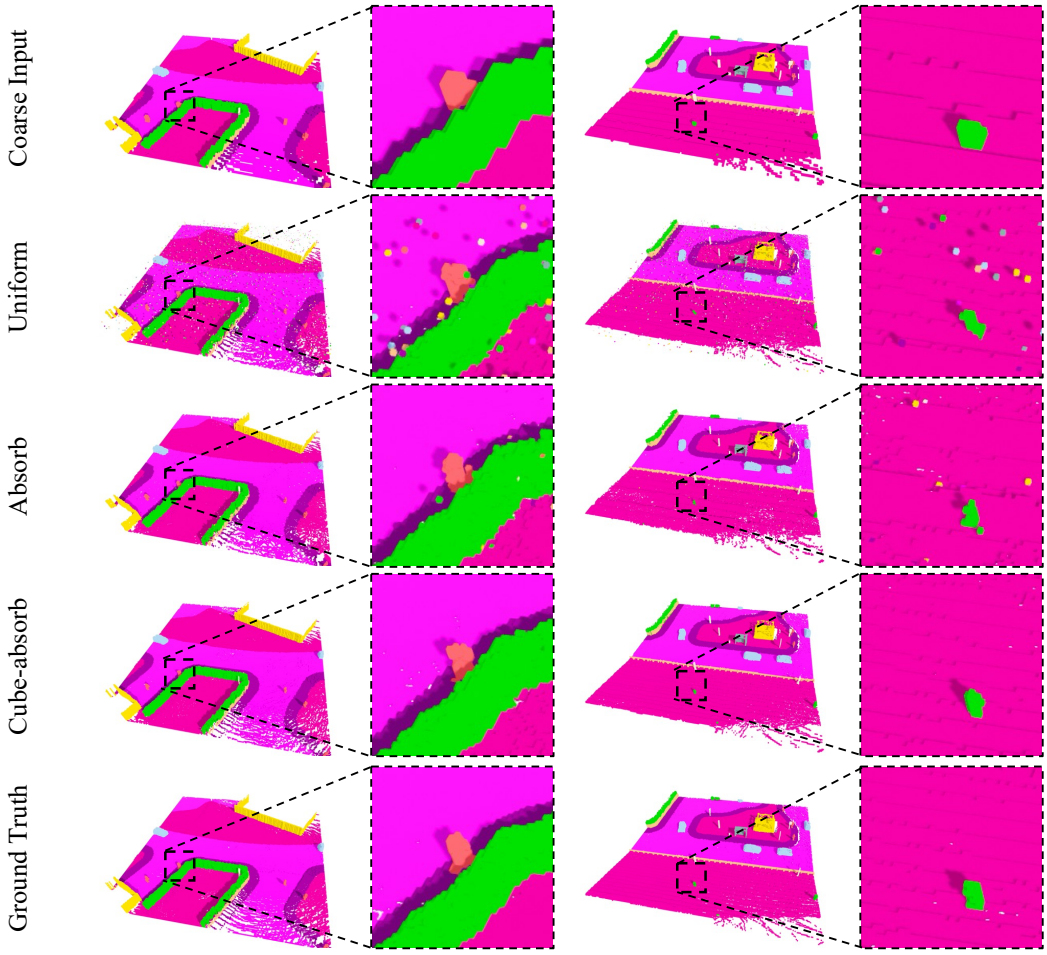

Figure 10. Coarse-to-fine generation results of different diffusion processes. The proposed cube-absorb diffusion demonstrates the closest alignment with the ground truth.

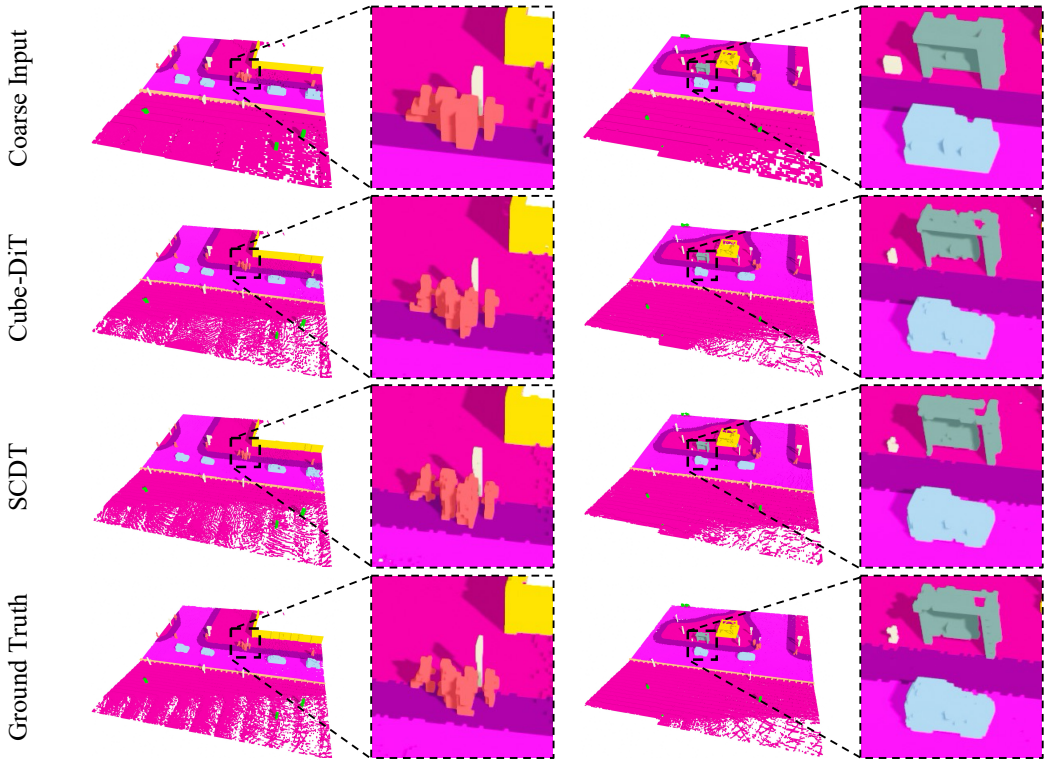

Figure 11. Coarse-to-fine generation results of Cube-DiT and the proposed SCDT. The proposed SCDT achieves generation performance on par with Cube-DiT while significantly reducing memory usage (6.38G versus 17.09G).

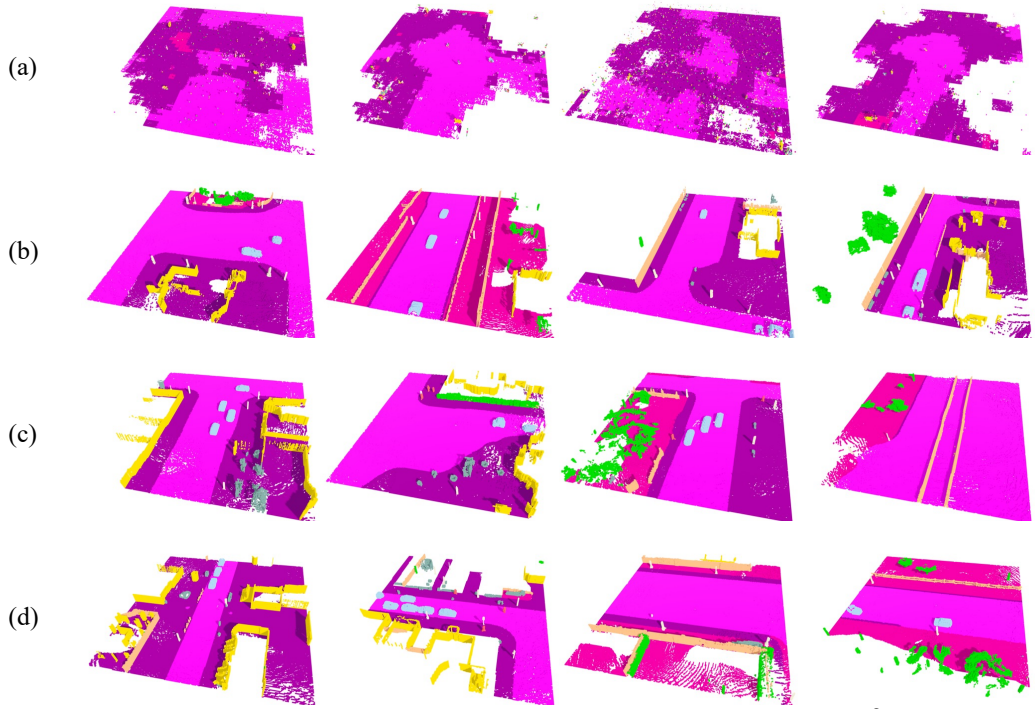

Figure 12. Unconditional generation results on various hierarchy resolutions and depths. Model (a): $256^2 \times 16$; Model (b): $64^2 \times 4 \rightarrow 256^2 \times 16$; Model (c): $64^2 \times 4 \rightarrow 128^2 \times 8 \rightarrow 256^2 \times 16$; Model (d): $32^2 \times 2 \rightarrow 64^2 \times 4 \rightarrow 128^2 \times 8 \rightarrow 256^2 \times 16$. Hierarchical models outperform single-level models, demonstrating their effectiveness in representing 3D structures. The model's performance is robust to both the initial resolution and hierarchy depth. .

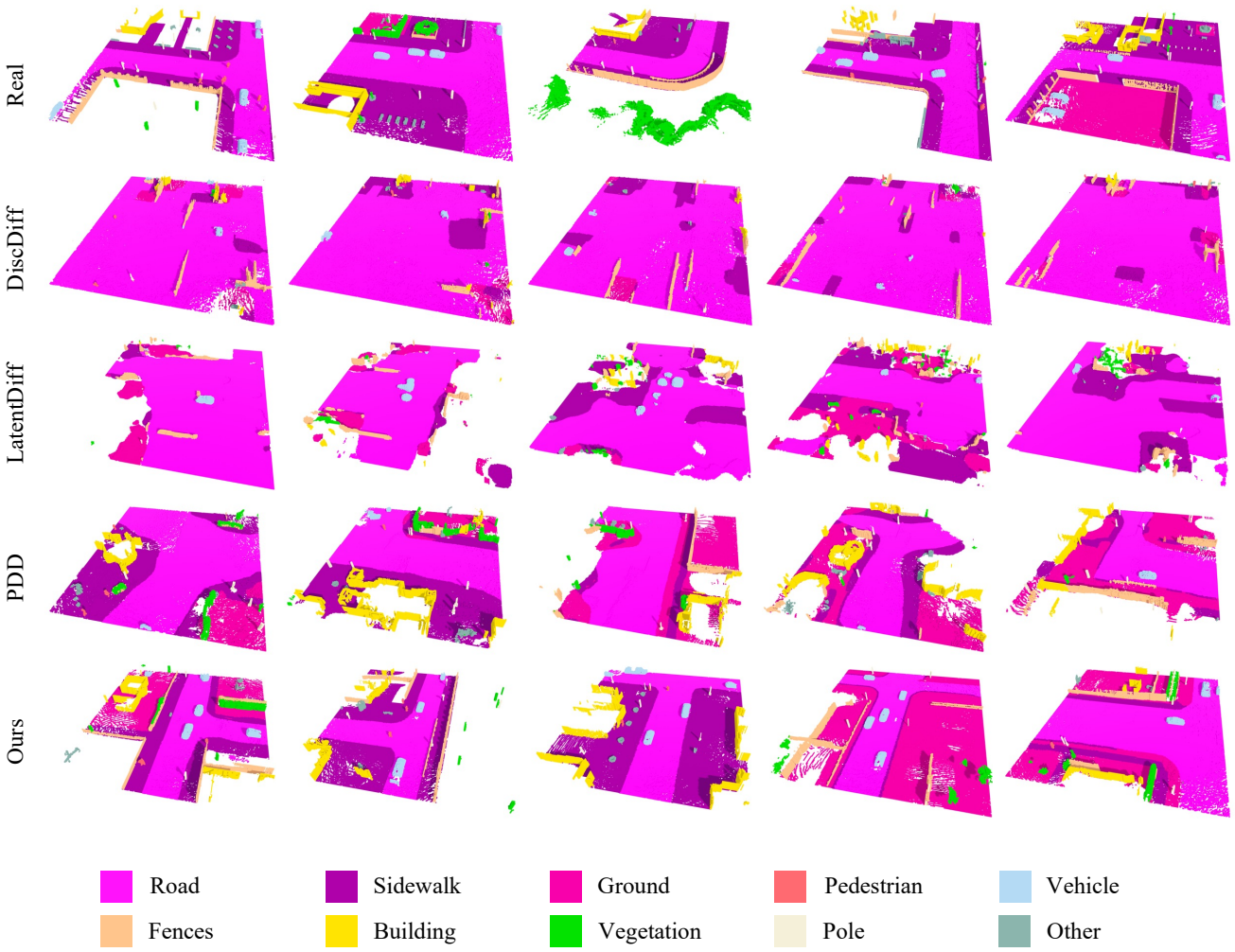

Figure 13. Additional visualization of unconditional generation results on CarlaSC. Real scenes are only for reference.

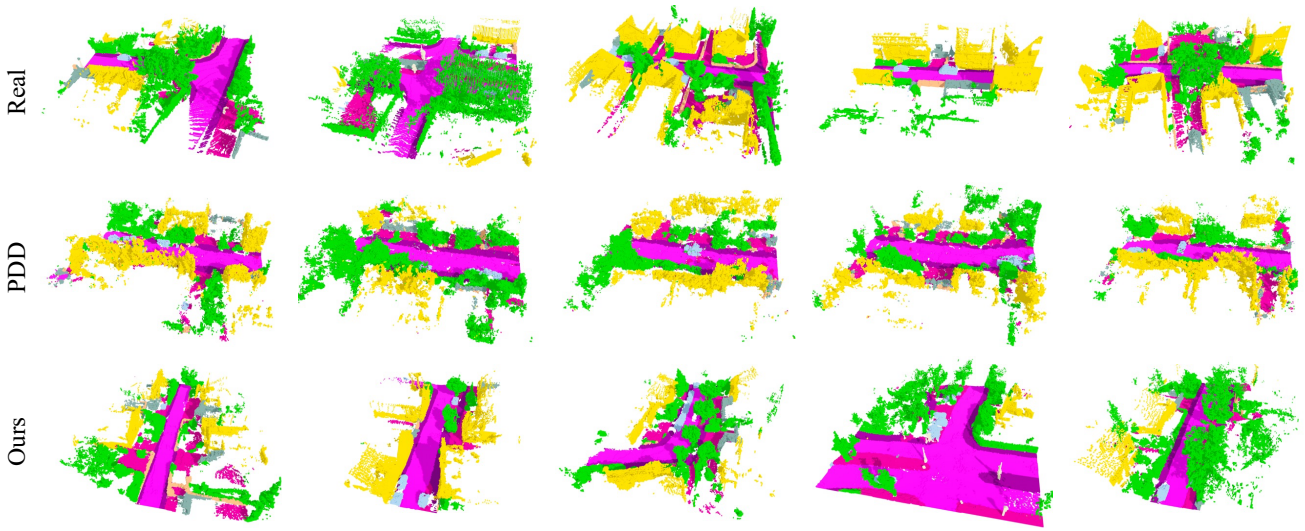

Figure 14. Additional visualization of unconditional generation results on KITTI-360. Real scenes are only for reference.

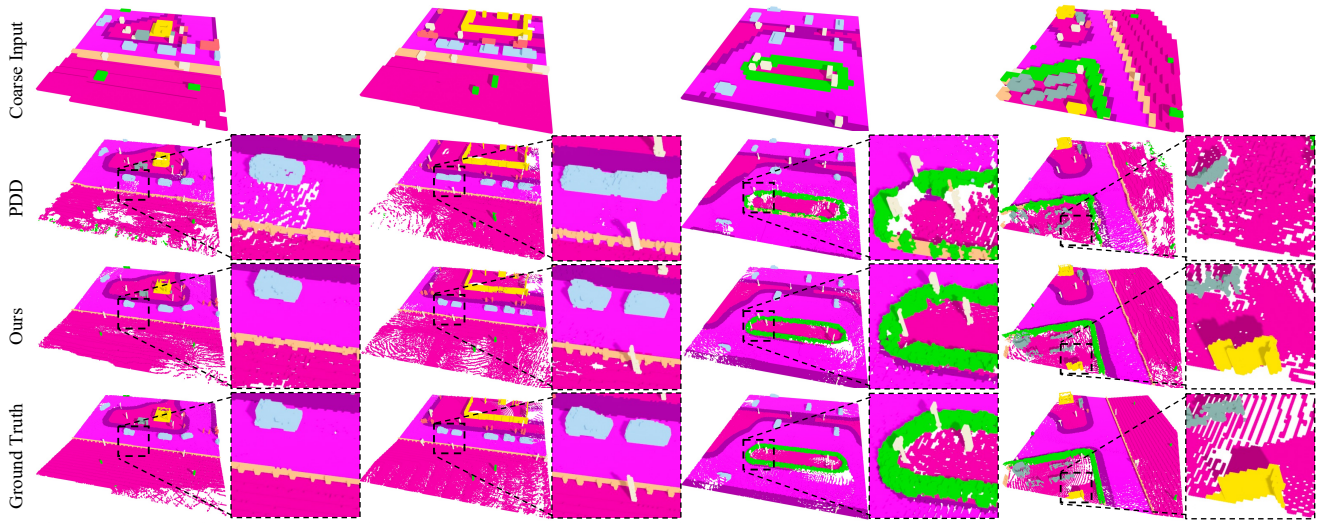

Figure 15. Additional visualization of conditional generation results on CarlaSC.

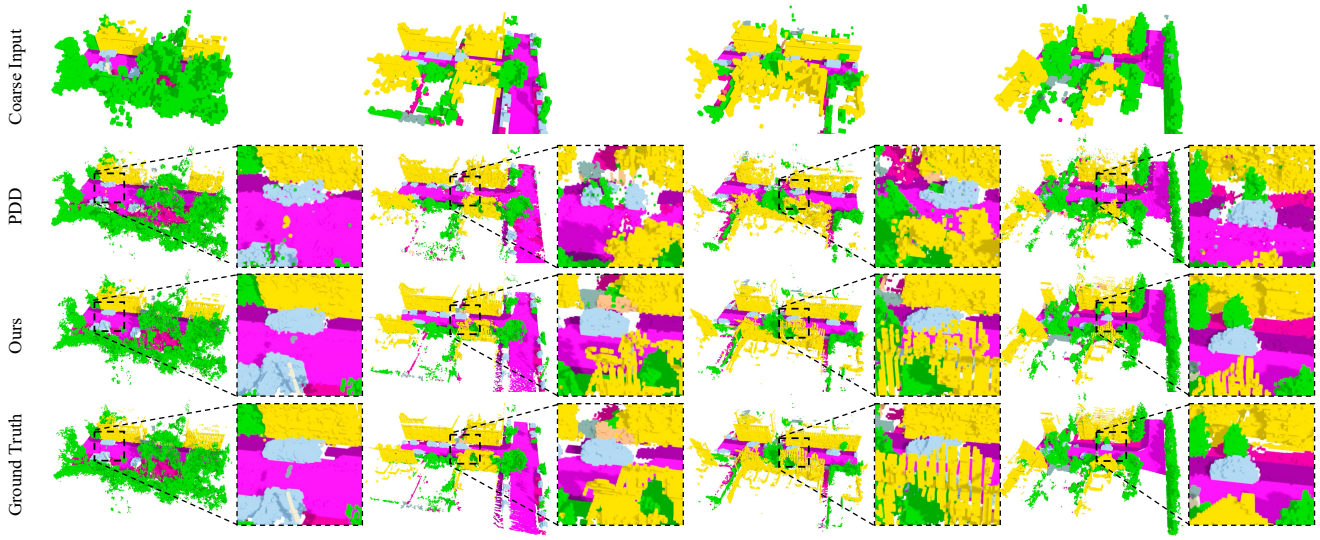

Figure 16. Additional visualization of conditional generation results on KITTI-360.