

SGFormer: Satellite-Ground Fusion for 3D Semantic Scene Completion

Supplementary Material

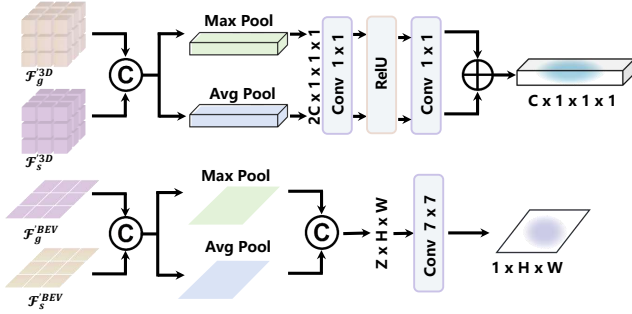


Figure 1. **Channel and Spatial Attention.** Our channel attention mechanism (**top part**) utilizes average and maximum pooling operations, followed by shared MLP layers. Similarly, our spatial attention mechanism (**bottom part**) employs the same pooling operations, paired with 7×7 convolution layer.

A. Extended Technical Details

A.1. Convolutional Enhancement Details

In the convolutional enhancement module, we utilize 3D and 2D U-Nets to improve prediction performance. For the ground branch, we employ a 3D U-Net with two encoder and decoder layers. In the satellite branch, we use a 2D U-Net with three encoder and decoder layers.

A.2. Fusion Module Details

Our channel and spatial attention mechanisms used in the dual-path weight generator are inspired by CBAM [8], with the detailed architecture illustrated in Figure 1.

Channel Attention. The channel attention operation is shown in the upper part of Figure 1. First, we compress the concatenated voxel features along the spatial dimensions using both average and maximum pooling operations, generating $\mathbf{F}_{channel}^{avg}$ and $\mathbf{F}_{channel}^{max} \in \mathbb{R}^{D \times 1 \times 1 \times 1}$. These compressed features are then passed through shared MLP layers consisting of two MLP layers with a ReLU activation function, producing two channel attention maps \mathbf{W}_c^{avg} and $\mathbf{W}_c^{max} \in \mathbb{R}^{D \times 1 \times 1 \times 1}$. Finally, we sum the two maps to compute the channel attention weight \mathbf{W}_c . Mathematically, the whole operation can be expressed as:

$$\mathbf{W}_c = \text{MLP}(\text{AvgPool}(\mathbf{F}_c^{3D})) + \text{MLP}(\text{MaxPool}(\mathbf{F}_c^{3D})) \quad (1)$$

where MLP is the shared MLP layers, AvgPool and MaxPool is average and maximum pooling operation, respectively.

Spatial Attention. The spatial attention operation is shown in the lower part of Figure 1. Similar to the channel attention operation, we compress the concatenated BEV features

along the channel axis to obtain $\mathbf{F}_{spatial}^{avg}$ and $\mathbf{F}_{spatial}^{max} \in \mathbb{R}^{1 \times H \times W}$. These features are concatenated and fed into a convolutional layer with a 7×7 kernel, producing the spatial attention weight \mathbf{W}_s . The equation of the spatial attention operation is shown as follows:

$$\mathbf{W}_s = \text{conv}^{7 \times 7}(\text{AvgPool}(\mathbf{F}_c^{BEV}) \textcircled{\text{}} \text{MaxPool}(\mathbf{F}_c^{BEV})) \quad (2)$$

where $\textcircled{\text{}}$ is concatenate operation, while $\text{conv}^{7 \times 7}$ is convolution operation with the 7×7 kernel.

Additionally, in our probability network, the spatial attention operation follows the same steps. However, since the input features are voxel features, the convolutional layer is replaced with a 3D dilated convolution. Moreover, the MLP used in the weight generator is a 2-layer MLP with ReLU activation. In the probability net, we utilize a basic ResNet block [3] combined with a 2-layer MLP.

A.3. Loss Function Detail

Scene Class Affinity Loss. Our scene class affinity loss is the same as previous work [2, 6]. The loss computes the class-wise derivable precision, recall, and specificity. The mathematical equation is:

$$\begin{aligned} P_c(\hat{p}, p) &= \log \frac{\sum_i \hat{p}_{i,c} [p_i = c]}{\sum_i \hat{p}_{i,c}} \\ R_c(\hat{p}, p) &= \log \frac{\sum_i \hat{p}_{i,c} [p_i = c]}{\sum_i [p_i = c]} \\ S_c(\hat{p}, p) &= \log \frac{\sum_i (1 - \hat{p}_{i,c}) (1 - [p_i = c])}{\sum_i (1 - [p_i = c])} \end{aligned} \quad (3)$$

where the $\hat{p}_{i,c}$ is the predicted probability for the class c , while p_i is the ground truth. P_c and R_c denote the precision and recall, respectively, evaluating the performance of voxels belonging to class c . And S_c is specificity, which measures the performance of voxels not belonging to class c . We get the scene class affinity loss \mathcal{L}_{scal} by summing the P_c , R_c , and S_c together as follows:

$$\mathcal{L}_{scal}(\hat{p}, p) = -\frac{1}{C} \sum_{c=1}^C (P_c(\hat{p}, p) + R_c(\hat{p}, p) + S_c(\hat{p}, p)). \quad (4)$$

In our paper, we use the semantic label and binary geometry label to obtain semantics affinity loss \mathcal{L}_{scal}^{sem} and geometry affinity loss \mathcal{L}_{scal}^{geo} , respectively, and add them together.

Cross-Entropy Loss. As mentioned in Section 3.5 of the paper, we use weighted cross-entropy loss to compute three losses: \mathcal{L}_{ce} , \mathcal{L}_{co} , and \mathcal{L}_{bev} . Our cross-entropy equation is

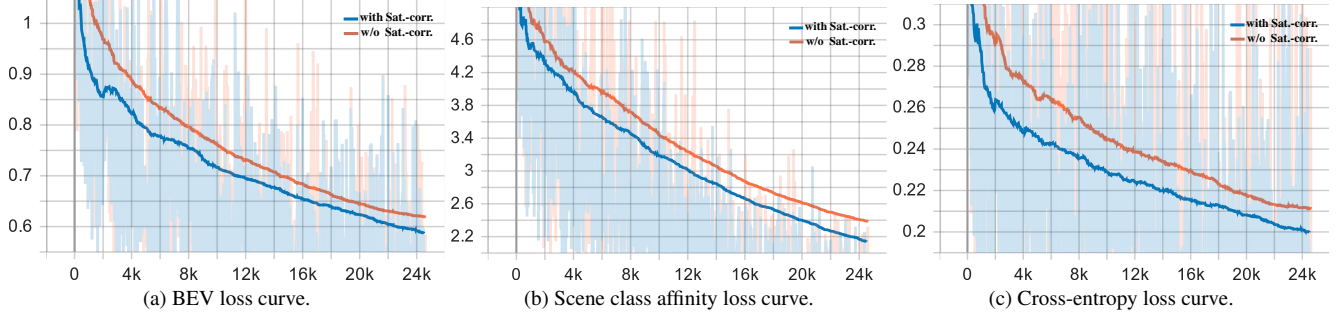


Figure 2. **Loss curve with/without satellite correction strategy.** The orange curves are the loss curves of SGFormer without the satellite correction, while blue curves are the loss curves with satellite correction.

| Channel Attention | Spatial Attention | Probability Net | IoU | mIoU | Global | Detail |
|-------------------|-------------------|-----------------|--------------|--------------|--------------|-------------|
| ✗ | ✗ | ✗ | 43.90 | 15.59 | 28.43 | 8.10 |
| ✓ | ✗ | ✗ | 44.85 | 16.25 | 29.19 | 8.71 |
| ✗ | ✓ | ✗ | 44.88 | 16.12 | 29.97 | 8.05 |
| ✗ | ✗ | ✓ | 44.95 | 16.45 | 29.06 | 9.12 |
| ✓ | ✓ | ✓ | 45.01 | 16.68 | 29.31 | 9.29 |

Table 1. Ablation on components of fusion module.

| Ground image backbone | IoU | mIoU |
|-----------------------|--------------|--------------|
| ResNet-50 | 44.30 | 16.14 |
| EfficientNet-B7 | 45.01 | 16.68 |

Table 2. Ablation on different backbones.

expressed as follows:

$$loss = - \sum_{c=1}^N w_c \log \frac{\exp(\hat{y}_c)}{\sum_{i=1}^N \exp(\hat{y}_i)} y_c \quad (5)$$

where \hat{y} is input, y is the ground truth, and N denotes the class number. w_c is the class weight that is inverse of class frequency.

B. Additional Ablation Study

In this section, we present the additional ablation analysis, which consists of three parts: adaptive fusion module, Backbone, and learning efficiency.

B.1. Ablation on Adaptive Fusion Module

Table 1 presents the ablation on detailed components of our adaptive fusion module. Overall, the integration of any component within our fusion module significantly enhances the prediction accuracy of SGFormer. This improvement arises because the two branches focus on different aspects of the task; thus, introducing an adaptive

weighting mechanism naturally brings substantial performance improvement. Specifically, incorporating the **channel attention network** simultaneously improves the accuracy for both scene layout and small objects: the mIoU for the "global" category increases from 28.43 to 29.19, while the mIoU for the "detail" category improves from 8.10 to 8.71. This enhancement is attributed to the module's capability to perform trade-offs at the object level, making the fusion strategy more inclined to rely on the satellite branch's outputs for objects in the "global" category and on the ground branch for the "detail" category.

The addition of the **spatial attention network** further increases the accuracy of our method in capturing the scene layout, increasing the mIoU from 28.43 to 29.97. This improvement is due to the spatial attention network's ability to balance the weights of the sensors across different regions based on each sensor's detection range, enabling our method to generate a more comprehensive scene layout.

Finally, the incorporation of the **probability net** enhances the prediction accuracy across all categories. The probability net adaptively identifies the more important voxels, thereby improving the overall learning efficiency of the model.

B.2. Ablation on Image Backbone

Table 2 shows the comparison between different ground image backbones, including EfficientNet-B7 [7] (used in SGFormer) and ResNet-50 [3] (used in baseline methods such as VoxFormer and Symphonize). The results highlight two

key observations. First, EfficientNet-B7 achieves higher performance due to its more efficient architecture. Second, even when using ResNet-50, SGFormer still outperforms baseline methods, demonstrating the superiority of our dual-branch framework.

B.3. Ablation on Learning Efficiency

Figure 2 shows the learning curves of our method with and without satellite correction strategy. We report all the losses we used in the training, including BEV loss, scene class affinity loss, and cross-entropy loss (summed with coarse loss). According to the figure, adding satellite correction operations can improve learning efficiency. This finding confirms what we mentioned about adding satellite correction to warm up learning.

C. Additional Visualization

Figure 3 and Figure 4 presents the more visualization results from SemanticKITTI [1]. Moreover, we include qualitative results on the SSCBench-KITTI-360 dataset [5]. Since VoxFormer [6] does not provide code for SSCBench-KITTI-360, we only compare our method to Symphonize [4] on this dataset.

D. Limitation

Although our work has achieved excellent results on two datasets, it still has several limitations. First, our method relies on appropriate satellite map inputs, and previous ablation experiments have shown that our approach is quite sensitive to localization noise. Therefore, if there is a significant deviation between the input satellite imagery and the actual location, the performance of our method will be greatly reduced. Additionally, we have not introduced many innovations targeting dynamic and small objects; thus, for these categories, our method does not show significant improvements compared to other approaches. Finally, adding a satellite branch brings additional computational resource, making our method less lightweight. We will address these issues in future work. Despite the above problems, we still believe that our method has made a valuable contribution and provided inspiration to the field of SSC task.

References

- [1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quen-
zel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Se-
mantickitti: A dataset for semantic scene understanding of li-
dar sequences. In *Proceedings of the IEEE/CVF international
conference on computer vision*, pages 9297–9307, 2019. 3, 4,
5
- [2] Anh-Quan Cao and Raoul De Charette. Monoscene: Monoc-
ular 3d semantic scene completion. In *Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern
Recognition*, pages 3991–4001, 2022. 1
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2
- [4] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 3
- [5] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023. 3, 6
- [6] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 1, 3
- [7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2
- [8] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

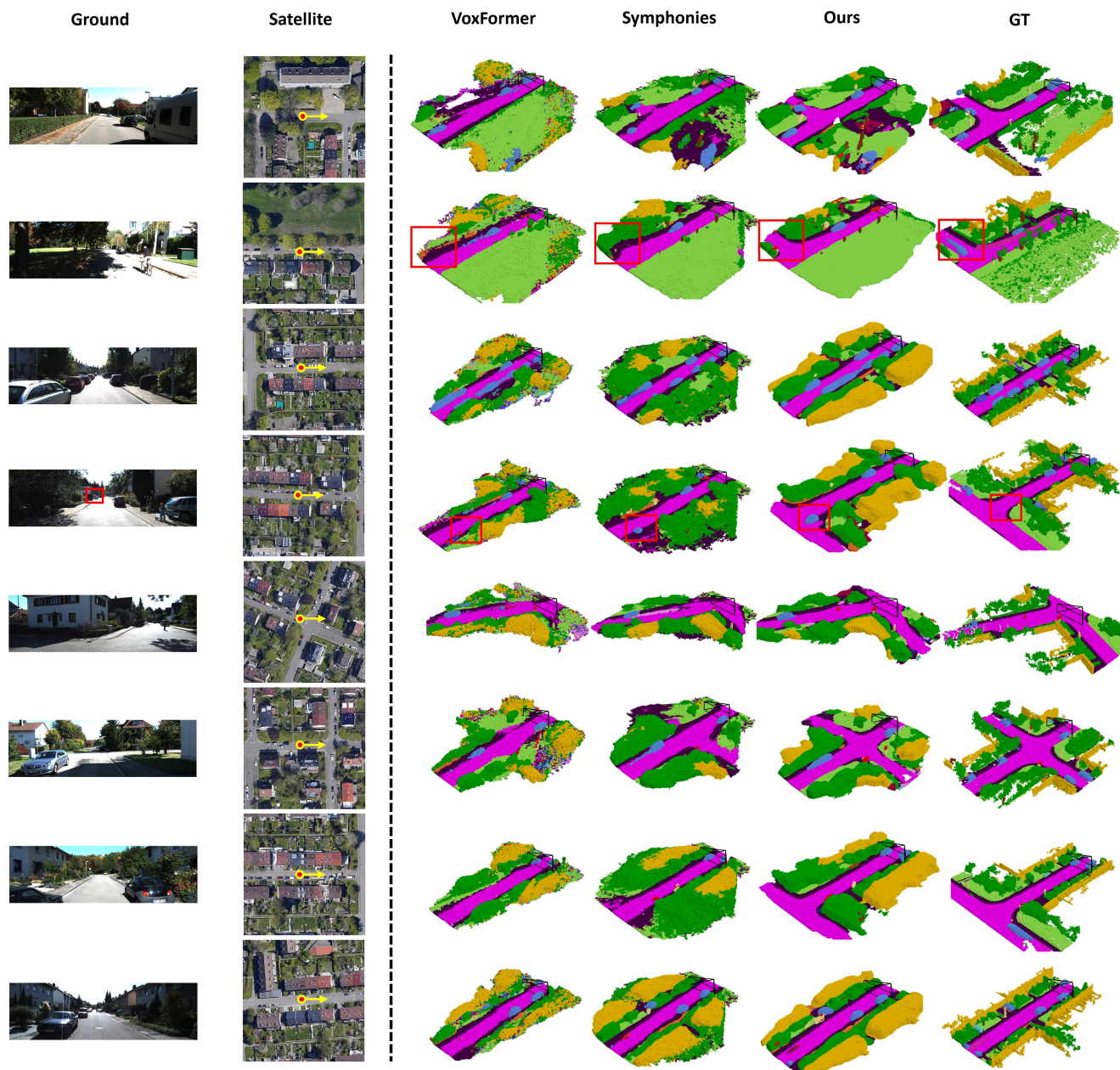


Figure 3. Additional Visualization on SemanticKITTI [1].

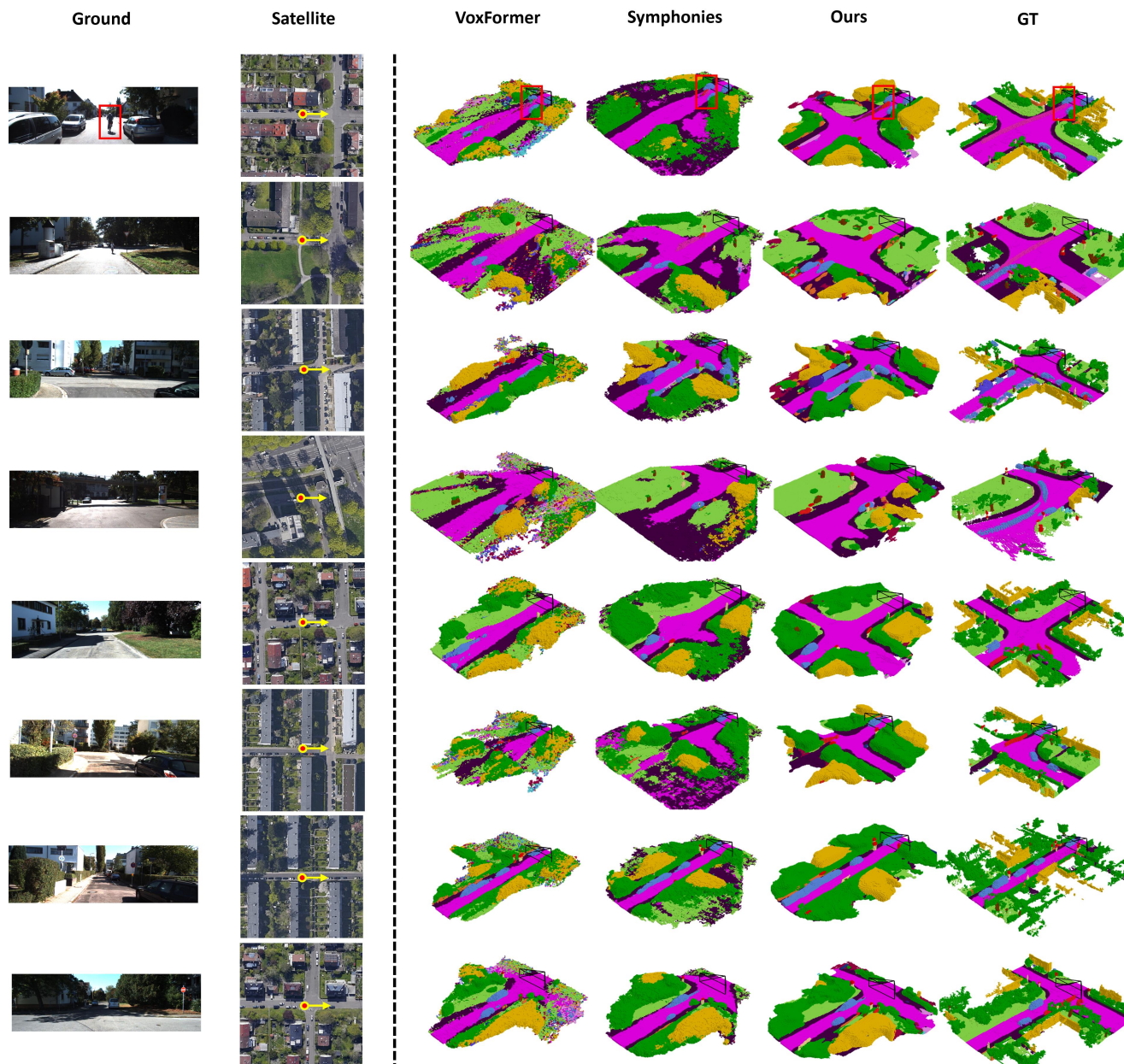


Figure 4. Additional Visualization on SemanticKITTI [1].

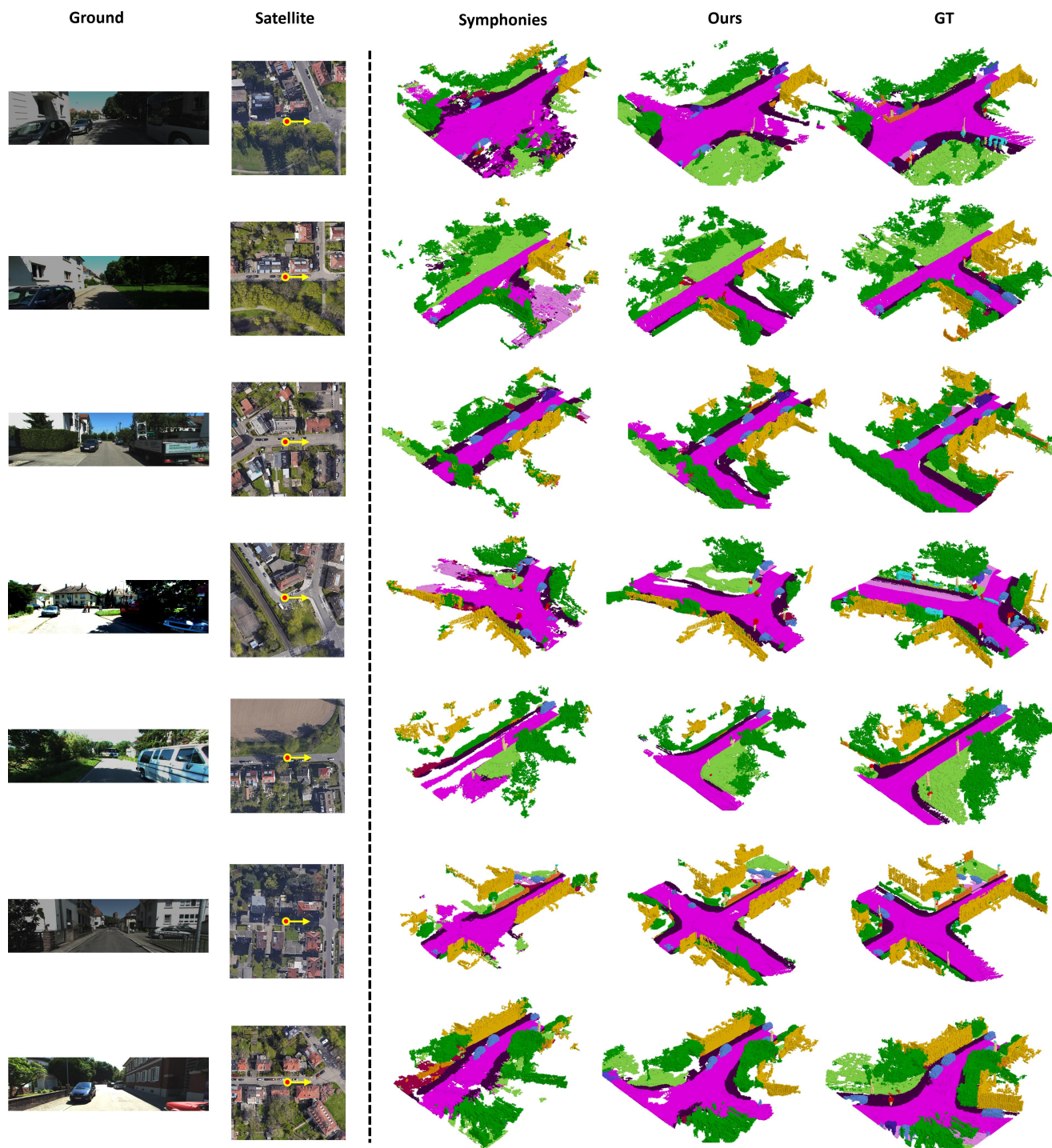


Figure 5. Additional Visualization on SSCBench-KITTI-360 [5].