# Supplementary Materials for SP3D

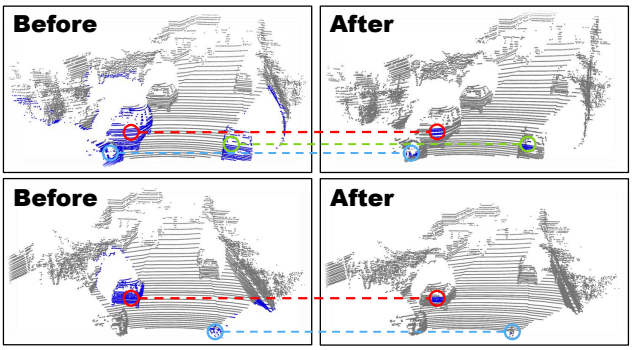


Figure 1. Visualization of semantic seed points transformed from LMMs-extracted foreground mask. The seed points are colored with blue. The left side represents the case before the mask shrink operation, while the right side shows the result after applying mask shrink. Mask shrink retains only the high-confidence core region of the foreground mask.

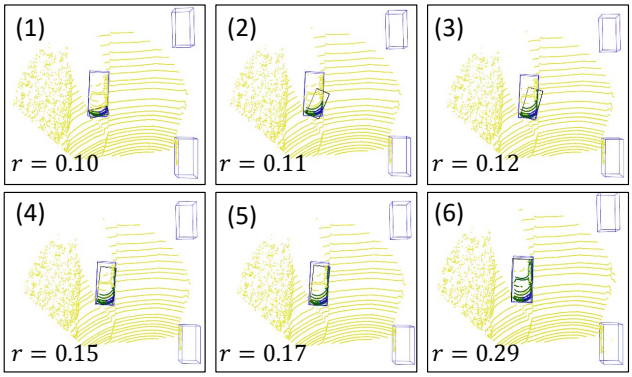


Figure 2. Visualization of the process of fitting bounding boxes with dynamic cluster radii in DCPG. As iterations proceed, the multi-scale neighborhood clustering mechanism can generate bounding boxes that more completely encompass foreground information.

## 1. The Visualization of Effectiveness of Mask Shrink

Fig. 1 illustrates the impact of the mask shrink operation on the accurate transmission of semantics. For ease of visualization, we have colored the transferred semantic seed points blue. The left column represents directly transferring semantic masks generated by LMMs, where uncertainty edge segmentation, coupled with the inherent one-to-many nature of the pixel-to-point cloud, often results in a significant number of background points being mistakenly classified as foreground. These pervasive noise exits in seed points significantly hinder the subsequent generation of high-quality pseudo-labels. At the same time, we observe that the noise is primarily concentrated at the edges of the mask. Based on this finding, we design a mask shrink strategy based on boundary constraints that only transfer the central region of the foreground masks onto the point cloud, eliminating edge semantic ambiguity and projection uncertainty. After incorporating this module, the effect on the seed points is shown on the right side of Fig. 1. It can be seen that we finally retained accurate seed points.

**$\gamma$ on mask shrink.** The core idea of mask shrink is to filter out the potentially ambiguous edge parts of the foreground mask, retaining only its core area. During the mask shrink process, we set a shrink factor $\gamma$ to control the size of the retained region. We qualitatively analyzed the $\gamma$ values in Fig. 3. As shown in the figure, when the $\gamma$ value is too high (e.g. 0.8 and 0.5), it generates a larger number of seed points, but this can lead to significant edge noise. When the $\gamma$ value is too small (e.g. 0.1), the number of seed points is significantly limited, which affects the fitting of subsequent pseudo-labels. Therefore, we have chosen a more balanced value of 0.3 for $\gamma$.

## 2. The Visualization of Effectiveness of DCPG

Fig. 2 demonstrates the pseudo-label fitting process of DCPG under different clustering radii. From this example, it can be seen that using a single fixed parameter for the clustering radius $r$ makes it difficult to fit the most appropriate bounding box pseudo-labels. In this case, our DCPG dynamically assigns different cluster radii $r$ to different seed points, which is capable of capturing multi-scale foreground information, thereby fitting higher-quality pseudo-labels. In addition, by integrating the DS score with the NMS strategy, we can eliminate the low-quality pseudo-labels effectively. Ultimately, it is the high-quality pseudo-labels that remain that provide the necessary support for the training of a high-performing initial 3D detector.

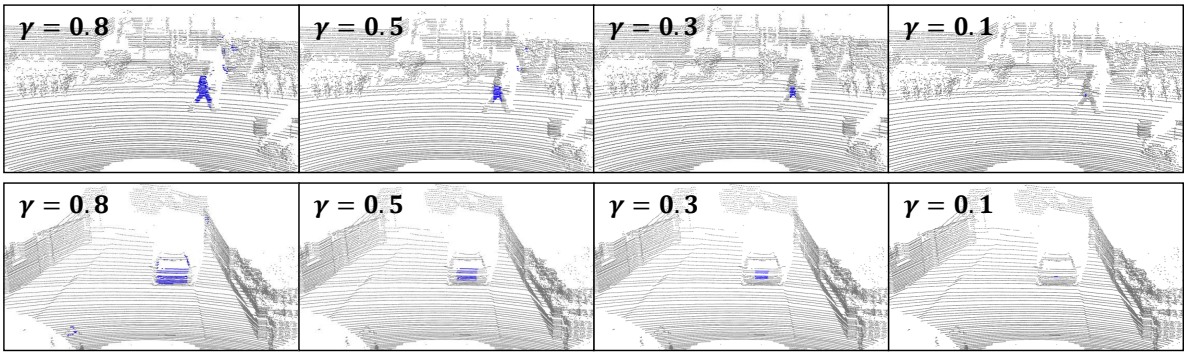


Figure 3. The impact of $\gamma$ on mask shrink. The seed points are colored with blue. When $\gamma$ is too large, mask shrink still retains some noise points that are not filtered out. When $\gamma$ is too small, it results in an insufficient number of generated seed points, thereby affecting the subsequent dynamic clustering pseudo-label generation.

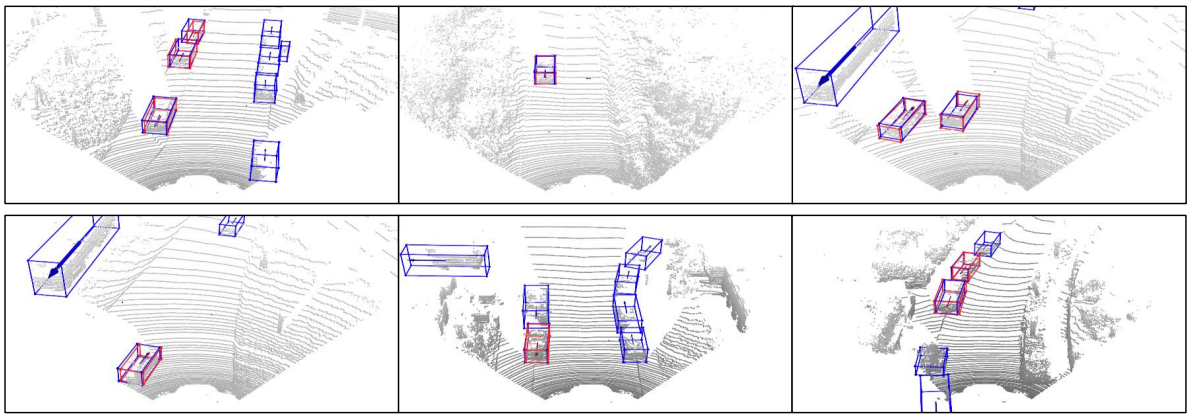

Figure 4. Visualization of pseudo-label quality assessment. Bounding boxes with red color represent the fitted pseudo-labels, while blue bounding boxes indicate ground truth boxes. SP3D can generate high-quality pseudo-labels close to the GT, but there is still significant room for improvement in the quantity of generated pseudo-labels.

## 3. Discussion of Failure Cases

The core contribution of SP3D is to utilize cross-modal semantic prompts to generate seed points, which are then used to dynamically fit the 3D point cloud bounding boxes. Finally, low-quality pseudo-labels are removed based on the DS score. Therefore, the quality and quantity of the seed points significantly affect the performance of SP3D. On one hand, the inherent prior bias of multimodal large models can lead to semantic acquisition errors in images. For example, "Cyclist" may be misclassified as "Pedestrian." In such cases, the noise interference is difficult to eliminate through mask shrink operations. On the other hand, when the number of points in the point cloud scene is extremely sparse, the number of seed points will also be relatively reduced. This makes it difficult for DCPG to perform as expected, and the fitted 3D bounding boxes may fail to correctly enclose the foreground objects. Moreover, due to the lack of corresponding semantic seed points, SP3D struggles to accurately generate pseudo-labels for 3D point clouds outside the camera's field of view. A promising approach is to utilize temporal information through tracking and other means to fill in the missing camera perspectives.

## 4. Comparison with Various Annotation Rates

To more intuitively demonstrate the impact of the proposed SP3D on the sparsely supervised algorithm, we take CoIn [1] as an example and conduct a group of comparative experiments under different annotation rates. Tab. 1 provides the variation in performance as annotation rates ranging from 10% to 0.1%. The experimental results indicate that the original sparsely-supervised 3D detector can significantly enhance performance upon integrating the proposed SP3D. For example, at a 2% labeling rate, the CoIn integrated with SP3D improved 3D AP by 15.41%, 14.42%, and 14.84% on easy, moderate, and hard difficulty levels, respectively. Also, this result represents an average improvement of 14.89% over the original detector. Besides, our SP3D significantly boosts the sparsely-supervised 3D

| Anno. Rate | Method | Car-3D @IoU 0.7 | | |
|---|---|---|---|---|
| | | Easy | Mod. | Hard |
| 100% | CenterPoint | 89.07 | 80.50 | 76.49 |
| 10% | CoIn | 85.95 | 71.80 | 62.64 |
| | + SP3D | 88.84 | 73.56 | 65.17 |
| 5% | CoIn | 81.64 | 67.48 | 58.32 |
| | + SP3D | 87.52 | 72.42 | 63.87 |
| 2% | CoIn | 72.03 | 54.82 | 43.77 |
| | + SP3D | 87.44 | 69.24 | 58.61 |
| 1% | CoIn | 70.39 | 51.31 | 41.31 |
| | + SP3D | 83.79 | 63.16 | 52.50 |
| 0.5% | CoIn | 66.77 | 47.68 | 38.38 |
| | + SP3D | 80.36 | 59.99 | 49.44 |
| 0.2% | CoIn | 45.47 | 31.20 | 23.52 |
| | + SP3D | 75.30 | 52.99 | 42.14 |
| 0.1% | CoIn | 6.84 | 4.65 | 3.61 |
| | + SP3D | 58.57 | 37.41 | 29.88 |

Table 1. Comparison with different annotation rates (10% $\rightarrow$ 0.1%). We report the results with 40 recall positions, under 0.7 IoU threshold.

detector's performance even at very low annotation rates, which achieves the 41.95% (36.92% higher than CoIn) average AP across different difficult levels under the annotation rate of 0.1%. The experimental results indicate that the performance of the original sparsely-supervised 3D detector can improve significantly after loading the SP3D-initialized model, even at low annotation rates.

## References

[1] Qiming Xia, Jinhao Deng, Chenglu Wen, Hai Wu, Shaoshuai Shi, Xin Li, and Cheng Wang. Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6254–6263, 2023. 2