

Supplementary Materials: Affinity3D: Propagating Instance-Level Semantic Affinity for Zero-Shot Point Cloud Semantic Segmentation

Anonymous Authors

1 MORE IMPLEMENTATION DETAILS

1.1 Visibility measurement module

The superpixel segmentation algorithm in the visibility measurement module was simple linear iterative clustering (SLIC) [1]. We set the approximate number of labels in the segmented output image to 4800 because we expected the boundary to be clearly segmented and adjacent local pixels to have similar depth.

1.2 Knowledge transfer module

Specifically, as illustrated in Fig. 1, we choose to transfer 2D knowledge to point clouds for the visible red points in images. We selected four corresponding scales of image and point cloud features to apply Kullback–Leibler loss for distillation. Both 2D and 3D semantic segmentation heads are composed of text embedding EMB^C as weights. We generate pseudo labels for 3D points belonging to unseen classes. For seen classes, the ground truth is directly used as supervision during training. Since pseudo labels for 3D points belonging to unseen classes are discriminated in images. Those invisible 3D points belonging to unseen classes are set to ignoring labels. For images, the pseudo labels are generated by perspective-projecting 3D points. Pixels with no corresponding 3D points are set to ignoring labels.

1.3 Training Details

The backbone of the 3D model was a SPVCNN [4] with a hidden size of 64. The SPVCNN consisted of 4 scales of layers. The initial spatial shape was $1000 \times 1000 \times 60$. The volume space was $[-50, 50]$ for X axis, $[-50, 50]$ for Y axis and $[-4, 2]$ for Z axis. The model was trained in 64 epochs with a learning rate of 0.24. The optimizer was Stochastic Gradient Descent (SGD) with momentum of 0.9 and weight decay of $1.0e - 4$. The learning rate scheduler was cosine annealing.

1.4 Inference Details

During test time, only the 3D model was available, and the image branch in the knowledge transfer module was removed. Therefore, our Affinity3D did not introduce additional parameters and inference time. Furthermore, all point cloud augmentations were deactivated during testing unless specifically noted for the utilization of Test Time Augmentation (TTA).

2 MORE EXPERIMENTS

2.1 Selection of propagation time in the affinity module

We evaluated the accuracy of pseudo labels for instances on the SemanticKITTI [2] train dataset. The instance generation module generated the instances, and the pseudo labels were obtained in the

Table 1: The ablation study of propagation time in affinity module. ‘GZS’ represents a generalized zero-shot setting.

Method	setting	β	Accuracy
CLIPInstance(without affinity)	GZS	\times	12723/15043 = 84.58%
CLIPInstance	GZS	1	12723/15043 = 84.58%
CLIPInstance	GZS	2	13084/15043 = 86.98%
CLIPInstance	GZS	3	12723/15043 = 84.58%
CLIPInstance	GZS	4	12621/15043 = 83.90%
CLIPInstance	GZS	5	12599/15043 = 83.75%
CLIPInstance	GZS	6	12592/15043 = 83.71%
CLIPInstance	GZS	7	12592/15043 = 83.71%

pseudo label generation module. The ground truth of an instance is defined as the class of most points belonging to the instance. Moreover, the accuracy was defined as the ratio of the true positives to the total number of instances. As shown in Table 1, the accuracy initially increased as the value of β rose, then decreased, eventually stabilizing at 83.71%. The maximum value was achieved when β was 2. Compared with affinity absence, the introduction of affinity consistently improves the quality of pseudo labels. It demonstrated the effectiveness of our affinity and the appropriate selection of the propagation time β .

2.2 Pseudo labels

Table 2: Comparison of pseudo labels on the nuScenes dataset. ‘GZS’ represents a generalized zero-shot setting.

Method	setting	Accuracy
MaskCLIP [5]	GZS	50805/107810 = 47.12%
CLIPInstance	GZS	54587/107810 = 50.63%

In Table 2, we conducted a comparison between our CLIPInstance and MaskCLIP [5] on the nuScenes train dataset [3] under the generalized zero-shot setting. As outlined in section 2.1, accuracy was determined as the ratio of true positives to the total instances. MaskCLIP generated 2D pseudo labels for images and mapped them to 3D points via perspective projection. Instance prediction in MaskCLIP was based on the class with the most points associated with it. Besides, CLIPInstance was derived from our pseudo label generation module. The results presented in Table 2 illustrated that CLIPInstance outperformed MaskCLIP, exhibiting an absolute improvement of 3.51%.

2.3 Visualization results for annotation-free setting

We presented more visualization results under the annotation-free setting in Fig. 2. It can be observed that compared with the baseline,

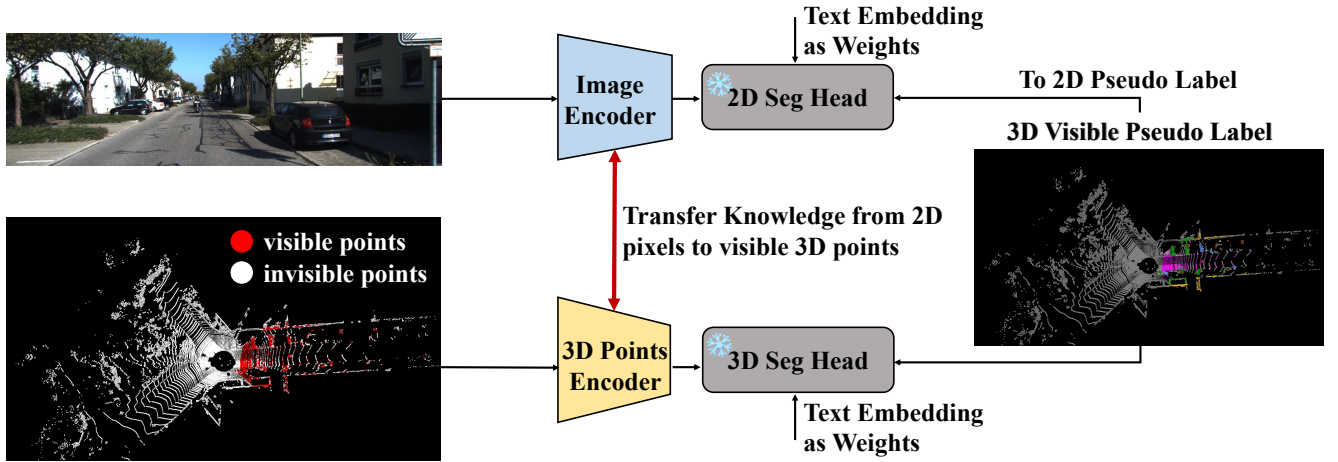


Figure 1: The illustration of knowledge transfer module.

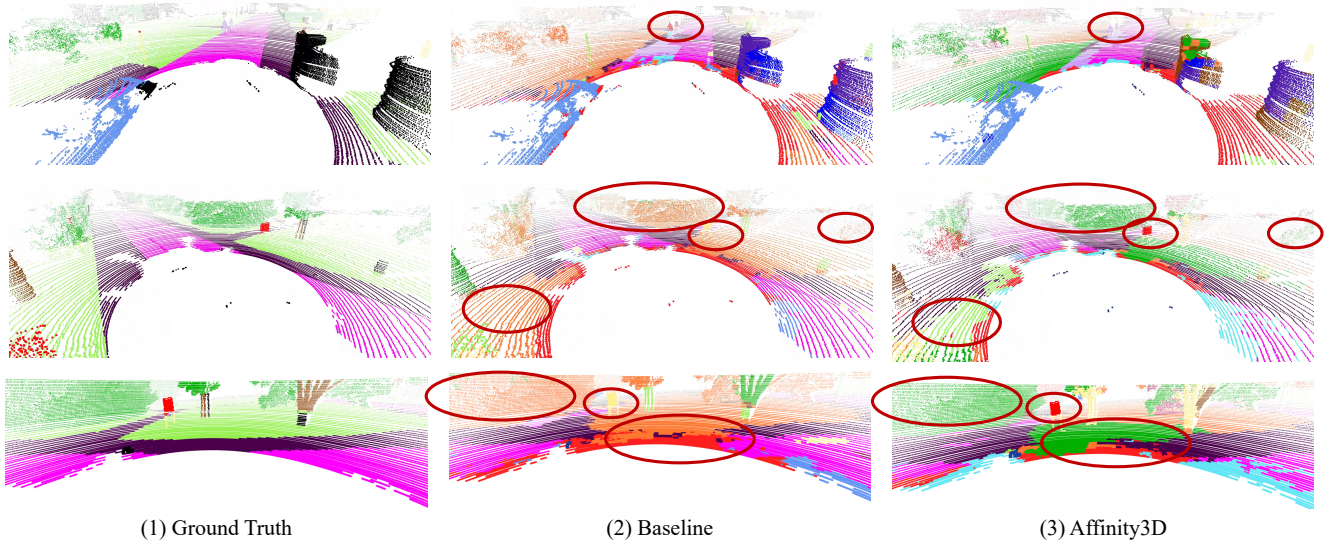


Figure 2: The visualization results of our Affinity3D, baseline, and ground truth under an annotation-free setting.

our method achieved more accurate predictions for ground and wall surfaces while exhibiting finer segmentation results along boundaries for traffic signs and bicyclists.

REFERENCES

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.
- [2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9297–9307.
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

- [4] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*. Springer, 685–702.
- [5] Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract Free Dense Labels from CLIP. In *European Conference on Computer Vision (ECCV)*.