7 Supplementary Materials

7.1 More implementation details

7.1.1 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.

7.1.2 Knowledge transfer module. Specifically, as illustrated in Fig. 6, 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.

7.1.3 Training Details. SPVCNN [37] with a hidden size of 64 was chosen as the backbone of the 3D model and 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.

7.1.4 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).

7.2 More Experiments

7.2.1 Experiments on the nuScenes dataset. We added generalized zero-shot experiments on the nuScenes dataset. The nuScenes dataset contains 1000 scenes with 40157 annotated samples. Each sample has 6 monocular camera images and a 32-beam LiDAR scan. It officially divides the data into 700/150/150 scenes for train/val/test and annotated for 17 classes in total. The motorcycle, construction vehicle, traffic cone, and trailer are selected as unseen classes, which are the same as in TCKZ [39]. As shown in Table 5, our method also achieved an hIoU of 71.93% in nuScenes, which is an 8.95% absolute improvement against TCKZ [39].

7.2.2 Selection of propagation time in the affinity module. We evaluated the accuracy of pseudo labels for instances on the training dataset of SemanticKITTI [2]. The instance generation module generated the instances, and the pseudo labels were obtained in the 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 6, 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.

Table 5: Comparisons on the nuScenes validation set in a generalized zero-shot setting. 'KT' means knowledge transfer module. Other abbreviations are the same in Table 1.

Setting	Ann.		Method	mIoU			hIoU
	S	U		Seen	Unseen	All	
FS	✓	✓	SPVCNN [37]	83.74	66.38	79.40	74.05
	✓		SPVCNN [37]	79.10	0	59.33	0
ZS	✓		MaskCLIP-3D+ [51]	52.88	35.79	48.61	42.79
	✓		3DGenZ [29]	55.28	20.52	46.59	29.93
	✓		TCKZ [39]	79.12	52.32	72.42	62.98
	✓		Affinity3D (without KT)	79.09	60.36	74.41	68.46
	✓		Affinity3D	**80.53**	**65.00**	**76.65**	**71.93**

Table 6: The ablation study of propagation time in affinity module. 'GZS' represents a generalized zero-shot setting.

Method	setting	β	Accuracy
CLIPInstance(without affinity)	GZS	×	12569/15043 = 83.55%
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%

Table 7: Comparison of pseudo labels on the nuScenes dataset. 'GZS' represents a generalized zero-shot setting.

Method	setting	Accuracy
MaskCLIP [53]	GZS	50805/107810 = 47.12%
CLIPInstance(with affinity)	GZS	70315/107810 = 65.22%

7.2.3 Pseudo labels. In Table 7, we conducted a comparison between our CLIPInstance and MaskCLIP [53] on the nuScenes train dataset [4] under the generalized zero-shot setting. As outlined in section 7.2.2, 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 7 illustrated that CLIPInstance outperformed MaskCLIP, exhibiting an absolute improvement of 18.10%.

We also compare our CLIPInstance with other point-language models. Experiments are conducted on the pseudo label generation

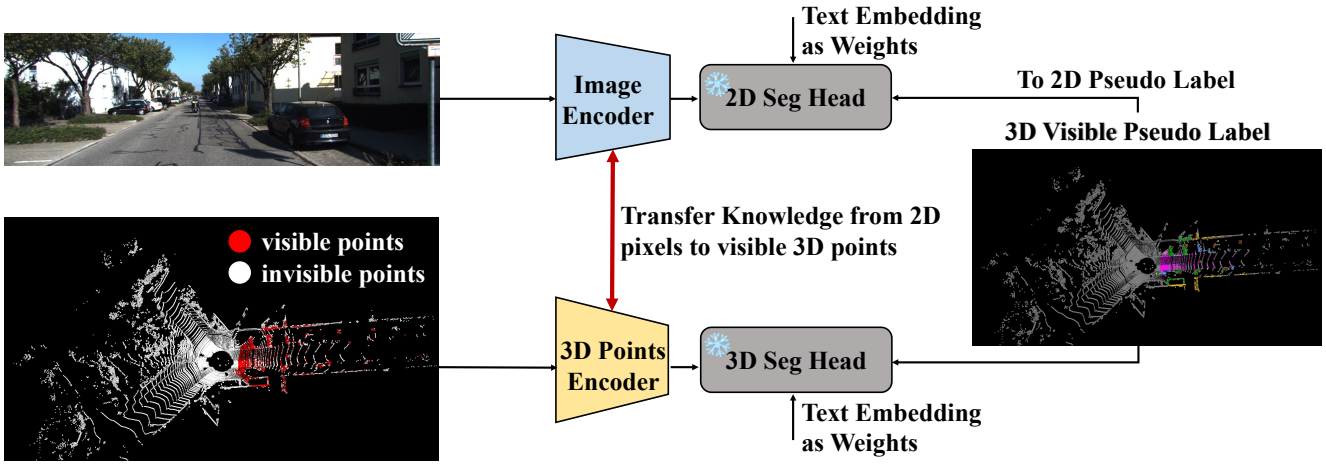
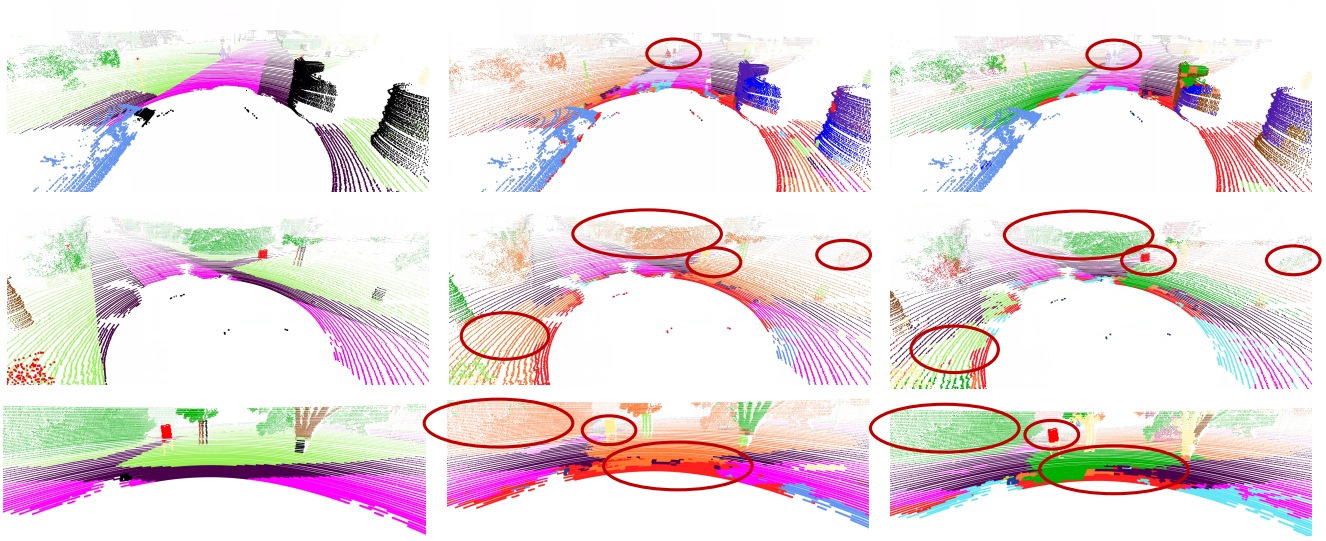


Figure 6: The illustration of the knowledge transfer module.



(1) Ground Truth (2) Baseline (3) Affinity3D

Figure 7: The visualization results of our Affinity3D, baseline, and ground truth under an annotation-free setting.

process. Table 8 shows the instance accuracies of pseudo labels of ULIP [43] and our method. On semanticKITTI, ULIP achieved 22.85% accuracy, and CLIP [35] achieved 83.55%. With the affinity module, our method reached 86.98%. The main reason for ULIP's bad performance is the domain gap. PointCLIP [55] and ULIP are trained on datasets like ModelNet40 [41] and ShapeNet [5], which differ from outdoor point clouds. Objects are partially visible in outdoor driving scenarios and differ in point count and distribution due to distance and occlusion. However, images have smaller morphological differences than point clouds, leading to better generalization. Therefore, we transfer semantic knowledge from CLIP instead of point-language models.

Table 8: Pseudo label quality on SemanticKITTI training set. Abbreviations are the same in Table 8.

Method	Affinity	Setting	Accuracy
ULIP Instance		GZS	3438/15043=22.85%
CI		GZS	12569/15043=83.55%
CI	✓	GZS	13084/15043=86.98%

7.2.4 Visualization results for annotation-free setting. We presented more visualization results under the annotation-free setting in Fig. 7. It can be observed that compared with the baseline, our method achieved more accurate predictions for ground and wall surfaces while exhibiting finer segmentation results along boundaries for traffic signs and bicyclists.