Temporal Overlapping Prediction: A Self-supervised Pre-training Method for Moving Object Segmentation

Supplementary Material

In this appendix, we describe a complementary formulation, the MOS labeling process, the cumulative distribution of moving objects and points, implementation details, the impact of ego-vehicle points on IoU, and potential negative social impacts. Finally, we present visual examples of the MOS results on the nuScenes and SemanticKITTI datasets in Fig. 2 and Fig. 3, respectively.

1. Formulation

As described in Sec. 3.2 Scenario-2 in the main paper, we provide the formula for the starting point $\mathbf{b}_i^{t_0}$ of the intersection segment. Since the spatial angle $\alpha_{i,j}$ is sufficiently small, we approximate that: (1) The two intersection segments are assumed to be equal in length, which is described as:

$$||\mathbf{q}_{i,j}|| - ||\mathbf{b}_i^{t_0}|| = ||\mathbf{q}_{i,j} - \mathbf{a}^t|| - ||\mathbf{b}_j^t - \mathbf{a}^t||. \tag{1}$$

(2) The distance between the two starting points $\mathbf{b}_i^{t_0}$ and $\mathbf{b}_j^t$ is equal to the sum of beam radii at the two points, represented as follows:

$$(||\mathbf{q}_{i,j}|| - ||\mathbf{b}_i^{t_0}||) \sin \frac{\alpha_{i,j}}{2} = ||\mathbf{b}_i^{t_0}|| \tan \frac{\theta_{\text{dvg}}}{2} + ||\mathbf{b}_j^t - \mathbf{a}^t|| \tan \frac{\theta_{\text{dvg}}}{2}. \tag{2}$$

We substitute Eq. 1 into Eq. 2 and solve for $||\mathbf{b}_i^{t_0}||$:

$$||\mathbf{b}_i^{t_0}|| = \frac{||\mathbf{q}_{i,j}||(\sin \frac{\alpha_{i,j}}{2} + \tan \frac{\theta_{\text{dvg}}}{2}) - ||\mathbf{q}_{i,j} - \mathbf{a}^t|| \tan \frac{\theta_{\text{dvg}}}{2}}{\sin \frac{\alpha_{i,j}}{2} + 2 \tan \frac{\theta_{\text{dvg}}}{2}}. \tag{3}$$

2. MOS Labeling

As discussed in Sec. 5.1 of the main paper, the nuScenes object attributes do not precisely describe the motion state, for example, "cycle.with_rider" can be either moving or static. Furthermore, a proportion of objects lack attributes: 0.466% of vehicles, 9.243% of cycles, and 1.913% of pedestrians. For our nuScenes MOS labeling, we calculate object speeds from their bounding box annotations. An object is classified as static if its speed is less than μ_{sta}, and as moving if its speed exceeds μ_{mov}. Objects with speeds between these thresholds are considered to have an unclear motion state and are thus classified as unknown. We use different speed thresholds for humans ($\mu_{\text{sta}}^{\text{hum}} = 0.375, \mu_{\text{mov}}^{\text{hum}} = 0.6$), cycles ($\mu_{\text{sta}}^{\text{cyc}} = 0.375, \mu_{\text{mov}}^{\text{cyc}} = 1.0$), and vehicles ($\mu_{\text{sta}}^{\text{veh}} = 0.5, \mu_{\text{mov}}^{\text{veh}} = 1.0$).

3. Cumulative Distribution

Fig. 1 presents a statistical analysis of moving objects from the nuScenes train-val split (discussed in Sec. 4), plotting the cumulative distribution functions of moving object instances and the number of their scanned points. The distributions are analyzed with respect to object size, defined as the number of points per object (x-axis). The significant gap between the two curves highlights a strong imbalance: the vast majority of moving objects consist of very few points. This is evidenced by the data: while moving objects with 19 or fewer points comprise 75.22% of all moving objects, they contribute a mere 15.49% of the total moving points. This disparity grows, with the smallest 90.06% of objects collectively accounting for only 37.02% of the total points. Such a skewed distribution implies that a small number of point-rich objects can disproportionately dominate the conventional IoU metric.

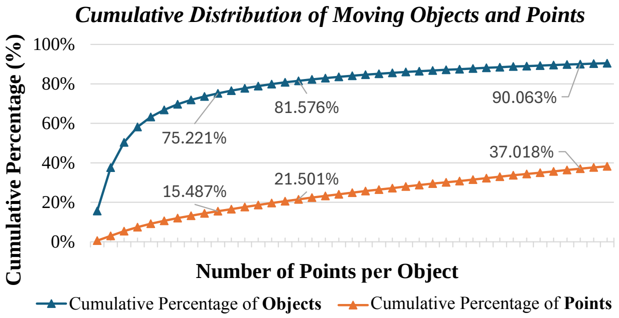

Figure 1. Cumulative distribution of moving objects and points.

4. Implementation Details

method	epochs	batch size	optimizer	lr start	lr max
TOP	50	8	AdamW	5.0×10^{-6}	5.0×10^{-5}
ALSO	200	4	AdamW	1.0×10^{-4}	1.0×10^{-3}
4DOcc	50	2	AdamW	8.0×10^{-5}	8.0×10^{-4}
UnO	50	16	AdamW	8.0×10^{-5}	8.0×10^{-4}

Pre-training Settings. The detailed pre-training settings are shown in the table above. The learning rate (lr) follows a warmup-cosine schedule, increasing from "lr start" to "lr max" during the first 2% of iterations (the warmup stage), and then follows a cosine schedule for the remaining iterations. All pre-training baseline methods maintain their original settings, while 4DOcc uses a very small batch size due to the high GPU memory consumption caused by the 0.1 m dense grid map.

Fine-tuning Settings. The Adam optimizer is used for all the fine-tuning; the learning rate follows a step-decay schedule, starting from 1.0×10^{-4} and decreasing by a factor of 0.99 after each iteration. For the nuScenes MOS experiments, the batch size is 4 for 5%, 10%, and 20% data subsets, while the batch size for the 50% data subset is 8. All methods are trained for over 400 epochs to ensure convergence. For SemanticKITTI cross-dataset transfer experiments, the batch size is 2 for all data subsets, and all methods are trained for over 300 epochs. For the nuScenes semantic segmentation experiment, all methods are trained for 300 epochs with a batch size of 4.

5. Impact of Ego-Vehicle Points on IoU Metric

As discussed in Sec. 4 of the main paper, the conventional IoU metric (denoted as IoU hereafter) includes ego-vehicle points, which inflates scores and masks actual perception performance on external moving objects. This issue is especially pronounced in the nuScenes dataset due to its high proportion of ego-vehicle points. To demonstrate this effect, we present nuScenes MOS results using IoU in Tab. 1. The results show that IoU scores are inflated by 1.5-2x compared to the ego-exclusive metric $\text{IoU}_{\text{w/o}}$. Furthermore, the apparent IoU performance gains are deceptive. These improvements are not from enhanced perception of the external environment, but from the far simpler task of ego-vehicle segmentation. Therefore, relying on this metric risks severely misrepresenting a model's actual capability.

Data	Pretrain	Best $\textbf{Recall}_{\textbf{obj}}$		
		$\text{Recall}_{\text{obj}}$	$\text{IoU}_{\text{w/o}}$	IoU
10%	No	24.98	36.44	68.89
	TOP	$28.03^{+3.04}$	$36.95^{+0.51}$	$70.28^{+1.39}$
20%	No	25.59	44.25	63.68
	TOP	$28.45^{+2.86}$	$44.30^{+0.05}$	$68.13^{+4.45}$

Table 1. nuScenes MOS results using the conventional IoU metric.

6. Potential Negative Social Impacts

While the proposed self-supervised method for MOS reduces annotation costs for autonomous systems, potential societal impacts should be considered. The technology could unintentionally prioritize detection accuracy for dominant object classes (e.g., vehicles) over vulnerable road users like cyclists/pedestrians if trained on imbalanced datasets, potentially compromising safety in edge cases. The temporal nature of our approach might propagate motion prediction errors in complex urban scenarios, leading to hazardous decisions by autonomous vehicles. Furthermore, while addressing sensor bias through our new metric helps, residual geographic/cultural biases in training data (e.g., urban vs rural environments) could limit global applicability. The method could also be repurposed for surveillance systems that infringe on privacy. To mitigate these risks, we recommend: (1) Rigorous testing across diverse operational domains. (2) Implementing fairness-aware data sampling strategies. (3) Establishing ethical guidelines for secondary applications.

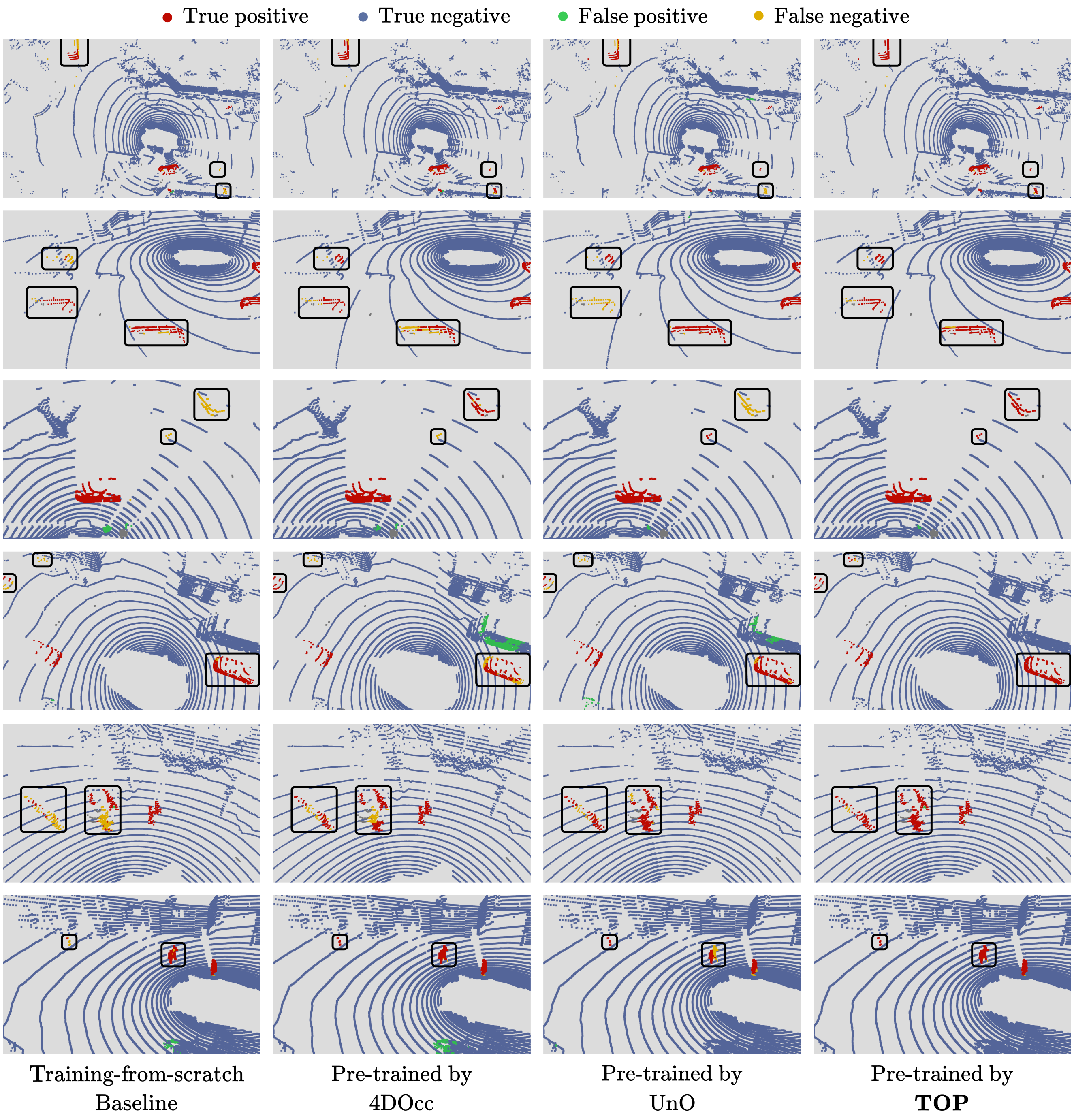

Figure 2. Qualitative results of the nuScenes MOS.

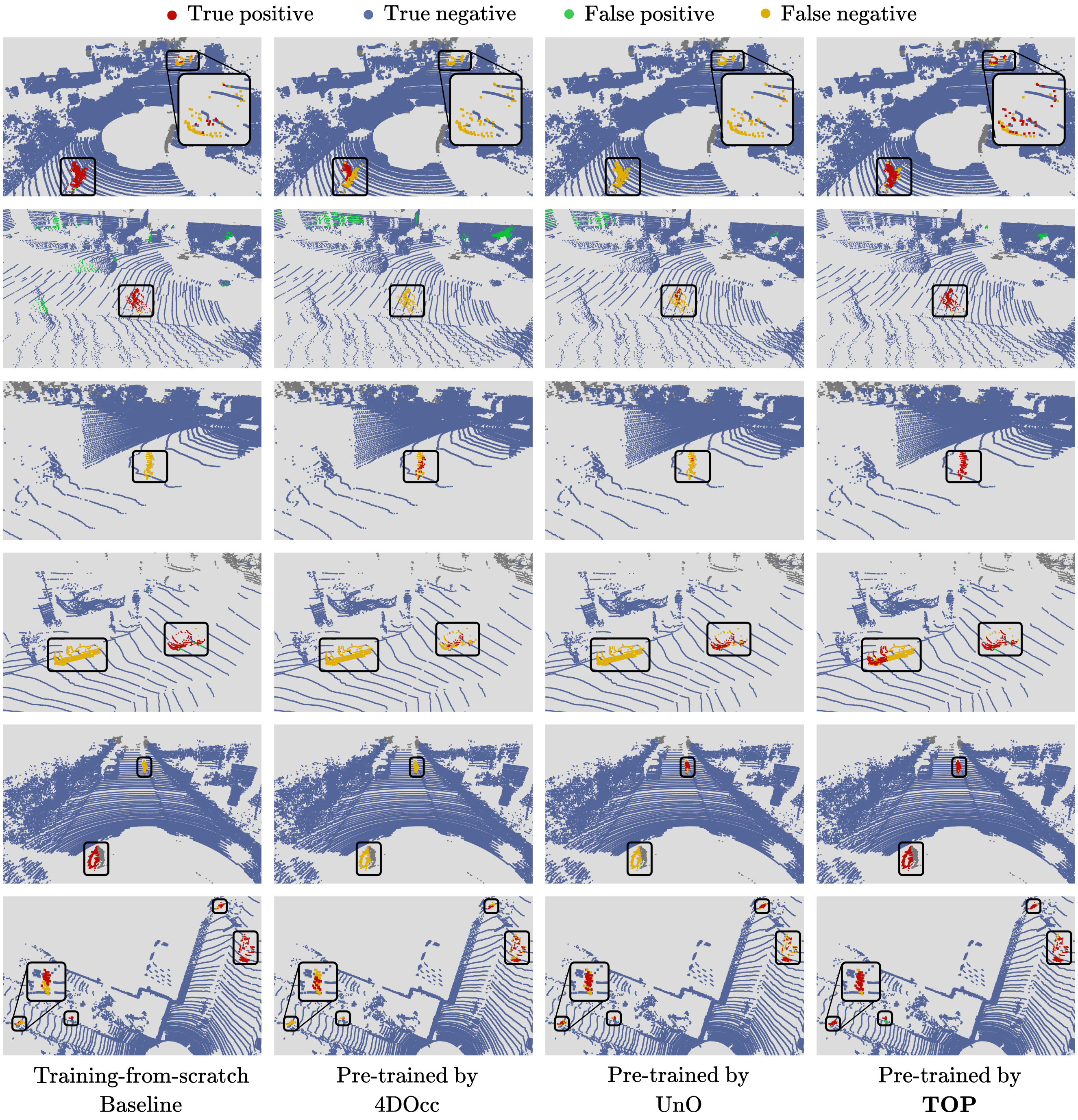

Figure 3. Qualitative results of the SemanticKITTI MOS.