Supplementary Material

A Implementation Details

We run the networks Centerpoint [36] and Transfusion-L [2] on 100×100m BEV grids around the ego vehicle. We use non-maximum suppression with a threshold of 0.1 (2D BEV IoU) for the detections. The optimizers, as well as their learning rate schedules are kept from the respective original implementations, but the schedules are shortened to match the lifecycle of the network weights during the iterative rounds of self-training. For the zero-shot generalization required by Oyster [38] after the first round, we found that starting from an initial BEV range of 50×50m, and then extending to 100×100m, gave the best results. For DBSCAN we used $\varepsilon = 1.0$, and minPts $= 5$. We optimize all tracks using Adam optimizer with learning rate 0.1 for 2000 steps for a complete point cloud sequence batched (at the same time), which takes less than 2s per nuScenes session on a Nvidia V100 GPU.

B Self-supervised lidar scene flow

As mentioned in Sec. 4, we extend the BEV range of SLIM [3] from 70×70m, 640×640 pixels to 120×120m and 920×920 pixels, but make no further modifications to the network. This results in the SOTA scene flow quality described in Table 5.

The small performance gap of our method between using ground truth and SLIM lidar scene flow (comparing the last and the second-to-last row of Table 4) demonstrates that SLIM lidar scene flow has suitable quality for our method, and also that our method does not require absolutely perfect lidar scene flow estimates to work well. Ground truth lidar scene flow is generated using the recorded vehicle egomotion for static points and the tracking information (bounding boxes) of moving objects.

Train Data	Val Data	AEE(moving)[m] ↓	AEE(static)[m]↓
AV2 Train	AV2 Val	0.079	0.075
KITTI Raw	KITTI Tracking	0.092	0.104
nuScenes Train	nuScenes Val	0.132	0.077
WOD Train	WOD Val	0.091	0.085

Table 5: Lidar scene flow metrics of SLIM [3] on the datasets (evaluated on val split), for a BEV range of 120×120m. Note that for KITTI, we only evaluate the forward-facing field of view (FoV) which has been annotated with tracked objects. Objects faster than 1m/s are considered *moving*. AEE refers to the average endpoint error across either all moving or static points.

C Additional Ablations

Box-size thresholds for initial pseudo ground truth: After fitting boxes to the initial lidar scene flow clusters, we discard abnormally-sized boxes, *i.e.* boxes that are smaller than a child or quite elongated *cf*. Section 3.2. Despite these thresholds being very permissive, having size limitations for the initial pseudo ground truth could potentially limit the applicability of our method. We therefore investigate how the performance of our method changes when omitting this constraint influences the performance in Table 6. While the impact on the overall performance is very minor and looks promising, further investigations would be required to exclude the possibility that small children and long but thin vehicles might potentially be underrepresented in the dataset, which would lead to a similar effect on the metrics.
Omitting weights dropping during self training: We also investigate the influence of weights dropping on the overall performance during self-training: I.e. we keep periodic regeneration of pseudo ground truth, but the network weights are never dropped between the rounds of incremental self training. The negative impact on performance is more significant here. We believe dropping the weights helps the network to escape the overfitting to noise from the previous iteration of pseudo ground truth more easily than via weak negative gradients.

Cluster Input	Method	Modification	AP@0.4 BEV	AP@0.4 3D
P, SF	LISO(K, SF)	-	0.380	0.308
P, SF	LISO(K, SF)	keep all cluster sizes	0.366	0.296
P, SF	LISO(K, SF)	never drop weights	0.334	0.261

Table 6: Additional ablations for **LISO**-CP on WOD (Movable). We investigate the influence of omitting the dropping of weights during self training ("never drop weights"), and the influence of omitting the discarding of clusters based on the size constraints as described in Section 3.2 ("keep all cluster sizes"). P: point cloud, SF: self-supervised lidar scene flow (SLIM), K: KISS-ICP.

D Performance of lidar scene flow clustering on nuScenes

In the evaluation on nuScenes (see Table 7), the worse performance of using DBSCAN [9] clustering on ground truth lidar scene flow compared to using DBSCAN on SLIM lidar scene flow is surprising. However, this peculiar effect is explained by Fig. 7, which shows the full precision-recall curves, generated using the official nuScenes protocol on the validation split [4]. The nuScenes protocol uses minimum precision and recall value thresholds of 0.1, discarding all results below these thresholds. As mentioned in Section 3.2, we assign confidence score

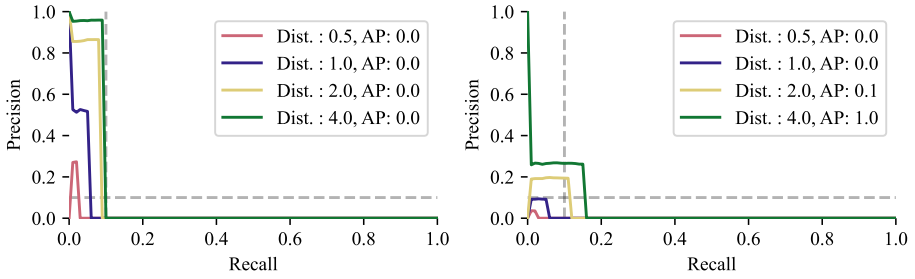

Fig. 7: Performance comparison of clustering ground truth lidar scene flow (left) and SLIM [3] lidar scene flow (right) on the nuScenes dataset. The methods are evaluated according to the official nuScenes protocol on the validation split. The dashed line represents the minimum threshold for precision and recall of 0.1, all results below these two thresholds are discarded. This leads to the surprising effect that the AP score is higher when using SLIM lidar scene flow, but this is only a result of the clipping dictated by the nuScenes evaluation protocol.

of 1.0 to all clusters discovered by DBSCAN. This causes all detections generated using DBSCAN on ground truth lidar scene flow to be discarded.

E Quality of pseudo ground truth during Iterative Self-Training

One critical aspect of iterative self-training is the quality of pseudo ground truth on the training dynamics, as depicted in Fig. 8. Finding the right balance between precision and recall in the pseudo ground truth is crucial for achieving optimal performance during self-training iterations: In our experiments, we find that having initially a small subset of high precision training samples is superior to having a larger set with higher recall but worse precision, because it allows the model to learn from a smaller but more reliable set of labeled data. A larger set of pseudo ground truth that is collected with less rigorous clustering, tracking and filtering, includes more noisy and mislabeled data. As discussed in [37, 38], the limited model capacity does prevent the model from overfitting to the inconsistent noises in the pseudo ground truth to some extent and the model generalizes mostly to the objects of interest, but in our experiments, higher quality pseudo ground truth with less noise ultimately leads to better performance. Motion cues (i.e. egomotion and lidar scene flow) are the superior clustering and tracking input signal, allowing our method to generate much cleaner initial pseudo ground truth when compared to Oyster, which we also demonstrate in our ablation in Table 4. Fig. 9 additionally visually demonstrates the difference between using lidar scene flow for initial pseudo ground truth creation and just using point clouds (Oyster) on an example point cloud: As expected, using lidar scene flow leads to fewer false positives in the initial pseudo ground truth.

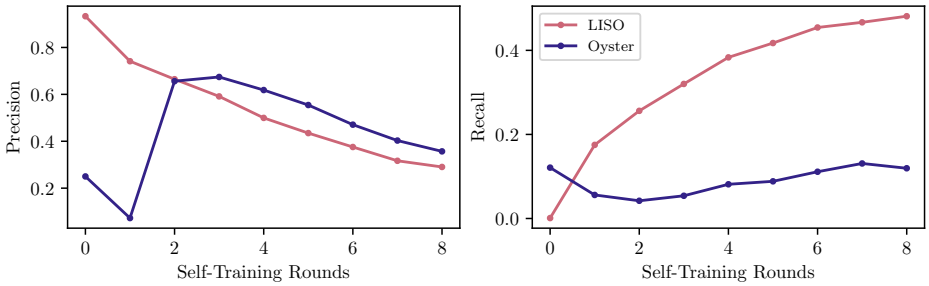

Fig. 8: Precision and recall of the (tracked) pseudo ground truth generated by Oyster and LISO over the course of self-training of Centerpoint on WOD (training split). Precision and recall are computed like in the AP metrics used in Fig. 4 and Table 2, i.e. true positives are occurences where the BEV IoU between ground truth and predicted boxes is greater than 0.4, but at a specific confidence threshold: For Oyster, we use the reported value from the publication $c = 0.4$ [38]. For LISO, we use $c = 0.3$ and only discard the learned weights every other round, as stated in Section 3.2. Note that the dip in Oyster's performance at round 1 stems from the zero-shot generalization, where the network is tasked to generalize from the training on the initial pseudo ground truth generated on the smaller BEV range to the full, previously unseen BEV range, going from 50×50m to 100×100m.

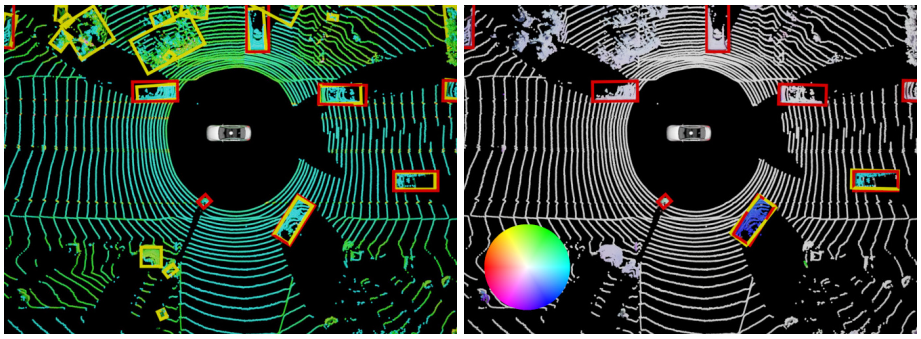

Fig. 9: Clustering results for the initial pseudo ground truth generation on WOD. Red boxes are ground truth boxes, yellow are predictions. **Left: Oyster** Clustering result on points, with high recall but low precision. **Right: LISO** Clustering result on points and SLIM lidar scene flow, resulting in in high precision pseudo ground truth (LISO). Points are colored according to flow direction and magnitude.

	method	AP↑	NDS↑	ATE↓	AOE ↓	ASE↓
GT	CP [36]	0.484	0.524	0.357	0.560	0.263
	TF [2]	0.627	0.614	0.287	0.501	0.207
Unsup.	DBSCAN [9]	0.008	0.109	0.987	0.171	0.962
	DBSCAN(SF)	0.003	0.106	1.186	**0.082**	0.952
	DBSCAN(GF)	0.000	0.000	1.000	1.0	1.0
	RSF [6]	0.019	0.183	0.774	1.003	0.507
Self Train	Oyster-CP [38]	0.091	0.215	0.784	1.514	0.521
	Oyster-TF [38]	0.093	0.233	0.708	1.564	0.448
	LISO-CP	0.109	0.224	0.750	1.062	0.409
	LISO-TF	**0.134**	**0.270**	**0.628**	0.938	**0.408**

Table 7: Full evaluation results on nuScenes dataset: We compare **LISO** with two different network architectures (TF [2], CP [36]) against different baselines and also give supervised training results as reference (two top rows). Along the AP score we report the nuScenes detection score NDS, which is a combination of the AP score, average translation, orientation, scale, attribute error/score (ATE, AOE, ASE, AEE respectively). All models get a high penalty on the Nuscenes Detection Score (NDS), because they cannot distinguish object classes and therefore score an Average Attribute Error of 1.0. Note that nuScenes uses a minimum precision and recall threshold of 0.1, and since the recall of GT flow clustering is lower than 0.1, all results are clipped away. SF: lidar scene flow by SLIM, GF: ground truth lidar scene flow.

F Qualitative Results

For more qualitative comparisons besides Fig. 10 or Fig. 5, we kindly refer the reader to the video accompanying this supplement.

G Quantitative Results

In Table 8 and Table 7 we have more detailed metrics for WOD and nuScenes. Please note that the models get a high penalty on the Nuscenes Detection Score (NDS), because they cannot distinguish object classes and therefore score an Average Attribute Error of 1.0.

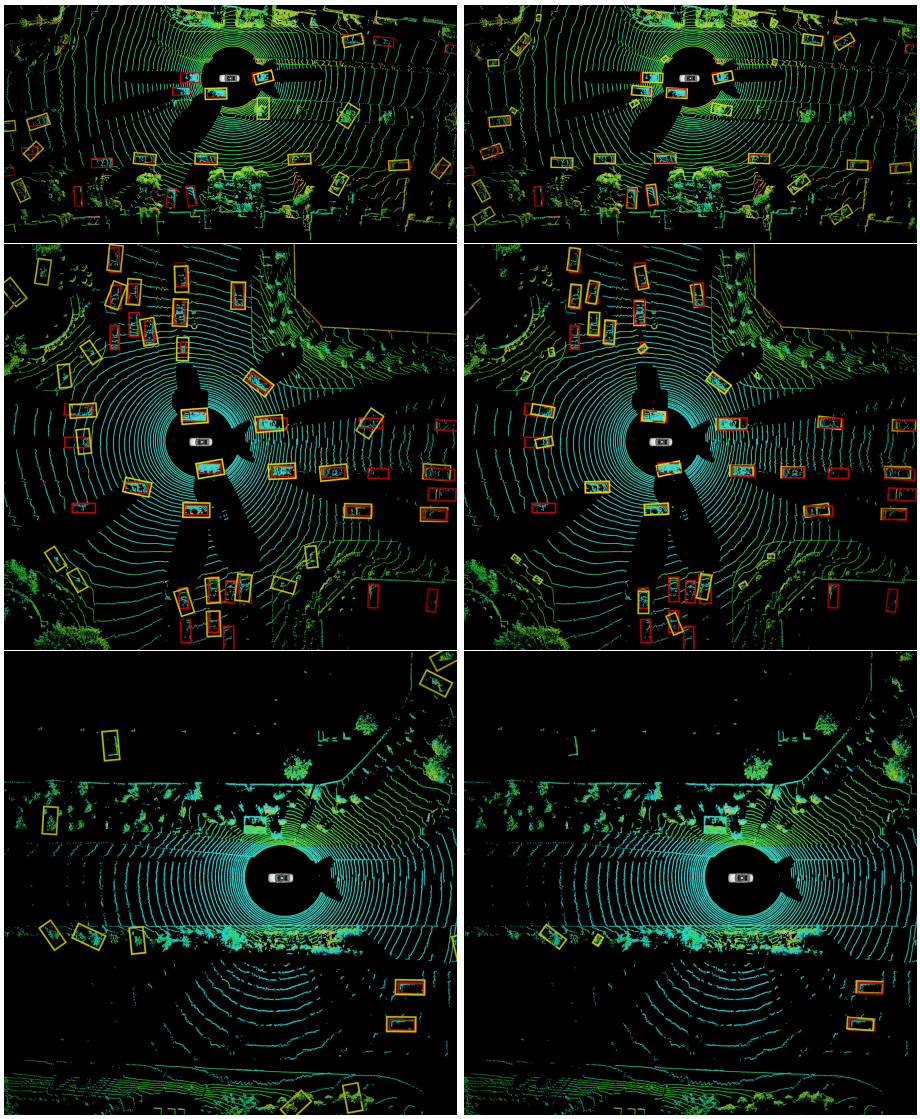

Fig. 10: Qualitative Results on WOD. Red boxes are ground truth boxes, yellow are predictions. **Left:** OYSTER-CP **Right:** LISO-CP Both methods struggle to some extent with false positive detections, but Oyster much more so, despite using the higher confidence threshold. We attribute this to the fact that Oyster has noisier initial pseudo ground truth, which leads to wrong training signals.

		Movable		Moving		Still		Vehicle		Pedestrian		Cyclist	
		AP@0.4		AP@0.4		AP@0.4		AP@0.4		AP@0.4		AP@0.4	
		BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D	BEV	3D
GT	CP [36]	0.765	0.684	0.721	0.624	0.735	0.657	0.912	0.841	0.513	0.413	0.134	0.094
	TF [2]	0.746	0.723	0.714	0.668	0.733	0.710	0.918	0.899	0.457	0.429	0.216	0.187
Unsupervised	DBSCAN [9]	0.027	0.008	0.009	0.000	0.027	0.006	0.184	0.048	0.002	0.000	0.001	0.000
	DBSCAN(SF)	0.026	0.010	0.064	0.041	0.000	0.000	0.073	0.046	0.010	0.006	0.009	0.006
	DBSCAN(GF)	0.114	0.071	0.318	0.120	0.000	0.000	0.113	0.075	0.111	0.063	0.240	0.151
	RSF [6]	0.030	0.020	0.080	0.055	0.000	0.000	0.109	0.074	0.000	0.000	0.002	0.000
	SeMoLi [19] †	-	0.195	-	**0.575**	-	-	-	-	-	-	-	-
	LISO-CP	0.292	0.211	0.272	0.204	0.208	0.140	0.607	0.440	0.029	0.009	0.010	0.004
Self Train	Oyster-CP [38]	0.217	0.084	0.151	0.062	0.176	0.056	0.562	0.204	0.000	0.000	0.000	0.000
	Oyster-TF [38]	0.121	0.015	0.051	0.007	0.098	0.010	0.475	0.058	0.000	0.000	0.000	0.000
	LISO-CP	**0.380**	**0.308**	**0.350**	0.296	**0.322**	**0.255**	**0.695**	**0.543**	**0.055**	**0.037**	**0.022**	**0.016**
	LISO-TF	0.327	0.208	0.349	0.245	0.233	0.126	0.669	0.408	0.024	0.008	0.012	0.005

Table 8: Full evaluation results on WOD dataset: We evaluate using the protocol of [15, 19], using an area of whole 100m×40m BEV grid around the ego vehicle, considering objects that move faster than 1m/s to be *moving* (difficulty level L2). CP, TF: network architecture, in the first two lines trained supervised for comparison. †: Results taken from [19]. SF: lidar scene flow by SLIM, GF: ground truth lidar scene flow.