Supplementary Material: Weakly Supervised Class-agnostic Motion Prediction for Autonomous Driving

A. More Details about Network Architecture

Pre-segmentation Network (PreSegNet). PreSegNet is a foreground/background (FG/BG) segmentation model, which takes a single 2D BEV map as input and predicts its FB/BG category map. The input size is $13 \times 256 \times 256$, where 13 and 256 are the channel and spatial sizes, respectively. The PreSegNet consists of two components: a backbone network and a FG/BG segmentation head. For the backbone network, we adopt the spatio-temporal pyramid network (STPN) proposed in MotionNet [6]. The STPN is designed for spatio-temporal feature extraction. To make it fit for the single frame segmentation in PreSegNet, we remove the 3D convolution in each spatio-temporal convolution block, and the rest architecture remains unchanged. For the FG/BG segmentation head, following MotionNet [6], we implement it using two-layer 2D convolutions with a channel size of 32.

Motion Prediction Network (WeakMotionNet) WeakMotionNet is a motion prediction network, which takes a sequence of synchronized BEV maps as input and predicts the future motion map and the FB/BG category map. Same as that of previous works [5, 6], the size of the input maps is $5 \times 13 \times 256 \times 256$, where 5, 13, and 256 are the temporal, channel, and spatial sizes, respectively. Specifically, the WeakMotionNet contains a backbone network, a motion prediction head, and an auxiliary FG/BG segmentation head. We implement the backbone network with the STPN proposed in MotionNet [6] and implement the two output heads using two-layer 2D convolutions with a channel size of 32.

B. More Experimental Details

The licenses of the assets used in our paper are shown in Table A.

More details about experiments on nuScenes. In the first part of Sec. 5 in our main paper, we introduce the nuScenes dataset [1] and the implementation details of our experiments on it. When generating foreground and background ground truth, we use the per-point semantic labels provided by nuScenes-lidarseg. In nuScenes-lidarseg, there are a total of 32 labels, 23 of which are the foreground classes (pedestrians, vehicles, cyclists, etc.) and the remaining 9 labels are background classes (nature, buildings, etc.). In our experiments, we directly merge the 23 foreground labels into a foreground category and merge the rest 9 background labels into a background category to generate per-point FG/BG ground truth.

More details about experiments on Waymo. In Sec. 5.1 of our main paper, we apply our method to Waymo Open Dataset [4] for further evaluation. Same to the motion data generation strategy for nuScenes, motion ground truth for Waymo is also bootstrapped from human-annotated object detection and tracking data. In Waymo, object boxes are dived into 4

Table A. Licenses of the assets used in our paper.

Assets	License websites
PyTorch [3]	https://github.com/pytorch/pytorch/blob/master/LICENSE
MotionNet [6]	https://github.com/pxiangwu/MotionNet
PillarMotion [2]	https://github.com/qcraftai/pillar-motion/blob/main/LICENSE.md
nuScenes [1]	https://www.nuscenes.org/terms-of-use-commercial
Waymo [4]	https://github.com/waymo-research/waymo-open-dataset/blob/master/LICENSE

Table B. Ablation study for Consistency-aware Chamfer Distance (CCD) loss under the FG/BG annotation ratio of 0.1%.

Loss function in WeakMotionNet	L2-norm	L1-norm	Future Frame	Past Frame	Confidence Reweight	Auxiliary FG/BG Segmentation	Static	Speed $\leq$ 5m/s	Speed $>$ 5m/s
							Mean Error $\downarrow$		
Chamfer loss (**Baseline**)	✓		✓				0.5522	0.7566	2.3569
Chamfer-L1		✓	✓				0.3480 (−37%)	0.5343 (−29%)	2.1475 (−9%)
Multi-frame Chamfer-L1		✓	✓	✓			0.3472 (−37%)	0.5443 (−28%)	1.8176 (−23%)
Consistency-aware Chamfer		✓	✓	✓	✓		0.2201 (−60%)	0.4782 (−37%)	**1.7510 (−26%)**
Consistency-aware Chamfer + Seg. (**Ours, 0.1%**)		✓	✓	✓	✓	✓	**0.0426 (−92%)**	**0.4009 (−47%)**	2.1342 (−9%)

Table C. Ablation study for Consistency-aware Chamfer Distance (CCD) loss under the FG/BG annotation ratio of 100%.

Loss function in WeakMotionNet	L2-norm	L1-norm	Future Frame	Past Frame	Confidence Reweight	Auxiliary FG/BG Segmentation	Static	Speed $\leq$ 5m/s	Speed $>$ 5m/s
							Mean Error $\downarrow$		
Chamfer loss (**Baseline**)	✓		✓				0.0900	0.4298	2.3898
Chamfer-L1		✓	✓				0.0437 (−51%)	0.3439 (−20%)	1.9370 (−19%)
Multi-frame Chamfer-L1		✓	✓	✓			0.0362 (−60%)	0.3267 (−24%)	**1.5621 (−35%)**
Consistency-aware Chamfer		✓	✓	✓	✓		0.0288 (−68%)	**0.3032 (−29%)**	1.6827 (−30%)
Consistency-aware Chamfer + Seg. (**Ours, 100%**)		✓	✓	✓	✓	✓	**0.0243 (−73%)**	0.3316 (−23%)	1.6422 (−31%)

categories: vehicle, pedestrian, cyclist, and sign. Accordingly, we treat the points belonging to the object boxes of vehicles, pedestrians, and cyclists as the foreground and the rest points in the scene as the background.

C. More Analysis and Visualization

C.1. More Analysis for Consistency-aware Chamfer loss

For robust self-supervised motion learning, we propose a Consistency-aware Chamfer distance (CCD) loss with L1-norm as distance metric, multi-frame point clouds for supervision, and multi-frame consistency for reweighting. In Sec. 5.2 of our main paper, we conduct an ablation study to analyze the effectiveness of the proposed CCD loss. The experiments in Table 3 of our main paper are performed under 1% FG/BG annotation. Here, we also analyze the performance of our CCD loss under the annotation ratio of 0.1% and 100%. The results are presented in Table B and Table C, respectively.

Under 0.1% and 100% annotation, compared with the Chamfer loss, our Chamfer-L1 loss with L1-norm as distance metric reduces the prediction error on the three groups by a large margin. By adding point clouds from the past frame as part of the target data, our multi-frame Chamfer-L1 loss further decreases the error of the fast speed group by around 16% under the two annotation ratios. Additionally, by using multi-frame consistency to estimate confidence and reweight, we also observe a significant reduction in error for static and slow groups. The above observations indicate that, under the FG/BG annotation ratio of 0.1% and 100%, our CCD loss still outperforms the Chamfer loss.

Moreover, to regularize the predicted motion, we add an auxiliary FG/BG segmentation head for WeakMotionNet and set the motion of predicted background areas to zero. As presented in Table B and Table C, by combining our Consistency-aware Chamfer loss with a FG/BG segmentation loss, we observe a significantly lower error in the static group, which is consistent with the performance under 1% annotation. However, as shown in Table B, under 0.1% annotation, using FG/BG predictions to regularize the predicted motion increases the error of the fast moving group from about 1.7m to about 2.1m. This is because the FG/BG predictions from the auxiliary FG/BG segmentation head are inaccurate. As presented in the Table 2 of our main paper, when using 0.1% FG/BG masks as weak supervision, the accuracy of predicted foreground areas is only 83.5%, which means that a lot of foreground points are wrongly treated as the background, and their motion is set to zero. The poor accuracy of FG/BG segmentation head is due to the foreground/background imbalance (about 1:10 in nuScenes) and the tiny annotation ratio (0.1% FG/BG masks).

C.2. More Qualitative Results on nuScenes

In Sec. 5.1 of our main paper, we train models by our weakly supervised motion prediction approach on nuScenes dataset. Fig. A presents more visualization results of our models under the annotation ratios of 100%, 1%, and 0.1%.

C.3. Qualitative Results of PreSegNet on nuScenes

In our two-stage weakly supervised approach, a FG/BG segmentation network, PreSegNet, in Stage1 is trained with incomplete masks and further generates dense FG/BG masks for the self-supervised motion learning of WeakMotionNet in

Table D. Results of FG/BG segmentation on Waymo Dataset.

Method	FG Acc. ↑	BG Acc. ↑	Overall Acc. ↑
Ours (0.1%)	94.5%	97.5%	97.3%
Ours (1.0%)	96.2%	**97.7%**	**97.6%**
Ours (100%)	**98.0%**	96.5%	96.6%

Stage2. Fig. B presents some qualitative results of PreSegNet on nuScenes validation set.

C.4. More Visualization Examples for PreSegNet and CCD loss

To study the effectiveness of the CCD loss, we present more visualization examples in Fig. C. In our weakly supervised motion prediction, outliers may be caused by occlusion of points (e.g., region A, C, and E) and inaccurate foreground predictions from PreSegNet (e.g., region B, D, and F). Specifically, in regions C and E, points in the future frame are occluded, and in region A, points in the past and current frames are occluded. By using multi-frame consistency as the confidence to reweight each data point, our CCD loss successfully reduces the number of outliers in the six regions, resulting in robust motion learning.

C.5. More Results on Waymo

In Sec. 5.1 of our main paper, we apply our weakly supervised approach to Waymo Open Dataset [4] for further evaluation. The motion prediction results of our weakly supervised models are presented in Table 7 of our main paper, and their FG/BG segmentation results are presented in Table D. Our models trained by 100%, 1%, and 0.1% FG/BG masks can reach an overall accuracy of over 95%. Qualitative results are shown in Fig. D.

D. More Discussions about Limitations

In our two-stage training approach, the self-supervised motion learning in Stage2 relies on the FG/BG segmentation generated from Stage1, which makes inaccurate segmentation likely to hinder motion learning. Although our CCD loss can mitigate the outliers caused by inaccurate segmentation, a good segmentation model is still vital. Therefore, more effective training of FG/BG segmentation model using incomplete masks is worth investigating in our future work. In addition, large displacements of objects are a long-standing challenge in motion tasks. Our CCD loss generates motion supervision by matching the warped points with the target. Large displacements will make the matching difficult. Therefore, handling large displacements is another focus of our future work.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2021. 1

[3] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[4] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 3

[5] Yunlong Wang, Hongyu Pan, Jun Zhu, Yu-Huan Wu, Xin Zhan, Kun Jiang, and Diange Yang. Be-sti: Spatial-temporal integrated network for class-agnostic motion prediction with bidirectional enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17093–17102, 2022. 1

[6] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020. 1

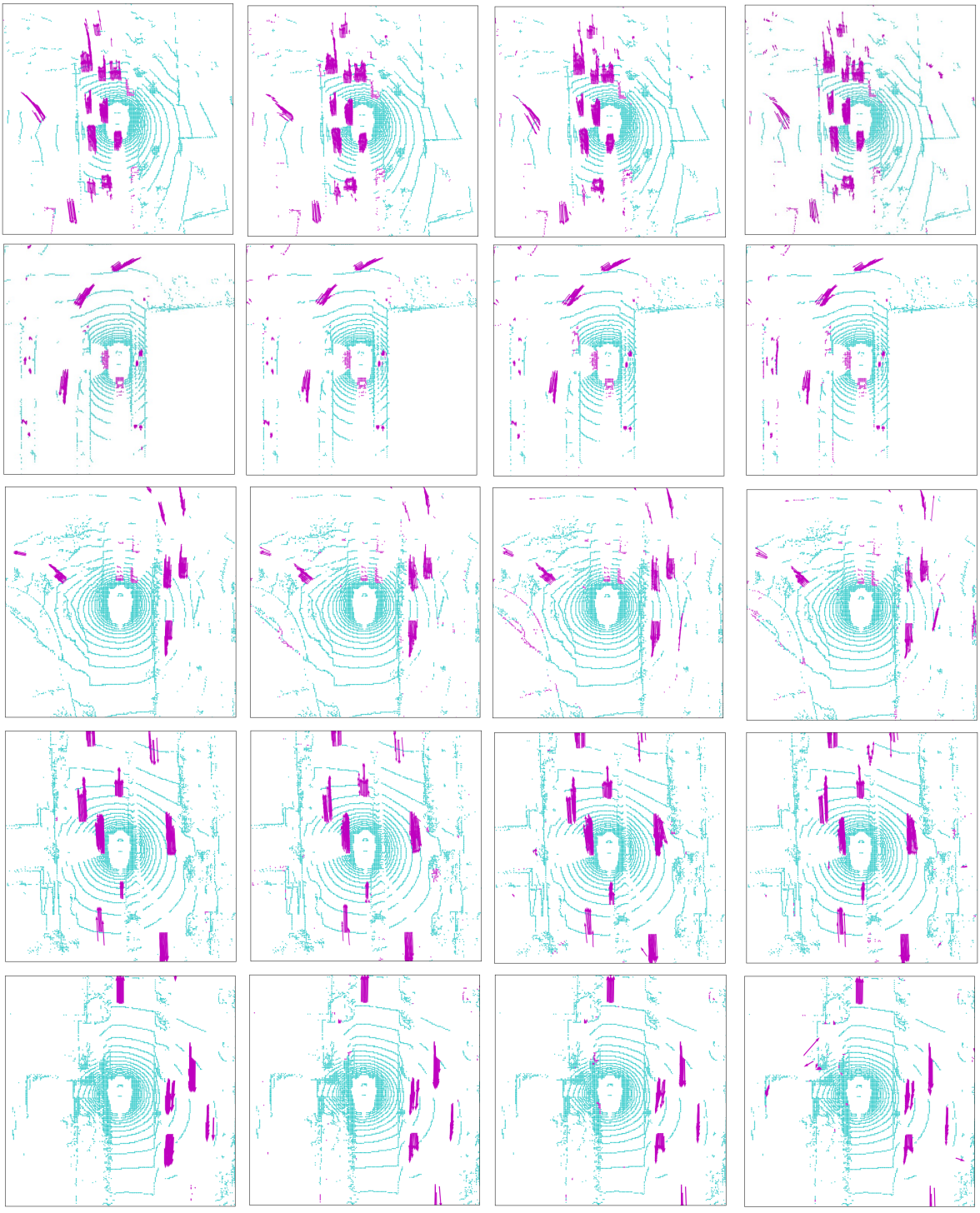
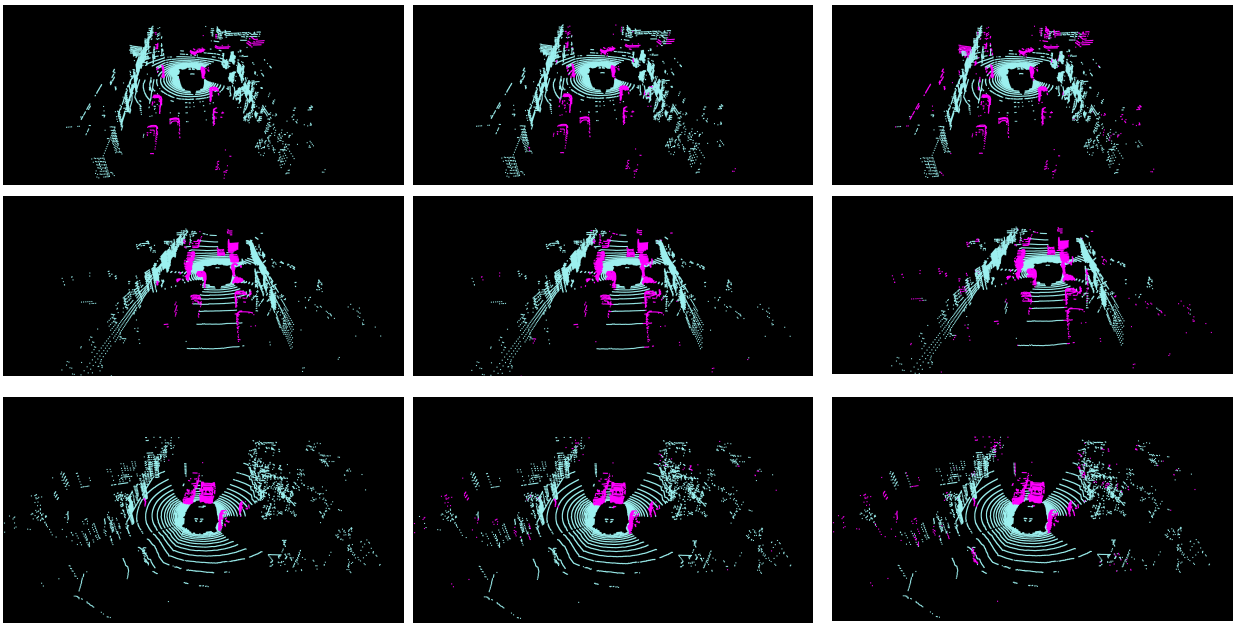


Figure A. Qualitative results of motion prediction and foreground/background segmentation on nuScenes. (a) Ground-truth. (b) Results of our method trained by 100% annotated FG/BG masks. (c) Results of our method trained by 1% annotated masks. (d) Results of our method trained by 0.1% annotated masks. We show motion with an arrow attached to each cell and represent different category with different color. **Purple**: Foreground; **Cyan**: Background.

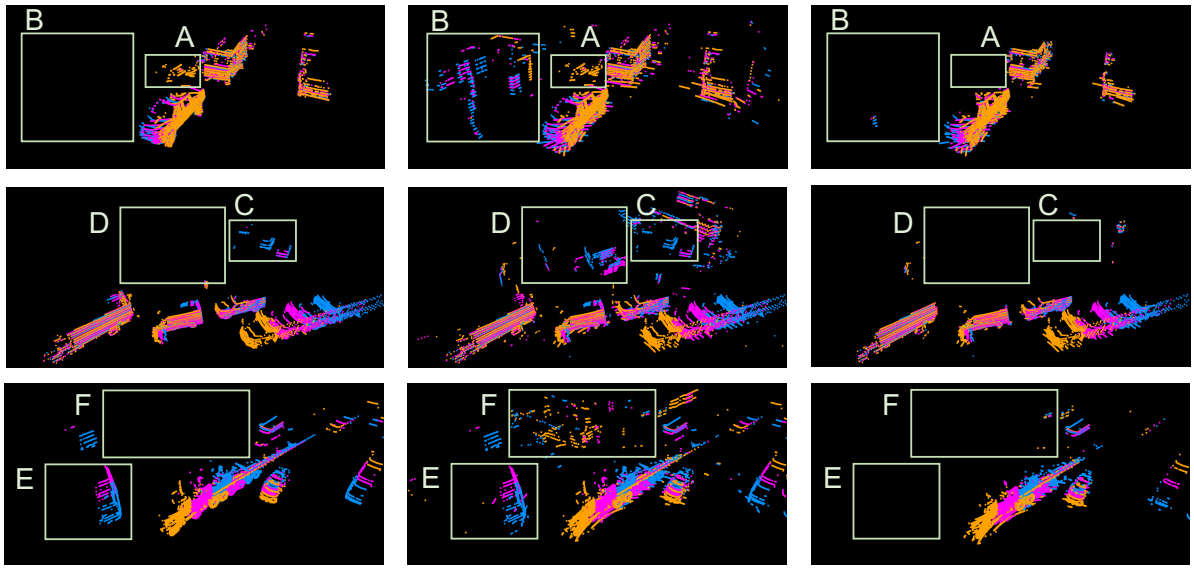


(a) FG/BG Ground Truth

(b) FG/BG Predictions (PreSegNet trained by 1% masks)

(c) FG/BG Predictions (PreSegNet trained by 0.1% masks)

Figure B. Qualitative results of foreground/background segmentation produced by PreSegNet in Stage1 on nuScenes validation set. (a) FG/BG Ground-truth. (b) Results of our PreSegNet trained by 1% annotated FG/BG masks. (c) Results of our PreSegNet trained by 0.1% annotated FG/BG masks. **Purple**: Foreground; **Cyan**: Background.



(a) Ground truth foreground points (b) Predicted foreground points from PreSegNet (0.1%) (c) Reweighted foreground points by CCD loss

Figure C. Visualization for PreSegNet and CCD loss. Outliers may be due to occlusions of points (e.g., region A, C, and E) and inaccurate foreground predictions from PreSegNet (e.g., region B, D, and F). In our CCD loss, we use multi-frame consistency to measure the confidence of points and assign uncertain points fewer weights, thereby suppressing potential outliers. For better visualization, we remove points with lower weights in (c). Different color represents point cloud in different frames. **Blue**: past frame; **Purple**: current frame; **Orange**: future frame.

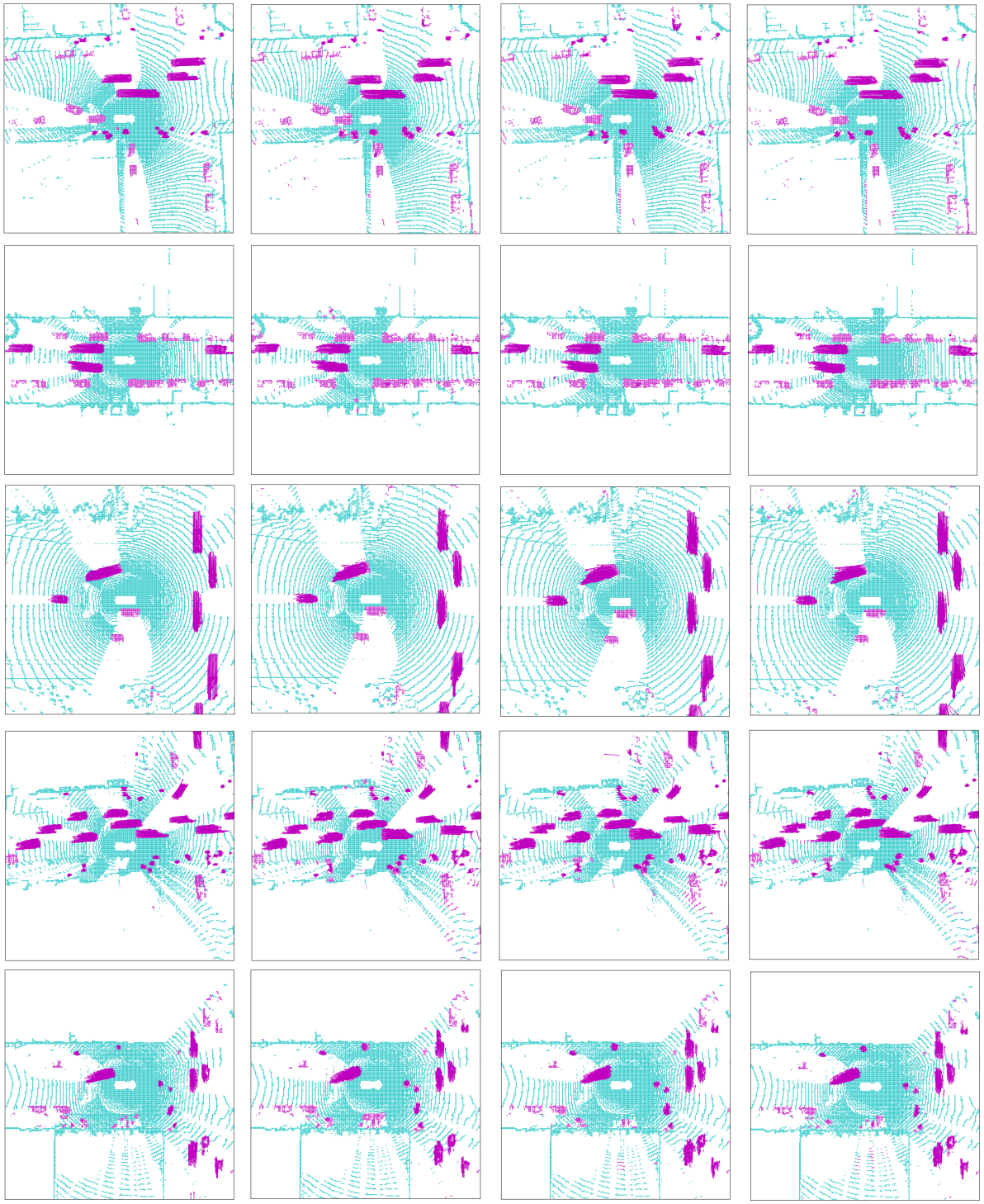


Figure D. Qualitative results of motion prediction and foreground/background segmentation on Waymo. (a) Ground-truth. (b) Results of our method trained by 100% annotated FG/BG masks. (c) Results of our method trained by 1% annotated masks. (d) Results of our method trained by 0.1% annotated masks. We show motion with an arrow attached to each cell and represent different category with different color. **Purple**: Foreground; **Cyan**: Background.