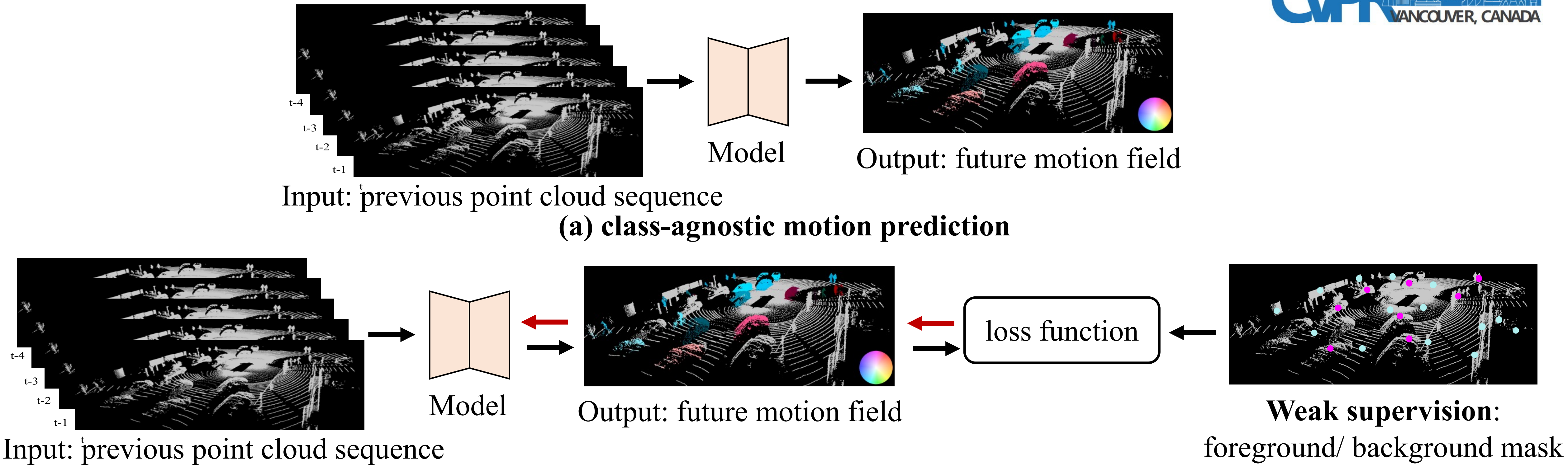
Weakly Supervised Class-agnostic Motion Prediction for Autonomous Driving

Ruibo Li[1,2], Hanyu Shi[2], Ziang Fu[3], Zhe Wang[3], Guosheng Lin[1,2]

[1]S-Lab for Advanced Intelligence, Nanyang Technological University
[2]School of Computer Science and Engineering, Nanyang Technological University
[3]SenseTime Research

Paper tag : THU-AM-107

Overview



(a) class-agnostic motion prediction

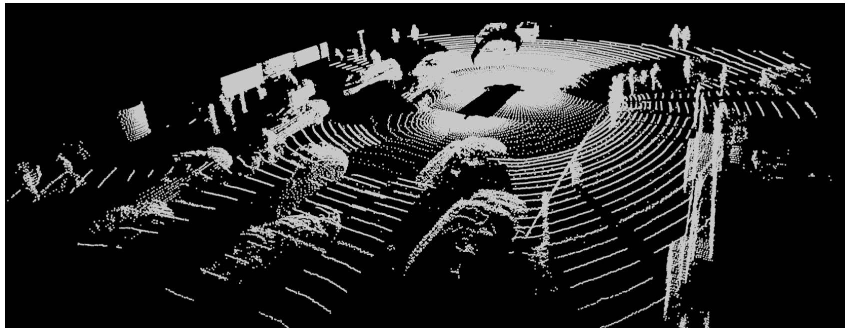
(b) Our: weakly supervised class-agnostic motion prediction

- A novel weakly supervised motion prediction paradigm with fully or partially annotated foreground/background (FB/BG) masks as supervision
- A two-stage weakly supervised motion prediction approach, where FG/BG segmentation from Stage1 will facilitate the self-supervised motion learning in Stage2.
- A novel Consistency-aware Chamfer Distance loss, where multi-frame information is used to suppress potential outliers for robust self-supervised motion learning.

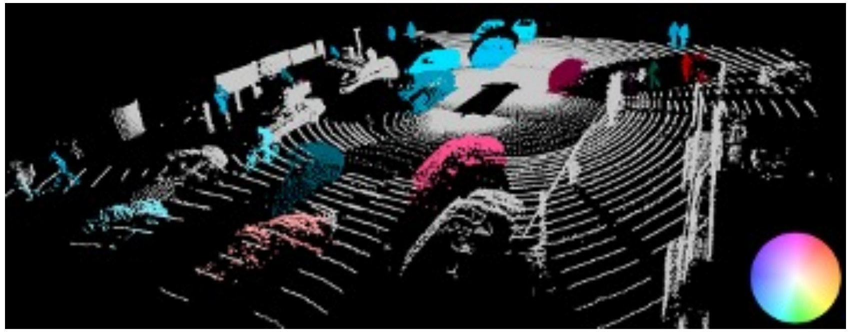
Motivation

- Ground truth motion data is scarce and expensive.

- There is still a large performance gap between self-supervised methods and fully supervised methods.



(a) Point cloud

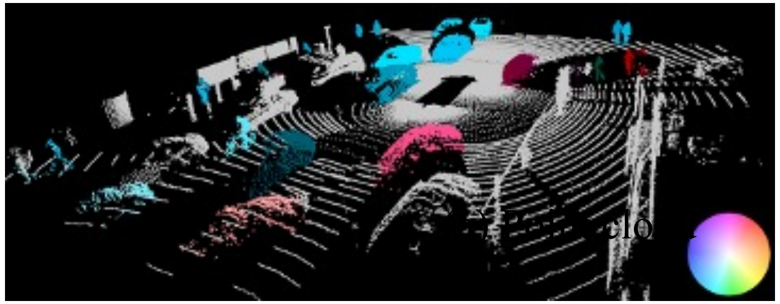


(b) Ground truth motion data
(moving points are colored by their motion, static points are **Gray**)

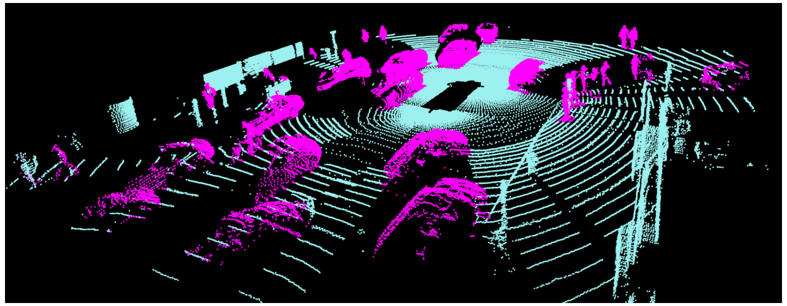
Motivation

- Outdoor scenes can often be decomposed into mobile foregrounds and static backgrounds, which enables us to associate motion understanding with scene parsing.

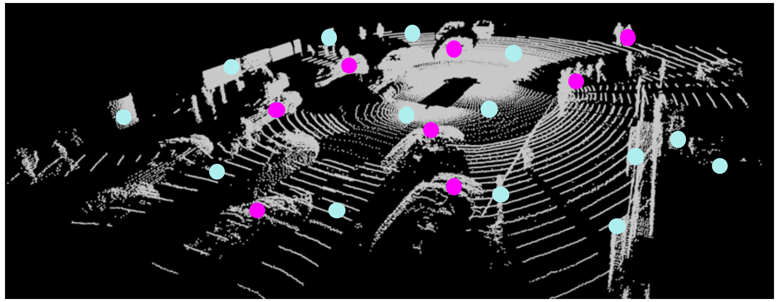
- We study a novel weakly supervised motion prediction paradigm, where fully or partially (1%, 0.1%) annotated foreground/background binary masks are used for supervision.



(b) Ground truth motion data
(moving points are colored by their motion, static points are **Gray**)

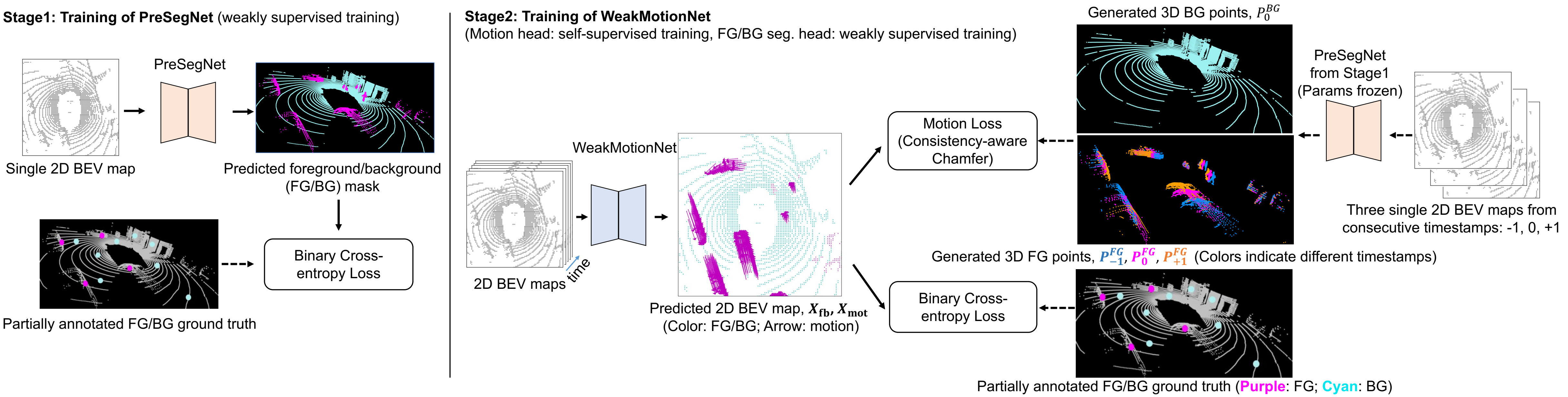


(c) Fully annotated
Foreground/Background masks
(**Purple**: FG; **Cyan**: BG)



(e) Partially annotated
Foreground/ Background masks

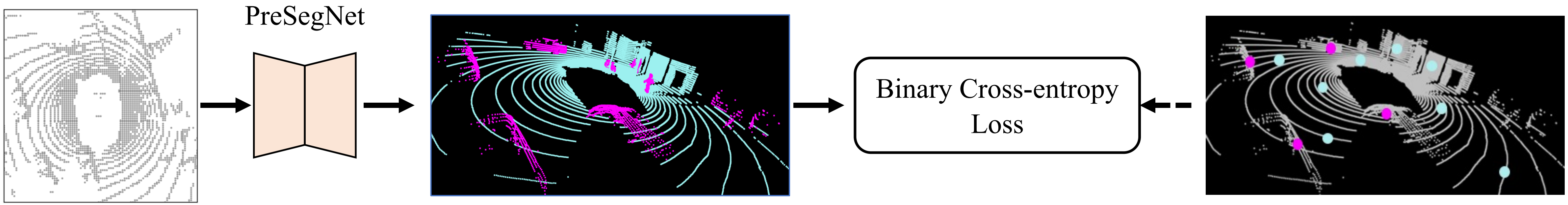
Method: two-stage weakly supervised motion prediction approach



- We propose a two-stage weakly supervised approach, where the segmentation model trained with the incomplete binary masks in Stage1 will facilitate the self-supervised learning of the motion prediction network in Stage2 by estimating possible moving foregrounds in advance.

Method: two-stage weakly supervised motion prediction approach

Stage1: Training of PreSegNet (weakly supervised training for FB/BG segmentation)



Input:
Single 2D BEV map

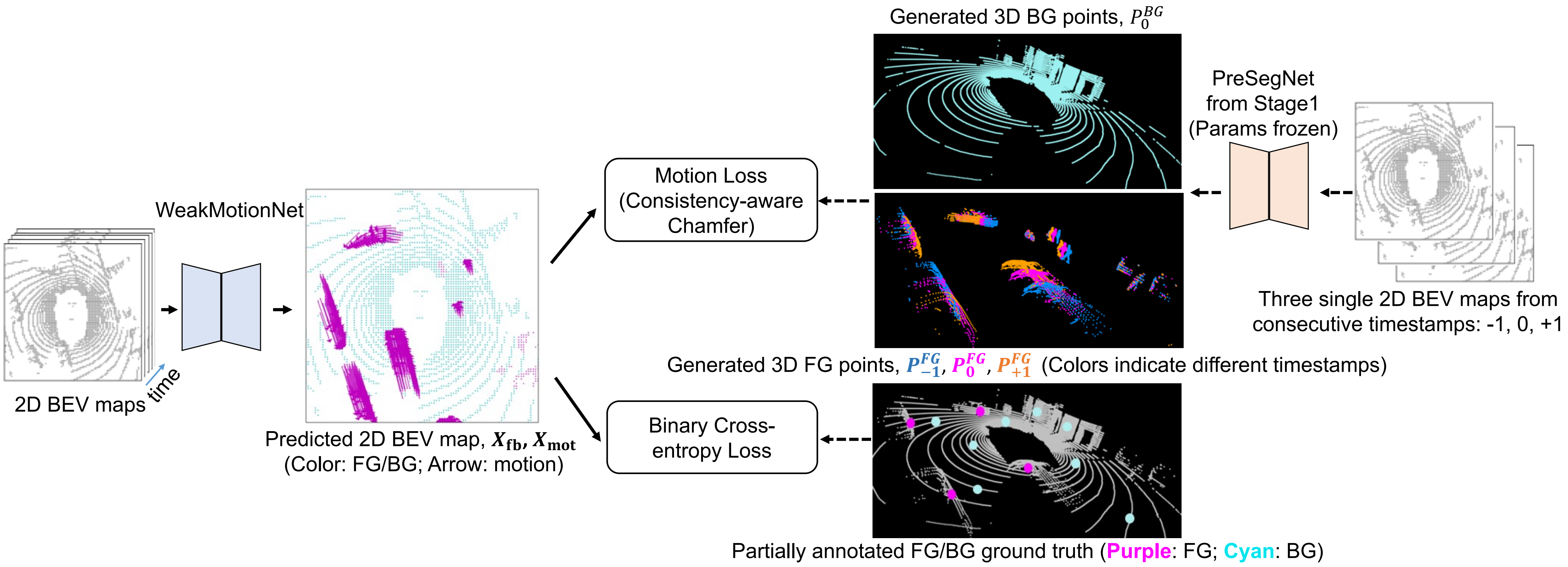
Output:
Foreground/background (FG/BG) mask

Supervision:
Partially annotated FG/BG ground truth

Method: two-stage weakly supervised motion prediction approach

Stage2: Training of Training of WeakMotionNet

(Motion head: self-supervised training, FG/BG seg. head: weakly supervised training)



Method: two-stage weakly supervised motion prediction approach

Consistency-aware Chamfer loss for self-supervised motion learning in Stage2

Consistency-aware Chamfer (CCD) loss:

$$\mathcal{L}_{CCD}(\boldsymbol{P}_{-1}, \boldsymbol{P}_0, \boldsymbol{P}_{+1}, \boldsymbol{F}) = \mathcal{L}_{SCCD}(\widehat{\boldsymbol{P}}_{0,b}, \boldsymbol{P}_{-1}, \boldsymbol{w}_0, \boldsymbol{w}_{-1}) + \mathcal{L}_{SCCD}(\widehat{\boldsymbol{P}}_{0,f}, \boldsymbol{P}_{+1}, \boldsymbol{w}_0, \boldsymbol{w}_{+1})$$

$\boldsymbol{P}_{-1}, \boldsymbol{P}_0, \boldsymbol{P}_{+1}$: point clouds from the past (-1), current (0) and future (+1) frames

$\boldsymbol{F}$: predicted motion from current frame (0) to future (+1) frame

$\widehat{\boldsymbol{P}}_{0,f} = \boldsymbol{P}_0 + \boldsymbol{F}, \quad \widehat{\boldsymbol{P}}_{0,b} = \boldsymbol{P}_0 - \boldsymbol{F}$: forward and backward warped point cloud

$\boldsymbol{w}_{-1}, \boldsymbol{w}_0, \boldsymbol{w}_{+1}$: confidence weights for the three point clouds

Each term in (CCD) loss:

$$\mathcal{L}_{SCCD}(\widehat{\boldsymbol{P}}_{0,f}, \boldsymbol{P}_{+1}, \boldsymbol{w}_0, \boldsymbol{w}_{+1}) = \frac{1}{\|\boldsymbol{w}_0\|_1} \sum_{i=1}^{N_0} w_0(i) \min_{\boldsymbol{s} \in \boldsymbol{P}_{+1}} \|\widehat{\boldsymbol{p}}_{0,f}(i) - \boldsymbol{s}\|_1 + \frac{1}{\|\boldsymbol{w}_{+1}\|_1} \sum_{j=1}^{N_{+1}} w_{+1}(j) \min_{\boldsymbol{s} \in \widehat{\boldsymbol{P}}_{0,f}} \|\boldsymbol{p}_{+1}(j) - \boldsymbol{s}\|_1$$

Improvements of CCD loss compared to typical Chamfer distance loss:

(1) exploit supervision from multi-frame point clouds

(2) employ multi-frame consistency to measure the confidence of points and assign uncertain points fewer weights to suppress potential outliers.

(3) adopt L1-norm to calculate the distance

Experiments: Comparison with State-of-the-Art Methods

Table1: Evaluation results of motion prediction on nuScenes test set

Method	Supervision	Modality	Static		Speed $\leq$ 5m/s		Speed $>$ 5m/s	
			Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓
FlowNet3D [25]	Full.	LiDAR	0.0410	0	0.8183	0.1782	8.5261	8.0230
HPLFlowNet [14]	Full.	LiDAR	0.0041	0.0002	0.4458	0.0960	4.3206	2.4881
PointRCNN [34]	Full.	LiDAR	0.0204	0	0.5514	0.1627	3.9888	1.6252
LSTM-ED [33]	Full.	LiDAR	0.0358	0	0.3551	0.1044	1.5885	1.0003
PillarMotion [26]	Full.	LiDAR+Image	0.0245	0	0.2286	0.0930	0.7784	0.4685
MotionNet [42]	Full.	LiDAR	0.0201	0	0.2292	0.0952	0.9454	0.6180
BE-STI [41]	Full.	LiDAR	0.0220	0	0.2115	0.0929	0.7511	0.5413
PillarMotion [26]	Self.	LiDAR+Image	0.1620	0.0010	0.6972	0.1758	3.5504	2.0844
Ours (0.1%)	Weak. (0.1% FG/BG masks)	LiDAR	<u>0.0426</u>	0	<u>0.4009</u>	<u>0.1195</u>	2.1342	1.2061
Ours (1%)	Weak. (1% FG/BG masks)	LiDAR	0.0558	0	0.4337	0.1305	<u>1.7823</u>	<u>1.0887</u>
Ours (100%)	Weak. (100% FG/BG masks)	LiDAR	**0.0243**	**0**	**0.3316**	**0.1201**	**1.6422**	**1.0319**

Table2: Motion prediction results on Waymo Dataset

Method	Supervision	Static	Speed $\leq$ 5m/s	Speed $>$ 5m/s
MotionNet [42]	Full.	0.0263	0.2620	0.9493
Ours (0.1%)	Weak.(0.1% FG/BG masks)	<u>0.0297</u>	0.3581	<u>1.6362</u>
Ours (1.0%)	Weak.(1.0% FG/BG masks)	0.0334	<u>0.3458</u>	**1.5655**
Ours (100%)	Weak.(100% FG/BG masks)	**0.0219**	**0.3385**	1.6576

Experiments: Ablation Studies

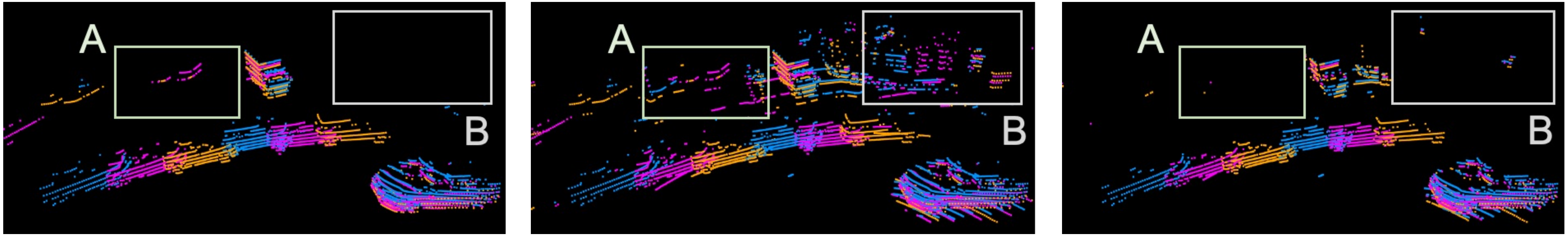
Table3: Effectiveness of two-stage training framework

Method	Static	Speed $\leq$ 5m/s	Speed > 5m/s
1% masks w/o Stage1	1.1976	3.1904	8.9025
1% masks with Stage1 (Ours)	**0.0558**	**0.4337**	**1.7823**

Table4: Ablation study for Consistency-aware Chamfer Distance (CCD) loss under the FG/BG annotation ratio of 1%.

Loss function in WeakMotionNet	L2-norm	L1-norm	Future Frame	Past Frame	Confidence Reweight	Auxiliary FG/BG Segmentation	Static	Speed $\leq$ 5m/s Mean Error ↓	Speed > 5m/s
Chamfer loss (**Baseline**)	✓		✓				0.4416	0.8087	2.3981
Chamfer-L1		✓	✓				0.2579 (−42%)	0.5110 (−37%)	2.1229 (−11%)
Multi-frame Chamfer-L1		✓	✓	✓			0.2677 (−39%)	0.5240 (−35%)	**1.7436 (−27%)**
Consistency-aware Chamfer		✓	✓	✓	✓		<u>0.1469</u> (−67%)	<u>0.4390</u> (−46%)	<u>1.7729</u> (−26%)
Consistency-aware Chamfer + Seg. (**Ours, 1%**)		✓	✓	✓	✓	✓	**0.0558 (−87%)**	**0.4337 (−46%)**	1.7823 (−26%)

Experiments: Ablation Studies



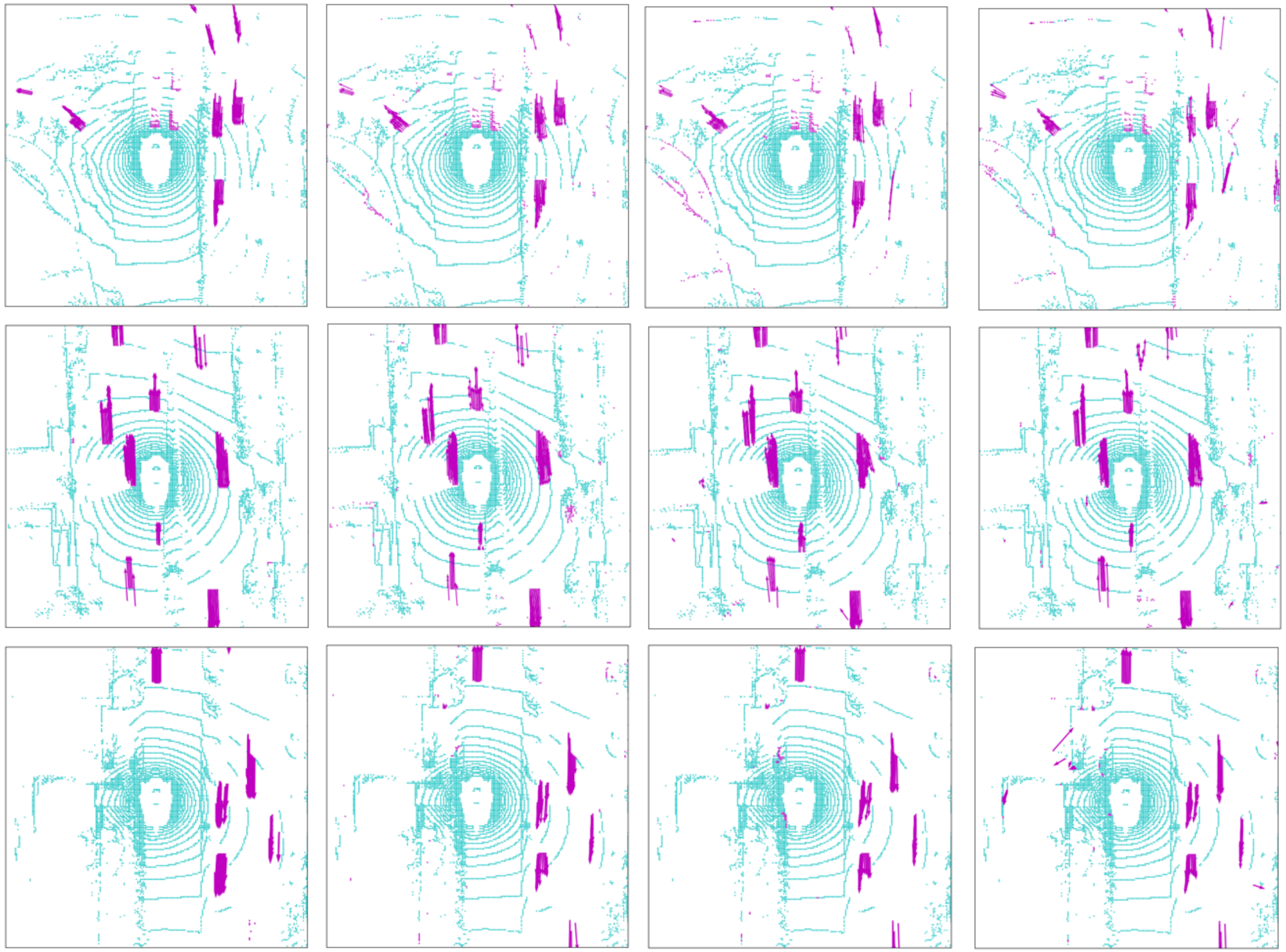
(a) Ground truth foreground points

(b) Predicted foreground points from PreSegNet (0.1%)

(c) Reweighted foreground points by CCD loss

Figure 1. Visualization for PreSegNet and CCD loss. Outliers may be due to occlusions of points (e.g., region A), and inaccurate foreground predictions from PreSegNet (e.g., region B). 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.

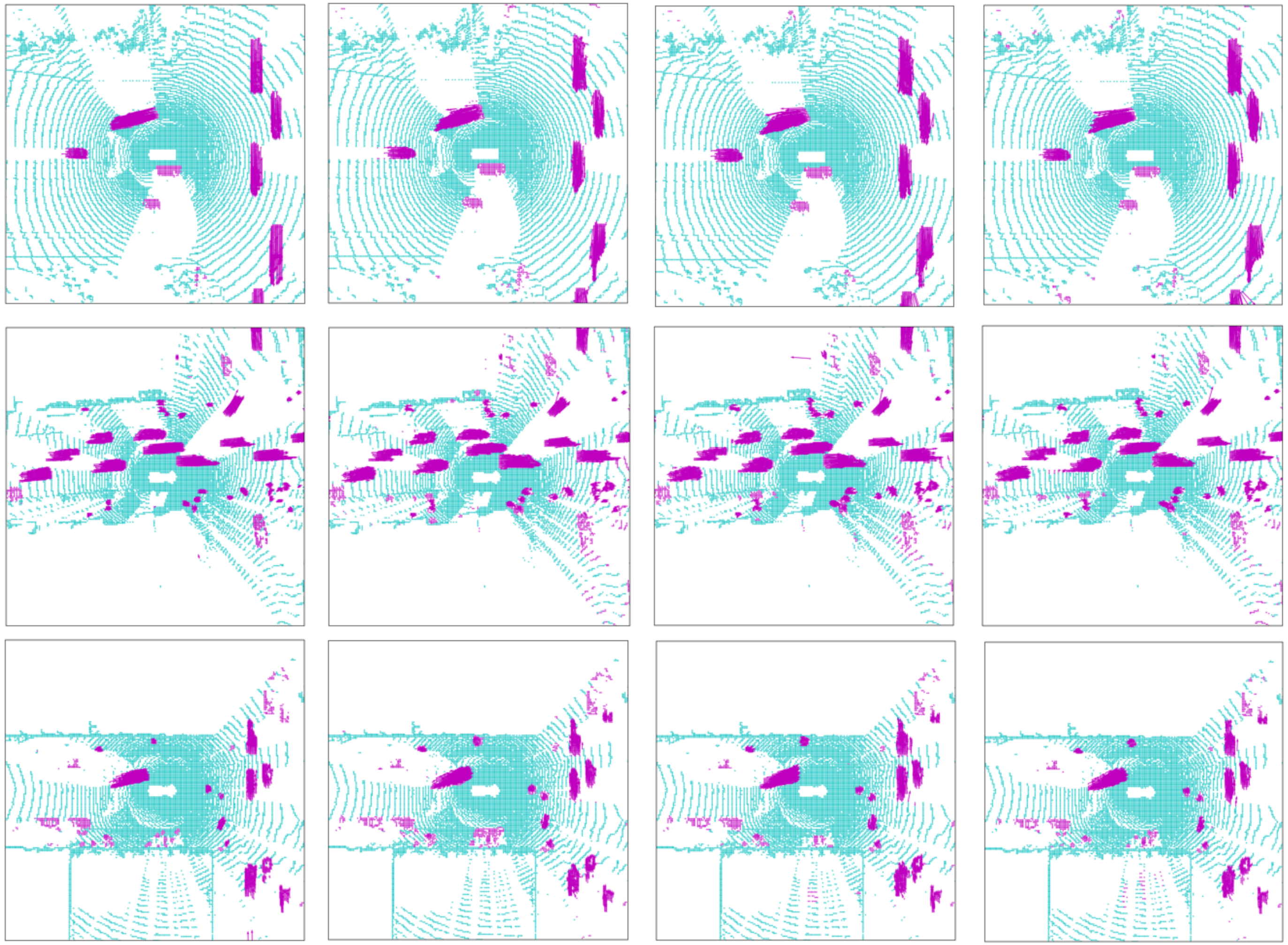
Experiments: Qualitative Results



(a) Ground Truth
(b) Predictions (Model trained by 100% masks)
(c) Predictions (Model trained by 1% masks)
(d) Predictions (Model trained by 0.1% masks)

Figure 2. Qualitative results of motion prediction and foreground/background segmentation on nuScenes. We show motion with an arrow attached to each cell and represent different category with different color. **Purple**: Foreground; **Cyan**: Background.

Experiments: Qualitative Results



(a) Ground Truth
(b) Predictions (Model trained by 100% masks)
(c) Predictions (Model trained by 1% masks)
(d) Predictions (Model trained by 0.1% masks)

Figure 3. Qualitative results of motion prediction and foreground/background segmentation on Waymo. We show motion with an arrow attached to each cell and represent different category with different color. **Purple**: Foreground; **Cyan**: Background.

Thanks for your attention!