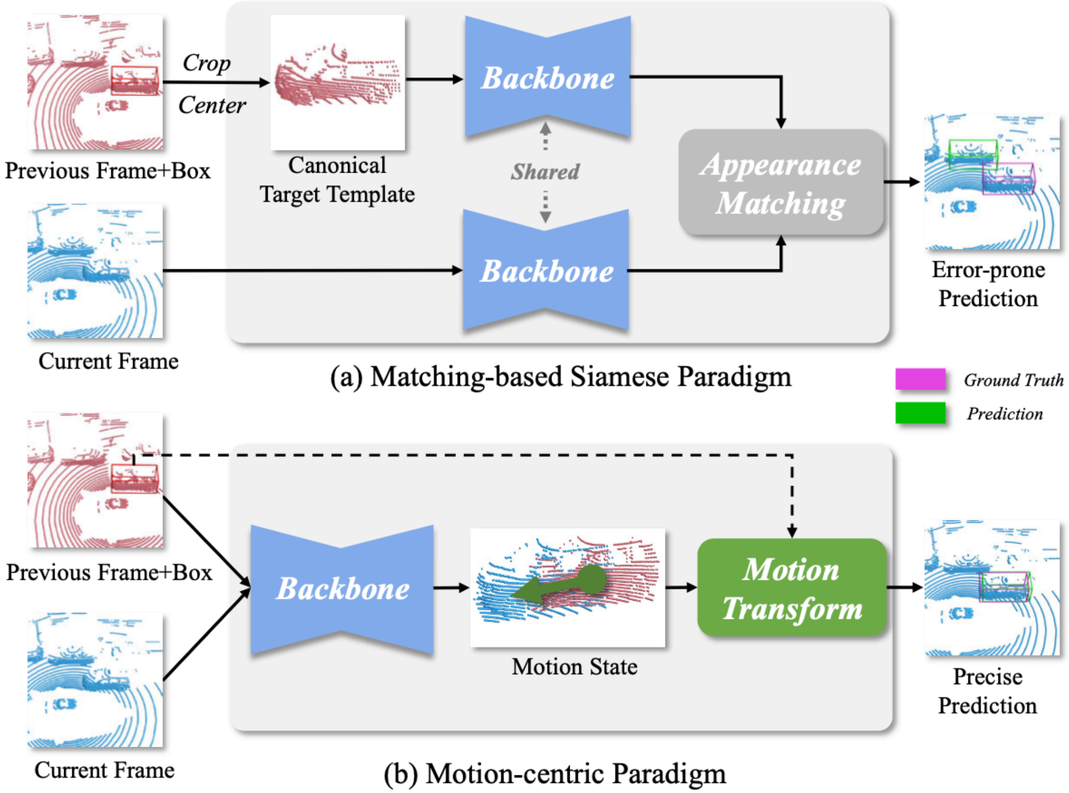
M²Track CVPR'22 Oral

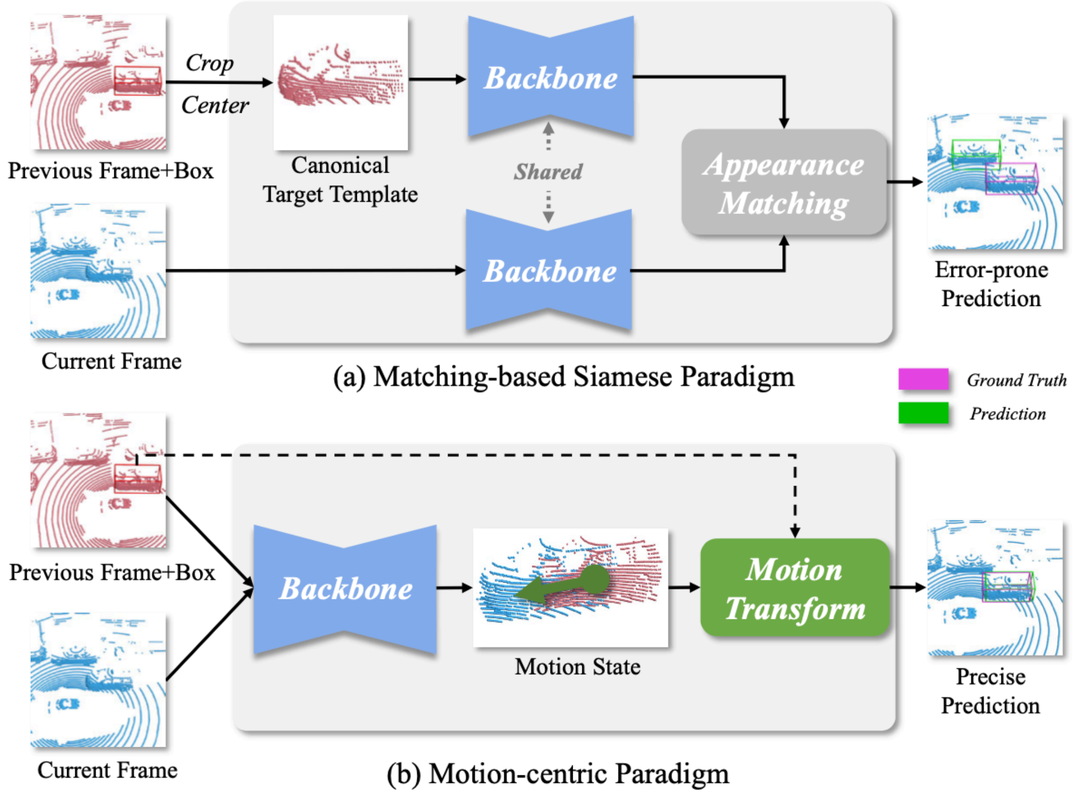
Beyond 3D Siamese Tracking: A Motion-Centric Paradigm for 3D Single Object Tracking in Point Clouds

Chaoda Zheng[123], Xu Yan[123], Haiming Zhang[123],
Baoyuan Wang[4], Shenghui Cheng[5], Shuguang Cui[123], Zhen Li[123*]
[1]The Chinese University of Hong Kong (Shenzhen) [2]The future network of intelligence institute (FNII)
[3]Shenzhen Research Institute of Big Data [4]Xiaobing.AI [5]Westlake University
{chaodazheng@link.,xuyan1@link., haimingzhang@link., lizhen@}cuhk.edu.cn



Beyond 3D Siamese Tracking: A Motion-Centric Paradigm for 3D Single Object Tracking in Point Clouds

Chaoda Zheng[123], Xu Yan[123], Haiming Zhang[123],
Baoyuan Wang[4], Shenghui Cheng[5], Shuguang Cui[123], Zhen Li[123*]
[1]The Chinese University of Hong Kong (Shenzhen) [2]The future network of intelligence institute (FNII)
[3]Shenzhen Research Institute of Big Data [4]Xiaobing.AI [5]Westlake University
{chaodazheng@link.,xuyan1@link., haimingzhang@link., lizhen@}cuhk.edu.cn

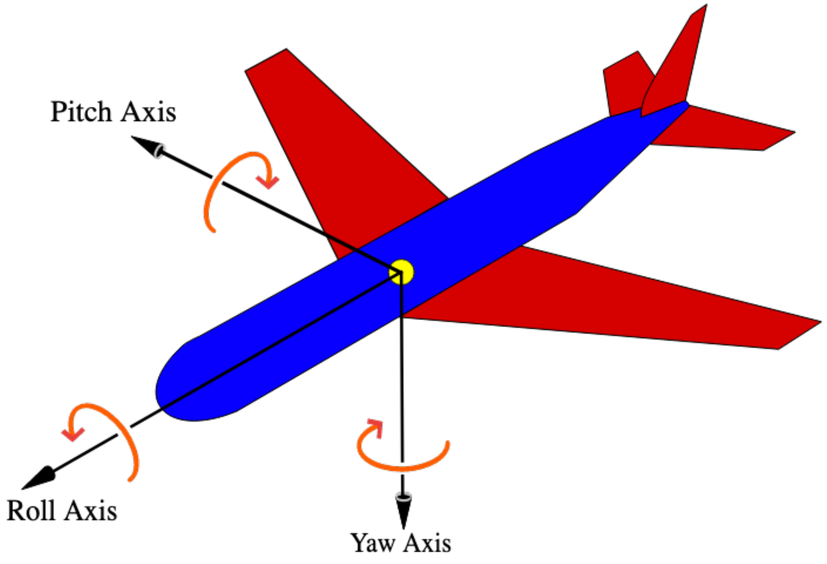


$$\mathcal{F}(\mathcal{P}_t, \mathcal{P}_{t-1}, \mathcal{B}_{t-1}) \mapsto (\Delta x, \Delta y, \Delta z, \Delta\theta);$$

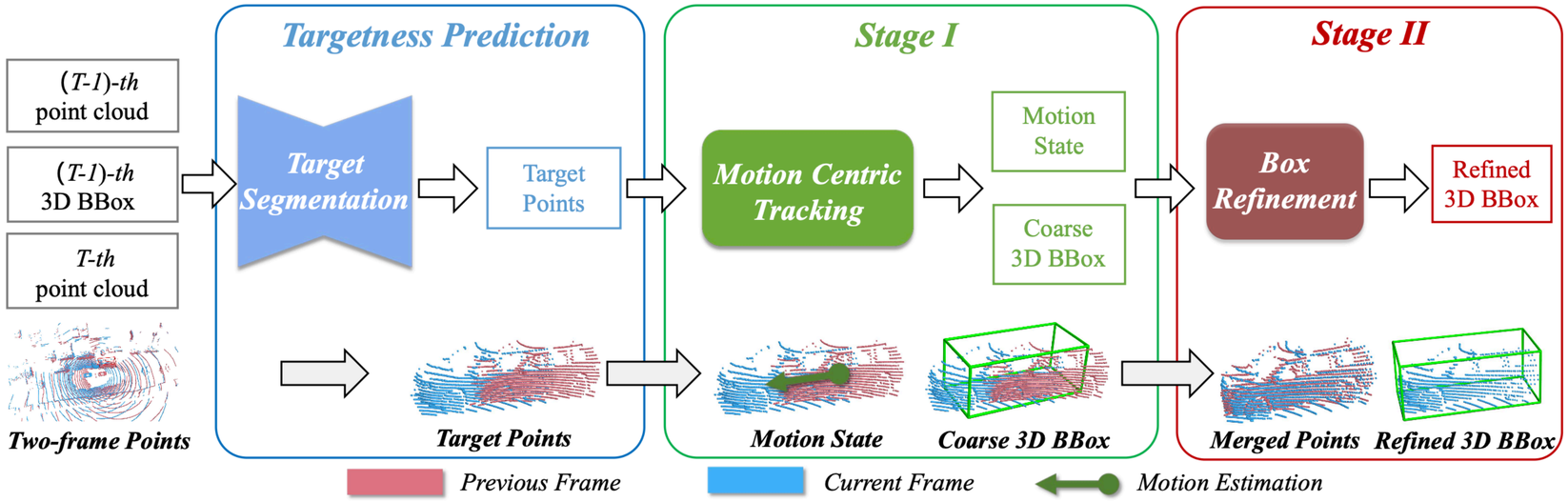
假设size不变

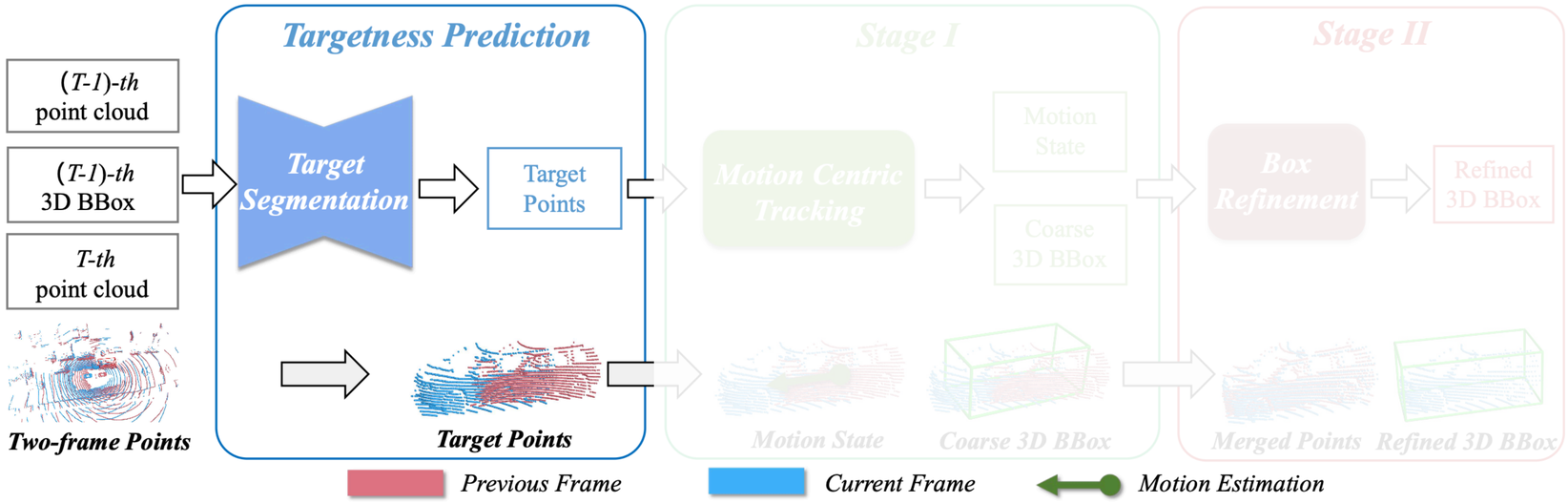
$$\mathcal{P}_t \in \mathbb{R}^{N_t \times 3}$$

$$\mathcal{B}_t \in \mathbb{R}^7$$ center(3), size(3), yaw(1)



M²Track CVPR'22 Oral

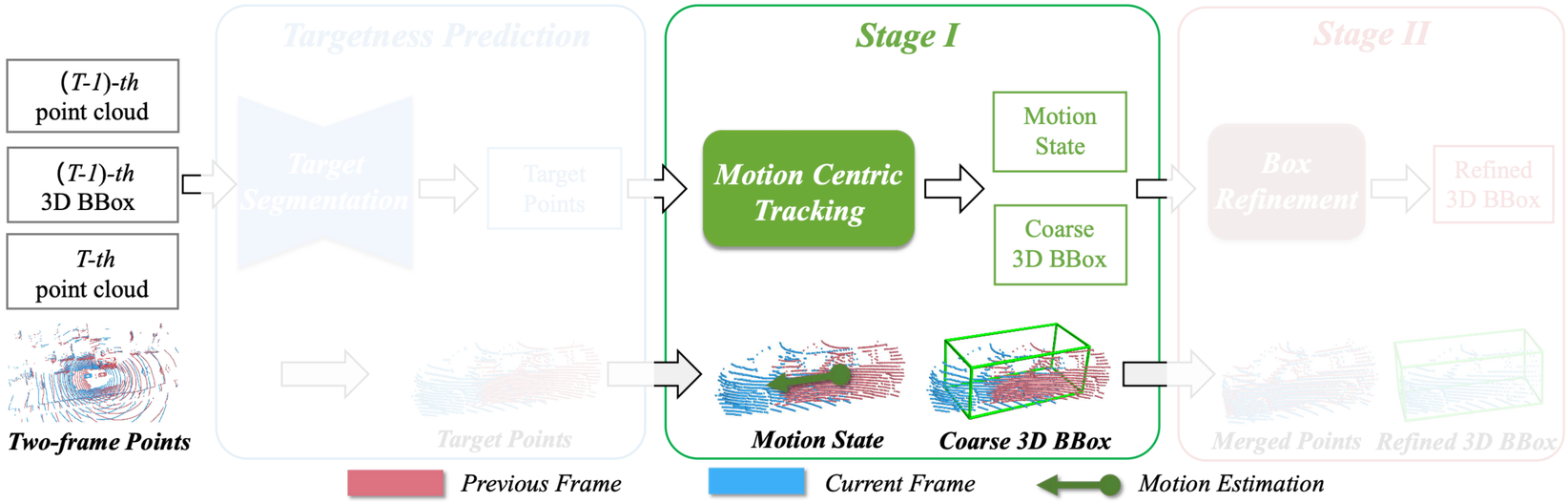




将两帧点云合并，并增加时间维度 $\mathcal{P}_{t-1,t} \in \mathbb{R}^{(N_{t-1}+N_t)\times 4}$

指定跟踪目标 $\mathcal{S}_{t-1,t} \in \mathbb{R}^{N_{t-1}+N_t}$ $s_i = \begin{cases} 0 & \text{if } p_i \text{ is in } \mathcal{P}_{t-1} \text{ and } p_i \text{ is not in } \mathcal{B}_{t-1} \\ 1 & \text{if } p_i \text{ is in } \mathcal{P}_{t-1} \text{ and } p_i \text{ is in } \mathcal{B}_{t-1} \\ 0.5 & \text{if } p_i \text{ is in } \mathcal{P}_t \end{cases}$

将二者concat，送入PointNet进行分割 $\widetilde{\mathcal{P}}_{t-1,t} \in \mathbb{R}^{(M_{t-1}+M_t)\times \hat{4}}$



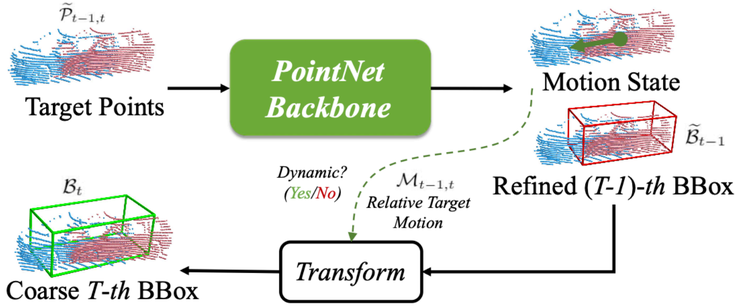
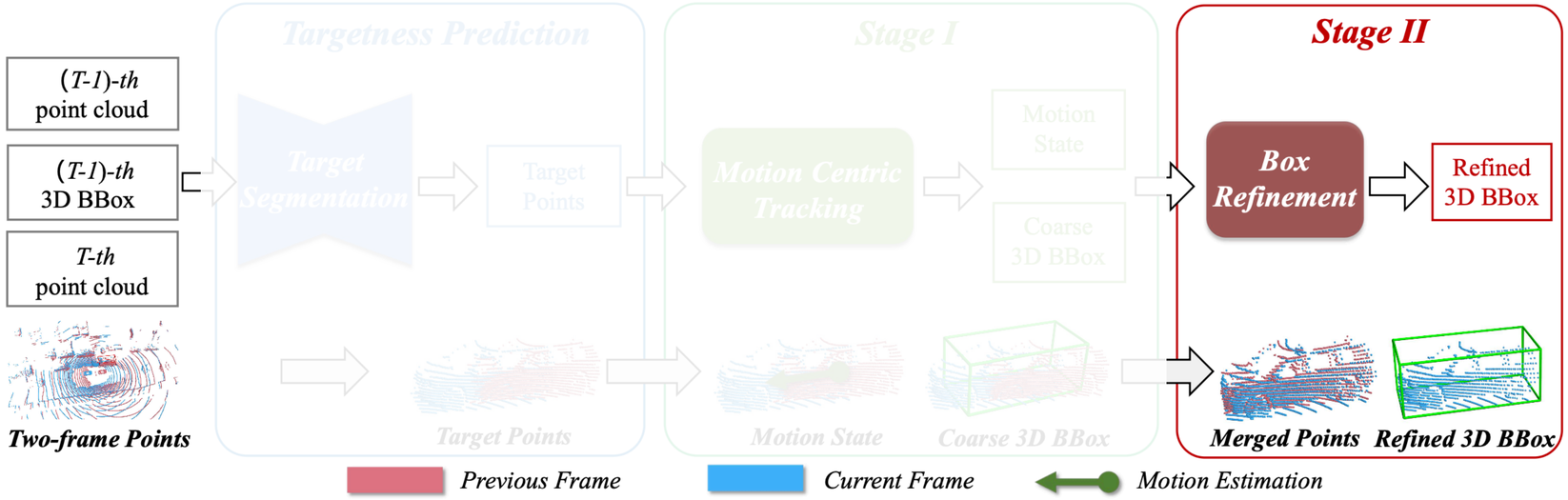


Figure 3. **Stage I.** Taking in the segmented target points $\widetilde{\mathcal{P}}_{t-1,t}$ and the target BBox $\mathcal{B}_{t-1}$ at the previous frame, the model outputs a relative target motion state (including a RTM $\mathcal{M}_{t-1,t}$ and 2D binary motion state logits), a refined target BBox $\widetilde{\mathcal{B}}_{t-1}$ at the previous frame, and a coarse target BBox $\mathcal{B}_t$ at the current frame.

PointNet + MLP，预测：

1) 对当前帧跟踪目标进行粗略估计的4D motion state
2) 2D binary logits indicating whether the target is dynamic
3) 对上一帧跟踪结果进行refine的4D motion state (减少累积误差)



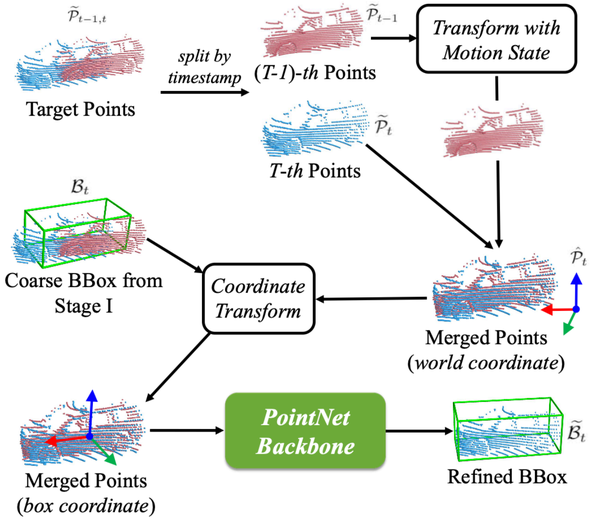


Figure 4. **Stage II.** Taking the segmented target points $\widetilde{\mathcal{P}}_{t-1,t}$ and the coarse target BBox $\mathcal{B}_t$ as inputs, the model regresses a refined target BBox $\widetilde{\mathcal{B}}_t$ on a denser point cloud, which is merged from two partial target point clouds according to their relative motion state.

Motivation: LiDAR point clouds suffer from great incompleteness, which hinders precise BBox regression.

1) 根据时间维度，将分割后的点云重新划分成两帧的点云
2) 根据Stage1预测的motion state，对前一帧点云进行变换，与当前帧合并
3) 将点云从世界坐标系变换到目标坐标系
4) PointNet回归motion state，进行refine

	Category Frame Number	Car *6424*	Pedestrian *6088*	Van *1248*	Cyclist *308*	Mean *14068*
Success	SC3D [10]	41.3	18.2	40.4	41.5	31.2
	SC3D-RPN [42]	36.3	17.9	-	43.2	-
	P2B [23]	56.2	28.7	40.8	32.1	42.4
	3DSiamRPN [8]	58.2	35.2	45.6	36.1	46.6
	LTTR [6]	65.0	33.2	35.8	66.2	48.7
	PTT [26]	67.8	44.9	43.6	37.2	55.1
	V2B [13]	**70.5**	48.3	50.1	40.8	58.4
	BAT [43]	65.4	45.7	52.4	33.7	55.0
	M^2-Track (Ours)	65.5	**61.5**	**53.8**	**73.2**	**62.9**
	Improvement	*↓5.0*	*↑13.2*	*↑1.4*	*↑7.0*	*↑4.5*
Precision	SC3D [10]	57.9	37.8	47.0	70.4	48.5
	SC3D-RPN [42]	51.0	47.8	-	81.2	-
	P2B [23]	72.8	49.6	48.4	44.7	60.0
	3DSiamRPN [8]	76.2	56.2	52.8	49.0	64.9
	LTTR [6]	77.1	56.8	45.6	89.9	65.8
	PTT [26]	**81.8**	72.0	52.5	47.3	74.2
	V2B [13]	81.3	73.5	58.0	49.7	75.2
	BAT [43]	78.9	74.5	67.0	45.4	75.2
	M^2-Track (Ours)	80.8	**88.2**	**70.7**	**93.5**	**83.4**
	Improvement	*↓ 1.0*	*↑13.7*	*↑3.7*	*↑3.6*	*↑8.2*

KITTI

	Dataset Category Frame Number	NuScenes Car *64,159*	Pedestrian *33,227*	Truck *13,587*	Trailer *3,352*	Bus *2,953*	Mean *117,278*	Waymo Open Dataset Vehicle *1,057,651*	Pedestrian *510,533*	Mean *1,568,184*
Success	SC3D [10]	22.31	11.29	30.67	35.28	29.35	20.70	-	-	-
	P2B [23]	38.81	28.39	42.95	48.96	32.95	36.48	28.32	15.60	24.18
	BAT [43]	40.73	28.83	45.34	52.59	35.44	38.10	35.62	22.05	31.20
	M^2-Track (Ours)	**55.85**	**32.10**	**57.36**	**57.61**	**51.39**	**49.23**	**43.62**	**42.10**	**43.13**
	Improvement	*↑15.12*	*↑3.27*	*↑12.02*	*↑5.02*	*↑15.95*	*↑11.14*	*↑8.00*	*↑20.05*	*↑11.92*
Precision	SC3D [10]	21.93	12.65	27.73	28.12	24.08	20.20	-	-	-
	P2B [23]	43.18	52.24	41.59	40.05	27.41	45.08	35.41	29.56	33.51
	BAT [43]	43.29	53.32	42.58	44.89	28.01	45.71	44.15	36.79	41.75
	M^2-Track (Ours)	**65.09**	**60.92**	**59.54**	**58.26**	**51.44**	**62.73**	**61.64**	**67.31**	**63.48**
	Improvement	*↑21.80*	*↑7.60*	*↑16.96*	*↑13.37*	*↑23.43*	*↑17.02*	*↑17.49*	*↑30.52*	*↑21.73*

NuScenes & Waymo

Table 3. Influence of Motion Augmentation. "aug" stands for motion augmentation.

Method	Success	Precision
BAT [43] w/o aug	65.37	78.88
BAT [43] w/ aug	63.59 ↓ 1.78	76.99 ↓ 1.89
P2B [23] w/o aug	56.20	72.80
P2B [23] w/ aug	55.21 ↓ 0.99	71.51 ↓ 1.29
M^2-Track w/o aug	65.29	77.12
M^2-Track w/ aug	65.49 ↑ 0.20	80.81 ↑ 3.69

Table 4. Integration with Appearance Matching.

Method	Success	Precision
PTT [26]	67.80	81.80
V2B [13]	**70.50**	81.30
M^2-Track	65.49	80.81
M^2-Track + BAT [43]	69.22 ↑ 3.73	81.09 ↑ 0.28
M^2-Track + P2B [23]	70.21 ↑ **4.72**	**81.80** ↑ **0.99**

Table 5. Results of M^2-Track when different modules are ablated. The last row denotes the full model. **Bold** denotes the largest change.

Box Aware Enhancement	Prev Box Refinement	Motion Classification	Stage-II	Kitti		NuScenes	
				Success	Precision	Success	Precision
	✓	✓	✓	62.00 ↓ 3.49	76.15 ↓ **4.66**	53.68 ↓ **2.17**	62.47 ↓ 2.62
✓		✓	✓	64.23 ↓ 1.26	78.12 ↓ 2.69	54.70 ↓ 1.15	61.94 ↓ **3.15**
✓	✓		✓	65.74 ↑ 0.25	80.29 ↓ 0.52	54.88 ↓ 0.97	64.40 ↓ 0.69
✓	✓	✓		61.29 ↓ **4.20**	77.31 ↓ 3.50	54.66 ↓ 1.99	64.15 ↓ 0.94
✓	✓	✓	✓	65.49	80.81	55.85	65.09

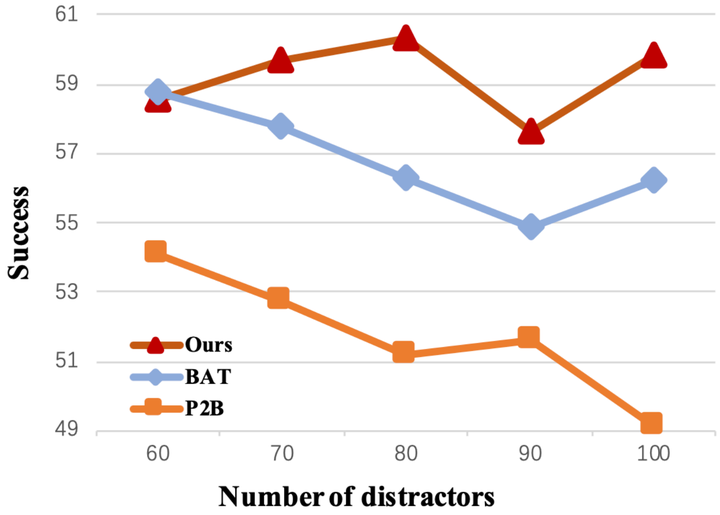


Figure 6. Robustness analysis with variant numbers of distractors.