MixCycle: Mixup Assisted Semi-Supervised 3D Single Object Tracking with Cycle Consistency
Supplementary

Qiao Wu[1] Jiaqi Yang[1*] Kun Sun[2] Chu'ai Zhang[1] Yanning Zhang[1] Mathieu Salzmann[3]
[1] Northwestern Polytechnical University [2] China University of Geosciences, Wuhan
[3] École Polytechnique Fédérale de Lausanne
qiaowu@mail.nwpu.edu.cn; jqyang@nwpu.edu.cn; sunkun@cug.edu.cn;
cazhang@mail.nwpu.edu.cn; ynzhang@nwpu.edu.cn; mathieu.salzmann@epfl.ch

A. Implementation Details

Framework Architecture. The overall pipeline with the grad flow of MixCycle is shown in Fig. 1. Due to the limitation of non maximum suppression (NMS) on gradient backpropagation, we only calculate the gradients of the directly supervised parts.

SOTMixup. Given the mix point cloud $P_A^m = P_A^b + \hat{P_A^o} + \hat{P_B^o}$ and Bounding-box B_A in label y_A at the SOTMixup, we only regard the points in B_A as the foreground points. Specifically, the points in $\hat{P_B^o}$ are considered as background noise if they are outside the B_A. We believe that modifying the size of the tracking target is incompatible with real tracking.

B. More Analysis

Training & Inference Time. We compare MixCycle and fully-supervised methods [1, 2, 4] in training time shown in Tab. 1. They are trained on Car in KITTI with 10% labels using 2 NVIDIA RTX-3090 GPUs. Our MixCycle takes around $2.0 \sim 2.5$ times as long as the fully-supervised methods. The experiments reveal that MixCycle requires a longer training time, but it is still in an acceptable range. Hence, we could expect a faster and more robust tracking network backbone for MixCycle.

Frame Number of Cycle Tracking and Unlabeled Data Losses Balance. **1)** Because of the limited memory of an NVIDIA RTX-3090 GPU, only a maximum of 2 cycle consistencies among 3 frames can be supervised. Therefore, we only present the losses for the self-supervised part. **2)** For the one without labels part, we have made experiments to balance those losses. We try to supervise different consistencies in a two-stage training by supervising $\mathcal{L}_{self}$ and $\mathcal{L}_{con0}$ in stage 1, and $\mathcal{L}_{self}$ and $\mathcal{L}_{con1}$ in stage 2, based on BAT with a 10% sampling rate on KITTI. Without SOTMixup, the cycle framework achieves 38.8/59.3 and 41.0/60.6 in Succ./Prec. in stage 1 & 2, respectively. The performance drops in stage 2 if we use SOTMixup. We conjecture this to be due to conflicts between the delicate losses set by SOTMixup in Self Cycle and the ambiguous losses in F.B. Cycle. We leave the design of a better training strategy for MixCycle as future work.

Table 1. Training time comparison of MixCycle and fully-supervised methods on Car in KITTI with 10% labels using 2 NVIDIA RTX-3090 GPUs. Decreases based on the same tracker is shown in red.

Method	Time	
P2B [1]	1h22m	
MLVSNet [2]	1h47m	
BAT [4]	1h22m	
Ours(P2B)	3h35m	2h13m↓
Ours(MLVSNet)	3h32m	1h45m↓
Ours(BAT)	3h40m	2h18m↓

Fairness of Comparison with Fully-supervised Method. Here we discuss fairness in the comparison experiments. The fully supervised method solely relies on labeled data, whereas our method utilizes both labeled and unlabeled data. **1)** The intention of our work is to reduce the effort in data annotation. While reducing the cost of collecting data is also important, we constitute a different research task on its own. **2)** We refer to a semi-supervised 3D object detection method SESS's [3] experimental setting for comparison experiments. SESS directly reduces the usage of data of fully supervised methods for comparison experiments because no other method shares the same semi-supervised settings with it, which is very similar to our situation. **3)** We present the performance comparison using the same amount of data but with different label usage in the Tab.3 and Tab.4

[*] Corresponding author

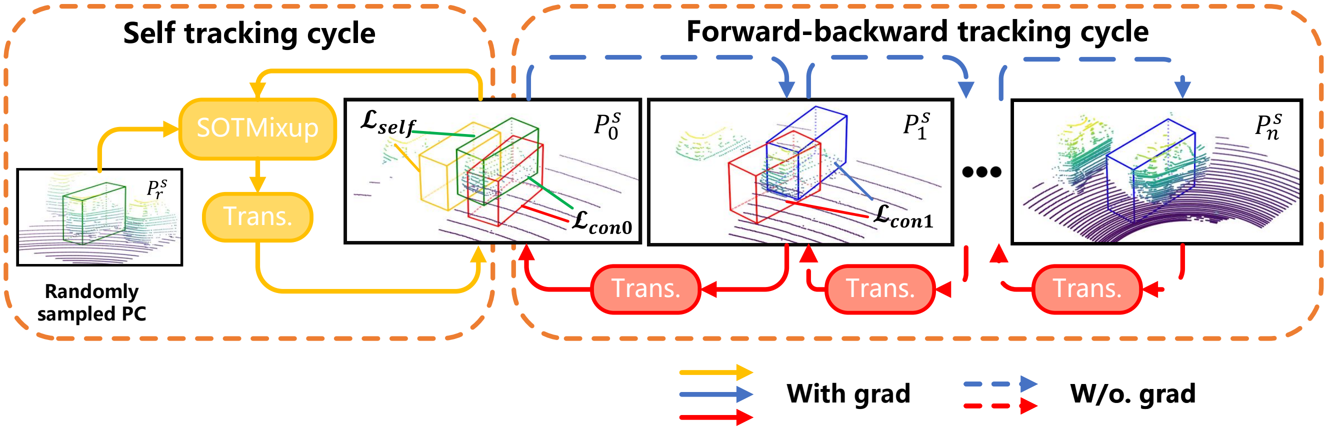

Figure 1. The framework of MixCycle. The gradient flow is represented by solid and dashed lines with arrows.

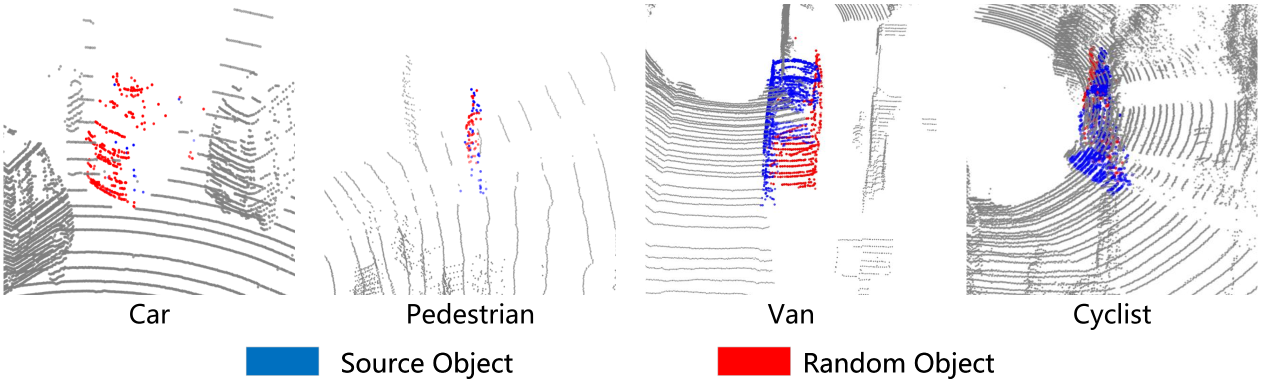

Figure 2. Visualization of SOTMixup.

of the paper.

Further Details. We further demonstrate the test result on each category and sample rate on KITTI and Nuscenes shown in Tab. 2 and Tab. 3. We achieve great success on Cyclist. The maximum improvement on the Cyclist class is up to **44.77%/75.83%** in success/precision based on P2B [1] with 10% labels. For the most important class Car in KITTI and NuScenes, MixCycle also achieves a remarkable improvement in every sample rate.

C. Visualization

SOTMixup. Our MixCycle leverages SOTMixup to supply diverse training samples. As shown in Fig. 2, we present SOTMixup in a variety of categories. Our SOTMixup completes the point cloud of the occluded area in the Van in Fig. 2, making the training samples more diverse. In the Car in Fig. 2, SOTMixup almost removes the point cloud of the source object, allowing the trackers to regress the correct target center by learning the distribution of object motion in extreme cases.

KITTI Results. We present the visualization results of the comparison between Our MixCycle and BAT [4] with 10% sample rate in Fig. 3. The visualization results further validate the superiority of our approach in sparse and complex scenarios.

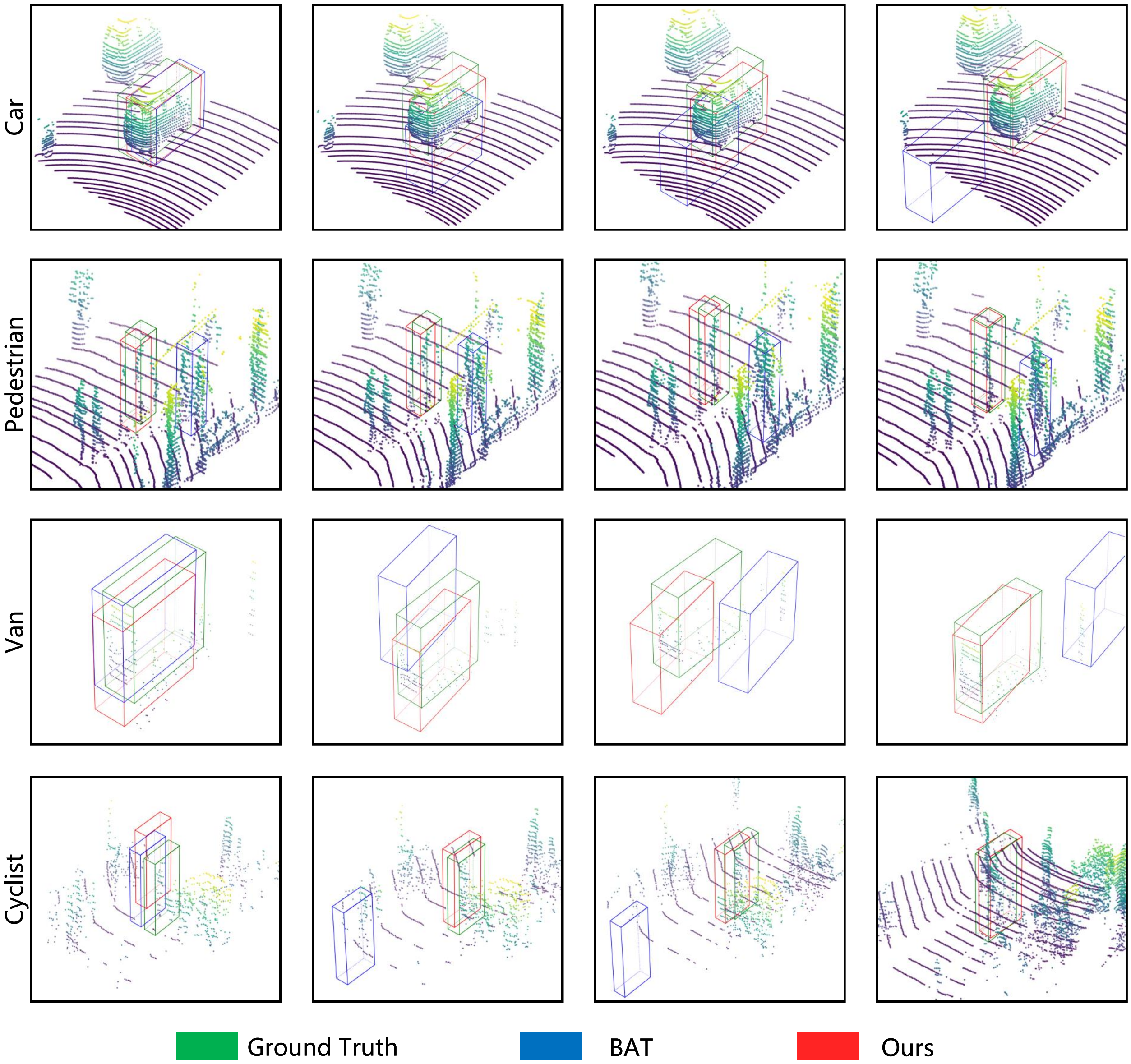

Figure 3. Visualization results. Our MixCycle and BAT are trained with 10% labels on KITTI.

Table 2. Comparsion of MixCycle against fully-supervised methods on each category in KITTI. Improvements and decreases based on the same tracker are shown in green and red, respectively. **Bold** and underline denote the best and the second-best performance, respectively.

		Category	Car		Pedestrian		Van		Cyclist		Mean	
		Frame Number	6424		6088		1248		308		14068	
Success	1%	P2B [1]	8.11		3.61		8.10		5.60		6.11	
		MLVSNet [2]	35.27		15.15		22.94		12.76		24.98	
		BAT [4]	16.69		3.81		7.17		6.77		10.05	
		Ours(P2B)	20.56	12.45↑	**22.76**	19.15↑	7.97	0.13↓	16.62	11.02↑	20.31	14.20↑
		Ours(MLVSNet)	**43.75**	8.48↑	20.68	5.53↑	**28.22**	5.28↑	**43.73**	30.97↑	**32.39**	7.41↑
		Ours(BAT)	32.63	15.94↑	6.08	2.27↑	16.33	9.16↑	34.12	27.35↑	19.73	9.67↑
	5%	P2B [1]	33.99		20.31		12.10		5.73		25.51	
		MLVSNet [2]	43.50		28.09		35.06		19.77		35.55	
		BAT [4]	24.30		21.00		13.17		13.25		21.62	
		Ours(P2B)	44.13	10.14↑	31.01	10.70↑	26.15	14.05↑	36.77	31.04↑	36.70	11.19↑
		Ours(MLVSNet)	**52.44**	8.94↑	24.04	4.05↓	**38.73**	3.67↑	46.54	26.77↑	38.80	3.26↑
		Ours(BAT)	49.24	24.94↑	**37.63**	16.63↑	26.08	12.91↑	**50.08**	36.83↑	**42.18**	20.56↑
	10%	P2B [1]	41.94		30.63		19.61		7.37		34.31	
		MLVSNet [2]	48.21		24.76		37.90		24.89		36.64	
		BAT [4]	43.96		28.84		18.12		35.84		34.95	
		Ours(P2B)	45.82	3.88↑	**41.59**	10.96↑	**42.59**	22.98↑	52.14	44.77↑	43.84	9.53↑
		Ours(MLVSNet)	54.08	5.87↑	30.39	5.63↑	41.29	3.39↑	49.95	25.06↑	42.60	5.97↑
		Ours(BAT)	**55.19**	11.23↑	38.62	9.78↑	34.92	16.8↑	**55.52**	19.68↑	**46.23**	11.28↑
Precision	1%	P2B [1]	7.39		2.24		6.07		4.42		4.98	
		MLVSNet [2]	46.54		28.80		25.41		16.62		36.33	
		BAT [4]	22.66		2.92		5.94		9.54		12.35	
		Ours(P2B)	29.97	22.58↑	**43.73**	41.49↑	6.08	0.01↑	11.12	6.70↑	33.39	28.41↑
		Ours(MLVSNet)	**59.24**	12.7↑	40.72	11.92↑	**31.08**	5.67↑	**79.03**	62.41↑	**49.16**	12.83↑
		Ours(BAT)	43.87	21.21↑	9.32	6.40↑	19.18	13.24↑	57.31	47.77↑	27.02	14.67↑
	5%	P2B [1]	45.99		40.26		10.82		5.43		39.50	
		MLVSNet [2]	57.53		52.07		42.30		28.77		53.19	
		BAT [4]	34.81		40.35		15.55		25.52		35.30	
		Ours(P2B)	56.94	10.95↑	58.04	17.78↑	30.92	20.1↑	67.33	61.90↑	55.34	15.83↑
		Ours(MLVSNet)	**66.61**	9.08↑	47.15	4.92↓	**45.26**	2.96↑	81.06	52.29↑	56.61	3.42↑
		Ours(BAT)	62.07	27.26↑	**68.05**	27.70↑	30.81	15.26↑	**82.63**	57.11↑	**62.33**	27.04↑
	10%	P2B [1]	56.11		57.70		21.73		7.35		52.68	
		MLVSNet [2]	63.63		48.31		44.65		35.08		54.69	
		BAT [4]	57.25		56.08		21.48		19.69		52.75	
		Ours(P2B)	58.30	2.19↑	**72.05**	14.35↑	**51.83**	30.1↑	83.18	75.83↑	64.22	11.54↑
		Ours(MLVSNet)	67.36	3.73↑	56.28	7.97↑	50.01	5.36↑	82.52	47.44↑	61.36	6.67↑
		Ours(BAT)	**70.02**	12.77↑	69.83	13.75↑	42.28	20.8↑	**85.37**	65.68↑	**67.81**	15.06↑

Table 3. Comparsion of MixCycle against fully-supervised methods on each category in NuScenes.

		Category	Car		Truck		Trailer		Bus		Mean	
		Frame Number	64159		13587		3352		2953		84051	
Success	0.1%	P2B [1]	15.77		13.09		12.81		16.12		15.23	
		MLVSNet [2]	20.99		25.16		22.46		13.53		21.46	
		BAT [4]	17.46		17.75		20.43		14.42		17.52	
		Ours(P2B)	23.01	7.24↑	25.22	12.13↑	22.37	9.56↑	17.65	1.53↑	23.15	7.92↑
		Ours(MLVSNet)	**29.67**	8.68↑	**42.43**	17.27↑	**31.34**	8.88↑	**19.22**	5.69↑	**31.43**	9.97↑
		Ours(BAT)	24.32	6.86↑	26.88	9.13↑	23.66	3.23↑	16.92	2.50↑	24.45	6.93↑
	0.5%	P2B [1]	24.42		19.21		20.30		12.38		22.99	
		MLVSNet [2]	29.82		32.25		27.40		22.74		29.87	
		BAT [4]	27.71		22.85		25.48		15.44		26.40	
		Ours(P2B)	**36.85**	12.43↑	28.23	9.02↑	21.75	1.45↑	21.14	8.76↑	34.30	11.31↑
		Ours(MLVSNet)	31.49	1.67↑	**46.75**	14.50↑	**48.49**	21.09↑	**28.47**	5.73↑	**34.53**	4.66↑
		Ours(BAT)	32.20	4.49↑	38.22	15.37↑	31.04	5.56↑	21.82	6.38↑	32.76	6.36↑
	1%	P2B [1]	23.95		27.83		25.84		14.57		24.32	
		MLVSNet [2]	33.23		39.08		39.62		22.23		34.04	
		BAT [4]	30.66		32.73		32.83		17.81		30.63	
		Ours(P2B)	34.80	10.85↑	35.24	7.41↑	30.40	4.56↑	22.61	8.04↑	33.43	9.10↑
		Ours(MLVSNet)	**40.61**	7.38↑	**45.43**	6.35↑	**58.09**	18.47↑	**35.38**	13.15↑	**41.90**	7.86↑
		Ours(BAT)	33.72	3.06↑	37.29	4.56↑	45.55	12.72↑	24.26	6.45↑	34.44	3.81↑
Precision	0.1%	P2B [1]	14.52		8.20		6.82		8.41		12.98	
		MLVSNet [2]	20.45		19.97		11.31		6.35		19.51	
		BAT [4]	16.31		12.16		9.19		12.22		15.21	
		Ours(P2B)	23.48	8.96↑	18.88	10.68↑	11.20	4.38↑	**13.99**	5.58↑	21.91	8.94↑
		Ours(MLVSNet)	**31.05**	10.60↑	**38.57**	18.60↑	**19.45**	8.14↑	11.53	5.18↑	**31.12**	11.60↑
		Ours(BAT)	24.10	7.79↑	21.07	8.91↑	13.81	4.62↑	9.67	2.55↓	22.69	7.48↑
	0.5%	P2B [1]	24.28		12.32		11.08		6.98		21.21	
		MLVSNet [2]	32.73		26.71		14.91		15.35		30.44	
		BAT [4]	28.69		18.06		15.09		8.89		25.73	
		Ours(P2B)	**39.22**	14.94↑	20.79	8.47↑	11.19	0.11↑	13.27	6.29↑	34.21	13.00↑
		Ours(MLVSNet)	34.17	1.44↑	**42.78**	16.07↑	**38.71**	23.8↑	**19.59**	4.24↑	**35.23**	4.80↑
		Ours(BAT)	33.35	4.66↑	32.21	14.15↑	18.62	3.53↑	14.49	5.60↑	31.92	6.18↑
	1%	P2B [1]	23.70		22.86		14.11		7.51		22.61	
		MLVSNet [2]	36.76		33.91		29.60		15.41		35.26	
		BAT [4]	32.47		28.36		20.42		11.19		30.58	
		Ours(P2B)	36.72	13.02↑	29.15	6.29↑	18.17	4.06↑	14.78	7.27↑	33.99	11.38↑
		Ours(MLVSNet)	**45.07**	8.31↑	**40.17**	6.26↑	**46.28**	16.68↑	**25.01**	9.60↑	**43.62**	8.36↑
		Ours(BAT)	35.29	2.82↑	32.30	3.94↑	32.63	12.21↑	17.15	5.96↑	34.06	3.49↑

References

[1] Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. P2b: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2020. 1, 2, 4, 5

[2] Zhoutao Wang, Qian Xie, Yu-Kun Lai, Jing Wu, Kun Long, and Jun Wang. Mlvsnet: Multi-level voting siamese network for 3d visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3101–3110, 2021. 1, 4, 5

[3] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 1

[4] Chaoda Zheng, Xu Yan, Jiantao Gao, Weibing Zhao, Wei Zhang, Zhen Li, and Shuguang Cui. Box-aware feature enhancement for single object tracking on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13199–13208, 2021. 1, 2, 4, 5