[Supplementary Material] Sequential Multi-View Fusion Network for Fast LiDAR Point Motion Estimation

Gang Zhang[2], Xiaoyan Li[1]✉, and Zhenhua Wang[3]

[1] Beijing Municipal Key Lab of Multimedia and Intelligent Software Technology,
Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing
University of Technology, Beijing 100124, China
[2] Damo Academy, Alibaba Group
[3] Cenozoic Robot
{zhanggang11021136,hblixy2,zhwang.me}@gmail.com

In the supplementary material, we present the implementation details in Section 1. Extensive experiments on the Waymo Open Dataset are shown in Section 2 to demonstrate the effectiveness of the proposed components. Finally, we present some representative visualization results in Section 3.

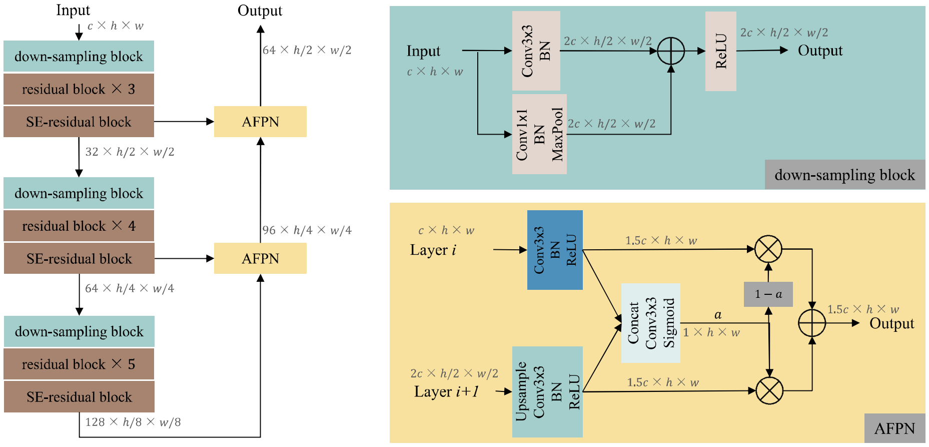


Fig. 1. The detailed architecture of 2D FCN.

1 Implementation Details

In this section, we illustrate the details of the 2D FCN architecture and the sequential instance copy-paste (SICP).

1.1 FCN Architecture

As shown in Fig. 1, the 2D FCN architecture used in the proposed SMVF has three downs-sampling stages and two up-sampling stages. The BEV branch and RV branch use a similar FCN framework, but the RV branch does not apply down-sampling along the height dimension. The down-sampling block fuses the down-sampling features from the 2D Convolution and 2D MaxPool, respectively. Subsequently, it adopts several standard residual blocks [3] and a standard SE-residual block [4] to extract features. For the up-sampling stage, inspired by AFF [2], the attention feature pyramid fusion (AFPN) module uses the spatial attention mechanism to automatically select features from different levels.

1.2 Sequential Instance Copy-Paste

The sequential instance copy-paste (SICP) aims to generate and insert reasonable object trajectories into the LiDAR sequences for more training samples. The SICP has five steps as follows.

(1) It constructs an object bank, consisting of the objects cropped from their original LiDAR scans according to the annotated 3D bounding boxes.
(2) It uniformly samples a category and an object from this category.
(3) It generates a trajectory for the sampled object, considering that the object is assumed to be moving along its 3D bounding box heading *yaw*. Its velocity $\tilde{v}$ is sampled uniformly within a predefined range of each category and the object is moving backward when $\tilde{v}$ is negative. On the Waymo Open Dataset, the predefined velocity ranges of *vehicle*, *pedestrian*, and *cyclist* are $[-40, 40]$, $[-5, 5]$, and $[-20, 20]$ m/s, respectively.
(4) It randomly selects a position from feasible position candidates for the generated object trajectory in the LiDAR sequences. The feasible position candidates satisfy that a) the distance r from the inserted object to the sensor in the current t^{th} scan should be the same as that in its original LiDAR scan; b) the inserted object should be placed on the ground in each scan; c) there is no occlusion between the inserted object and any existing object in each scan.
(5) The occluded background points are removed.

2 Experiments On the Waymo Open Dataset

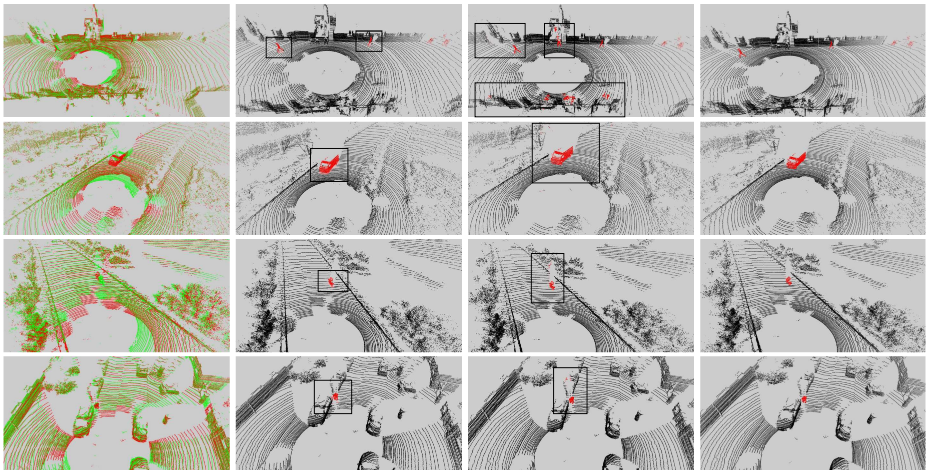
In this section, we report some ablative experiments on the validation split of the Waymo Open Dataset for the motion velocity prediction task.

2.1 Analysis of the number of history LiDAR scans

On the Waymo Open Dataset, we evaluate the effect of the number of history LiDAR scans. As shown in Table 1, more history LiDAR scans ($k \leqslant 3$) lead to higher performance, but the performance saturates when $k > 3$. These observations are similar to those of the SemanticKITTI.

Table 1. Analysis of the number of history LiDAR scans on the validation split of the Waymo Open Dataset. k denotes the number of history LiDAR scans. A: All. M: Moving. S: Stationary.

Method	k	SICP	Metric	*Vehicle*			*Pedestrian*			*Cyclist*			*Background*	Runtime
				A	M	S	A	M	S	A	M	S		
SMVF [ours]	1	✓	mean (m/s) ↓	0.16	0.46	0.05	0.23	0.28	0.11	0.35	0.38	0.11	0.028	**30 ms**
			⩽ 0.1 m/s ↑	67.8%	10.2%	**89.8%**	29.4%	11.7%	67.9%	15.3%	7.7%	73.2%	99.0%	
			⩽ 1.0 m/s ↑	97.9%	92.8%	99.7%	98.1%	97.4%	99.5%	97.0%	96.8%	99.3%	99.5%	
SMVF [ours]	2	✓	mean (m/s) ↓	**0.14**	**0.41**	0.05	0.21	0.26	0.11	**0.29**	0.32	0.09	0.027	32.8 ms
			⩽ 0.1 m/s ↑	69.1%	11.7%	89.2%	33.3%	15.4%	69.1%	18.5%	10.7%	77.1%	99.1%	
			⩽ 1.0 m/s ↑	**98.5%**	**94.8%**	**99.8%**	98.6%	98.1%	**99.7%**	**98.1%**	**97.8%**	99.7%	99.5%	
SMVF [ours]	3	✓	mean (m/s) ↓	**0.14**	0.42	**0.04**	0.20	**0.24**	**0.10**	**0.29**	**0.31**	**0.08**	**0.026**	36.3 ms
			⩽ 0.1 m/s ↑	**69.6%**	12.0%	**89.7%**	35.4%	**18.2%**	**70.0%**	**19.2%**	**11.4%**	**78.5%**	**99.2%**	
			⩽ 1.0 m/s ↑	98.4%	94.4%	**99.8%**	98.8%	98.3%	**99.7%**	97.9%	97.7%	**99.9%**	**99.6%**	
SMVF [ours]	4	✓	mean (m/s) ↓	**0.14**	**0.41**	**0.04**	**0.19**	**0.24**	**0.10**	**0.29**	**0.31**	0.09	0.027	38.4 ms
			⩽ 0.1 m/s ↑	**69.6%**	**12.1%**	89.6%	**35.5%**	**18.2%**	**70.0%**	19.1%	11.2%	78.2%	**99.2%**	
			⩽ 1.0 m/s ↑	**98.5%**	94.7%	**99.8%**	**98.9%**	**98.4%**	**99.7%**	97.9%	97.6%	99.8%	**99.6%**	



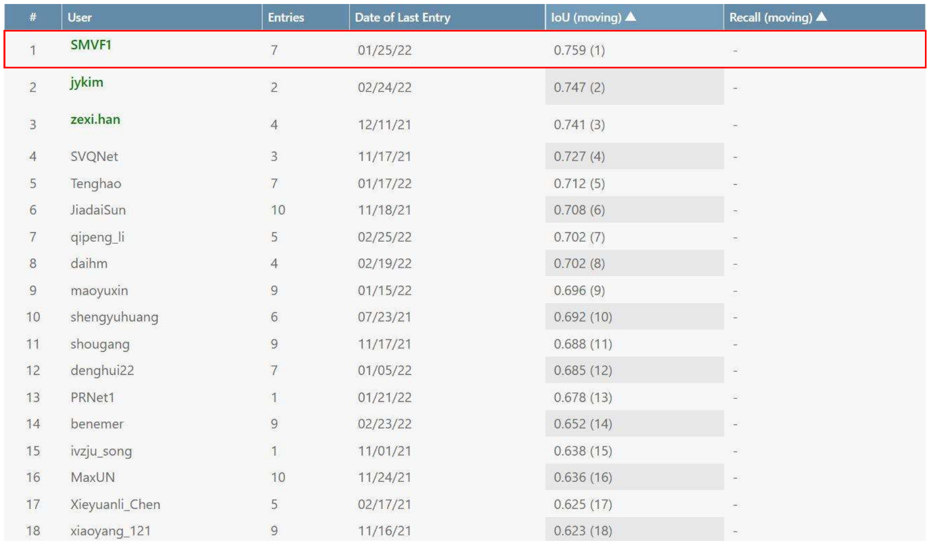
a) LiDAR Sequence b) BEV c) RV d) SMVF (ours)

Fig. 2. Visualization comparison results on the SemanticKITTI test set. The left column (a) shows the two consecutive LiDAR scans, where the red points mark the current t^{th} scan and the green ones mark the history $(t-1)^{\text{th}}$ scan. The middle two columns (b,c) present the moving object segmentation results of the BEV-based MotionNet [7] and the RV-based Chen *et al.* [1], respectively. The right column shows the prediction results of our SMVF. The predicted moving points are shown in red, while the stationary ones are shown in black.

3 Visualization

3.1 Results On the SemanticKITTI

The visualization results of the BEV-based MotionNet [7], the RV-based Chen *et al.* [1], and our SMVF are shown in Fig. 2. The prediction results of Chen *et al.* [1] are acquired by its official code. The MotionNet is re-implemented on



#	User	Entries	Date of Last Entry	IoU (moving) ▲	Recall (moving) ▲
1	SMVF1	7	01/25/22	0.759 (1)	-
2	jykim	2	02/24/22	0.747 (2)	-
3	zexi.han	4	12/11/21	0.741 (3)	-
4	SVQNet	3	11/17/21	0.727 (4)	-
5	Tenghao	7	01/17/22	0.712 (5)	-
6	JiadaiSun	10	11/18/21	0.708 (6)	-
7	qipeng_li	5	02/25/22	0.702 (7)	-
8	daihm	4	02/19/22	0.702 (8)	-
9	maoyuxin	9	01/15/22	0.696 (9)	-
10	shengyuhuang	6	07/23/21	0.692 (10)	-
11	shougang	9	11/17/21	0.688 (11)	-
12	denghui22	7	01/05/22	0.685 (12)	-
13	PRNet1	1	01/21/22	0.678 (13)	-
14	benemer	9	02/23/22	0.652 (14)	-
15	ivzju_song	1	11/01/21	0.638 (15)	-
16	MaxUN	10	11/24/21	0.636 (16)	-
17	Xieyuanli_Chen	5	02/17/21	0.625 (17)	-
18	xiaoyang_121	9	11/16/21	0.623 (18)	-

Fig. 3. Screenshot of the public leaderboard on the SemanticKITTI moving object segmentation (MOS) benchmark at 2022-03-07. Our SMVF ranks **1st**.

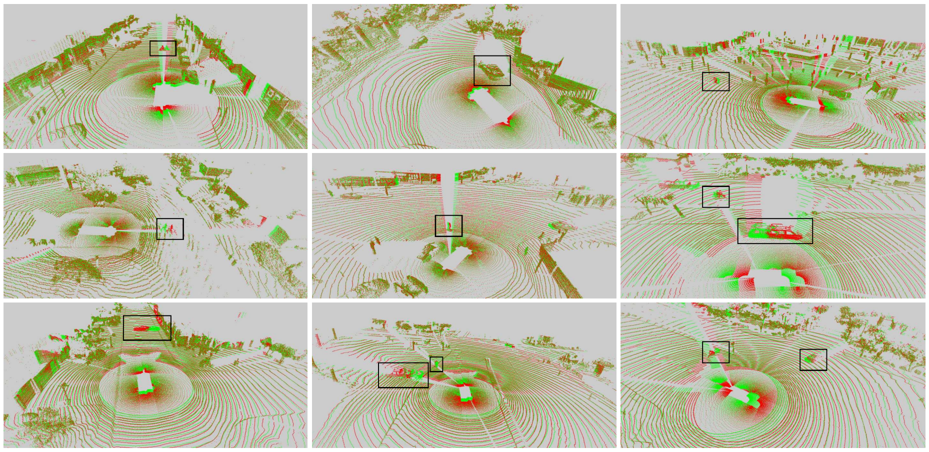


Fig. 4. Visualization results of the proposed SICP on the Waymo Open Dataset. The red points mark the current t^{th} scan and the green ones mark the history $(t-1)^{\text{th}}$ scan. The black boxes show the inserted objects.

the SemanticKITTI based on its official code and the settings can be seen in the main paper. The two single-view methods frequently confuse points projected to the same or nearby 2D grids, while our SMVF solves the problem by the cooperation between different views. Our SMVF achieves the **1st** place on the public SemanticKITTI moving object segmentation (MOS) leaderboard in Fig. 3.

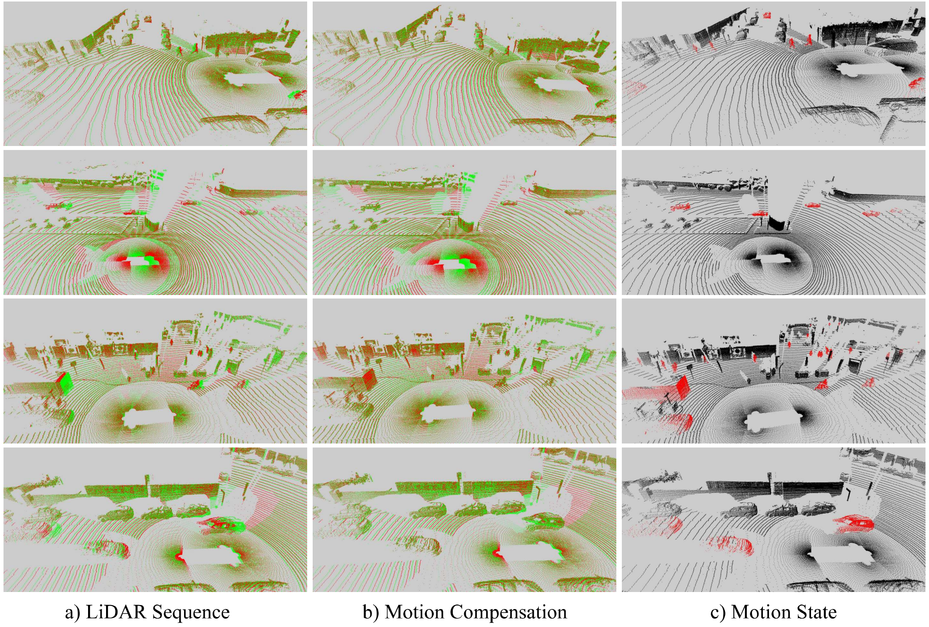


Fig. 5. Visualization results of our SMVF on the validation split of the Waymo Open Dataset. The first column (a) shows two consecutive LiDAR scans, where the red points mark the current t^{th} scan and the green ones mark the history $(t-1)^{\text{th}}$ scan. For visualizing the velocity, the positions of the 3D points are compensated by their predicted velocities in the middle column (b). The last column (c) presents the point-level motion state, where the points with their speed $\geqslant 0.5\ m/s$ are shown in red, and the others are shown in black.

3.2 Results On the Waymo Open Dataset

Sequential Instance Copy-Paste. The visualization results of the proposed SICP on the Waymo Open Dataset are shown in Fig. 4. The SICP can be used for more than two LiDAR scans, although we present only two consecutive LiDAR scans for better visualization. We find that the SICP can generate realistic LiDAR sequences for all categories except some abnormal cases. For example, the first column of the 1st and 3rd rows in Fig. 4 presents that a *cyclist* and a *vehicle* break the traffic rules and are crossing the road. For the safety of the autonomous driving system, these abnormal cases make a motion estimator generalize well to various scenarios since these cases may occur in real-world applications.

Motion Velocity Prediction. We visualize the predicted velocity results of the proposed SMVF on the Waymo Open Dataset. For convenience, we use the velocity predictions to compensate the current t^{th} LiDAR scan, assuming that the speed of an object is constant in a short time interval (*e.g.* a LiDAR

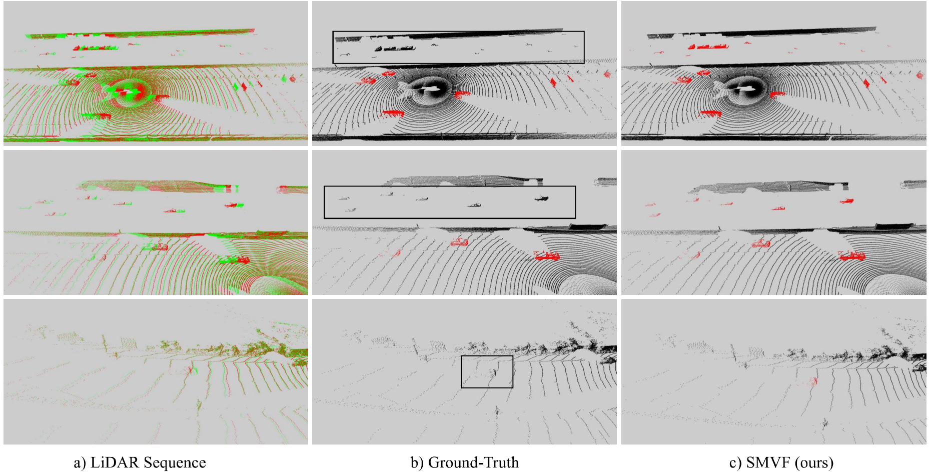


Fig. 6. Generalization to the unlabeled objects of our SMVF on the validation split of the Waymo Open Dataset. The middle and last columns present the results from the ground truth and our SMVF. The points with their speed $\geqslant 0.5\ m/s$ are shown in red, and the others are shown in black.

scan period). If the motion velocity prediction is perfect, the compensated scan should completely overlap with the $(t-1)^{th}$ LiDAR scan. As shown in Fig. 5, our SMVF can predict accurate velocities for all categories. Moreover, our SMVF accurately classifies the road points near a moving object as stationary, while the FastFlow3D [5] frequently classifies these points as moving as shown in Fig.6 of the paper [5]. Although the FastFlow3D integrates the point features with the BEV motion features, the point features cannot be fully explored by a shared MLP and fail to compensate for the 2D projection information loss. Our SMVF solves the problem by introducing a single-scan RV branch.

Generalization to the Unlabeled Objects. As illustrated in FastFlow3D [5], the point-level ground-truth velocities are obtained by the tracked 3D bounding boxes. Without a tracked box, an object is regarded as stationary. For a reliable autonomous driving system, a motion estimator should generalize well to these unlabeled objects. In the main paper, we have quantitatively evaluated the generalization ability by ablating *pedestrian* and *cyclist* during training. Here, we present some visualization results in Fig. 6. Our SMVF can predict accurate results for these unlabeled objects.

References

1. Chen, X., Li, S., Mersch, B., Wiesmann, L., Gall, J., Behley, J., Stachniss, C.: Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data. IEEE Robotics and Automation Letters **6**(4), 6529–6536 (2021)
2. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3560–3569 (2021)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. Jund, P., Sweeney, C., Abdo, N., Chen, Z., Shlens, J.: Scalable scene flow from point clouds in the real world. IEEE Robotics and Automation Letters (2021)
6. Lim, H., Oh, M., Myung, H.: Patchwork: concentric zone-based region-wise ground segmentation with ground likelihood estimation using a 3d lidar sensor. IEEE Robotics and Automation Letters **6**(4), 6458–6465 (2021)
7. Wu, P., Chen, S., Metaxas, D.N.: 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. pp. 11385–11395 (2020)