Bring Event into RGB and LiDAR: Hierarchical Visual-Motion Fusion for Scene Flow

Supplementary Material

Hanyu Zhou[1], Yi Chang[1]*, Zhiwei Shi[1]
[1] National Key Lab of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
{hyzhou, yichang, shizhiwei}@hust.edu.cn

In this supplementary material, we will further describe the complementary nature of motion correlation distributions between RGB, event and LiDAR in Sec. 1. We further demonstrate the effectiveness of various homogeneous feature fusions in Sec. 2.1. We conduct the ablation experiments on the influence of training data and flow backbone on the final results in Sec. 2.2. Then, we discuss the importance of visual structure fusion in Sec. 3.1. We also verify the robustness of the proposed method for various illumination conditions in Sec. 3.3. Next, we compare the inference time of the proposed method with other state-of-the-art methods in Sec. 3.2. We further provide the implementation of training details in Sec. 3.4. Finally, we provide more visualization results of comparison on the synthetic Event-KITTI dataset in Sec. 4.1 and the real DSEC dataset in Sec. 4.2.

1. Complementary Motion Correlation

Our insight is that RGB provides x, y-axis spatial-dense correlation, event provides x, y-axis temporal-dense correlation, and LiDAR promises x, y, z-axis accurate but sparse correlation for scene flow. In order to further illustrate the complementarity of the correlation features between RGB, event and LiDAR, we calculate the multimodal correlation feature distributions along the x, y, and z dimensions within a time period. Note that, we just temporally visualize the statistical distribution of the x, y, z-axis correlation features to facilitate understanding of the correlation complementarity in the motion space. We have two conclusions. First, RGB, event and LiDAR share similar correlation distributions in the x and y axes, which motivates us to align the cross-modal x, y-axis correlation features for complementary motion knowledge fusion. Second, LiDAR contains sparse z-axis features, event has temporal-dense features and RGB has spatial-dense features. Therefore, there is a complementary knowledge of the correlation between these modalities in x, y, z dimensions for spatiotemporal-dense 3D motion.

*Corresponding author.

Training data	Method	EPE	ACC
RGB, LiDAR	CamLiFlow	0.113	55.69%
	Ours w/ RAFT	0.107	59.85%
	Ours w/ GMA	0.107	60.03%
	Ours w/ FlowFormer	**0.105**	**61.17%**
RGB, Event, LiDAR	RPEFlow	0.103	60.81%
	Ours w/ RAFT	0.086	69.52%
	Ours w/ GMA	0.084	70.18%
	Ours w/ FlowFormer	**0.084**	**70.34%**

Table 1. Ablation study on training data and flow backbones.

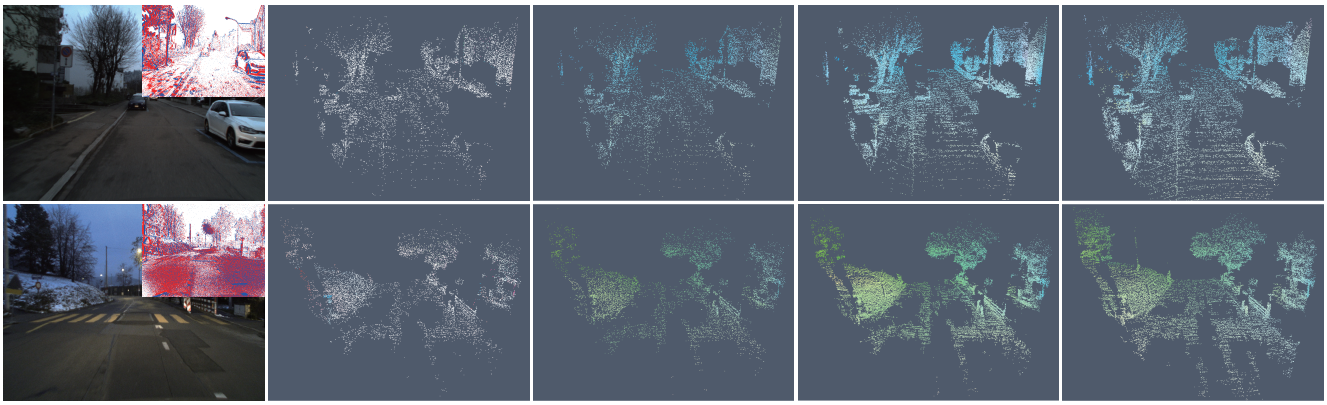
2. Ablation Study

2.1. Effectiveness of Homogeneous Feature Fusion

In Fig. 1, we visualize the scene flow results of different homogeneous feature fusions (*i.e.*, visual luminance fusion (VLF), visual structure fusion (VSF) and motion correlation fusion (MCF)). We input RGB, event and LiDAR, and output the scene flows from different homogeneous feature fusions. Without any homogeneous feature fusion, we only implicitly fuse the features of various modalities, but the estimated scene flow contains many outliers. With only VLF, most outliers in the scene flow are removed. With VLF and VSF, the structure integrity of the scene flow is improved and the mismatched features are reduced. With VLF, VSF and MCF, the scene flow is further improved. Therefore, the proposed hierarchical visual-motion fusion framework can explicitly fuse multimodal complementary knowledge to progressively improve scene flow from visual to motion space.

2.2. Influence of Training Data and Backbone

In Table 1, we compare the influence of various flow backbones (*e.g.*, RAFT [1], GMA [2] and FlowFormer [3])



(a) RGB / Event (b) LiDAR (c) w/ VLF (d) w/ VLF + VSF (e) w/ VLF + VSF + MCF

Figure 1. Visualization of scene flows from different homogeneous fusion stages. "VLF" denotes the Event-RGB visual luminance fusion. "VSF" denotes the Event-LiDAR visual structure fusion. "MCF" denotes the RGB-Event-LiDAR motion correlation fusion.

and training data (*e.g.*, RGB, event and LiDAR) on the final scene flow results. As for the training data, the scene flow performance from RGB-Event-LiDAR three modalities is indeed significantly better than that of RGB-LiDAR two modalities. As for the flow backbone, there is no obvious difference in the improvement of scene flow results from various flow backbones. This shows that multimodal training data is more conducive to the model than the flow backbone to learn the intrinsic motion patterns of the scene. It is worth that the proposed method that uses the homogeneous features to fuse the knowledge can significantly improve scene flow performance, indicating that the homogeneous space can more explicitly model the multimodal complementary knowledge to improve scene flow. Therefore, under the whole framework, the event serves as a bridge to effectively close the modality gap due to the intrinsic heterogeneous nature between RGB and LiDAR, and the cross-modal homogeneous feature space can further make the scene flow fusion process more physically interpretable.

3. Discussion

3.1. Importance of Structure Fusion

In order to study the effect of event in visual structure fusion of LiDAR, we visualize the entire fusion process in Fig. 2. We input the sparse point cloud (x, y, z) (seeing Fig. 2 (b)) and the corresponding event stream (t, x, y) (seeing Fig. 2 (a)). We temporally project the event stream into the 2D (x, y) image coordinate along the t-axis dimension for boundary map in Fig. 2 (c), and spatially project the point cloud into the 2D (x, y) image coordinate along the z-axis dimension for depth map in Fig. 2 (d). Then, we cluster the 2D boundary map and 2D depth map into a neighbor space like an image superpixel in Fig. 2 (e), which transforms the pixel-level structure into a region-level structure. Next, we use KNN algorithm to associate the points of 2D depth map with those of 2D boundary map in the same neighbor for obtaining a fused depth map in Fig. 2 (f). Finally, we use the intrinsic parameters to inversely project the 2D depth map into the 3D coordinate system for the final point cloud in Fig. 2 (g). Compared with the input LiDAR point cloud in Fig. 2 (b), the enhanced point cloud (seeing Fig. 2 (g)) has a relatively complete structure and clearer contours, which can reduce the subsequent motion feature matching error.

Method	Inference time (ms)	EPE	ACC
RAFT-3D [6]	400.8	0.167	13.16%
PV-RAFT [7]	2545.7	0.183	37.28%
CamLiFlow [8]	285.6	0.113	55.69%
RPEFlow [9]	328.8	0.103	60.81%
VisMoFlow	335.2	**0.084**	**70.34%**

Table 2. Discussion on inference time.

3.2. Inference Time

In Table 2, we choose inference time as the efficiency metric for scene flow estimation, and RTX 3090 as the inference platform. Note that, during the inference stage, the spatial resolution of the RGB image and event is 640×480, and the sample number of the LiDAR point cloud is set as 17000. We can observe that the proposed method does not have an advantage in inference time, but its performance is much superior to other methods. This is because the proposed framework performs multimodal fusion in both visual and motion spaces, sacrificing computational resources but significantly improving the performance of the scene flow. In the future, we will achieve network lightweighting to make the entire multimodal fusion more efficient using model quantization [4] and pruning [5] techniques.

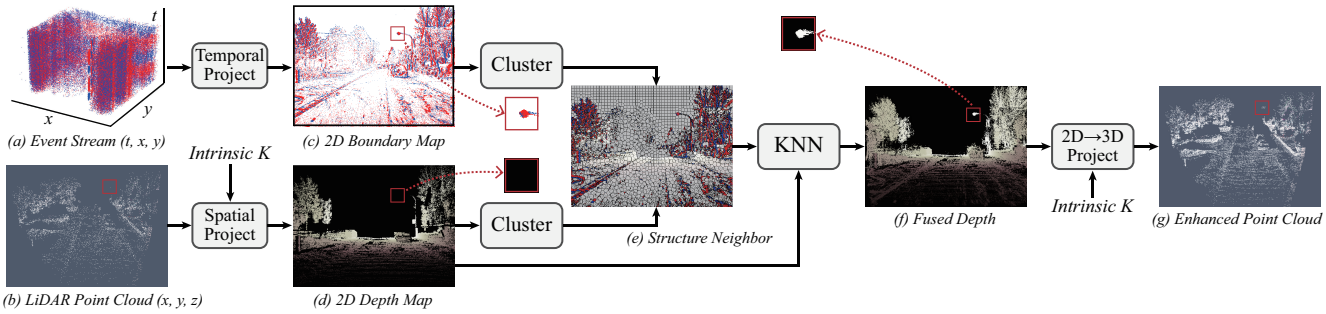


Figure 2. Detailed process of Event-LiDAR visual structure fusion. First, we project (a) Event stream and (b) LiDAR point cloud to the 2D image coordinate system for obtaining the (c) 2D boundary map and (d) 2D depth map, respectively. Then, we cluster the boundary map and the depth map to obtain the (e) structure neighbor, which represents a region-level structure. Next, we fuse the (e) structure neighbor and the (d) depth map to obtain a (f) fused depth map via K-Nearest Neighbor. Finally, we inversely project the fused depth into a 3D point cloud. (g) Enhanced point cloud has a relatively complete structure and clearer contours than (b) input point cloud.

Time slice number T	EPE	ACC
3	0.091	67.95%
5	0.086	69.72%
10	**0.084**	**70.34%**
15	0.085	69.80%
20	0.091	68.17%

Table 3. Discussion on the number of event time slices.

Selection number k	EPE	ACC
1	0.096	64.14%
3	0.086	68.17%
5	**0.084**	**70.34%**
8	0.087	69.58%
10	0.094	67.81%

Table 4. The choice of selection number of K-Nearest Neighbor.

3.3. Robustness for Various Illumination

In our framework, only RGB modality is sensitive to illumination, while event modality with high dynamic range is robust to illumination, and LiDAR is naturally insensitive to ambient illumination due to its own imaging mechanism. In order to verify the robustness for various illumination conditions, we conduct experiments to compare the scene flow performance of other state-of-the-art methods and our method under daytime, dusk and low-light conditions in Fig. 3. For daytime scenes, all scene flow methods perform well. When applied to dusk scenes, the RGB-based unimodal method RAFT-3D exhibits obvious artifacts, while the multimodal methods are still able to maintain the performance. For low-light scenes, the RGB-based unimodal method cannot work, and the scene flow visualization of the multimodal method RPEFlow [9] also shows obvious color deviation. In contrast, the proposed multimodal method VisMoFlow can still maintain the overall smoothness of 3D motion.

Sample number N	EPE	ACC
100	0.094	66.92%
500	0.088	69.85%
1000	**0.084**	70.34%
2000	0.084	**70.36%**

Table 5. Impact of correlation sample number on scene flow.

3.4. Implementation of Training Details

Setup in Time Slices of Events. Event time slice is to divide the event stream over a period of time into multiple small time periods. The larger the number of time slices, the denser the temporal dimension motion information but the sparser the spatial dimension visual information. In Table 3, we study the impact of different numbers of time slices on the final scene flow results. We choose [3, 5, 10, 15, 20] as the candidate values of the number T of time slices. When T is less than 10, the scene flow is gradually improved. When T is equal to 15, the scene flow performance shows a slight decline. The main reason is that, the larger the number of time slices, the visual features of the single event stream become fewer, resulting in invalid motion feature matching, thus interfering with the scene flow. Therefore, choosing a reasonable number of time slices is very important.

Selection Number of K-Nearest Neighbor. In Table 4, we study the impact of the selection number of K-Nearest Neighbor (KNN) on the final scene flow results. We choose [1, 3, 5, 8, 10] as the candidate value of the selection number k of K-Nearest Neighbor. We can observe that, the larger the selection number k, the better the scene flow performance. However, when k is larger than 5, the scene flow performance decreases slightly. The main reason is that, KNN filters the LiDAR coordinates that match an event coordinate from near to far according to the distance. If k is too large, some LiDAR coordinates that are too far away may be introduced into the intel-modal depth fusion, affecting the visual

Figure 3. Visualization of scene flows under various illumination conditions.

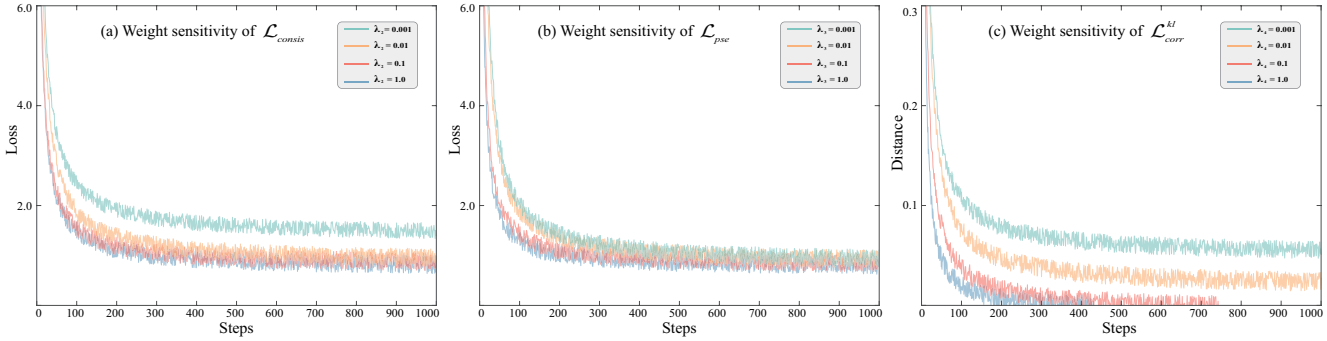
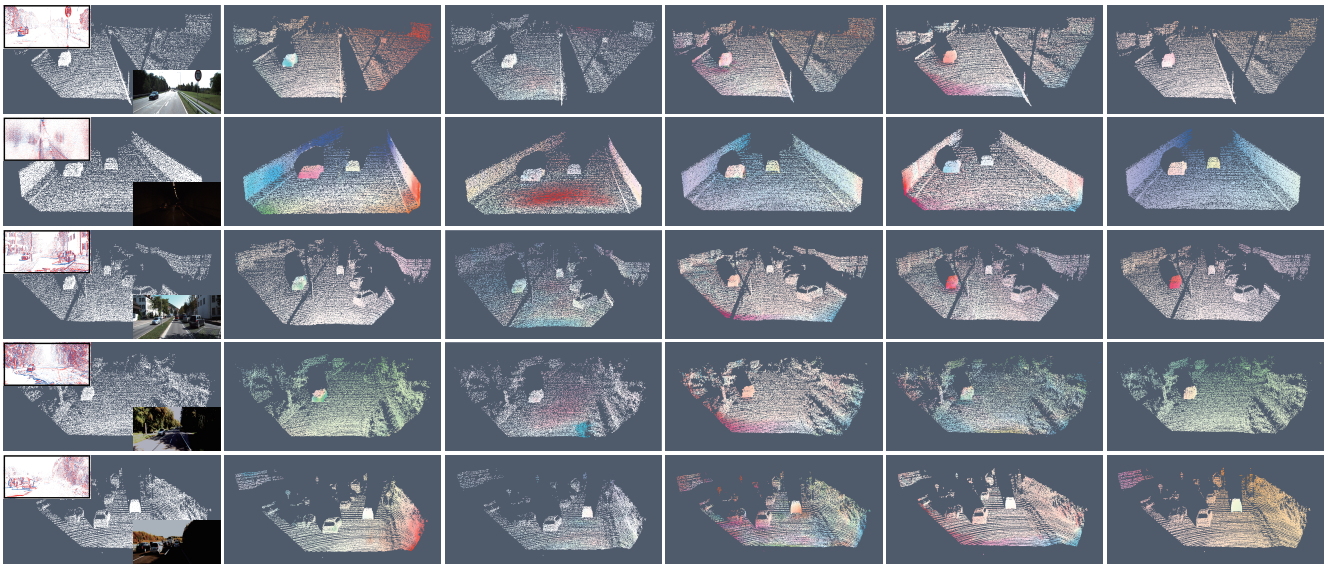


Figure 4. The balance weight sensitivity of model fusion losses.



(a) Multimodal data (b) RAFT-3D (c) PV-RAFT (d) CamLiFlow (e) RPEFlow (f) VisMoFlow

Figure 5. Comparison of scene flows on the synthetic Event-KITTI dataset.

structure fusion, and thus interfering with the scene flow.

Sample Number of Correlation. In Table 5, we study the influence of the correlation sample number on the final scene flow. We choose [100, 500, 1000, 2000] as the candidate values of the sample number N. We observe that, as the sample number N becomes larger, the scene flow metric is also improved. However, when N is increased to 2000, the scene flow performance is basically not improved, but instead, the computational cost increases. Hence, to make a trade-off between performance and cost, we set 1000 as the correlation sample number N for the distribution alignment of the multimodal correlation features.

Weight Sensitivity of Model Losses. To choose the optimal balance weights for the total loss, we conduct the study on the weight sensitivity of the typical fusion losses in Fig. 4, such as, $\mathcal{L}_{consis}$, $\mathcal{L}_{pse}$ and $\mathcal{L}_{corr}^{kl}$. The purpose of $\mathcal{L}_{consis}$ is to guarantee that the Event-RGB visual luminance fusion not only enhances the visual quality but also improves the capability of motion feature matching. $\mathcal{L}_{pse}$ is to ensure the cross-attention transformer to learn the Event-LiDAR visual structure fusion process. $\mathcal{L}_{corr}^{kl}$ aims to align the feature distributions in the motion space between RGB, event and LiDAR, which is beneficial to subsequent correlation fusion. In Fig. 4 (a), the larger the balance weight of $\mathcal{L}_{consis}$, the more rapidly the proposed framework converges. In Fig. 4 (b), the balance weight of $\mathcal{L}_{pse}$ is robust for the framework training. In Fig. 4 (c), the K-L divergence loss $\mathcal{L}_{corr}^{kl}$ is sensitive to the framework training. If the weight is too large, the gradient will disappear. Therefore, we set the balance weights of the fusion losses as $[\lambda_2, \lambda_3, \lambda_4]$ as [1.0, 1.0, 0.01].

4. Comparison

4.1. Comparison on Synthetic Dataset

The visualization of the scene flows predicted by the VisMoFlow and other state-of-the-art methods on the synthetic Event-KITTI dataset are presented in Fig. 5. It can be clearly observed that RGB-based unimodal method RAFT-3D [6] suffers slight artifacts. LiDAR-based unimodal method PV-RAFT [7] performs well in the background regions of scene flow, but there exist obvious outliers in the independent moving object regions. RGB-LiDAR multimodal method CamLiFlow [8] performs well in daytime scenes, but is difficult in nighttime scenes. In contrast, RGB-Event-LiDAR multimodal methods (*e.g.*, RPEFlow [9] and VisMoFlow) perform relatively well in all-day conditions. However, the scene flow visualization of RPEFlow has local color discontinuities, while our method VisMoFlow is overall smooth. This shows that the VisMoFlow can better fuse the cross-modal complementary knowledge from visual to motion space for the spatiotemporal continuity of 3D motion.

4.2. Comparison on Real Dataset

We also show the visualization results of the scene flows predicted by the proposed method VisMoFlow and other state-of-the-art methods on the real DSEC dataset in Fig. 6. Note that, LiDAR used in the DSEC dataset is a 16-channel LiDAR, and the acquired point cloud is relatively sparse. We can observe that, the scene flows estimated by other competing methods cannot work well. This is because sparse point clouds have incomplete structures, leading to mismatching the motion features. On the contrary, the proposed method introduces the event to enhance RGB for high dynamic imag-

(a) Multimodal data (b) PV-RAFT (c) CamLiFlow (d) RPEFlow (e) VisMoFlow

Figure 6. Comparison of scene flows on the real DSEC dataset.

ing and LiDAR for physical contour integrity in the visual space, and fuse the multimodal spatiotemporal complementary correlation in the motion space for 3D motion continuity, thus effectively improving scene flow in all-day scenes.

References

[1] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419, 2020. 1

[2] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Int. Conf. Comput. Vis.*, pages 9772–9781, 2021. 1

[3] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Eur. Conf. Comput. Vis.*, pages 668–685. Springer, 2022. 1

[4] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2704–2713, 2018. 2

[5] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Int. Conf. Learn. Represent.*, 2019. 2

[6] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8375–8384, 2021. 2, 5

[7] Yi Wei, Ziyi Wang, Yongming Rao, Jiwen Lu, and Jie Zhou. Pv-raft: Point-voxel correlation fields for scene flow estimation of point clouds. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6954–6963, 2021. 2, 5

[8] Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, and Lijun Chen. Camliflow: Bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5791–5801, 2022. 2, 5

[9] Zhexiong Wan, Yuxin Mao, Jing Zhang, and Yuchao Dai. Rpeflow: Multimodal fusion of rgb-pointcloud-event for joint optical flow and scene flow estimation. In *Int. Conf. Comput. Vis.*, pages 10030–10040, 2023. 2, 3, 5