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

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

Abstract

*Single RGB or LiDAR is the mainstream sensor for the challenging scene flow, which relies heavily on visual features to match motion features. Compared with single modality, existing methods adopt a fusion strategy to directly fuse the cross-modal complementary knowledge in motion space. However, these direct fusion methods may suffer the modality gap due to the visual intrinsic **heterogeneous** nature between RGB and LiDAR, thus deteriorating motion features. We discover that event has the **homogeneous** nature with RGB and LiDAR in both visual and motion spaces. In this work, we bring the event as a bridge between RGB and LiDAR, and propose a novel hierarchical visual-motion fusion framework for scene flow, which explores a homogeneous space to fuse the cross-modal complementary knowledge for physical interpretation. In visual fusion, we discover that event has a complementarity (relative v.s. absolute) in luminance space with RGB for high dynamic imaging, and has a complementarity (local boundary v.s. global shape) in scene structure space with LiDAR for structure integrity. In motion fusion, we figure out that RGB, event and LiDAR are complementary (spatial-dense, temporal-dense v.s. spatiotemporal-sparse) to each other in correlation space, which motivates us to fuse their motion correlations for motion continuity. The proposed hierarchical fusion can explicitly fuse the multimodal knowledge to progressively improve scene flow from visual space to motion space. Extensive experiments have been performed to verify the superiority of the proposed method.*

1. Introduction

Scene flow aims to model the correspondence between adjacent visual RGB or LiDAR features to estimate 3D motion features, which has been applied in various vision tasks, *e.g.*, 3D object detection [1] and motion segmentation [2]. The key to scene flow estimation is the highly discriminative
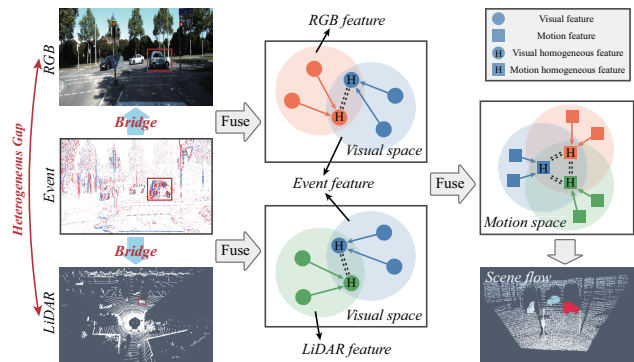
---

*Corresponding author.

Figure 1. Illustration of the main idea. There exists a large modality gap due to the visual intrinsic heterogeneous nature between RGB and LiDAR, thus deteriorating the motion features. We discover that the event has the homogeneous nature with RGB and LiDAR in both visual and motion spaces. In this work, we bring the event as a bridge between RGB and LiDAR, and propose a novel hierarchical visual-motion fusion framework for scene flow, which explores a homogeneous feature space to explicitly fuse the cross-modal complementary knowledge for physical interpretation.

visual feature for the valid motion feature matching.

Existing scene flow methods can be divided into two categories: unimodal direct learning methods [3–7] and multimodal motion fusion methods [8, 9]. Unimodal direct learning methods mainly resort to the ideal visual feature to directly learn the motion feature from single RGB or LiDAR modality. Multimodal motion fusion methods further exploit the complementary knowledge between RGB and LiDAR to fuse their features in the motion space. For example, Liu *et al.* [9] fused dense RGB motion features and sparse LiDAR motion features for scene flow. However, these methods neglect the visual intrinsic heterogeneous nature, *e.g.*, degraded visual RGB features with weakened texture under low light due to low dynamic range and limited visual LiDAR features with incomplete contour due to non-uniform laser beam. These can enlarge the modality gap and deteriorate the motion features. Therefore, *we suggest that introducing an auxiliary modality as the bridge to assistantly enhance the*

RGB and LiDAR features in both visual and motion spaces.

To our investigation, event camera [10] is a neuromorphic sensor, which can sense the dynamic scene boundary structure with high dynamic range and high temporal resolution. In visual space, RGB records the absolute value of luminance while the event triggers the relative change of luminance, and LiDAR reflects the global shape of the scene while the event senses the local boundary. In motion space, we transform the visual features of these modalities into correlation space, and we figure out that RGB provides the spatial-dense 2D features, event provides the temporal-dense 2D features, and LiDAR provides the spatiotemporal-sparse 3D features. Therefore, the event has the complementarity with RGB and LiDAR in visual homogeneous (*i.e.*, luminance and structure) and motion homogeneous (*i.e.*, correlation) spaces.

In this work, we bring the auxiliary event as a bridge between RGB and LiDAR in Fig. 1, and propose a novel hierarchical visual-motion fusion framework for scene flow (VisMoFlow), which *focuses on fusing the cross-modal complementary knowledge in the homogeneous space.* In visual luminance fusion, we fuse the relative luminance of event and the absolute luminance of RGB for high dynamic imaging with the theoretically derived spatiotemporal gradient consistency. In visual structure fusion, we fuse the local boundary of event into the global shape of LiDAR for physical structure integrity using the self-similarity clustering strategy. In motion correlation fusion, we map event, RGB and LiDAR into the same motion manifold, and fuse the x, y-axis spatial-dense correlation features of RGB, the x, y-axis temporal-dense correlation features of event and the x, y, z-axis sparse correlation features of LiDAR for 3D motion spatiotemporal continuity by aligning the motion distributions between various modalities. The proposed hierarchical visual-motion fusion can explicitly learn the multimodal complementary knowledge from visual to motion space for scene flow. Overall, the main contributions are as follows:

- We bring the auxiliary event as a bridge between RGB and LiDAR modalities, and propose a novel hierarchical visual-motion fusion framework for scene flow, which can explicitly fuse the multimodal knowledge to progressively improve scene flow from visual space to motion space.
- We reveal that RGB, event and LiDAR modalities have the complementary knowledge in both visual homogeneous (*i.e.*, luminance and structure) and motion homogeneous (*i.e.*, correlation) spaces, which can make the entire multimodal fusion process more physically interpretable.
- We conduct extensive experiments on daytime and nighttime scenes, which verify that the VisMoFlow achieves state-of-the-art performance for all-day scene flow.

2. Related Work

Unimodal Scene Flow Estimation. Since scene flow is the 3D version of optical flow, the previous scene flow methods

[3–7, 11–13] mainly directly follow the optical flow framework [14, 15]. They take monocular [5, 16, 17] or stereo sequences [6, 12] as input, and calculate cost volume with warp operation through feature pyramid [18] or recurrent network [14] for 3D motion. However, since 2D RGB images lack the intrinsic 3D geometry property, LiDAR becomes a trend in scene flow estimation due to its 3D scene perception capability. Most LiDAR-based scene flow methods [7, 13, 19, 20] also follow the architecture of the optical flow framework, while the difference is that they use MLP or voxel to represent the point cloud features. However, due to the potential incomplete physical contour of LiDAR, these single LiDAR-based scene flow methods are still limited. Instead, there is a strong complementarity between the two modalities, which inspires us to fully exploit the cross-modal complementary knowledge for scene flow.

Multimodal Scene Flow Estimation. As unimodal scene flow methods provide the partial information, multimodal fusion strategies [8, 9, 21, 22] have gradually received attention. Liu *et al.* [9] proposed a bidirectional 2D-3D motion feature fusion framework, which can exploit the complementary features between RGB and LiDAR. Wan *et al.* [22] introduced the event camera to enhance the RGB and LiDAR features in motion space via a cross-attention mechanism. However, these methods may suffer the potential degraded visual RGB features with weakened texture and limited visual LiDAR features with incomplete contour, thus leading to a large modality gap and deteriorating the valid motion features. In this work, we will introduce the auxiliary event as a bridge to hierarchically enhance the RGB and LiDAR features in both visual and motion spaces.

Multimodal Fusion. Multimodal fusion focuses on exploiting the inter-modality complementarity for the target task. To our knowledge, most existing multimodal fusion methods learn the mutual knowledge between different modalities via cross-attention mechanism [23–26] or similarity representation learning [27–30]. For example, Cai *et al.* [26] proposed to utilize transformer [31] to directly fuse the proposal features of RGB images and LiDAR for 3D object detection. However, these methods neglect the modality gap due to the intrinsic heterogeneous nature of feature representation between various modalities, limiting the complementary knowledge fusion. To alleviate this issue, we introduce an auxiliary modality as the bridge to mitigate the modality gap, and explore the homogeneous space to explicitly fuse the cross-modal complementary knowledge for scene flow.

3. Hierarchical Visual-Motion Fusion

3.1. Overall Framework

Multimodal scene flow framework aims to fuse the cross-modal complementary knowledge for enhancing the discriminative visual and motion features that scene flow relies
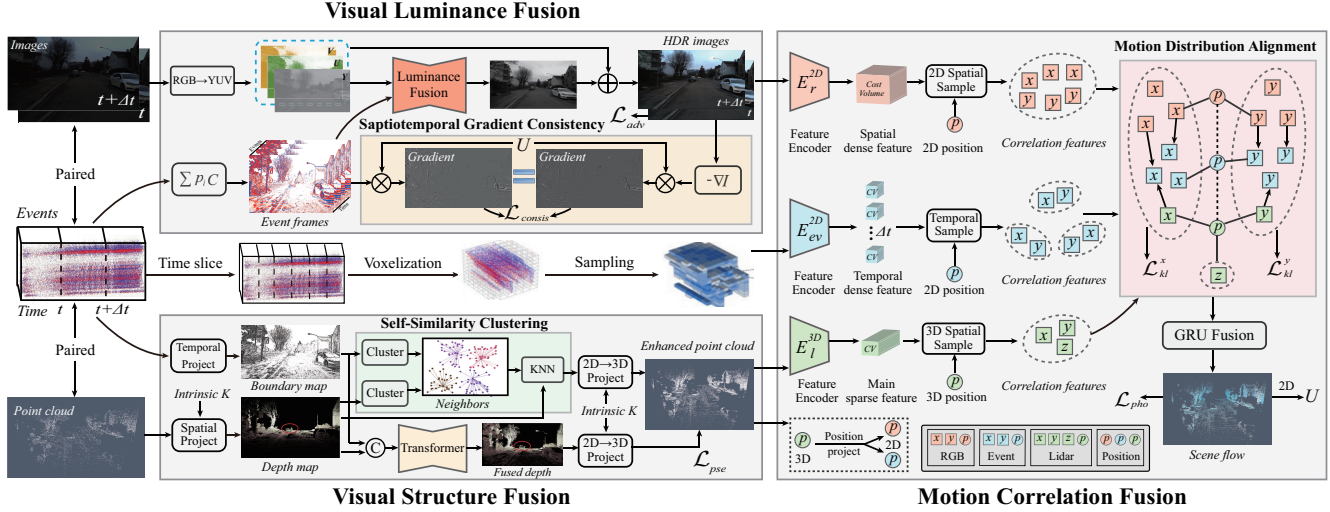
Figure 2. The architecture of the VisMoFlow mainly contains visual luminance fusion, visual structure fusion and motion correlation fusion. In visual luminance fusion, we fuse the relative luminance of event and the absolute luminance of RGB for high dynamic imaging. In visual structure fusion, we fuse the local boundary structure of event and the global shape structure of LiDAR for structure integrity. In motion correlation fusion, we fuse the spatiotemporal complementary correlation knowledge of the three modalities for 3D motion continuity.

on. However, there exists a modality gap due to the visual intrinsic heterogeneous nature between RGB and LiDAR, deteriorating motion features. To this end, we introduce the auxiliary event as a bridge to mitigate the gap between RGB and LiDAR, and propose a novel multimodal hierarchical visual-motion fusion framework for scene flow in Fig. 2. The whole architecture looks complicated but is simply regarded as a task that builds a homogeneous space to fuse the cross-modal complementary knowledge for physical interpretation. In visual luminance fusion, we transform the event and RGB into the luminance space, and fuse the complementary (relative v.s. absolute) knowledge for high dynamic imaging under the constraint of the similar spatiotemporal gradient. In visual structure fusion, we represent the event and LiDAR into the spatial structure space, and fuse the complementary (local boundary v.s. global shape) knowledge for physical structure integrity using the self-similarity clustering strategy. In motion correlation fusion, we map the visual features of RGB, event and LiDAR to the same correlation space, and fuse the complementary (x, y-axis spatial-dense correlation, x, y-axis temporal-dense correlation and x, y, z-axis sparse correlation) knowledge for 3D motion spatiotemporal continuity via motion distribution alignment. Under the unified framework, the proposed method can explicitly fuse the cross-modal complementary knowledge to progressively improve scene flow from visual space to motion space. Next, we will describe the hierarchical visual-motion fusion framework in the form of "what to fuse" and "how to fuse".

3.2. Visual Luminance Fusion

Under oversaturated and low-light conditions, the RGB image may loss the texture, thus mismatching the motion features. The main reason is that RGB imaging mechanism limits the dynamic range. Therefore, we introduce event camera with the advantage of high dynamic range to assistantly enhance the visual RGB features.

Relative *v.s.* Absolute Luminance. Event camera can asynchronously capture the events triggered by the brightness change, recording the relative luminance, while RGB camera records the absolute luminance via global scan. Motivated by this, we argue that event and RGB share the homogeneous luminance space, where the relative value of event and the absolute value of RGB are complementary to each other. Specifically, given the RGB image I and the corresponding event stream, we transform the RGB image into YUV color space $[I^Y, I^U, I^V]$, where I^Y denotes the luminance, I^U, I^V is the chrominance information. We also accumulate the event stream to generate the intensity frame:

$$I^X = \sum p_i C, \quad (1)$$

where p_i denotes the event polarity. C is the event trigger threshold. Therefore, the absolute value I^Y from the RGB image and the relative value I^X from the event are complementary in the homogeneous luminance space. This makes us naturally transform RGB and event into a unified luminance space to achieve complementary knowledge fusion.

Spatiotemporal Gradient Consistency. We set various weights to fuse the absolute luminance of the RGB and the relative luminance of the event as follows:

$$H^Y = \mathcal{W}(I^Y, I^X) = (\omega_y I^Y + \omega_x I^X)/(\omega_y + \omega_x), \quad (2)$$

where ω_x, ω_y denote the weights of event and RGB, respectively. The fused luminance image H^Y has a high dynamic range, but without color information. The color can be compensated from U, V channels I^U, I^V of the input RGB
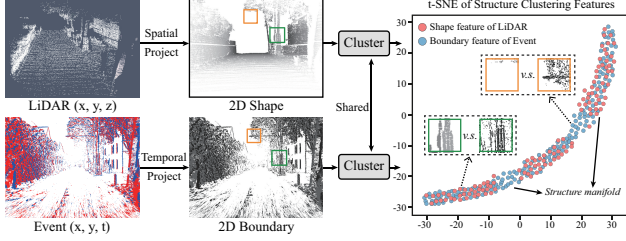
Figure 3. Clustering feature distribution of event and LiDAR. Event and LiDAR share the same structure manifold, where the boundary distribution of event is continuous while the shape distribution of LiDAR is truncated. This motivates us to take the structure as a homogeneous space to fuse the boundary and shape knowledges.

images, namely $H^{YUV} = \mathcal{C}(H^Y, I^U, I^V)$. We then merge this equal and Eq. 2 as follows:

$$H = \mathcal{C}(\mathcal{W}(I^Y, I^X), I^U, I^V). \quad (3)$$

The fused YUV image H should be further inversely transformed to the RGB color space. We follow [32] to utilize the luminance fusion network to model the Eq. 3. We take the adjacent RGB images $[I_t, I_{t+\Delta t}]$ at the timestamps t and $t + \Delta t$, and the corresponding event stream as input, and output the fused RGB images $[\tilde{I}_t, \tilde{I}_{t+\Delta t}]$. Note that, the fused images have two limitations: slightly unnatural color appearance and potential motion discontinuity. To recover the realistic color appearance, we force the adversarial loss [33] to make the fused images look natural:

$$\mathcal{L}_{adv} = \mathbb{E}[log(1 - D(\tilde{I}_t))] + \mathbb{E}[log(1 - D(\tilde{I}_{t+\Delta t}))], \quad (4)$$

where D is the discriminator. To prevent the motion discontinuity, we explore a motion-related prior knowledge to constrain the fusion process. We start ideally to yield the optical flow basic model as follows via Taylor expansion:

$$I(x + dx, t + dt) = I(x, t) + (\frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial t} dt) + O(dx, dt). \quad (5)$$

Then, we remove the high-order error term $O(dx, dt)$ to approximate Eq. 5 as follows:

$$I_t = -\nabla I \cdot U, \quad (6)$$

where U denotes optical flow, $-\nabla I \cdot U$ is the spatiotemporal gradient of current frame. I_t denotes temporal brightness change, which can be approximated as the accumulated events warped by optical flow in a certain time, namely $I_t = U \cdot \sum_{e_i \in \Delta t} p_i C$. Therefore, we take the spatiotemporal gradient consistency to constrain the fusion:

$$\mathcal{L}_{consis} = \sum ||U \cdot \sum_{e_i \in \Delta t} p_i C + \nabla I \cdot U||_1 \odot V / \sum V, \quad (7)$$

where V is the valid mask generated by optical flow U. The visual luminance fusion that obeys the basic motion model can guarantee the valid motion feature matching.

3.3. Visual Structure Fusion

LiDAR point cloud has the incomplete contour of the scene due to a non-uniform laser beam, limiting the scene flow performance. We aim to introduce the event to assist in completing the physical structure of the LiDAR point cloud. **Local Boundary *v.s.* Global Shape.** LiDAR emits uniform

laser beams in the horizontal direction, and the reflected laser can show the global shape of $360°$ scene. However, LiDAR just emits the non-uniform laser beams in the vertical direction due to the vertical angle. The larger the vertical angle, the fewer the number of the reflected point cloud, resulting in the incomplete contour of LiDAR. In contrast, event can sense the rich motion boundaries of the scene. Therefore, the local boundary of event and the global shape of LiDAR can jointly complete the intrinsic structure of the scene. To illustrate the complementary nature between LiDAR and event in the structure space, as shown in Fig. 3, we cluster the spatially-projected 2D LiDAR frame and the temporally-projected 2D event frame into the same structure manifold, and visualize the clustering feature distributions of LiDAR and event. First, the distribution trend of LiDAR is similar to the event. Second, the distribution of event boundary is continuous while the distribution of LiDAR shape is truncated, indicating that the contours of the 2D LiDAR frame are incomplete but those of the 2D event frame are relatively complete. This motivates us to take the scene structure as a homogeneous space to fuse the local boundary of event into the global shape of LiDAR for physical structure integrity.

Self-Similarity Clustering. Structure fusion mainly depends on a self-similarity clustering strategy, which is divided into two steps: intra-modal structure neighbor and inter-modal data fusion. The former is to properly cluster LiDAR shape points and event boundary points into the same neighbor. The latter is to map the local boundary coordinates of event to the global shape coordinates of LiDAR in the neighbor, thus completing the physical contour. Given the aligned LiDAR point cloud $pc = \{x_i, y_i, z_i | i \in M\}$, the corresponding event stream $ev = \{x_j, y_j, t_j, p_j | j \in N\}$, intrinsic parameter $K = \{f, c_x, c_y\}$, we first normalize the event stream along the timestamps to obtain the 2D coordinates of event $P_e = \{u_j = x_j, v_j = y_j, p_j | \{x_j, y_j, p_j\} \in ev\}$, and use the intrinsic parameter to project the LiDAR point cloud into the event camera coordinate system $P_l = \{u_i, v_i, d_i\}$:

$$u_i = f \cdot x_i/z_i + c_x, v_i = f \cdot y_i/z_i + c_y, d_i = z_i, \\ s.t.\{x_i, y_i, z_i\} \in pc, 0 < u_i < w, 0 < u_i < h, \quad (8)$$

where w, h is the size of event frame, f is focal length, c_x, c_y are x, y-axis offsets. In the intra-modal structure neighbor stage, we construct the distance function as the self-similarity measurement metric for 2D coordinates of the event:

$$d_p = \sqrt{(p_1 - p_2)^2}, d_s = \sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}, \\ D_e = \sqrt{d_p^2 + (d_s/N_s)^2}, s.t.\{u_{1,2}, v_{1,2}, p_{1,2}\} \in P_e, \quad (9)$$

where N_s is the max distance of 2D coordinates. We also define the same distance function D_l for the 2D coordinates of LiDAR. We iteratively optimize the distance functions D_e, D_l to obtain the neighbors U, which represent the region-level structure information like the image superpixel. In the inter-modal data fusion stage, we propose a similarity-based bi-direction fusion strategy: event→LiDAR structure fusion
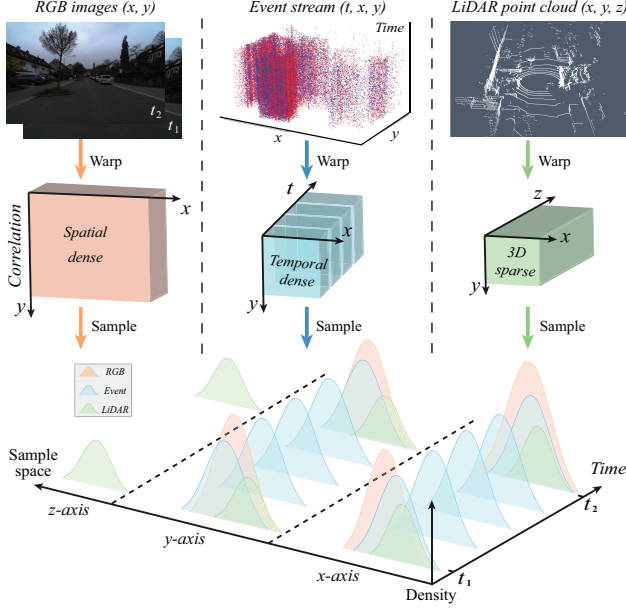
Figure 4. Correlation distributions of RGB, event and LiDAR. The three modalities have similar distributions in x, y-axis correlation, with z-axis correlation unique to LiDAR. RGB correlation is spatially dense, event is temporally dense, while LiDAR is spatiotemporally sparse. This inspires us to build the homogeneous correlation space to fuse the complementary motion knowledge.

and LiDAR→event depth fusion. As for the event→LiDAR structure fusion, we take the various 2D LiDAR coordinates of each neighbor as the center, and use the K-nearest neighbor to select the top k 2D event coordinates with the highest similarity and fill them into the 2D LiDAR coordinates in the same neighbor. As for the LiDAR→event depth fusion, we filter the top k 2D LiDAR coordinates in the same way, and weightedly fuse their depths to the selected event coordinates to obtain the final depth map. To ensure the back-propagation of the whole framework, we introduce a cross-attention transformer [31] to model the visual structure fusion process between event and LiDAR. We use the weighted-fused depth maps as the pseudo labels d_t^{pse}, $d_{t+\Delta t}^{pse}$ to train the cross-attention transformer:

$$\mathcal{L}_{pse} = \sum ||\tilde{d}_t - d_t^{pse}||_1 + ||\tilde{d}_{t+\Delta t} - d_{t+\Delta t}^{pse}||_1, \quad (10)$$

where $\tilde{d}_t$, $\tilde{d}_{t+\Delta t}$ are the fused adjacent depth maps estimated from the transformer network, and then are inversely transformed to the 3D point clouds $\tilde{pc}_t$, $\tilde{pc}_{t+\Delta t}$.

3.4. Motion Correlation Fusion

Visual luminance and structure fusions do ensure the visualization of RGB and LiDAR, while how to physically and complementarily fuse the mapped motion features of various modalities to improve scene flow is a challenge.

Spatiotemporal Dense v.s. Sparse Correlation. Motion space that scene flow relies on consists of two types: temporal correlation subspace [14] and decoded motion subspace

[15]. The former is to model the feature similarity between adjacent frames, and the latter contains the features decoded by recurrent network (*e.g.*, GRU [34]). Since the correlation subspace has more physical meaning, it has the potential to be a homogeneous space. As far as we know, RGB provides texture information with high spatial resolution, event provides boundary information with high temporal resolution, and LiDAR provides 3D (x, y, z) scene structure information. This inspires us to consider that, in the correlation space, RGB may promise the x, y-axis spatial-dense correlation, event may ensure the x, y-axis temporal-dense correlation and LiDAR may guarantee the x, y, z-axis accurate but sparse correlation. To verify our insight, as shown in Fig. 4, we transform the visual RGB, event and LiDAR features into the correlation subspace of motion space via warp operator [18], and calculate their distributions in x, y, z three dimensions, respectively. Note that, the correlation distribution of event is analyzed along the timestamps. On one hand, the correlation distribution trends of these three modalities are quite similar along the x and y axes, with z-axis correlation unique to LiDAR. On the other hand, in terms of the complementarity in the correlation space, RGB is spatial-dense, event is temporal-dense, and LiDAR is spatiotemporal-sparse. Therefore, we treat the correlation space as a homogeneous space to fuse the cross-modal knowledge for motion continuity.

Motion Distribution Alignment. Correlation fusion is first to fuse the dense-sparse complementary knowledge between various modalities in the single-axis correlation, and then merge the fused correlation distributions across all three axes. First, we voxelize the event stream into multiple-slice voxels. We encode the fused RGB images, the voxelized events and the fused LiDAR point clouds to the RGB correlation cv_r, the event correlation cv_e and the LiDAR correlation cv_l using the warp operator. Then, we randomly sample N 3D points from LiDAR. From these points, we obtain 2D coordinates for RGB and event with intrinsic parameters. Next, we use the sampled coordinates as the center, and perform 2D spatially-sampling for the RGB correlation to obtain the corresponding spatial-dense x, y-axis correlations cv_r^x, cv_r^y. As for the event, we temporally sample the correlation features to get the temporal-dense x, y-axis correlations cv_e^x, cv_e^y. For the LiDAR, we spatially sample the correlations into the relatively sparse x, y, z-axis correlations cv_l^x, cv_l^y, cv_l^z. Before the correlation fusion, we apply K-L divergence to align the multimodal motion correlation distributions:

$$\mathcal{L}_{corr}^{kl} = \sum \Phi(cv_l^{x,y}) \cdot log \frac{\Phi(cv_l^{x,y})}{\Phi(cv_r^{x,y})} + \Phi(cv_l^{x,y}) \cdot log \frac{\Phi(cv_l^{x,y})}{\Phi(cv_e^{x,y})}, \quad (11)$$

where Φ is the softmax function. We first fuse the aligned x, y-axis correlations between various modalities and then concatenate the z-axis correlations of LiDAR:

$$corr = Concat\{\frac{1}{T}\sum_{i=0}^{T}(cv_r^x + cv_e^{x,i} + cv_l^x)/3,$$
$$\frac{1}{T}\sum_{i=0}^{T}(cv_r^y + cv_e^{y,i} + cv_l^y)/3, cv_l^z\}, \quad (12)$$

where T denotes the number of the temporal slices of events.

Method	Scene flow methods						Optical flow methods			
	RAFT-3D	RAFT-3D w/ e	PV-RAFT	CamLiFlow	RPEFlow	**VisMoFlow**	SMURF	FlowFormer	RPEFlow	**VisMoFlow**
Input	RGB	RGB	PC	RGB+PC	RGB+PC+EV	**RGB+PC+EV**	RGB	RGB	RGB+PC+EV	**RGB+PC+EV**
Day EPE	0.095	–	0.055	0.033	0.060	**0.012**	2.01	0.607	0.556	**0.198**
Day ACC	73.48%	–	79.96%	91.40%	81.73%	**98.55%**	83.34%	89.08%	88.98%	**97.11%**
Night EPE	0.112	0.104	0.055	0.047	0.056	**0.027**	11.360	2.085	0.716	**0.353**
Night ACC	65.65%	72.65%	79.97%	85.62%	81.47%	**95.62%**	55.12%	77.05%	78.85%	**96.28%**

Table 1. Quantitative results on synthetic Event-KITTI dataset. "PC" is LiDAR point cloud, "EV" is event. "w/ e" is image enhancement.
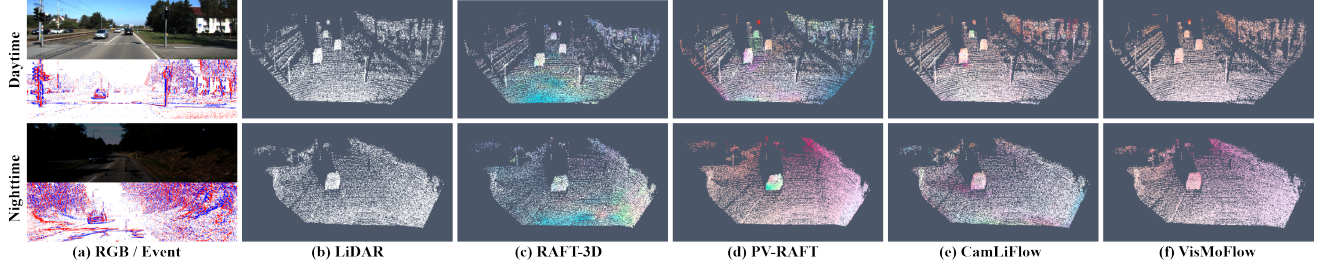

(a) RGB / Event   (b) LiDAR   (c) RAFT-3D   (d) PV-RAFT   (e) CamLiFlow   (f) VisMoFlow

Figure 5. Visual comparison of scene flows on synthetic Event-KITTI dataset.

Method	RAFT-3D	PV-RAFT	CamLiFlow	RPEFlow	**VisMoFlow**
Input	RGB	PC	RGB+PC	RGB+PC+EV	**RGB+PC+EV**
Day EPE	0.167	0.183	0.113	0.103	**0.084**
Day ACC	13.16%	37.28%	55.69%	60.81%	**70.34%**
Night EPE	0.359	0.190	0.125	0.094	**0.090**
Night ACC	5.04%	40.98%	53.10%	66.49%	**68.31%**

Table 2. Quantitative results on real DSEC dataset.

Furthermore, we introduce GRU to recursively decode the fused correlation features to estimate the final scene flow. In addition, we learn the scene flow and the corresponding optical flow with photometric loss [35]:

$$\mathcal{L}_{pho} = \sum \psi(\tilde{I}_t - w_{2D}(\tilde{I}_{t+\Delta t})) \odot V_{2D} / \sum V_{2D}$$
$$+ \sum \psi(\tilde{pc}_t - w_{3D}(\tilde{pc}_{t+\Delta t})) \odot V_{3D} / \sum V_{3D}, \quad (13)$$

where ψ is a L_p norm ($p = 0.4$). V_{2D}, V_{3D} is the 2D and 3D occlusion masks by checking forward-backward consistency [36]. We also apply the photometric loss to train the event motion feature encoder. The proposed motion correlation fusion can ensure 3D motion spatiotemporal continuity.

3.5. Optimization and Implementation Details

Consequently, the total objective for the proposed hierarchical fusion framework is written as follows:

$$\mathcal{L} = \mathcal{L}_{pho} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{consis} + \lambda_3 \mathcal{L}_{pse} + \lambda_4 \mathcal{L}_{corr}^{kl}, \quad (14)$$

where $[\lambda_1, ..., \lambda_4]$ are the weights that control the importance of the related losses. The first term is to maintain the capability of motion feature encoders of RGB, event and LiDAR. The second and third terms are to constrain the Event-RGB visual luminance fusion, the fourth term is to promise that the cross-attention transformer can learn to fuse the Event-LiDAR structure, and the intention of the last term aims to promote the RGB-Event-LiDAR motion correlation fusion. We set the KNN selection number k as 5, the event temporal slices T as 10, and the sample number N as 1000. During the training phase, we only need three steps.

First, we train Event-RGB visual luminance fusion for high dynamic imaging. Second, we train Event-LiDAR visual structure fusion for physical structure integrity. Finally, we train RGB-Event-LiDAR motion correlation fusion for 3D motion spatiotemporal continuity. Note that, the proposed method is first pre-trained to initialize the motion estimators in a supervised manner. During the testing phase, the whole framework consists of the luminance fusion network, the cross-attention transformer and motion estimation networks, thus achieving the end-to-end estimation of scene flow.

4. Experiments

4.1. Experiment Setup

Dataset. We conduct extensive experiments on one synthetic (*e.g.*, Event-KITTI) and one real (*e.g.*, DSEC) datasets, including daytime and nighttime scenes.

• **Event-KITTI.** The origin KITTI [37] is set in daytime scenes. We take the V2E model [38] to generate the event stream from daytime images, and use the noise model [39] to synthesize nighttime images corresponding to daytime.

• **DSEC.** DSEC is a multimodal dataset [40], which covers the multimodal data of RGB camera, LiDAR and event camera in various outdoor scenes and various times.

Comparison Methods. We choose optical flow and scene flow methods for comparison in both daytime and nighttime scenes. As for optical flow comparison, we choose unsupervised (SMURF [41]), supervised (FlowFormer [42]), and multimodal fusion (RPEFlow [22]) methods. As for scene flow comparison, we choose RGB-based unimodal (RAFT-3D [6]), LiDAR-based unimodal (PV-RAFT [7]), RGB-LiDAR multimodal fusion (CamLiFlow [9]) and RGB-Event-LiDAR multimodal fusion (RPEFlow) methods. Note that, in nighttime scenes, we add an additional comparison: first performing image enhancement (*e.g.*, KinD++ [43]) and then estimate motion on the enhanced results (named as "w/
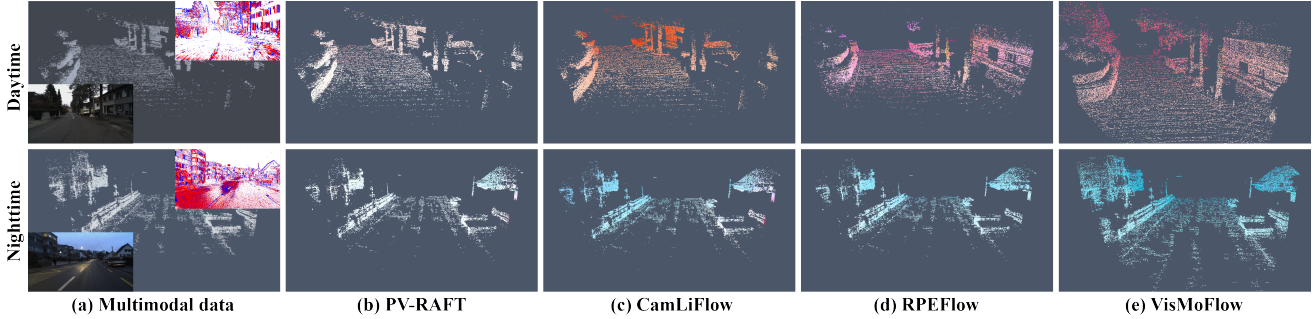
Figure 6. Visual comparison of scene flows on real DSEC dataset.

Fusion strategy	EPE	ACC
w/o any fusion	0.151	49.20%
w/ visual fusion	0.118	54.35%
w/ motion fusion	0.098	62.87%
w/ visual-motion fusion	**0.084**	**70.34%**

Table 3. Effectiveness of hierarchical visual-motion fusion.

Figure 7. Effectiveness of homogeneous spaces in fusion. (a) Input data. (b-d) Scene flow. "FC" is implicit feature concatenate. "L" is luminance space. "S" is structure space. "C" is correlation space.

e"). We choose the average end-point error (EPE [37]) and the percentage of flow accuracy within 1px for optical flow or 5cm for scene flow (ACC [9]) as the evaluation metrics.

4.2. Comparison Experiment

Comparison on Synthetic Dataset. In Table 1 and Fig. 5, we compare different motion methods on the synthetic Event-KITTI dataset. First, the proposed method significantly outperforms the competing methods in optical flow and scene flow. Second, in nighttime scenes, the performance of those methods that use RGB images decreases slightly. Note that, image enhancement can indeed promote the RGB-based motion methods in nighttime scenes, but the improvement is limited. Third, multimodal methods are more effective than unimodal methods in improving motion estimation due to the complementary knowledge between various modalities. Notably, the proposed visual-motion fusion surpasses the RPEFlow that relies solely on motion fusion.

Comparison on Real Scenes. In Table 2, we compare the scene flow methods in real daytime and nighttime scenes on the DSEC dataset. We have two conclusions. First, multimodal methods are far superior to unimodal methods. This is because the complementary knowledge of various modalities in real scenes can make up for their own disadvantage, conducive to deeply exploiting the characteristic of scene flow. Second, event camera can further effectively improve nighttime scene flow. The main reason is that, event camera have the advantage of high dynamic range to sense scene motion at night. In Fig. 6, our scene flow visualization results show more complete structures and perform better than other competing methods. The main reason is that, the visual structure fusion does fuse the local boundary of event into the global shape of LiDAR for point cloud physical integrity, thus improving the matching capability of motion features.
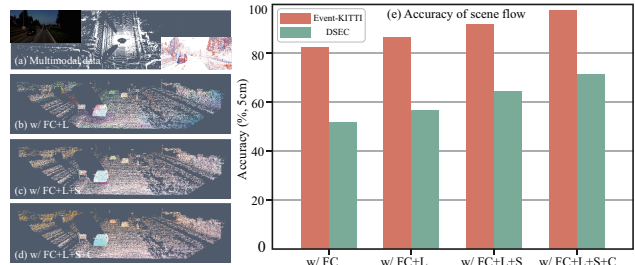
4.3. Ablation Study

How does Hierarchical Visual-Motion Fusion Work? In Table 3, we demonstrate the effectiveness of the proposed hierarchical visual-motion fusion framework. With only visual fusion, there is an improvement in scene flow. With only motion fusion, scene flow is significantly improved. When we apply visual fusion and motion fusion, the hierarchical fusion is superior to the motion fusion alone. Therefore, motion fusion is the key to promoting scene flow, and visual fusion can update the upper limit of scene flow performance.

The Role of homogeneous Space in Fusion. We illustrate the impact of different homogeneous spaces on scene flow in Fig. 7. When only implicit feature fusion (*i.e.*, feature concatenate (FC)), more errors of the scene flow are caused by feature mismatching. With visual luminance space (w/ FC+L), scene flow is improved due to the enhanced visual RGB features. Visual structure space (w/ FC+L+S) further promotes the integrity of the scene flow structure. With motion correlation space (w/ FC+L+S+C), the motion metric is greater, and the scene flow visualization is spatiotemporally smoother. Therefore, the homogeneous space can serve as a guide to physically facilitate the fusion process.

Effectiveness of Fusion Losses. In Table 4, we conduct ablation studies on the main fusion losses. Without any fusion losses, the whole framework still has the visual-motions fusion process. Each loss improves scene flow to a certain extent, while the motion distribution alignment loss $\mathcal{L}_{corr}^{kl}$ plays a key role in scene flow. The three losses can jointly promote the visual→motion progressive fusion process.

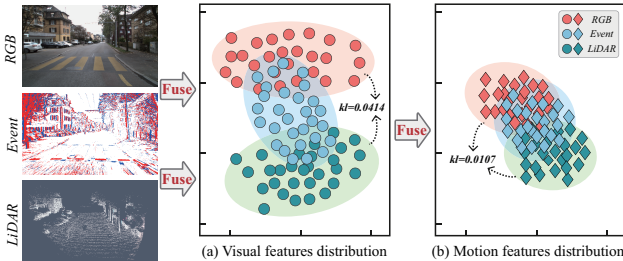| $\mathcal{L}_{consis}$	$\mathcal{L}_{pse}$	$\mathcal{L}_{corr}^{kl}$	EPE	ACC
×	×	×	0.122	53.17%
√	×	×	0.112	56.41%
×	√	×	0.107	58.25%
×	×	√	0.092	65.43%
√	√	√	**0.084**	**70.34%**

Table 4. Ablation study on fusion losses.

Figure 8. t-SNE visualization of visual and motion features. In (a) visual space, there exists a gap between RGB and LiDAR, while event serves as a bridge across the two modalities. In (b) motion space, the three modalities are further close to each other.

4.4. Discussion

How does Event Play the Bridge. In Fig. 8, we illustrate the importance of event as a bridge between RGB and LiDAR in visual and motion feature spaces via t-SNE. In the visual space, there exists a large gap between RGB and LiDAR features, while event features can just serve as a bridge across the two modalities. When the fused visual features are mapped to the motion space, the motion features of RGB and LiDAR are close to each other and have a certain intersection, which contains most of the event motion features. Therefore, the event can reasonably bridge the RGB and LiDAR.

Effectiveness of Spatiotemporal Gradient Consistency. We analyze the impact of spatiotemporal gradient consistency on the Event-RGB fused image result and the optical flow result from the training process in Fig. 9. Without the consistency loss, the training curve fluctuates violently, the fused RGB image looks clear while there are some artifacts in the optical flow. With the consistency loss, the fluctuation of the training curve is smoother, the fused RGB image is also clear and the optical flow boundaries become sharp. This shows that spatiotemporal gradient consistency does constrain the mapping relationship between the Event-RGB fused visual features and the motion features.

Why Structure Neighbor is Necessary? Structure neighbor is to extract the region-level structure from the pixel-level structure, which has the local self-similarity characteristics. In Table 5, without any neighbor, the scene flow is slightly limited. This is because the full pixel association between event and LiDAR increases the probability of local-global structure feature mismatching. With grid-based neighbor, the scene flow performance is improved but is lower than the clustering-based neighbor. The main reason is that, the

Structure neighbor strategy	EPE	ACC
w/o any neighbor	0.095	64.70%
w/ grid-based neighbor	0.088	69.05%
w/ clustering-based neighbor	**0.084**	**70.34%**

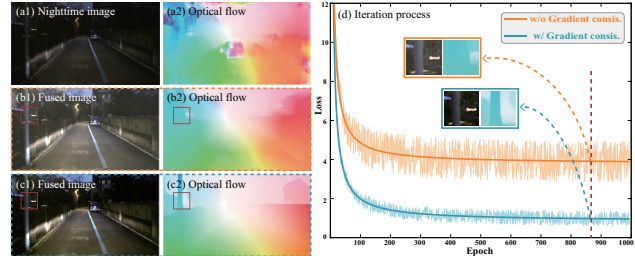Table 5. Choice of different structure neighbor strategies.

Figure 9. Effect of spatiotemporal gradient consistency on the fused image and optical flow. (a) Without fusion (b) Fusion without gradient consistency. (c) Fusion with gradient consistency. The gradient consistency can smooth training curve, conducive to constraining the mapping relationship between the fused image and optical flow.

grid-based neighbor may contain partial information of different objects, leading to the structure fusion bias, while the clustering-based neighbor can reduce this error.

Limitation. The proposed method can indeed cope with the all-day (*i.e.*, daytime and nighttime) scene flow estimation, but usually fails for adverse weather, such as rain and fog. Rain will lead to varying degrees of dynamic streak noise to RGB, event and LiDAR, and fog will attenuate atmospheric light, limiting the imaging of these sensors. In the future, we will further introduce the near-infrared sensor into these modalities to handle adverse weather scene flow.

5. Conclusion

In this work, we bring the event as a bridge between RGB and LiDAR, and propose a novel hierarchical visual-motion fusion framework to fuse the multimodal complementary knowledge in the homogeneous space for scene flow. To the best of our knowledge, we are the first to investigate the inter-modal homogeneous space to explicitly learn the complementary knowledge for physical interpretation. We construct the Event-RGB luminance space, Event-LiDAR structure space and RGB-Event-LiDAR correlation space. The three homogeneous spaces progressively improve scene flow from visual space to motion space. We also conduct extensive experiments on daytime and nighttime scenes, and verify the superiority of the proposed method. We believe that the proposed method could not only facilitate the development of scene flow but also enlighten the researchers of the broader field, *e.g.*, adverse scene multimodal perception.

References

[1] Fabian Duffhauss and Stefan A Baur. Pillarflownet: A real-time deep multitask network for lidar-based 3d object detection and scene flow estimation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10734–10741. IEEE, 2020. 1

[2] Stefan Andreas Baur, David Josef Emmerichs, Frank Moosmann, Peter Pinggera, Björn Ommer, and Andreas Geiger. Slim: Self-supervised lidar scene flow and motion segmentation. In *Int. Conf. Comput. Vis.*, pages 13126–13136, 2021. 1

[3] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Int. Conf. Comput. Vis.*, pages 2574–2583, 2017. 1, 2

[4] Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, and Raquel Urtasun. Deep rigid instance scene flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3614–3622, 2019.

[5] Junhwa Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2684–2694, 2021. 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, 6

[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. 1, 2, 6

[8] Rishav Rishav, Ramy Battrawy, René Schuster, Oliver Wasenmüller, and Didier Stricker. Deeplidarflow: A deep learning architecture for scene flow estimation using monocular camera and sparse lidar. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10460–10467. IEEE, 2020. 1, 2

[9] 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. 1, 2, 6, 7

[10] Guillermo Gallego, Tobi Delbrück, G. Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44:154–180, 2019. 2

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2462–2470, 2017. 2

[12] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12240–12249, 2019. 2

[13] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 529–537, 2019. 2

[14] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419, 2020. 2, 5

[15] 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. 2, 5

[16] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester. Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *Int. Conf. Comput. Vis.*, pages 2780–2790, 2019. 2

[17] Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, 7(2):3491–3498, 2022. 2

[18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8934–8943, 2018. 2, 5

[19] Yaqi Shen, Le Hui, Jin Xie, and Jian Yang. Self-supervised 3d scene flow estimation guided by superpoints. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5271–5280, 2023. 2

[20] Itai Lang, Dror Aiger, Forrester Cole, Shai Avidan, and Michael Rubinstein. Scoop: Self-supervised correspondence and optimization-based scene flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5281–5290, 2023. 2

[21] Fangqiang Ding, Andras Palffy, Dariu M Gavrila, and Chris Xiaoxuan Lu. Hidden gems: 4d radar scene flow learning using cross-modal supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9340–9349, 2023. 2

[22] 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, 6

[23] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Int. Conf. Comput. Vis.*, pages 4193–4202, 2017. 2

[24] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10941–10950, 2020.

[25] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Eur. Conf. Comput. Vis.*, pages 412–428. Springer, 2022.

[26] Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Objectfusion: Multi-modal 3d object detection with object-centric fusion. In *Int. Conf. Comput. Vis.*, pages 18067–18076, 2023. 2

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. 2

[28] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Int. Conf. on Machine Learning*, pages 1558–1567. PMLR, 2017.

[29] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Int. Conf. Comput. Vis.*, pages 4338–4347, 2021.

[30] Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, and Yuexin Ma. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *Int. Conf. Comput. Vis.*, pages 21674–21684, 2023. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 2, 5

[32] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. 45(7):8553–8565, 2023. 4

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2223–2232, 2017. 4

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 5

[35] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Eur. Conf. Comput. Vis.*, pages 3–10, 2016. 6

[36] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Eur. Conf. Comput. Vis.*, pages 1–18. Springer, 2018. 6

[37] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3061–3070, 2015. 6, 7

[38] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1312–1321, 2021. 6

[39] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6749–6757, 2020. 6

[40] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 6

[41] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3887–3896, 2021. 6

[42] 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. 6

[43] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM Int. Conf. Multimedia*, pages 1632–1640, 2019. 6