x^2-Fusion: Cross-Modality and Cross-Dimension Flow Estimation in Event Edge Space

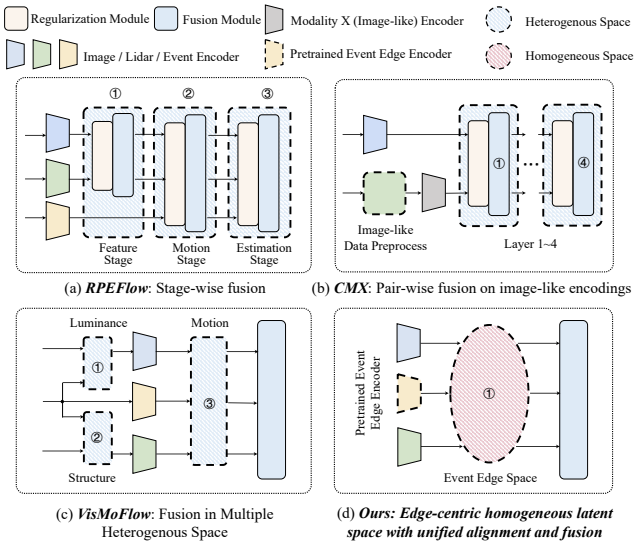
Ruishan Guo[1,*], Ciyu Ruan[1,*], Haoyang Wang[1,*], Zihang Gong[2], Jingao Xu[3], Xinlei Chen[1,†]
[1]Shenzhen International Graduate School, Tsinghua University, [2]Harbin Institute of Technology,
[3]The University of Hong Kong
{grs24, rcy23, haoyang-22}@mails.tsinghua.edu.cn, gongzihang0201@gmail.com
jingaoxu@hku.hk, chen.xinlei@sz.tsinghua.edu.cn

Abstract

Estimating dense 2D optical flow and 3D scene flow is essential for dynamic scene understanding. Recent work combines images, LiDAR, and event data to jointly predict 2D and 3D motion, yet most approaches operate in separate heterogeneous feature spaces. Without a shared latent space that all modalities can align to, these systems rely on multiple modality-specific blocks, leaving cross-sensor mismatches unresolved and making fusion unnecessarily complex. Event cameras naturally provide a spatiotemporal edge signal, which we can treat as an intrinsic edge field to anchor a unified latent representation, termed the ***Event Edge Space****. Building on this idea, we introduce* ***x^2-Fusion****, which reframes multimodal fusion as representation unification: event-derived spatiotemporal edges define an edge-centric homogeneous space, and image and LiDAR features are explicitly aligned in this shared representation. Within this space, we perform reliability-aware adaptive fusion to estimate modality reliability and emphasize stable cues under degradation. We further employ cross-dimension contrast learning to tightly couple 2D optical flow with 3D scene flow. Extensive experiments on both synthetic and real benchmarks show that x^2-Fusion achieves state-of-the-art accuracy under standard conditions and delivers substantial improvements in challenging scenarios.*

1. Introduction

Optical and scene flow estimate dense 2D and 3D motion correspondences across images, depth maps, and point clouds, forming a core tool for dynamic scene understanding in autonomous driving, tracking, and 3D reconstruction [1–3]. Recent works fuse image, LiDAR, and events via advanced architectures, coupling dense photometry with accurate geometry to exploit complementarity and outperform single-modality baselines [4–7].



(a) ***RPEFlow***: Stage-wise fusion (b) ***CMX***: Pair-wise fusion on image-like encodings

(c) ***VisMoFlow***: Fusion in Multiple Heterogenous Space (d) ***Ours: Edge-centric homogeneous latent space with unified alignment and fusion***

Figure 1. **Fusion paradigms for multimodal perception.** (a) RPEFlow [8]: stage-wise fusion across 2D/3D spaces. (b) CMX [9]:image-like encodings with pairwise rectification/attention at each backbone stage. (c) VisMoFlow [10]: separate luminance/structure/correlation spaces with dedicated modules. (d) Ours: preserves native domains and learns an *Event Edge Space*, a shared edge-centric latent space guided by a frozen event teacher, that aligns image, LiDAR, and events before fusion.

Despite these advances, existing multimodal flow estimation works keep each modality in its native format (images as 2D grids, LiDAR as point clouds, events as asynchronous streams) and fuse them spatially across separate, *heterogeneous* feature spaces. This brings three issues:
(i) High Complexity: Without a shared channel-wise basis, fusion requires pairwise alignment between every modality pair, leading to module-heavy architectures with stage-wise fusion blocks [8], pair-wise rectification/attention units [9], or multiple hand-crafted physical spaces [10] (Fig. 1(a-c)). This makes models cumbersome, difficult to train, and hard to scale to additional modalities.
(ii) Information Erosion: Processing features in sepa-

[*]Equal Contribution. [†]Corresponding author.

rate heterogeneous spaces delays fusion to late stages, where signals have already been corrupted by early-stage modality-specific distortions that become difficult to correct through cross-modal interaction.

(iii) High Fragility: Without a common representational foundation, modalities cannot provide stable priors for one another. Under perception degradations such as exposure extremes, LiDAR sparsity, or motion blur, the alignment itself breaks down, causing catastrophic failures [11–14]. Overall, these methods do not align all sensors into a *homogeneous* channel-wise latent space, preventing simple, robust, and efficient cross-modal interaction.

Multimodel Fusion in Homogeneous Event Edge Space. We address this limitation by introducing ***Event Edge Space***, the first homogeneous latent space that unifies image, LiDAR, and event representations in shared edge-centric domain. This design is motivated by two insights:

• *Why Edge?* Edges capture object boundaries and scene discontinuities in a modality-agnostic manner. They represent consistent structural information across sensors, irrespective of appearance variations, sampling density differences, or sensor-specific noise patterns [15–18]. An edge-centric latent space thus provides a natural common language for aligning heterogeneous inputs before interaction.

• *Why Edge using Event?* Event camera records per-pixel brightness changes at ultra-high temporal resolution, firing precisely where strong image gradients persist under motion, i.e., on moving edges [19, 20]. Spatial aggregation traces pixel-aligned edge curves; temporal aggregation yields continuous edge trajectories, forming a natural spatiotemporal edge signal. Moreover, events share 2D pixel coordinates with images, while their asynchronous sparse activations mirror LiDAR's irregular spatiotemporal sampling [21, 22]. This dual correspondence—geometric alignment with image and structural similarity with LiDAR—makes events an ideal anchor for a truly homogeneous edge-centric space.

To translate this insight into a practical architecture, we introduce three key technical components. First, we pretrain an ***Event Edge Encoder*** to distill stable, motion-aware edge embeddings, then freeze it and use its features as edge prototypes to *symmetrically regularize* RGB and LiDAR encoders, aligning both modalities into the shared Event Edge Space. Second, within this homogeneous space, ***Reliability-aware Adaptive Fusion*** estimates global and local reliability from spatiotemporal cues and performs cross-attention to produce unified 2D/3D features. Third, ***Cross-dimension Contrastive Learning*** explicitly enforces inter-frame motion coherence and 2D–3D geometric consistency, enabling mutual reinforcement between optical flow and scene flow tasks to fully exploit complementary cues. Extensive experiments on both synthetic and real-world benchmarks demonstrate that our approach achieves state-of-the-art performance under normal conditions. Moreover, it significantly outperforms existing methods under challenging scenarios such as extreme lighting and LiDAR sparsity. Our main contributions are summarized as follows:

- To our knowledge, we introduce Event Edge Space, the first edge-centric homogeneous space unifying image, LiDAR, and events representations in a common domain.
- Within the homogeneous space, we couple reliability-aware adaptive fusion with cross-dimension contrastive learning, enabling consistent 2D-3D flow estimation.
- We achieve state-of-the-art performance on both synthetic and real-world benchmarks, demonstrating the effectiveness of our approach across diverse scenarios.

2. Related Work

2.1. Multimodel Fusion Paradigms

Multimodal sensing enhances scene understanding by combining dense appearance, precise geometry, and fine-grained dynamics [23–28]. In recent 2D/3D motion tasks, coupling photometric cues with 3D structure [8–10, 29] significantly outperforms single-modality baselines [30–35].

However, most prior work models pairwise interactions across heterogeneous feature spaces via bespoke fusion blocks: RPEFlow [8] adopts stage-wise attention and mutual-information regularization at the feature, motion, and estimation stages, requiring dedicated fusion for each branch and stage. VisMoFlow [10] attempts to mitigate modality heterogeneity by constructing three hand-crafted physical spaces:luminance, structure, correlation. Although these are described as *homogeneous spaces*, the three domains are processed by distinct modules (luminance fusion, structural cross-attention, correlation alignment) and rely on different physical assumptions. Consequently, interaction occurs sequentially across these domains rather than within a unified channel-wise latent space, and the overall fusion mechanism remains multi-stage and module-intensive. Unified RGB–X methods like CMX [9] convert all inputs to image-like encodings and fuse via shared blocks. While providing a common 2D representation, this grid-based confinement overlooks modality-native properties such as 3D topology and asynchronous event dynamics.

In contrast, we embed images, LiDAR, and events into a homogeneous edge-centric space. This enables fusion via lightweight weighting and unified cross-attention, replacing stacked modality-wise modules to a simpler, robust design.

2.2. Cross-modality and Cross-dimension Learning

Existing cross-modality learning methods primarily focus on: 1) modeling inter-modal relationships [36–40] and 2) designing effective fusion strategies [41–49].

For relationship modeling, co-learning approaches such as cross-modal distillation [38] learn shared representations

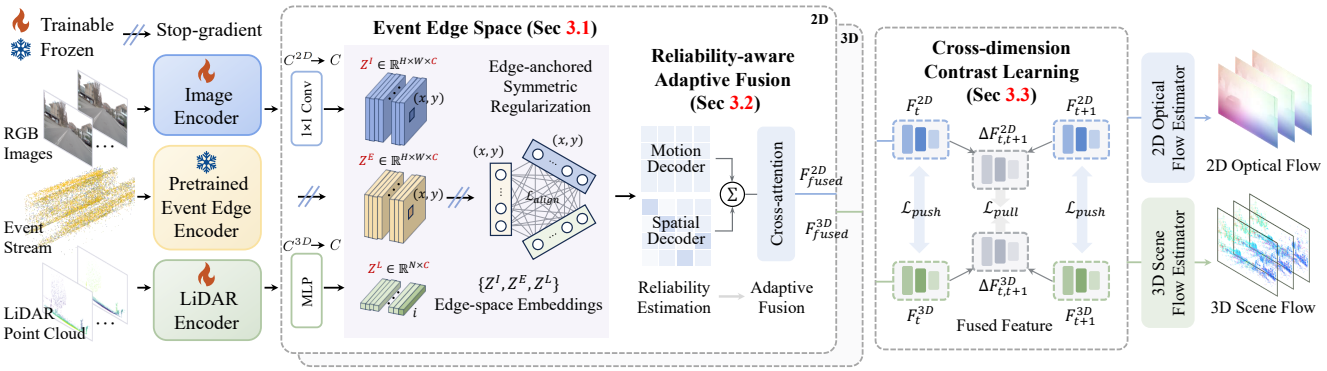


Figure 2. **Overview of x^2-Fusion.** Given image, events and LiDAR, we first pretrain an *Event Edge Encoder* to distill motion-aware edge features. We then freeze the encoder and use its embeddings as edge prototypes to *symmetrically regularize* channel-wise representations across modalities, aligning them into a shared ***Event Edge Space*** (Sec. 3.1). Within this space, ***Reliability-aware Adaptive Fusion*** estimates global and local reliability and fuses modalities via a cross-attention block to produce 2D/3D features (Sec. 3.2). Finally, ***Cross-dimension Contrast Learning*** enforces inter-frame coherence and 2D–3D consistency, and the task heads output optical and scene flow (Sec. 3.3).

but are prone to failure propagation when any sensor degrades. Modality dropout [39] improves robustness by simulating sensor failures in training. Knowledge-guided methods inject domain priors (e.g., geometric constraints [40]) but often lack adaptability to unseen scenarios.

Fusion strategies generally fall into two categories: attention-based approaches [42, 43], which dynamically weight modalities via gating or cross-attention [44], and representation-learning methods [45, 46], which align feature spaces by enforcing embedding consistency through contrastive objectives[47], canonical correlation[48], or mutual information regularization [49].

We propose a unified fusion framework on a shared, edge-centric latent space. Within this space, cross-modality adaptive attention and cross-dimension contrastive learning combine to enable robust multimodal integration.

3. Method

Overview. To enable robust optical and scene flow estimation under all-day, dynamic conditions, we propose a unified framework named x^2**-Fusion**, as illustrated in Fig. 2, that uses stable scene-edge representations to unify image, events and LiDAR in a shared edge-centric latent space, and then performs reliability-aware adaptive fusion with joint 2D–3D motion representation learning.

For robust multimodal fusion, we first embed image, LiDAR, and events into a homogeneous ***Event Edge Space***, where frozen event embeddings provide edge-centric prototypes to align modality features. Within this shared space, ***Reliability-aware Adaptive Fusion*** estimates modality reliability from stable edge motion and structural cues and drives a single cross-attention block, yielding resilient 2D/3D representations under partial sensor degradations.

For joint estimation, the ***Cross-dimension Contrast Learning*** module jointly optimizes 2D and 3D motion representations by enforcing intra-frame geometric discrimination and inter-frame consistency, enabling cross-dimension synergy that surpasses independent learning.

3.1. Event Edge Space

We propose the ***Event Edge Space*** to unify multimodal features based on the intrinsic geometric commonality of motion edges across all three modalities. Leveraging the fact that event streams are inherently composed of fine-grained spatiotemporal edge points, we exploit their temporally dense and edge-salient nature to construct this homogeneous latent space. It serves as a unified feature domain, anchored by the semantics of motion edges.

To faithfully preserve fine-grained spatiotemporal edge cues within this space, we first explicitly pretrain an event edge encoder (Sec. 3.1.1) to distill stable, motion-aware edge representations from raw events. We then freeze the event encoder, treating its features as Edge Prototypes to learn bidirectional mappings I $\Leftrightarrow$ E and L $\Leftrightarrow$ E for aligning image and LiDAR features (Sec. 3.1.2).

3.1.1. Event Edge Encoder Pretraining

We process event stream by first voxelizing it in space-time. The resulting voxel tensor is then fed into a sparse 3D convolutional neural network, which outputs a multi-scale event feature pyramid $\{F^E\}$, where $F_s^E \in \mathbb{R}^{H_s \times W_s \times C_s}$ denotes the event edge feature map at scale s.

To explicitly bias F_s^E towards edge-aware representations, we define an event edge strength from raw events. For each pixel (x, y) in a window, we measure the normalized event activity $\tilde{A}^E(x, y)$ and the normalized temporal variance $\tilde{\sigma}_t(x, y)$ of its events, and combine them into

$$e^E(x, y) = \tilde{A}^E(x, y)\,\big(1 - \tilde{\sigma}_t(x, y)\big) \in [0, 1], \qquad (1)$$

where larger e^E indicates stronger and more temporally coherent motion edges. Scale-specific edge maps e_s^E are ob-

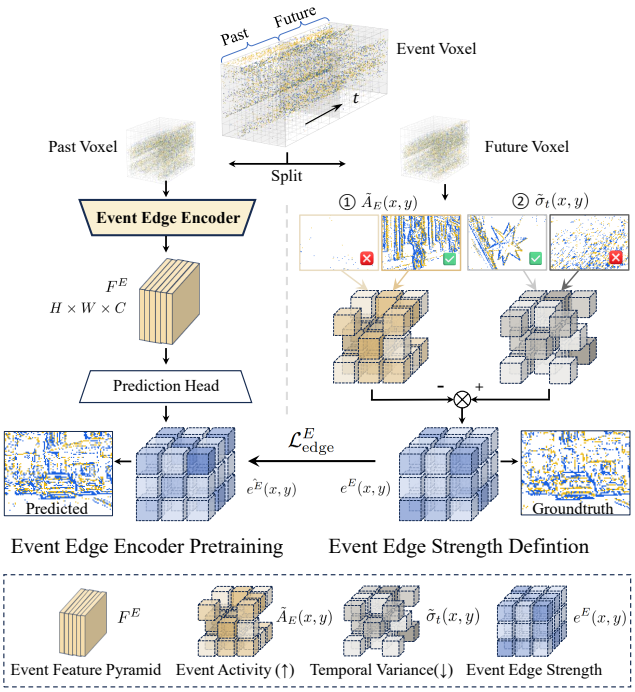


Figure 3. Our proposed event edge encoder pretraining learns explicit, high-fidelity motion-aware edge representations by predicting edge strength from voxelized event streams.

tained by spatial pooling of e_E to the resolution (H_s, W_s). The definitions of $\tilde{A}^E(x,y)$ and $\tilde{\sigma}_t(x,y)$ are in the supplementary material.

We pretrain the event encoder self-supervisedly by predicting future edge strength from past events. Specifically, for each temporal window, we split it into two halves: the first half is voxelized as the encoder input, and the latter provide the edge groundtruth. The encoder consumes the past voxels, generating $F_s^{E,\text{past}}$, on top of which a lightweight prediction head g_s regresses $e_s^{E,\text{future}}$ with scale weights λ_s:

$$\mathcal{L}^E_{\text{edge}} = \sum_s \lambda_s \left\| g_s\big(F_s^{E,\text{past}}\big) - e_s^{E,\text{future}} \right\|_1. \quad (2)$$

This pretraining step distills stable, motion-aware edge features into the multi-scale event pyramid $\{F_s^E\}$, which subsequently define the event-guided prototype space used for Image–Event–LiDAR alignment in Sec. 3.1.2.

3.1.2. Image–LiDAR Alignment in Event Edge Space

Image and LiDAR encoders. We employ modality-specific backbones to extract multi-scale features from images and LiDAR point clouds. The image Encoder processes image frames and derived cues to produce a 2D feature pyramid $F^I \in \mathbb{R}^{H\times W\times C^{2D}}$. The LiDAR Encoder processes consecutive point clouds, augmenting point-wise features with geometric and depth descriptors, yielding 3D point-wise feature pyramids $F^L \in \mathbb{R}^{N\times C^{3D}}$. Full architectural details and input augmentations are provided in the supplementary material.

Projection. To map all modalities into the shared Event Edge Space, we attach lightweight projection heads (h_s^I, h_s^L) to the Image (F^I) and LiDAR (F^L) feature pyramids. These heads, implemented as a 1×1 convolution and a shared MLP respectively, project features channel-wise into the common embedding dimension C_s:

$$Z_s^I = h_s^I(F_s^I), \quad Z_s^L = h_s^L(F_s^L), \quad Z_s^E \equiv F_s^E. \quad (3)$$

In this homogeneous space Z_s, all features share the same dimension C_s and are used for alignment. Crucially, the pretrained event features Z_s^E are treated as fixed edge prototypes; we freeze the event encoder and stop the gradient at Z_s^E during multimodal training.

Edge-anchored symmetric regularization. We employ the frozen event embeddings Z^E as stable edge prototypes to symmetrically regularize image (Z^I) and LiDAR (Z^L) features in the homogeneous space. This process effectively transfers the refined edge semantics to both the 2D (optical flow) and 3D (scene flow) task branches.

For the 2D Branch (pixel-wise optical flow), features are aligned on the image plane. We follow an interpolation strategy (based on Bi-CLFM [23]) to obtain the pixel-wise LiDAR embeddings $Z^{L,2D} \in \mathbb{R}^{H\times W\times C}$ from Z^L, thus aligning $Z^{L,2D}$ with the 2D grid features Z^I and Z^E.

For the 3D Branch (point-wise scene flow), features are aligned in the point cloud domain. We sample the image (Z^I) and event (Z^E) embeddings at the projected point locations to obtain point-wise features $Z^{I,3D}, Z^{E,3D} \in \mathbb{R}^{N\times C}$. Similarly, the event edge map e^E is sampled to yield point-wise weights $e^{E,3D}(i)$.

We measure per-location alignment by the ℓ_1 distance as

$$D_s^{2/3D}(p) = \sum_{(m,n)\in\{I,E,L\}} \left\| Z_s^m(p) - Z_s^n(p) \right\|_1, \quad (4)$$

where p denotes either a 2D pixel (x,y) or a 3D point i, yielding $D_s^{2D}(x,y)$ and $D_s^{3D}(i)$, respectively. Note that we stop the gradient at $Z_s^E(p)$ to fix edge prototypes.

The weighted 2D and 3D alignment losses are then defined using the event edge map e^E as weighting prior:

$$\mathcal{L}^{2/3D}_{\text{align}} = \sum_s \sum_p e_s^{E/\{E,3D\}}(p)\, D_s^{2/3D}(p), \quad (5)$$

and the total alignment loss is

$$\mathcal{L}_{\text{align}} = \lambda_{2D} \cdot \mathcal{L}^{2D}_{\text{align}} + \lambda_{3D} \cdot \mathcal{L}^{3D}_{\text{align}}. \quad (6)$$

3.2. Reliability-aware Adaptive Fusion

Benefiting from the tri-modal alignment in the Event Edge Space, we fuse features directly within this shared latent domain, eliminating the need for bespoke modality-specific interaction modules. We leverage event data's strengths, temporal fine-grained information and high dynamic range,

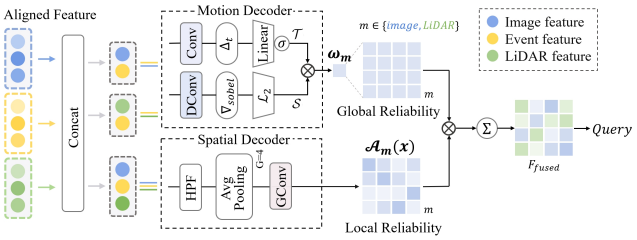


Figure 4. The proposed reliability-aware adaptive fusion module adaptively integrates image, LiDAR, and event features through hierarchical reliability weighting and cross-modal attention.

to achieve robust adaptive fusion. Our method employs a symmetric dual-branch decoder (2D/3D), where each branch utilizes an reliability-aware adaptive fusion module to dynamically weight all three modalities based on sensor reliability for robust fusion.

Adaptive fusion. Leveraging the homogeneous embeddings $\{Z_s^I, Z_s^L, Z_s^E\}$, we utilize two lightweight feature decomposition streams: a motion stream aggregating features along temporal dimension to capture global motion consistency, and a spatial stream aggregating features within local neighborhoods to capture local structural agreement.

To facilitate adaptive integration, we estimate a global reliability score ω_m for each modality $m \in \{\mathrm{I}, \mathrm{L}\}$ by measuring its consistency with the event-derived motion signal. Concretely, this score is derived via a spatiotemporal decomposition applied to the concatenated feature map $\hat{Z} = [Z_M, Z_E]$:

$$\mathcal{T}(\hat{Z}) = \sigma(\mathbb{L}(\Delta_t(\mathrm{Conv}(\hat{Z})))), \mathcal{S}(\hat{Z}) = \|\nabla(\mathrm{DConv}(\hat{Z}))\|_2. \tag{7}$$

Here, $\mathcal{T}(\cdot)$ captures fine-grained temporal changes using feature differencing Δ_t, followed by convolution and a linear projection layer $\mathbb{L}$; $\mathcal{S}(\cdot)$ encodes spatial structure via dilated convolutions and gradient magnitude. The resultant global reliability score is computed as:

$$\omega_m = \mathrm{softmax}_m((\mathcal{T} \otimes \mathcal{S})\hat{Z}). \tag{8}$$

To further account for local spatial variations in reliability, we design a modality-specific attention mechanism that processes the tri-modal feature volume $\tilde{Z} = [Z_I, Z_L, Z_E]$ via a lightweight network:

$$\mathcal{A}_m(x) = \mathrm{softmax}((\mathcal{H} \oplus \mathcal{P} \oplus \mathcal{G})\tilde{Z})_m, \tag{9}$$

where $\mathcal{H}$, $\mathcal{P}$, and $\mathcal{G}$ denote high-pass filtering (HPF), average pooling, and grouped 1×1 convolutions, respectively. The output $\mathcal{A}_m(x)$ reflects the spatially-varying reliability of modality m at location x.

Finally, the fused feature $F_{fused}(x)$ is computed as a weighted sum over modalities. By combining the global reliability score and local attention, this mechanism adaptively emphasizes key modalities at both scales to enhance reliability of the model.

$$F_{fused}(x) = \sum_{m \in \mathcal{M}} \frac{\omega_m \mathcal{A}_m(x)}{\sum_n \omega_n \mathcal{A}_n(x)} Z_m(x). \tag{10}$$

Cross-attention transformer. To enhance multimodal interaction, we introduce a cross-attention transformer for flow estimation. For 2D and 3D branch, the module takes fused features F_{fused} as queries and projects auxiliary features $[Z^{L,2D}, Z^E]$ and $[Z^{I,3D}, Z^{E,3D}]$ via learnable MLPs.

The cross-attention is implemented using a transformer block, where queries, keys, and values are computed as:

$$Q = W_q F_{fused}, \quad K = W_k F_{aux}, \quad V = W_v F_{aux}. \tag{11}$$

Here, W_q, W_k, and W_v are implemented as 1×1 convolutions for 2D and linear layers for 3D, enabling modality- and domain-specific adaptation. The attention output is given by:

$$\mathrm{Attention}(Q, K, V) = V \cdot \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{12}$$

where d is the feature dimension. The attended features are refined via a feed-forward MLP and combined with the input using residual connections:

$$F_{out}^{2/3D} = \mathrm{MLP}(\mathrm{Attention}(Q, K, V)^{2/3D}) + F_{fused}. \tag{13}$$

3.3. Cross-dimension Contrast Learning (CCL)

Inspired by mutual information optimization in multimodal fusion, we extend to cross-temporal and cross-task constraints. Specifically, we regularize the 2D and 3D features from cross-attention module, enhancing their intra-frame distinctiveness and inter-frame consistency.

Cross-temporal contrast (pull). To capture inter-frame dynamics, we compute temporal differences of 2D and 3D fused features. As 3D features lie in a different spatial domain, we project them to 2D space to obtain F_{proj}^{3D}. Temporal motion vectors M^{2D} and M^{3D} are then computed via global average pooling:

$$M^{2/3D} = \mathcal{P}(\Delta F_{out}^{2/3D}). \tag{14}$$

A cosine similarity loss encourages geometric alignment between normalized motion embeddings:

$$\mathcal{L}_{pull} = 1 - \frac{\langle \phi(M^{2D}), \psi(M_{proj}^{3D}) \rangle}{\|\phi(M^{2D})\|_2 \cdot \|\psi(M_{proj}^{3D})\|_2}, \tag{15}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote normalization functions.

Cross-task contrast (push). To encourage complementary intra-frame representations, we minimize the mutual information between 2D and 3D features via variational encoding. Each modality is encoded into a latent distribution:

$$\begin{aligned} q_\phi(\mathbf{z}^{2D}|F^{2D}) &= \mathcal{N}(\boldsymbol{\mu}^{2D}, \boldsymbol{\sigma}^{2D}), \\ q_\psi(\mathbf{z}^{3D}|F_{proj}^{3D}) &= \mathcal{N}(\boldsymbol{\mu}^{3D}, \boldsymbol{\sigma}^{3D}), \end{aligned} \tag{16}$$

Modality	Method	#Params(M)	EKubric						DSEC					
			EPE_{2D} ↓	ACC_{1px} ↑	Fl ↓	EPE_{3D} ↓	$ACC_{.05}$ ↑	$ACC_{.10}$ ↑	EPE_{2D} ↓	ACC_{1px} ↑	Fl ↓	EPE_{3D} ↓	$ACC_{.05}$ ↑	$ACC_{.10}$ ↑
Img	RAFT [29]	5.3M	0.838	93.31%	2.36%	-	-	-	0.586	88.98%	1.47%	-	-	-
Img	FlowFormer [4]	16.2M	0.702	92.58%	2.07%	-	-	-	0.567	89.82%	1.33%	-	-	-
PC(Point Cloud)	PV-RAFT [50]	239.2K	-	-	-	0.093	82.42%	92.60%	-	-	-	0.190	32.74%	55.62%
EV	E-RAFT [51]	5.3M	-	-	-	-	-	-	0.481	91.75%	1.31%	-	-	-
Img+Depth	RAFT-3D [52]	7.7M	0.714	94.39%	-	0.049	92.88%	-	0.572	89.55%	-	0.144	49.30%	-
Img+PC	CamLiFlow [23]	7.7M	0.770	95.11%	1.80%	0.035	94.90%	95.86%	0.399	94.94%	1.33%	0.129	51.08%	68.17%
Img+EV	Diff-ABFlow [53]	17.2M	-	-	-	-	-	-	1.460	50.00%	7.43%	-	-	-
Img+EV	DCEIFlow [54]	7.1M	3.109	56.37%	18.40%	-	-	-	0.970	73.01%	4.90%	-	-	-
Img+EV	STFlow [55]	9.2M	-	-	-	-	-	-	0.630	92.07%	1.45%	-	-	-
Img+PC+EV	†VisMoFlow [10]	-	-	-	-	0.026	95.98%	96.77%	-	-	-	0.100	61.78%	75.82%
Img+PC+EV	RPEFlow [8]	9.8M	0.439	95.99%	1.48%	0.027	95.33%	96.32%	0.326	95.28%	1.15%	0.103	60.80%	74.97%
Img+PC+EV	x^2-Fusion (Ours)	8.2M	**0.430**	**96.86%**	**1.43%**	**0.024**	**96.78%**	**97.62%**	**0.322**	**95.80%**	**1.13%**	**0.094**	**64.39%**	**78.27%**
			EKubric-Img (Under-Exposure Degradation)						DSEC-Img (Under-Exposure Degradation)					
Img+PC+EV	RPEFlow [8]		3.663	33.96%	27.06%	0.043	89.01%	92.27%	0.817	83.24%	5.06%	0.117	54.97%	69.94%
Img+PC+EV	x^2-Fusion (Ours)		**1.143**	**68.87%**	**9.11%**	**0.033**	**94.14%**	**95.31%**	**0.394**	**92.73%**	**1.14%**	**0.107**	**58.87%**	**71.43%**
Absolute Improvement (vs. RPEFlow)			(↓ 2.520)	(↑ 34.91%)	(↓ 17.95%)	(↓ 0.010)	(↑ 5.13%)	(↑ 3.04%)	(↓ 0.423)	(↑ 9.49%)	(↓ 3.92%)	(↓ 0.010)	(↑ 3.90%)	(↑ 1.49%)
			EKubric-Img (Over-Exposure Degradation)						DSEC-Img (Over-Exposure Degradation)					
Img+PC+EV	RPEFlow [8]		2.801	53.75%	17.63%	0.039	91.32%	95.85%	0.565	89.59%	2.73%	0.109	54.91%	70.02%
Img+PC+EV	x^2-Fusion (Ours)		**0.994**	**80.19%**	**7.05%**	**0.032**	**93.86%**	**95.93%**	**0.376**	**93.03%**	**1.13%**	**0.107**	**58.67%**	**71.09%**
Absolute Improvement (vs. RPEFlow)			(↓ 1.807)	(↑ 26.44%)	(↓ 10.58%)	(↓ 0.007)	(↑ 2.54%)	(↑ 0.08%)	(↓ 0.189)	(↑ 3.44%)	(↓ 1.6%)	(↓ 0.002)	(↑ 3.76%)	(↑ 1.07%)
			EKubric-PC (Sparse Degradation)						DSEC-PC (Sparse Degradation)					
Img+PC+EV	RPEFlow [8]		0.569	94.62%	2.02%	0.027	95.37%	96.34%	0.493	90.36%	1.19%	0.056	81.18%	88.69%
Img+PC+EV	x^2-Fusion (Ours)		**0.439**	**95.93%**	**1.49%**	**0.022**	**96.99%**	**97.43%**	**0.331**	**94.49%**	**1.18%**	**0.047**	**86.40%**	**90.31%**
Absolute Improvement (vs. RPEFlow)			(↓ 0.130)	(↑ 1.31%)	(↓ 0.53%)	(↓ 0.005)	(↑ 1.62%)	(↑ 1.09%)	(↓ 0.162)	(↑ 4.13%)	(↓ 0.01%)	(↓ 0.009)	(↑ 5.22%)	(↑ 1.62%)
			EKubric-PC (Drifting Degradation)						DSEC-PC (Drifting Degradation)					
Img+PC+EV	RPEFlow [8]		1.164	62.44%	9.92%	0.113	66.47%	73.76%	1.051	68.05%	3.82%	0.457	0.51%	2.46%
Img+PC+EV	x^2-Fusion (Ours)		**0.763**	**88.52%**	**6.06%**	**0.088**	**79.34%**	**82.80%**	**0.409**	**91.39%**	**2.13%**	**0.285**	**19.36%**	**28.87%**
Absolute Improvement (vs. RPEFlow)			(↓ 0.401)	(↑ 26.08%)	(↓ 3.86%)	(↓ 0.025)	(↑ 12.77%)	(↑ 9.04%)	(↓ 0.642)	(↑ 23.34%)	(↓ 1.69%)	(↓ 0.172)	(↑ 18.85%)	(↑ 26.41%)

Table 1. Quantitative evaluation on both standard and degraded scenarios of EKubric and DSEC datasets. † Reproduced per original methodology (code unavailable). Extended results are provided in the supplementary material.

where $\mathbf{z}^{2D}$ and $\mathbf{z}^{3D}$ represent the latent embeddings sampled from these distributions.

The mutual information is approximated by a symmetric binary cross-entropy(BCE) loss over sampled latent pairs:

$$\mathcal{L}_{push} = \frac{1}{2} \sum_{t \in \{t_1, t_2\}} \text{BCE}(\sigma(\mathbf{z}_t^{2D}), \sigma(\mathbf{z}_t^{3D})). \quad (17)$$

The total contrastive loss is a weighted sum of the above:

$$\mathcal{L}_{contra} = \mathcal{L}_{pull} + \gamma \cdot \mathcal{L}_{push}. \quad (18)$$

3.4. Total Objective Function

Our training objective is a weighted sum of three losses: flow supervision, feature stabilization, and cross-modal alignment.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{contra}} \mathcal{L}_{\text{contra}}. \quad (19)$$

Task loss ($\mathcal{L}_{\text{task}}$). For joint flow estimation, we adopt PWC-style coarse-to-fine supervision, upsampling and warping the level-l estimate to initialize level $l-1$.

$$\mathcal{L}_{\text{task}} = \sum_{2/3D} \lambda_{2/3D} \sum_l \omega_l^{2/3D} \left\| \mathbf{f}_{\text{pred},l}^{2/3D} - \mathbf{f}_{\text{gt},l}^{2/3D} \right\|_2. \quad (20)$$

Alignment loss ($\mathcal{L}_{\text{align}}$). Defined in Sec. 3.1.2, this loss ensures Image and LiDAR features are consistently aligned to the fixed event prototypes in the homogeneous space.

Contrastive loss ($\mathcal{L}_{\text{contra}}$). Introduced in Sec. 3.3, this auxiliary term regularizes the 2D and 3D features, enhancing their inter-frame consistency ($\mathcal{L}_{\text{pull}}$) and intra-frame complementarity ($\mathcal{L}_{\text{push}}$).

The event encoder is initially optimized using the self-supervised edge loss $\mathcal{L}_{\text{edge}}^{E}$ (Sec. 3.1.1) and is subsequently frozen during the optimization of $\mathcal{L}_{\text{total}}$.

4. Evaluation

4.1. Experiments Setups

Datasets and metrics. We conduct extensive experiments on both ***synthetic*** (EKubric) and ***real-world*** (DSEC) dataset. EKubric provides 15,367 annotated Image-LiDAR-Event triplets, while DSEC captures diverse urban driving scenarios with complex traffic elements. To stress-test robustness under realistic extreme conditions, we synthesize four degradations (low/high-exposure image and sparse/drifting LiDAR) using empirically grounded noise models[56, 57], producing artifacts that mirror common degradation modes. For quantitative evaluation, we em-

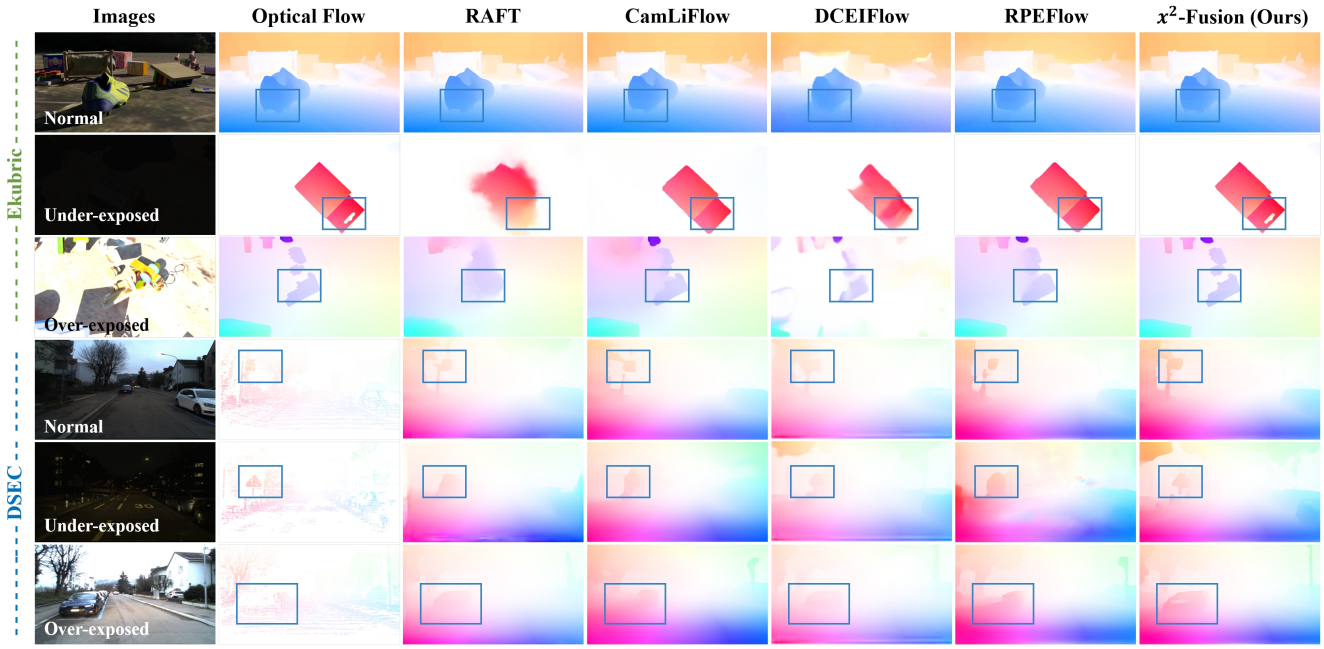


Figure 5. Visual comparison of optical flows on EKubric and DSEC dataset. x^2-Fusion achieves the state-of-the-art performance across various exposure degradation scenarios with clearer motion boundaries and finer details. Please zoom in for details.

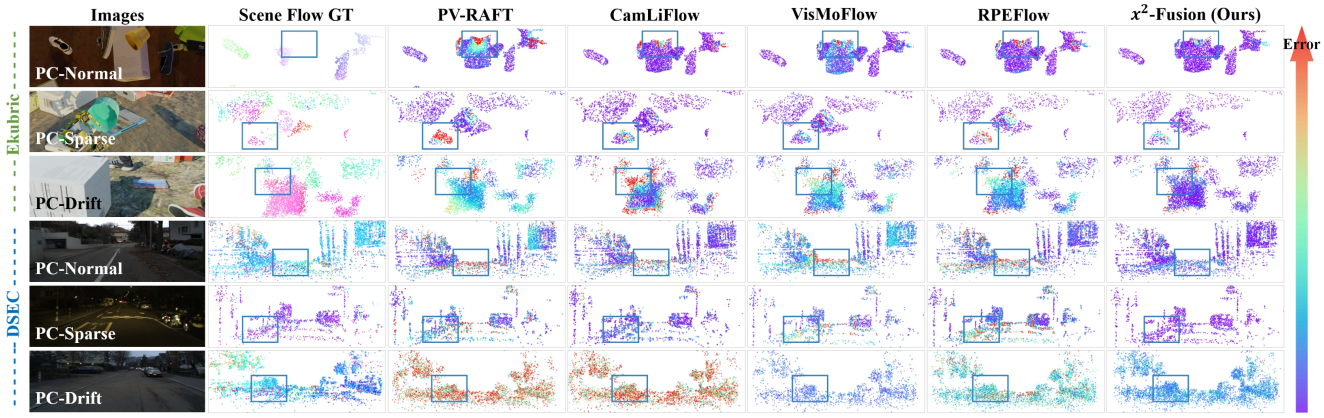


Figure 6. Visual comparison of scene flows on EKubric and DSEC dataset. x^2-Fusion demonstrates achieves lower end-point errors across diverse degradation scenarios. Comprehensive results and analyses are provided in the supplementary material. Please zoom in for details.

ploy EPE_{2D} (End-Point Error), ACC_{1px} (accuracy within 1 pixel), and Fl (outlier rate with $\text{EPE}_{2D}>$ 3px & error>5%) for 2D Flow estimation; and EPE_{3D}, and $\text{ACC}_{.05}$/$\text{ACC}_{.10}$ (accuracy within 5/10 cm) for 3D Flow estimation.

Implementation details. Our model is implemented in PyTorch, trained on four NVIDIA RTX A6000 GPUs and evaluated on a single GPU. We use the Adam optimizer with an initial learning rate of 10^{-4}, weight decay of 10^{-6}, and batch size of 8, using MultiStepLR scheduling with $0.5\times$ decay at specified milestones. All experiments leverage synchronized batch normalization and mixed-precision training for efficient optimization.

4.2. Comparison Experiment

In Tab. 1 and Fig. 5–6, we evaluate methods on synthetic and real-world datasets under normal and degraded conditions (e.g., extreme exposure, LiDAR corruption). First, x^2-Fusion achieves state-of-the-art optical and scene flow estimation, consistently improving EPE and accuracy metrics over recent baselines. Second, it remains highly robust against various degradations (including under/over-exposure and sparse/drifting point clouds), effectively mitigating partial data failure. Third, unified tri-modal fusion surpasses bi/uni-modal variants, with the largest gains in degraded conditions, proving that complementary modalities provide more stable and reliable motion references.

Settings	$EPE_{2D}\downarrow$	$EPE_{3D}\downarrow$
w/o Event Edge Space	0.491	0.146
w/o Edge-anchored Regularization	0.393	0.119
w/o Event Edge Encoder	0.378	0.114
w/ Event Edge Space	**0.322**	**0.094**

Table 2. Effects of the Event Edge Space.

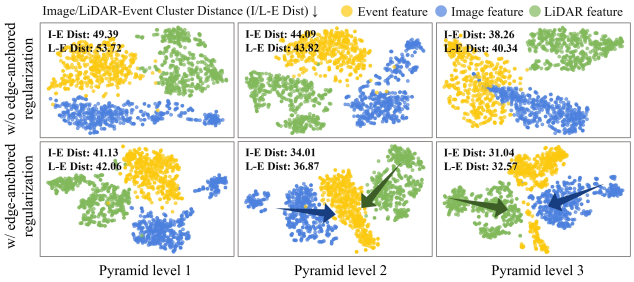


Figure 7. t-SNE visualization of cross-modal feature alignment. The edge-anchored symmetric regularization reduces tri-model feature divergence through pyramid-level alignment.

4.3. Ablation Study

How does Event Edge Space work? In Tab. 2 and Fig. 7, we demonstrate the effectiveness of the proposed Event Edge Space. All variants are trained on the DSEC training split and evaluated on its validation split. When the Event Edge Space (EES) is removed(w/o EES), cross-attention must simultaneously repair heterogeneous feature geometries and model motion. As shown in Tab. 2, performance drops on both optical and scene flow, indicating that an explicit edge-centric latent space offloads alignment and allows the decoder to focus on motion modeling over already-aligned embeddings. When the symmetric regularization is ablated(w/o Reg), accuracy improves over w/o EES yet remains below the full model; t-SNE in Fig. 7 shows that the regularizer pulls image/LiDAR embeddings toward event-defined clusters, forming tighter cross-modal neighborhoods and providing better-aligned inputs. When the event edge encoder is left unpretrained(w/o Edge), results lie between w/o EES and the full model, suggesting that an event pathway already helps unify the space, while reliability-aware edge pretraining is crucial for a meaningful prototype and for transferring edge semantics. Overall, edge prototypes provide a fixed alignment anchor and symmetric regularization pulls image/LiDAR toward it, producing better-aligned inputs.

Why does cross-dimension learning and joint task outperform independent tasks? To further assess the necessity of cross-dimensional learning and joint task estimation, we compare models trained with separate supervision for optical and scene flow (Model#A) in Tab. 3 and Fig. 8. Although all models incorporate three-modal inputs, the joint estimation variant (Model#B) achieves better performance by leveraging the inherent correlations between 2D and 3D branches. Building upon this, our model integrates early-stage event-guided fusion, which facilitates cross-dimensional interactions (Model#C) while boosting estimation accuracy. The t-SNE visualization in Fig. 8 reveals that our framework enhances feature representation in two complementary ways: it yields more compact intra-class distributions to improve task-specific discriminability, while preserving inter-class relationships that support consistent scene understanding across dimensions. These improvements are further supported by optical and scene flow results, where reduced error demonstrate the benefits of unified estimation and cross-dimensional learning.

Variants	Setting	$EPE_{2D}\downarrow$	$EPE_{3D}\downarrow$
Model#A	Indep. 2D + 3D	0.404	0.119
Model#B	Joint 2D & 3D	0.386	0.113
Model#C	Joint 2D & 3D + CCL	**0.325**	**0.103**

Table 3. Effects of cross-dimension learning and joint task.

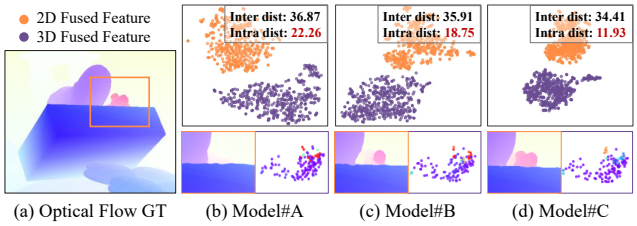


Figure 8. Qualitative analysis of joint task performance and contrastive learning effects. (Top) t-SNE visualization of fused 2D-3D feature distributions; (Bottom) Visualization of (left) 2D optical flow and (right) 3D scene flow error maps.

5. Conclusion

We introduce ***Event Edge Space(EES)***, the first edge-centric homogeneous latent space that unifies images, LiDAR, and events in a common feature domain. Building on it, we present x^2***-Fusion*** , which reframes multimodal fusion as representation unification. Within this shared edge-centric space, x^2-Fusion couples *reliability-aware adaptive fusion* and *cross-dimension contrastive learning*, promoting task-specific discrimination and enforcing 2D–3D consistency. On both synthetic and real-world benchmarks, our method achieves state-of-the-art flow accuracy and remains robust under challenging conditions. Beyond flow estimation, this framework offers broader insights for tasks such as text–image–video fusion and general cross-domain feature alignment, where bridging heterogeneous data is critical.

Although our study focuses on the image–LiDAR–event setting for flow estimation, EES is modality-agnostic. In principle, any modality with stable edge cues can be unified in it. In future work, we will extend it to additional sensors and validate generality beyond image–LiDAR–event.

Acknowledgments This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Low-Altitude Airspace Strategic Program Portfolio (Grant No. Z25306110), Shenzhen Science and Technology Program (Grant No. ZDCYKCX20250901094203005 and No. ZDCY202517012) and Meituan Academy of Robotics Shenzhen.

References

[1] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 1

[2] Haoyang Wang, Jingao Xu, Xinyu Luo, Xuecheng Chen, Ting Zhang, Ruiyang Duan, Yunhao Liu, and Xinlei Chen. Ultra-high-frequency harmony: mmwave radar and event camera orchestrate accurate drone landing. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 15–29, 2025.

[3] Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, and Min Chen. Flownet3d++: Geometric losses for deep scene flow estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 91–98, 2020. 1

[4] 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 *European conference on computer vision*, pages 668–685. Springer, 2022. 1, 6

[5] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[6] Jiehao Luo, Jintao Cheng, Xiaoyu Tang, Qingwen Zhang, Bohuan Xue, and Rui Fan. Mambaflow: A novel and flow-guided state space model for scene flow estimation. *arXiv preprint arXiv:2502.16907*, 2025.

[7] Ciyu Ruan, Ruishan Guo, Zihang Gong, Jingao Xu, Wenhan Yang, and Xinlei Chen. Pre-mamba: A 4d state space model for ultra-high-frequent event camera deraining. *arXiv preprint arXiv:2505.05307*, 2025. 1

[8] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10030–10040, 2023. 1, 2, 6

[9] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023. 1, 2

[10] Hanyu Zhou, Yi Chang, and Zhiwei Shi. Bring event into rgb and lidar: Hierarchical visual-motion fusion for scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26477–26486, 2024. 1, 2, 6

[11] Haoyang Wang, Jingao Xu, Xinyu Luo, Ting Zhang, Xuecheng Chen, Ruiyang Duan, Yunhao Liu, Jianfeng Zheng, Weijie Hong, Xiaoqiang Ji, et al. mme-loc: Facilitating accurate drone landing with ultra-high-frequency localization. *IEEE Transactions on Mobile Computing*, 2026. 2

[12] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36:21298–21342, 2023.

[13] Xinyu Luo, Haoyang Wang, Ciyu Ruan, Chenxin Liang, Jingao Xu, and Xinlei Chen. Eventtracker: 3d localization and tracking of high-speed object with event and depth fusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1974–1979, 2024.

[14] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In *European Conference on Computer Vision*, pages 179–195. Springer, 2022. 2

[15] Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pages 190–205. Springer, 2024. 2

[16] Yuhan Liu, Yongjian Deng, Hao Chen, Bochen Xie, Youfu Li, and Zhen Yang. Event-based video frame interpolation with edge guided motion refinement. *arXiv preprint arXiv:2404.18156*, 2024.

[17] Ciyu Ruan, Zihang Gong, Ruishan Guo, Jingao Xu, and Xinlei Chen. Edmamba: A simple yet effective event denoising method with state space model. *arXiv preprint arXiv:2505.05391*, 2025.

[18] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3571–3579, 2022. 2

[19] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[20] Haoyang Wang, Ruishan Guo, Pengtao Ma, Ciyu Ruan, Xinyu Luo, Wenhua Ding, Tianyang Zhong, Jingao Xu, Yunhao Liu, and Xinlei Chen. Event camera meets mobile embodied perception: abstraction, algorithm, acceleration, application. *ACM Computing Surveys*, 58(8):1–41, 2026. 2

[21] Yiheng Li, Hongyang Li, Zehao Huang, Hong Chang, and Naiyan Wang. Sparsefusion: Efficient sparse multi-modal fusion framework for long-range 3d perception. *arXiv preprint arXiv:2403.10036*, 2024. 2

[22] Haoyang Wang, Ruishan Guo, Pengtao Ma, Ciyu Ruan, Xinyu Luo, Wenhua Ding, Tianyang Zhong, Jingao Xu, Yunhao Liu, and Xinlei Chen. Towards mobile sensing with

event cameras on high-agility resource-constrained devices: A survey. *arXiv e-prints*, pages arXiv–2503, 2025. 2

[23] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5801, 2022. 2, 4, 6

[24] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17182–17191, 2022.

[25] Shuang Guo, Friedhelm Hamann, and Guillermo Gallego. Unsupervised joint learning of optical flow and intensity with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2025.

[26] Haoyang Wang, Xinyu Luo, Wenhua Ding, Jingao Xu, Xuecheng Chen, Ruiyang Duan, Jialong Chen, Haitao Zhang, Yunhao Liu, and Xinlei Chen. Enabling high-frequency cross-modality visual positioning service for accurate drone landing. *arXiv preprint arXiv:2510.00646*, 2025.

[27] Zhexiong Wan, Bin Fan, Le Hui, Yuchao Dai, and Gim Hee Lee. Instance-level moving object segmentation from a single image with events. *International Journal of Computer Vision*, 133(7):4042–4063, 2025.

[28] Wenhua Ding, Zhengli Zhang, Xin Xu, Haoyang Wang, Yinan Zhu, Shilong Ji, Xin Zhou, Jingao Xu, Dongyue Huang, and Xinlei Chen. Hawkeye: Practical in-flight obstacle avoidance with event camera and lidar fusion. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*, pages 1213–1215, 2025. 2

[29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2, 6

[30] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 2

[31] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3254–3263, 2019.

[32] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in Neural Information Processing Systems*, 34:7838–7851, 2021.

[33] Pengjie Zhang, Lin Zhu, Xiao Wang, Lizhi Wang, and Hua Huang. Ematch: A unified framework for event-based optical flow and stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5845–5855, 2025.

[34] Xinglong Luo, Ao Luo, Kunming Luo, Zhengning Wang, Ping Tan, Bing Zeng, and Shuaicheng Liu. Learning efficient meshflow and optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[35] Gokul Raju Govinda Raju, Nikola Zubic, Marco Cannici, and Davide Scaramuzza. Perturbed state space feature encoders for optical flow with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5111–5120, 2025. 2

[36] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021. 2

[37] Hao Chen, Feihong Shen, Ding Ding, Yongjian Deng, and Chao Li. Disentangled cross-modal transformer for rgb-d salient object detection and beyond. *IEEE Transactions on Image Processing*, 33:1699–1709, 2024.

[38] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18268–18278, 2023. 2

[39] Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 378–386, 2020. 3

[40] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1477–1485, 2023. 2, 3

[41] Weichen Zhang, Ruiying Peng, Chen Gao, Jianjie Fang, Xin Zeng, Kaiyuan Li, Ziyou Wang, Jinqiang Cui, Xin Wang, Xinlei Chen, et al. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*, 2025. 2

[42] 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 *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017. 3

[43] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 3

[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[45] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pages 1558–1567. PMLR, 2017. 3

[46] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018. 3

[47] Honggu Zhou, Xiaogang Peng, Yikai Luo, and Zizhao Wu. Pointcmc: cross-modal multi-scale correspondences learning for point cloud understanding. *Multimedia Systems*, 30(3):138, 2024. 3

[48] Qi Lyu and Xiao Fu. Finite-sample analysis of deep cca-based unsupervised post-nonlinear multimodal learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9568–9574, 2022. 3

[49] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2, 3

[50] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6954–6963, 2021. 6

[51] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 6

[52] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8375–8384, 2021. 6

[53] Haonan Wang, Hanyu Zhou, Haoyue Liu, and Luxin Yan. Injecting frame-event complementary fusion into diffusion for optical flow in challenging scenes. *arXiv preprint arXiv:2510.10577*, 2025. 6

[54] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing*, 31:7237–7251, 2022. 6

[55] Qianang Zhou, Junhui Hou, Meiyi Yang, Yongjian Deng, Youfu Li, and Junlin Xiong. Spatially-guided temporal aggregation for robust event-rgb optical flow estimation. *IEEE Transactions on Multimedia*, 2026. 6

[56] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024. 6

[57] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023. 6