

Bidirectional Cross-Modal Prompting for Event-Frame Asymmetric Stereo

Ninghui Xu^{1,2,†}Fabio Tosi²Lihui Wang^{1,*}Jiawei Han^{2,4}Luca Bartolomei²Zhiting Yao³Matteo Poggi²Stefano Mattoccia²

¹School of Instrument Science and Engineering, Southeast University,
State Key Lab of Comprehensive PNT Network and Equipment Technology,
Key Lab of Micro-Inertial Instrument and Advanced Navigation Technology, MOE

²Department of Computer Science and Engineering, University of Bologna

³College of Computer Science and Software Engineering, Hohai University

⁴Beijing Institute of Technology

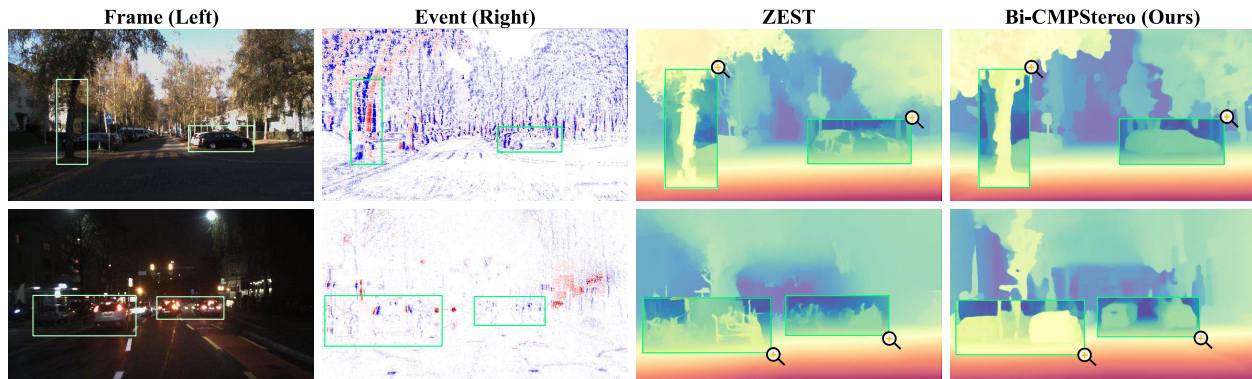


Figure 1. **Qualitative comparison on event-frame asymmetric stereo.** Compared to the state-of-the-art ZEST [41], our method achieves higher accuracy and higher-quality structural details in both complex-texture (first row) and sparse low-light scenes (second row).

Abstract

Conventional frame-based cameras capture rich contextual information but suffer from limited temporal resolution and motion blur in dynamic scenes. Event cameras offer an alternative visual representation with higher dynamic range free from such limitations. The complementary characteristics of the two modalities make event-frame asymmetric stereo promising for reliable 3D perception under fast motion and challenging illumination. However, the modality gap often leads to marginalization of domain-specific cues essential for cross-modal stereo matching. In this paper, we introduce Bi-CMPStereo, a novel bidirectional cross-modal prompting framework that fully exploits semantic and structural features from both domains for robust matching. Our approach learns finely aligned stereo representations within a target canonical space and integrates complementary representations by projecting each modality into both event and frame domains. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods in accuracy and generalization.

1. Introduction

Stereo matching aims to establish pixel-wise correspondences between stereo images, which is a fundamental technique in computer vision to compute dense disparity maps for depth estimation [55], with critical applications in robotics, autonomous driving, and AR/VR. Deep learning has driven remarkable progress for conventional RGB camera stereo in recent years [32, 52, 61], with iterative refinement-based methods demonstrating particular prominence [33, 39, 66, 68].

As novel bio-inspired neuromorphic sensors, event cameras asynchronously detect per-pixel illumination changes (i.e., events) with microsecond temporal resolution [37], embodying a paradigmatic shift from traditional visual modalities [72]. By rapidly capturing the dynamic information, event-based sensors exhibit compelling advantages over frame-based vision: extremely low latency ($\leq 3 \mu\text{s}$ vs. 30 ms) and high dynamic range ($\geq 120 \text{ dB}$ vs. 60 dB) [19, 73]. These properties have motivated their use in stereo matching for highly-dynamic scenarios with challenging illumination or high-speed motion [22, 48, 62, 84].

However, the inherent sparsity of events, which responds exclusively to temporal intensity variations, limits symmet-

[†] Work done while visiting the University of Bologna

*Corresponding author

ric event stereo in extracting sufficient contextual information for dense disparity estimation [1, 13]. Moreover, the high cost of neuromorphic sensors further limits the deployment of dual-event systems in practical settings. Consequently, conventional frame-based cameras, offering rich contextual information and cost efficiency [47], are complementary to event sensors, motivating growing interest in the stereo setups combining an event camera with a standard RGB sensor [9, 15, 18, 38, 41, 65, 79]. Such stereo configurations with heterogeneous binocular inputs are defined as asymmetric systems, since they do not support a fully symmetric processing pipeline for the two views.

Despite the potential benefits of cross-modal asymmetric stereo, the significant modality gap between events and frames undermines the feature-space alignment assumption underlying stereo models to infer cross-view feature similarities. Several attempts have been made to mitigate the discrepancy through either 1) domain-level alignment that unifies events and frames into a common representation for a Siamese feature extractor [30, 41, 65] or 2) feature-level alignment employing separate extractors for each domain to derive shared embeddings [15, 18, 38]. However, focusing on cross-modal commonalities can marginalize discriminative domain-specific features that are salient in one domain but sparsely captured in the other. For instance, color cues readily available in images require complex networks to be extracted from events, making them prone to marginalization during alignment. This trade-off limits these models from achieving performance competitive with event symmetric counterparts using homogeneous binocular inputs. Therefore, the key challenge is learning expressive representations without information-lossy marginalization.

In this paper, we address this challenge by proposing Cross-Modal Prompting Stereo (CMPStereo) network, a novel network that preserves domain-specific semantic and structural cues crucial for robust stereo matching. Our key idea is to alternately designate the asymmetric stereo inputs as the target-domain and source-domain modalities, using the target domain as the canonical space for feature alignment. To this end, we introduce a Stereo Canonicalization Constraint (SCC) that regularizes the network to learn target-domain discriminative features from both modalities, promoting high-fidelity cross-modal alignment. A Cross-Domain Embedding Adapter (CDEA) further reinforces target-domain cues weakly encoded in source-domain representations. Through two complementary CMPStereo variants with events and images as target domains, we present Bidirectional Cross-Modal Prompting (Bi-CMPStereo) framework that simultaneously exploits bidirectional cost volumes across domains to achieve robust asymmetric stereo. To prevent shortcut learning in con-

text features, we adopt Hierarchical Visual Transformation (HVT) [75] for enhanced robustness and generalization.

Fig. 1 shows the compelling performance of our proposals, and our contributions can be summarized as follows:

- We propose CMPStereo with SCC and CDEA to preserve domain-specific features and achieve high-fidelity cross-modal alignment in a target canonical space.
- We introduce Bi-CMPStereo, a bidirection framework integrating complementary information from event and image domains to enable robust disparity estimation.
- Our method significantly outperforms state-of-the-art approaches in both accuracy and generalization on DSEC and MVSEC benchmarks.

2. Related Work

Frame-based Stereo. Stereo matching has evolved from classical hand-crafted methods [55] to deep learning paradigms [24, 32, 52, 61]. Early CNN-based networks [42, 78] replaced individual pipeline components, while end-to-end architectures diverged into 2D [36, 43, 70, 77] and 3D approaches [7, 29, 56, 80] processing cost volumes through 3D convolutions. RAFT-Stereo [39] pioneered iterative refinement with correlation volume-based updates, followed by subsequent works [10, 33, 63, 68, 69, 86]. Transformer-based architectures [26, 35, 45, 71] leverage self-attention for long-range dependencies, while [25] propose data-driven MRF models. To improve generalization, various strategies have been explored, including domain-invariant feature learning [8, 17, 40, 81, 83], geometric priors [2, 60], and self-supervised learning with photometric losses [23, 50, 64] or pseudo-labels [51, 58, 59]. Recently, foundation stereo models [5, 11, 28, 66, 76] combine pre-trained monocular depth with stereo matching, achieving superior generalization through large-scale diverse training.

Event-based Stereo. Symmetric event stereo leverages binocular event cameras to estimate disparity. Early studies [62, 84] modeled the asynchronous nature of events through temporal convolutions, with [84] introducing continuous and discrete time operations to capture intrinsic dynamics. Recent works exploit the high temporal resolution of events via temporal aggregation: [48] learns from past and future events to produce compact representations, while [16] introduces stereoscopic flow to continuously exploit information across timesteps, and [85] unifies event-based flow and stereo in a shared correspondence space. However, event sparsity in static or low-texture regions hinders dense disparity estimation. To overcome this, [1, 21] reconstruct image features from events to estimate dense disparity maps, while others transfer knowledge from image to event domains through unsupervised adaptation [14] or cross-sensor distillation [82]. Complementary efforts integrate additional sensing modalities, such as LiDAR [3] or active projection [34]. Despite notable progress, symmetric

Code is available at: <https://github.com/xnh97/Bi-CMPStereo>.

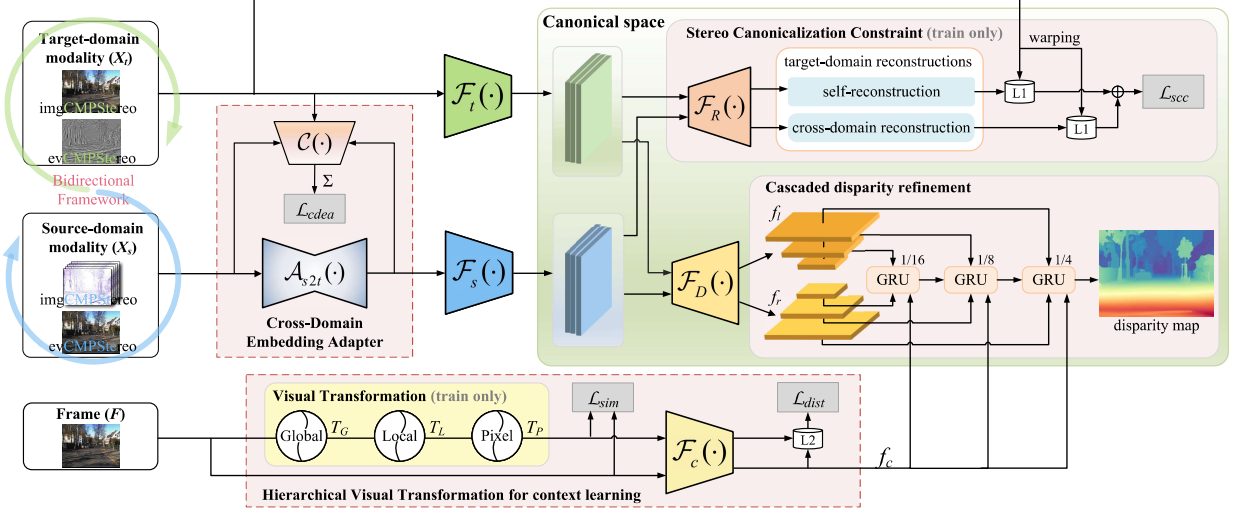


Figure 2. **Overview of the proposed CMPStereo network.** In our bidirectional framework, the asymmetric stereo inputs are alternately designated as the target-domain and source-domain modalities. The Cross-Domain Embedding Adapter (CDEA) operates on the source domain to achieve initial source-to-target adaptation. CMPStereo employs domain-specific encoders $\mathcal{F}_t(\cdot)$ and $\mathcal{F}_s(\cdot)$ combined with Stereo Canonicalization Constraint (SCC) to map the source and target domains onto a canonical space of the target domain. The multi-scale stereo features produced by a shared decoder $\mathcal{F}_D(\cdot)$ are fed to cascaded ConvGRUs for iterative disparity refinement. The context is extracted from the image frame F with the Hierarchical Visual Transformation (HVT) [75] to avoid shortcut-learning.

event stereo remains constrained by inherent sparsity, motivating asymmetric setups that combine events with frames.

Cross-Modal Event-Frame Stereo. Cross-modal stereo leverages the complementary strengths of events and frames. Event-frame fusion stereo employs binocular pairs of both modalities, using sequential fusion [47], differentiable event selection with cross-similarity [13], and cross-modality propagation [12] to complement sparse event features with dense frame information. However, requiring both modalities per view increases complexity and cost. Asymmetric stereo alleviates this by deploying an event camera in one view and a frame camera in the other, but the cross-view modality gap significantly intensifies matching challenges. Early handcrafted approaches [30, 65] rely on edge-based matching, while [9] decomposes the problem via temporal fusion and [79] proposes a two-stage learning framework for event-frame association. More recently, Zhuang *et al.* [89] address the domain gap by gating frame features with temporal cues from frame sequences, and [41] achieves zero-shot generalization through visual prompting. However, existing strategies risk marginalizing discriminative domain-specific cues essential for stereo matching. Other recent works exploit event-frame disparity as an intermediate representation for tasks such as video interpolation [18] and motion deblurring [15, 38].

3. Method Overview

We aim to estimate disparity from a calibrated event-frame asymmetric stereo setup. To address the inherent perceptual gap between the two sensing modalities, we introduce

Bi-CMPStereo, a bidirectional framework that enforces semantic and structural consistency across domains. At its core, CMPStereo learns aligned stereo representations by mapping both target modality X_t and source modality X_s into a canonical space of the target domain for disparity estimation. To exploit complementary cues, we instantiate this architecture in two configurations that alternate the domain roles: *evCMPStereo*, which treats events (represented as event concentration E [48]) as X_t and the frame F as X_s , and *imgCMPStereo*, which reverses this assignment using F as X_t and events (encoded as voxel grid V [88]) as X_s , following established practices in event-based vision. The key components of CMPStereo are detailed below.

3.1. CMPStereo

The overall architecture of CMPStereo is illustrated in Fig. 2. The network employs domain-specific encoders to extract stereo features, which are aligned in the target canonical space through the Cross-Domain Embedding Adapter (CDEA) and Stereo Canonicalization Constraint (SCC). Context features are learned via Hierarchical Visual Transformation (HVT) to prevent shortcut learning, and disparity is iteratively refined through cascaded ConvGRUs.

Cross-Domain Embedding Adapter. CMPStereo begins with the Cross-Domain Embedding Adapter (CDEA) for the source-domain modality X_s , which employs a U-shaped [54] adapter $\mathcal{A}_{s2t}(\cdot)$ to map X_s into an embedding space aligned with the target domain. This process explicitly activates discriminative target-domain cues latent in the source representations, facilitating fine-grained feature alignment in the target domain.

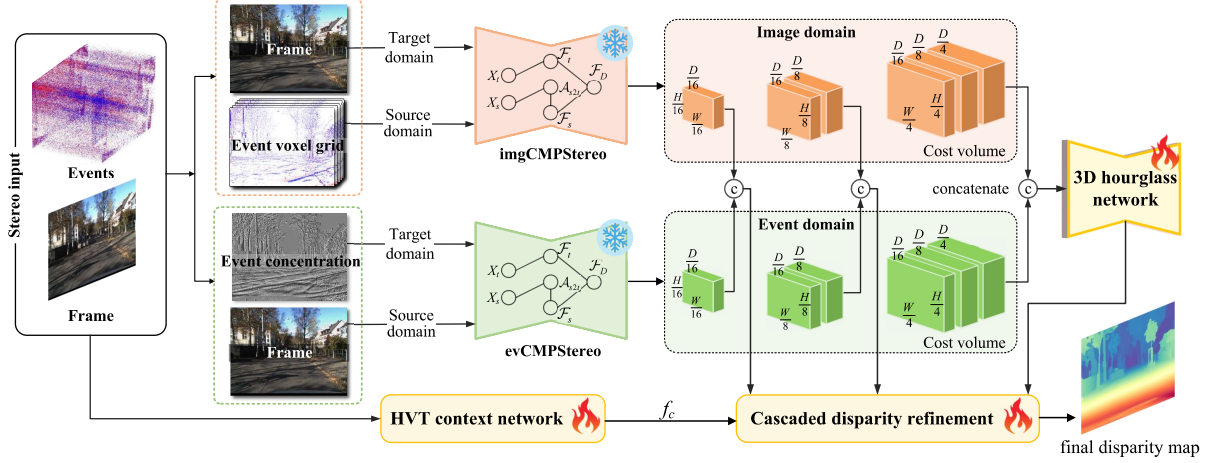


Figure 3. **Overview of the Bi-CMPStereo framework.** Bi-CMPStereo integrates complementary representations from imgCMPStereo and evCMPStereo in the image and event domains for reliable disparity estimation. The two pre-trained CMPStereo networks are frozen as stereo feature extractors to construct multi-scale cost volumes, which are concatenated and fused via a 3D hourglass network at 1/4 scale, followed by the same cascaded architecture for iterative disparity refinement.

To guide the adapter in modeling the target-domain embedding distribution and ensure the independence of two domain-specific CMPStereo variants, we employ a domain classifier $\mathcal{C}(\cdot)$ that distinguishes between event and frame embeddings. The classifier learns discriminative cues from both domains and supervises the source-to-target adaptation by minimizing the following loss:

$$\mathcal{L}_{cdea} = \ell_{ce}(\mathcal{C}(E), 1) + \ell_{ce}(\mathcal{C}(F), 0) + \ell_{ce}(\mathcal{C}(\mathcal{A}_{s2t}(X_s)), \mathcal{Y}_t) \quad (1)$$

where ℓ_{ce} denotes the binary cross-entropy loss, E and F are event and frame representations, and $\mathcal{Y}_t$ is the target domain label. By using a single shared domain classifier, each adapter learns to map its source domain specifically toward its designated target domain, ensuring that evCMPStereo and imgCMPStereo learn complementary representations in their respective canonical spaces.

Notably, CDEA performs domain-level adaptation rather than explicit pixel-level translation, as the latter typically results in blurred representations [44, 46, 53] that may diminish stereo-specific features. In addition, CDEA is designed based on a single domain classifier, which avoids adversarial learning and results in better training stability.

Stereo Canonicalization Constraint. To extract cross-modal stereo representations, CMPStereo adopts two domain-specific encoders $\mathcal{F}_t(\cdot)$ and $\mathcal{F}_s(\cdot)$ to account for the modality discrepancy between the source and target domains, while employing a shared decoder $\mathcal{F}_D(\cdot)$ to accommodate the feature-space alignment assumption underlying stereo matching, producing multi-scale features $f_t^i, f_r^i \in \mathbb{R}^{C_i \times \frac{H}{i} \times \frac{W}{i}}$, $i \in \{4, 8, 16\}$ for cascaded disparity estimation. While stereo matching over the correlation volume implicitly drives the encoders to converge on a canonical latent space through joint optimization, the latent space is

prone to collapsing into overly similar representations that marginalize discriminative cues. To mitigate this issue, we introduce the Stereo Canonicalization Constraint (SCC) in the bottleneck layer to achieve high-fidelity cross-domain alignment in the target canonical space.

Our goal is to train the domain-specific encoders to learn a canonical latent space that faithfully represents both modalities without marginalizing discriminative cues. The key insight is that an expressive intermediate representation should be able to faithfully reconstruct the original inputs in the target-domain space. We therefore enforce this reconstruction constraint during training using a shared lightweight decoder $\mathcal{F}_R(\cdot)$ that maps both source and target representations from the canonical space to their target-domain reconstructions:

$$\mathcal{L}_{scc} = \|\mathcal{F}_R(\mathcal{F}_s(\mathcal{A}_{s2t}(X_s))) - X_s^{(t)}\|_1 + \|\mathcal{F}_R(\mathcal{F}_t(X_t)) - X_t\|_1 \quad (2)$$

where $X_s^{(t)} := \mathcal{W}(X_t, d_{gt})$ is the source-domain modality expressed in the target-domain space, obtained by warping the target-domain counterpart using the ground-truth disparity d_{gt} . The deliberately lightweight design of $\mathcal{F}_R$ prevents it from hallucinating missing details, compelling the encoders to preserve fine-grained cues in the latent space.

This constraint provides three-fold benefits: (1) the self-reconstruction of the target-domain modality preserves domain-specific discriminative features essential for stereo matching; (2) the cross-domain reconstruction of the source-domain modality enforces fine-grained source-to-target alignment, driving CMPStereo to operate within the target space; and (3) the shared reconstruction decoder effectively regularizes both domain-specific encoders toward a unified canonical latent space, promoting stereo-consistent feature generation. It is worth noting that SCC is

employed only in the training stage to promote stereo representation learning while not involved during inference, thus imposing no additional runtime overhead at test time.

Hierarchical Visual Transformation for context learning. Context features are essential for the ConvGRU updater, serving both to initialize its hidden state with a global scene prior and to inform each iteration [57]. We extract context features from the image frame, which provides rich semantic content, using a context network $\mathcal{F}_c(\cdot)$ with residual blocks and downsampling layers [39], producing multi-scale context features $f_c^i \in \mathbb{R}^{C \times \frac{H}{i} \times \frac{W}{i}}$, $i \in \{4, 8, 16\}$. However, in cross-modal stereo, the network may over-rely on the frame context to circumvent challenging cross-modal alignment, learning shortcut heuristics rather than exploiting stereo correspondenc, resulting in poor generalization.

To alleviate this issue, we adopt the Hierarchical Visual Transformation (HVT) [8, 75] to learn shortcut-invariant context features and facilitate the generation of robust stereo representations. HVT synthesizes a spectrum of augmented views by applying three hierarchical visual transformations $\{T_G(F), T_L(F), T_P(F)\}$ to the original frame F from three levels: Global, Local, Pixel, enforcing features invariance across views to prevent superficial shortcuts. To ensure that the transformations induce substantial visual discrepancy, the visual similarity between original and transformed frames is minimized:

$$\mathcal{L}_{sim} = \sum_J \text{Cos}(\phi(T_J(F)), \phi(F)) \quad (3)$$

where $J \in \{G, L, P\}$ and $\text{Cos}(\cdot, \cdot)$ denotes the cosine similarity function. $\phi(\cdot)$ extracts low-level visual features to measure pixel-level similarity.

Consistency between context features of transformed and original frames is enforced by minimizing:

$$\mathcal{L}_{dist} = \sum_J \|\mathcal{F}_c(T_J(F)) - \mathcal{F}_c(F)\|_2 \quad (4)$$

The combined objective is defined as:

$$\mathcal{L}_{HVT} = \lambda_{hvt,1} \mathcal{L}_{sim} + \lambda_{hvt,2} \mathcal{L}_{dist} \quad (5)$$

where $\lambda_{hvt,1}$ and $\lambda_{hvt,2}$ are loss weights. This objective guides the model to learn shortcut-invariant context features for robust stereo matching. Similar to SCC, HVT is applied solely as a training-time regularization on the context network without affecting inference efficiency.

Cascaded Disparity Refinement. We estimate disparity through a cascaded coarse-to-fine architecture. At coarse resolutions, large receptive fields capture abstract scene structures with higher cross-modal coherence, yielding robust matching. To leverage this coarse-level robustness while exploiting fine-level details, the framework mitigates

high-resolution ambiguity via coarse disparity priors and enables hierarchical cross-modal alignment: semantic consistency enforced at the coarse scale—where the modality gap is narrower—propagates through the shared decoder to yield fine-grained structural alignment at higher resolutions.

Given multi-scale stereo features $f_l^i, f_r^i, i \in \{4, 8, 16\}$, we construct the cost volumes $C_{gwc}^i \in \mathbb{R}^{N_c \times \frac{D}{i} \times \frac{H}{i} \times \frac{W}{i}}$ at each scale using group-wise correlation [27]:

$$C_{gwc}^i(g, d, x, y) = \frac{1}{N_c/N_g} \langle f_{l,g}^i(x, y), f_{r,g}^i(x-d, y) \rangle \quad (6)$$

where N_c is the feature channels; $\langle \cdot, \cdot \rangle$ denotes the inner product; $g \in \{1, 2, \dots, N_g\}$ is the group index among the total $N_g=8$ groups; $d \in \{1, 2, \dots, \frac{D}{i}\}$ is the disparity index.

To further capture matching cues across scales, we construct multi-layer pyramids by pooling each cost volume: 3-, 2-, and 1-layer pyramids are built for the 1/4, 1/8, and 1/16 scales, respectively, all downsampled to the common resolution of 1/16, as shown in Fig. 3. At each cascade level, a single-level ConvGRU iteratively updates the hidden state through lookups on the corresponding pyramid. The hidden state is initialized from context features and in each iteration a convolutional projection from the hidden state produces a residual disparity to update the current estimate.

The disparity estimate from the previous scale is propagated as the initialization for the next cascade level via convex upsampling [57], and the final prediction is similarly upsampled to the full resolution.

3.2. Bi-CMPStereo

Within Bi-CMPStereo framework, evCMPStereo and imgCMPStereo are first trained independently end-to-end. We use the Smooth L1 loss with exponentially increasing weights to compute the disparity loss $\mathcal{L}_d$ on all updated disparities $\{d_i\}_{i=1}^N$ at all scales:

$$\mathcal{L}_d = \sum_{i=1}^N \gamma^{N-i} \|d_i - d_{gt}\|_1 \quad (7)$$

The overall training objective for CMPStereo is defined as a weighted combination of all losses:

$$\mathcal{L}_{pre} = \mathcal{L}_d + \lambda_1 \mathcal{L}_{cdea} + \lambda_2 \mathcal{L}_{scc} + \lambda_3 \mathcal{L}_{HVT} \quad (8)$$

where λ_1, λ_2 , and λ_3 are loss weights.

After training, both networks learn well-aligned stereo representations within their respective domains. To leverage complementary features from both modalities, Bi-CMPStereo integrates these representations in a unified network, as shown in Fig. 3. Specifically, for a given scene, each modality is processed by the pre-trained frozen evCMPStereo and imgCMPStereo networks, producing intra- and cross-modal representations to construct two

Method	MAE ↓				IPE ↓				2PE ↓				RMSE ↓			
	Zu	In	Th	All	Zu	In	Th	All	Zu	In	Th	All	Zu	In	Th	All
Event-frame asymmetric stereo																
ZEST[41]	2.483	1.282	1.339	2.168	48.325	39.415	38.127	46.920	20.195	13.847	13.303	19.288	3.779	2.319	2.608	3.394
ZEST [†] [41]	0.785	0.580	0.797	0.763	21.003	12.913	19.922	20.382	4.845	3.303	4.615	4.646	1.474	1.211	1.713	1.438
SEVFI [18]	0.729	0.623	0.875	0.711	17.374	14.558	24.225	16.932	4.497	3.408	5.970	4.307	1.552	1.309	1.844	1.509
evCMPStereo	0.594	0.496	0.761	0.577	12.752	10.031	18.508	12.309	3.097	2.060	4.116	2.909	1.352	1.070	1.697	1.310
imgCMPStereo	0.582	0.482	0.747	0.565	11.989	8.706	17.530	11.432	3.000	1.852	3.977	2.790	1.324	1.132	1.703	1.292
Bi-CMPStereo	0.547	0.450	0.715	0.532	11.112	7.992	16.444	10.613	2.606	1.507	3.587	2.415	1.244	1.026	1.651	1.210
Event-based symmetric stereo																
SE-CFF [48]	0.627	0.555	0.771	0.612	13.123	10.397	17.745	12.477	3.506	2.536	4.408	3.288	1.462	1.325	1.862	1.445
DTC [84]	0.631	0.563	0.768	0.621	12.809	10.183	17.569	12.069	3.153	2.481	4.129	3.018	1.501	1.357	1.736	1.508

Table 1. Comparison of our method with state-of-the-art stereo methods on the DSEC dataset [20]. Zu, In, and Th denote the Zurich City, Interlaken, and Thun sequences in the dataset, respectively.

complementary cost volumes at multiple scales. Subsequently, at 1/16 and 1/8 scales, the two cost volumes are directly concatenated to efficiently combine dual-modal information. At 1/4 scale, they are fused through a 3D hourglass network [67] to aggregate complementary matching cues. The resulting multi-scale cost volumes, together with context features from HVT, are used for cascaded disparity refinement, as in CMPStereo. In this stage, only the 3D hourglass network, the ConvGRU updater, and the HVT context network are optimized with:

$$\mathcal{L}_{final} = \mathcal{L}_d + \lambda_3 \mathcal{L}_{HVT} \quad (9)$$

4. Experiments

4.1. Implementation and Experimental Settings

In this section, we describe our experimental setup including datasets, implementation details, evaluation metrics, and baseline methods.

Datasets. We evaluate our method on DSEC [20], a high-quality event stereo dataset captured in outdoor driving scenarios. It employs two Prophesee Gen 3.1 event cameras (640×480) to provide stereo event streams, while synchronized intensity frames are captured by a separate pair of RGB cameras. DSEC contains 41 driving scenarios with 26K samples, covering diverse lighting and environmental conditions. Following [4], we split the dataset into 31 training sequences with 19K samples and 10 testing sequences with 7K samples. For asymmetric stereo, we use the left intensity frame and the right event stream, where the intensity frames are warped to the left event camera using the provided calibration parameters. We also evaluate the generalization of our method on MVSEC [87] and M3ED [6] after training on DSEC [20]. MVSEC utilizes two DAVIS346B (346×260) to provide stereo events and frames, and M3ED delivers synchronized data via Prophesee EVK4 HD event cameras (1280×720). Following [41], we evaluate on three MVSEC subsets: indoor_flying1, 2, 3 (with frames 500-

Method	MAE↓	IPE↓	2PE↓	RMSE↓
SE-CFF [48]	0.519	9.583	2.620	1.231
DTC [84]	0.527	9.277	2.416	1.291
Zhuang <i>et al.</i> [89]	0.488	8.881	2.335	1.168
Bi-CMPStereo	0.475	8.557	2.253	1.101

Table 2. Comparison on the DSEC online benchmark.

1500, 500-2000, and 500-2500, denoted as S1, S2, and S3), and one M3ED subset: car_urban_day_horse (300-700).

Implementation Details. We implement our networks in PyTorch [49] and use two NVIDIA RTX 3090 GPUs (each with 24 GB memory) for training and inference. We use the Adam [31] optimizer with an initial learning rate of $5e^{-4}$, which decays following a cosine annealing schedule. We first pre-train evCMPStereo and imgCMPStereo independently for 100 epochs each with a batch size of 8, and then train the final Bi-CMPStereo network for another 100 epochs with the same batch size. The loss weights $\lambda_{hvt,1}$ and $\lambda_{hvt,2}$ in Eq 5 are set as 1 and 0.5 respectively. λ_1 , λ_2 and λ_3 in Eq 8 and Eq 9 are set as 0.5, 2 and 1. γ in Eq 7 is set to 0.9. On our DSEC training split with 18948 samples, both evCMPStereo and imgCMPStereo take about 35 minutes per epoch, while the Bi-CMPStereo requires around 46 minutes. During inference, Bi-CMPStereo runs at about 0.1s per frame with the 640×480 resolution.

Metrics. We use the standard metrics for evaluation, including mean absolute error (MAE) and root mean square error (RMSE) of the disparity, as well as the n-pixel error (nPE) that is the percentage of pixels whose disparity error exceeds n pixels. We report 1PE and 2PE for DSEC, and 2PE and 3PE for MVSEC and M3ED in the generalization evaluation. The first, second and third-best are highlighted with different colors in tables.

Baseline Methods. We compare our method with the state-of-the-art event-frame asymmetric stereo method ZEST [41]. ZEST aligns the representations of frames and events, enabling the use of foundation monocular [74] and stereo [33] models pretrained on diverse image datasets for monocular cue-guided cross-modal stereo. To make ZEST

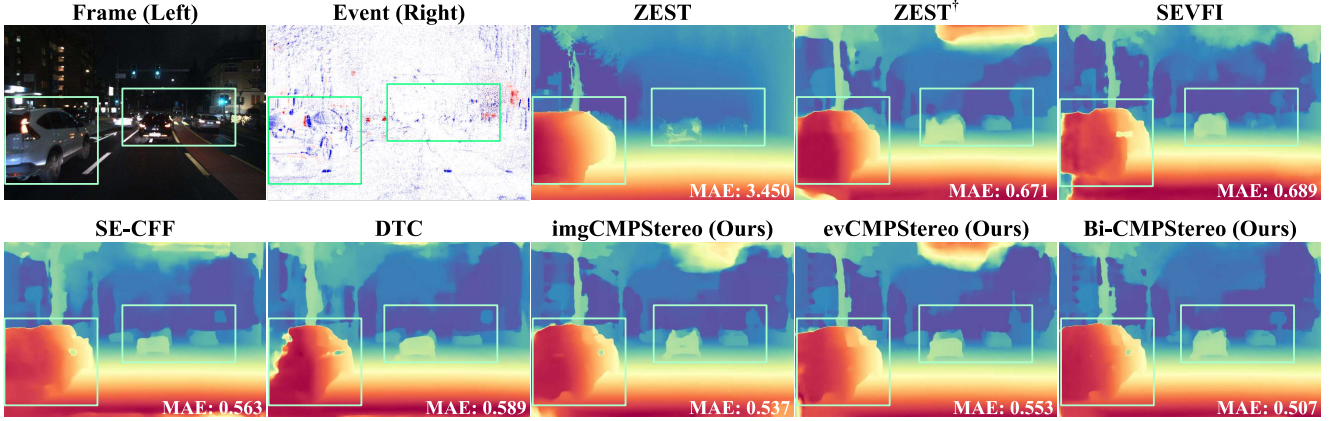


Figure 4. **Qualitative comparison with state-of-the-art methods on a low-light scene from the DSEC dataset [20].** The mean absolute error (MAE) is shown at the bottom right of each estimated disparity map. As can be seen in the highlighted regions, our method more accurately reconstructs the fine structural details of the cars.

more competitive on DSEC, we retrain it under the same training configuration as our method, denoted as $ZEST^\dagger$. We also compare with SEVFI-Net [18], a deformable convolution-based network to estimate cross-modal stereo disparities. For a comprehensive assessment, we include two state-of-the-art event-based symmetric stereo models: SE-CFF [48] and DTC-SPADE [84] as baseline methods.

4.2. Results on DSEC Dataset

We compare our imgCMPStereo, evCMPStereo, and Bi-CMPStereo with baseline methods on DSEC [20]. Table 1 shows that our method consistently outperforms all baselines across all metrics. Notably, even our single-domain variants achieve superior performance, demonstrating the effectiveness of CMPStereo for cross-modal stereo. By integrating the complementary cues from both branches, Bi-CMPStereo achieves substantial further gains. It can be seen that existing asymmetric stereo methods remain inferior to event-based symmetric approaches, primarily because their alignment schemes suppress modality-specific cues. In contrast, our approach achieves high-fidelity alignment by effectively preserving discriminative features from each modality, surpassing even state-of-the-art event-based symmetric methods. Visual comparisons in Fig. 4 highlight the superior quality of our approach, producing disparity maps with sharper edges and richer scene content.

Furthermore, we evaluate our method on the official DSEC online benchmark, where ground-truth disparities are withheld. We additionally compare with Zhuang *et al.* [89], a recent asymmetric approach utilizing temporal feature gating from frame sequences. Since their source code is unavailable, we directly compare with their reported metrics. Results in Table 2 show that we outperform both symmetric baselines (SE-CFF [48], DTC [84]) and the multi-frame-based approach [89], further demonstrating the remarkable capability and robustness of our method.

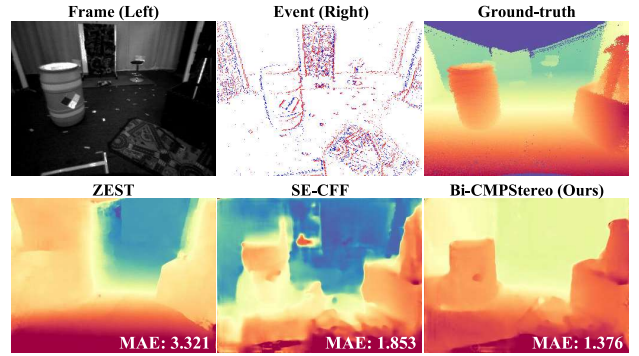


Figure 5. **Generalization Result.** Qualitative comparison of cross-dataset generalization on the MVSEC [87] dataset.

4.3. Cross-Dataset Generalization on MVSEC

We further evaluate the generalization of our method by directly applying the DSEC-trained model to MVSEC [87] and M3ED [6]. We compare it against ZEST [41], which serves as a representative baseline, as it is designed for zero-shot generalization via visual prompting from the image domain. As shown in Table 3 and Table 4, our Bi-CMPStereo exhibits superior generalizability, consistently outperforming any competing methods. Leveraging models pretrained on large-scale image data, ZEST [41] achieves competitive generalization, outperforming its DSEC-trained variant $ZEST^\dagger$ in most cases. However, even with its external image-domain knowledge, ZEST still lags behind event-based symmetric baselines. In contrast, our Bi-CMPStereo, trained solely on DSEC, achieves superior performance across all metrics, demonstrating the robust generalization to unseen scenes. Qualitative results in Fig. 5 further confirm the effectiveness of our method across datasets.

4.4. Ablation Study

In this section, we perform a series of ablation studies to validate the effectiveness of each component in our method.

Method	MAE ↓				2PE ↓				3PE ↓				RMSE ↓			
	S1	S2	S3	All	S1	S2	S3	All	S1	S2	S3	All	S1	S2	S3	All
Event-frame asymmetric stereo																
ZEST[41]	3.340	7.190	5.130	5.220	33.220	61.380	43.970	46.190	19.060	41.720	28.660	29.820	3.830	7.830	5.600	5.763
ZEST [†] [41]	4.923	6.022	4.267	4.998	44.676	60.123	45.688	50.275	32.255	46.015	31.699	36.594	6.626	8.023	5.799	6.724
SEVFI [18]	5.537	4.728	5.779	5.375	63.568	60.108	68.703	64.697	48.482	45.459	53.147	49.548	10.143	8.352	10.440	9.678
Bi-CMPStereo	1.743	2.181	1.674	1.858	29.247	37.041	29.869	32.121	14.924	20.889	15.336	17.095	2.708	3.419	2.466	2.837
Event-based symmetric stereo																
SE-CFF [48]	2.428	2.867	2.518	2.614	52.993	53.451	52.837	53.076	30.102	32.864	32.238	31.972	5.729	6.241	5.807	5.934
DTC [84]	2.175	2.502	2.204	2.297	48.767	50.646	50.237	50.047	27.382	29.928	29.530	29.185	5.313	5.571	5.210	5.353

Table 3. Cross-dataset generalization results on the MVSEC [87] dataset. S1, S2 and S3 refers to single MVSEC subsets.

	Method	MAE↓	2PE↓	3PE↓	RMSE↓
Event-frame Asymmetric Stereo	ZEST[41]	2.060	29.020	19.040	3.390
	ZEST [†] [41]	2.653	36.011	23.020	5.069
	SEVFI[18]	3.634	52.107	36.615	5.282
	Bi-CMPStereo	1.557	17.044	11.420	2.877
Event Stereo	SE-CFF[48]	1.615	17.364	12.182	2.976
	DTC[84]	1.902	22.375	15.661	3.068

Table 4. Generalization results on the M3ED [6] dataset.

Method	MAE↓	1PE↓	2PE↓	RMSE↓
w/o CDEA & SCC	0.594	12.232	3.040	1.351
w/o CDEA	0.583	12.118	2.968	1.325
w/o SCC	0.589	12.054	2.945	1.354
w/o cascades	0.588	11.927	2.994	1.347
full imgCMPStereo	0.565	11.432	2.790	1.292

Table 5. Ablation study of imgCMPStereo. Results on the DSEC [20] dataset.

To assess the impact of each module on CMPStereo network, we evaluate imgCMPStereo with different settings on the DSEC dataset [20] and report the results in Table 5. The w/o cascades variant performs disparity optimization only at the finest 1/4 scale with 10 iterations to maintain the same total iteration count. The drop in its accuracy demonstrates the importance of the cascaded architecture. Removing the CDEA module weakens modality alignment by failing to sufficiently activate potential target-domain cues in the source-domain modality, leading to a clear performance drop and confirming its effectiveness. The effectiveness of the SCC module is evident from the substantial accuracy gain it brings to CMPStereo, highlighting its pivotal role in promoting high-fidelity cross-modal alignment.

We validate the effectiveness of CDEA and SCC for bidirectional estimation in Table 6, showing that they significantly enhance Bi-CMPStereo’s performance. Beyond facilitating modality alignment, CDEA and SCC further promote the two single-domain CMPStereo branches toward their respective target domains, thereby reinforcing their mutual complementarity and leading to more precise stereo predictions. HVT for context learning to enhance generalization is validated on Bi-CMPStereo through DSEC-to-MVSEC [87] cross-dataset evaluation. As presented in Table 7, HVT-based context learning achieves notable im-

Method	MAE↓	1PE↓	2PE↓	RMSE↓
w/o CDEA	0.546	11.128	2.610	1.239
w/o SCC	0.551	11.290	2.614	1.257
full Bi-CMPStereo	0.532	10.613	2.415	1.210

Table 6. Ablation study of Bi-CMPStereo. Results on the DSEC [20] dataset.

Method	MAE↓	2PE↓	3PE↓	RMSE↓
w/o HVT	2.093	36.005	22.478	3.060
full Bi-CMPStereo	1.858	32.121	17.095	2.837

Table 7. Impact of HVT on Bi-CMPStereo. Cross-dataset generalization results on the MVSEC [87] dataset.

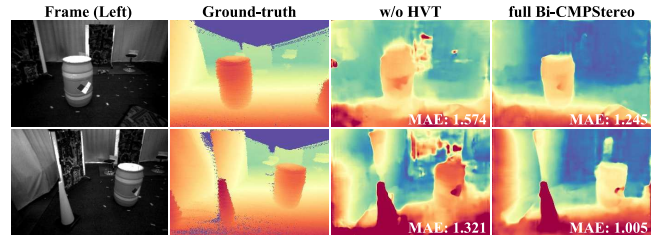


Figure 6. Visual comparison between Bi-CMPStereo with and without HVT. Generalization results on MVSEC [87].

provements in generalization, and the visual comparison in Fig. 6 substantiates its enhanced robustness across datasets.

5. Conclusion

We introduced Bi-CMPStereo, a novel bidirectional framework for event-frame asymmetric stereo that preserves domain-specific features through cross-modal prompting. Our CMPStereo network, equipped with Stereo Canonicalization Constraint and Cross-Domain Embedding Adapter, learns high-fidelity aligned representations in target canonical spaces. Applying Hierarchical Visual Transformation in context learning further boosts the model’s robustness and generalization. By combining complementary features from both event and image domains, our approach effectively addresses the modality gap and achieves robust disparity estimation that surpasses state-of-the-art methods. Extensive experiments demonstrate strong performance and generalization on DSEC, MVSEC and M3ED datasets.

Acknowledgement. This work was supported by Major Science and Technology Major Project of Jiangsu Province (BG2025025).

References

- [1] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 882–890, 2021. 2
- [2] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, pages 207–217. IEEE, 2021. 2
- [3] Luca Bartolomei, Matteo Poggi, Andrea Conti, and Stefano Mattoccia. Lidar-event stereo fusion with hallucinations. In *European Conference on Computer Vision*, pages 125–145. Springer, 2024. 2
- [4] Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. *arXiv preprint arXiv:2509.15224*, 2025. 6
- [5] Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2
- [6] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4016–4023, 2023. 6, 7, 8
- [7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2
- [8] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9559–9568, 2023. 2, 5
- [9] Xihao Chen, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Depth from asymmetric frame-event stereo: A divide-and-conquer approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3045–3054, 2024. 2, 3
- [10] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [11] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2
- [12] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 454–462, 2022. 3
- [13] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 2, 3
- [14] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 2
- [15] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 2, 3
- [16] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2024. 2
- [17] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13022–13032, 2022. 2
- [18] Chao Ding, Mingyuan Lin, Haijian Zhang, Jianzhuang Liu, and Lei Yu. Video frame interpolation with stereo event and intensity cameras. *IEEE Transactions on Multimedia*, 26: 9187–9202, 2024. 2, 3, 6, 7, 8
- [19] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conrath, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1
- [20] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954, 2021. 6, 7, 8
- [21] Dipon Kumar Ghosh and Yong Ju Jung. Two-stage cross-fusion network for stereo event-based depth estimation. *Expert Systems with Applications*, 241:122743, 2024. 2
- [22] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [23] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2
- [24] Ziren Gong, Xiaohan Li, Fabio Tosi, Jiawei Han, Stefano Mattoccia, Jianfei Cai, and Matteo Poggi. Ov3r: Open-vocabulary semantic 3d reconstruction from rgb videos. *arXiv preprint arXiv:2507.22052*, 2025. 2

- [25] Tongfan Guan, Jiaxin Guo, Chen Wang, and Yun-Hui Liu. Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, Hawaii, USA, 2025. ICCV 2025 Highlight. 2
- [26] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pages 263–279. Springer, 2022. 2
- [27] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 5
- [28] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2
- [29] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2
- [30] Haram Kim, Sangil Lee, Junha Kim, and H Jin Kim. Real-time hetero-stereo matching for event and frame camera with aligned events using maximum shift distance. *IEEE Robotics and Automation Letters*, 8(1):416–423, 2022. 2, 3
- [31] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [32] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1738–1764, 2020. 1, 2
- [33] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 1, 2, 6
- [34] Jianing Li, Yunjian Zhang, Haiqian Han, and Xiangyang Ji. Active event-based stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 971–981, 2025. 2
- [35] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 2
- [36] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2
- [37] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1
- [38] Mingyuan Lin, Chi Zhang, Chu He, and Lei Yu. Learning parallax for stereo event-based motion deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2, 3
- [39] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1, 2, 5
- [40] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13012–13021, 2022. 2
- [41] Hanyue Lou, Jinxiu Liang, Mingguo Teng, Bin Fan, Yong Xu, and Boxin Shi. Zero-shot event-intensity asymmetric stereo via visual prompting from image domain. *Advances in Neural Information Processing Systems*, 37:13274–13301, 2024. 1, 2, 3, 6, 7, 8
- [42] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 2
- [43] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2
- [44] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 4
- [45] Junhong Min, Youngpil Jeon, Jimin Kim, and Minyong Choi. S²M²: Scalable Stereo Matching Model for Reliable Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, page to appear, Honolulu, Hawai’i, 2025. IEEE. Accepted at ICCV 2025, Hawai’i Convention Center, Oct 19–23. 2
- [46] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, 129(4):900–920, 2021. 4
- [47] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021. 2, 3
- [48] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6114–6123, 2022. 1, 2, 3, 6, 7, 8
- [49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Ad-*

- vances in neural information processing systems, 32, 2019. 6
- [50] Matteo Poggi and Fabio Tosi. Federated online adaptation for deep stereo. In *CVPR*, 2024. 2
 - [51] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4713–4729, 2021. 2
 - [52] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314–5334, 2021. 1, 2
 - [53] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 4
 - [54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3
 - [55] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 1, 2
 - [56] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnets: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2021. 2
 - [57] 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. 5
 - [58] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1605–1613, 2017. 2
 - [59] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 855–866, 2023. 2
 - [60] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
 - [61] Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *arXiv preprint arXiv:2407.07816*, 2024. Extended version of CVPR 2024 Tutorial ”Deep Stereo Matching in the Twenties” (<https://sites.google.com/view/stereo-twenties>). 1, 2
 - [62] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 1, 2
 - [63] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2
 - [64] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
 - [65] Ziwei Wang, Liyuan Pan, Yonhon Ng, Zheyu Zhuang, and Robert Mahony. Stereo hybrid event-frame (shef) cameras for 3d perception. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9758–9764. IEEE, 2021. 2, 3
 - [66] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *CVPR*, 2025. 1, 2
 - [67] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022. 6
 - [68] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2
 - [69] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *arXiv preprint arXiv:2409.00638*, 2024. 2
 - [70] Haoifei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020. 2
 - [71] Haoifei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
 - [72] Ninghui Xu, Lihui Wang, Jiajia Zhao, and Zhiting Yao. Denoising for dynamic vision sensor based on augmented spatiotemporal correlation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4812–4824, 2023. 1
 - [73] Ninghui Xu, Lihui Wang, Zhiting Yao, and Takayuki Okatani. Mets: Motion-encoded time-surface for event-based high-speed pose tracking. *International Journal of Computer Vision*, 133(7):4401–4419, 2025. 1
 - [74] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 6
 - [75] Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision*, 132(11):4823–4849, 2024. 2, 3, 5
 - [76] Chengtang Yao, Lidong Yu, Zhidan Liu, Jiayi Zeng, Yuwei Wu, and Yunde Jia. Diving into the fusion of monocular priors for generalized stereo matching. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 14887–14897, 2025. [2](#)
- [77] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6044–6053, 2019. [2](#)
- [78] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. [2](#)
- [79] Dehao Zhang, Qiankun Ding, Peiqi Duan, Chu Zhou, and Boxin Shi. Data association between event streams and intensity frames under diverse baselines. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. [2](#), [3](#)
- [80] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 185–194, 2019. [2](#)
- [81] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*, 2020. [2](#)
- [82] Haihao Zhang, Yunjian Zhang, Jianing Li, Lin Zhu, Meng Lv, Yao Zhu, Yanwei Liu, and Xiangyang Ji. Enhanced event-based dense stereo via cross-sensor knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5437–5447, 2025. [2](#)
- [83] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R. Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13001–13011, 2022. [2](#)
- [84] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8676–8686, 2022. [1](#), [2](#), [6](#), [7](#), [8](#)
- [85] 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. [2](#)
- [86] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. [2](#)
- [87] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi-vehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. [6](#), [7](#), [8](#)
- [88] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. [3](#)
- [89] Hao Zhuang, Yan Yang, and Liyuan Pan. Asymmetric event-image stereo with temporal feature gating and iterative structure-detail refinement. 2025. [3](#), [6](#), [7](#)