Event-Driven Dynamic Scene Depth Completion

Zhiqiang Yan[1] **Jianhao Jiao**[2] **Zhengxue Wang**[3] **Gim Hee Lee**[1]
[1]National University of Singapore [2]University College London
[3]Nanjing University of Science and Technology
{yanzq, gimhee.lee}@nus.edu.sg

Abstract

Depth completion in dynamic scenes poses significant challenges due to rapid ego-motion and object motion, which can severely degrade the quality of input modalities such as RGB images and LiDAR measurements. Conventional RGB-D sensors often struggle to align precisely and capture reliable depth under such conditions. In contrast, event cameras with their high temporal resolution and sensitivity to motion at the pixel level provide complementary cues that are beneficial in dynamic environments. To this end, we propose **EventDC**, the first event-driven depth completion framework. It consists of two key components: Event-Modulated Alignment (EMA) and Local Depth Filtering (LDF). Both modules adaptively learn the two fundamental components of convolution operations: offsets and weights conditioned on motion-sensitive event streams. In the encoder, EMA leverages events to modulate the sampling positions of RGB-D features to achieve pixel redistribution for improved alignment and fusion. In the decoder, LDF refines depth estimations around moving objects by learning motion-aware masks from events. Additionally, EventDC incorporates two loss terms to further benefit global alignment and enhance local depth recovery. Moreover, we establish the first benchmark for event-based depth completion comprising one real-world and two synthetic datasets to facilitate future research. Extensive experiments on this benchmark demonstrate the superiority of our EventDC. Project page.

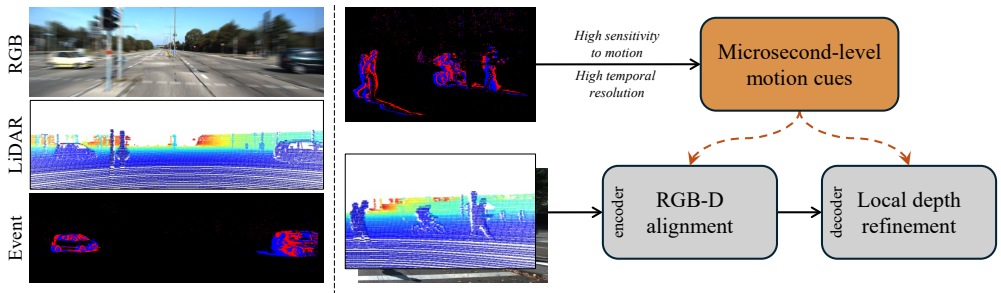
1 Introduction

Depth completion [48, 33, 36, 47, 65] aims to predict dense depth from sparse measurements, typically using auxiliary modalities such as RGB images. As a cornerstone of 3D perception, it plays a crucial role in a wide range of downstream applications including self-driving [70, 57, 26, 52], augmented reality [45, 60, 54, 66, 71], scene understanding [56, 40, 78, 53, 72], *etc*. Although recent methods have demonstrated impressive results in static scenes, dynamic environments remain highly challenging. As illustrated in Fig. 1(a), the rapid ego-motion results in blurry RGB images and misalignment with LiDAR measurements, while fast-moving objects further exacerbate depth inaccuracies in their vicinity. These challenges make precise depth completion even more difficult.

The unique characteristics of event cameras [10, 9, 11] provide a compelling complement to conventional RGB-D sensors in dynamic scenes. Their microsecond-level temporal resolution enables the reliable capture of rapid ego-motion without introducing motion blur, and their asynchronous change-driven operation makes them inherently well-suited for detecting fast-moving objects. These properties help mitigate the limitations of RGB-D measurements by offering temporally consistent and low-latency signals, particularly in regions where traditional sensors often fail. As a result, event-based sensing proves especially advantageous for depth completion in highly dynamic environments.

In this work, we present EventDC, a novel depth completion framework that leverages event data to tackle the challenges posed by dynamic scenes. As shown in Fig. 1(b), the core idea is to exploit the unique properties of event streams to guide depth completion especially in motion-affected regions.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Challenging dynamic scene (b) Our event-driven solution

Figure 1: Data example and our solution for depth completion in dynamic environments. Leveraging high temporal resolution and motion sensitivity, event provides valuable complementary information for depth completion in dynamic scenes. Multiple event streams are aggregated for clear visualization.

To this end, our EventDC incorporates two key components: Event-Modulated Alignment (EMA) and Local Depth Filtering (LDF). EMA is an encoder-side module that adaptively adjusts convolutional sampling positions using event information to achieve pixel redistribution for enhanced global alignment and more effective multi-modal fusion between RGB and LiDAR features. Furthermore, it incorporates a structure-aware loss to mitigate the RGB-D inconsistency caused by rapid ego-motion. LDF is a decoder-side module that focuses on refining depth around moving objects. It first learns motion masks from event streams to get the regions influenced by object motion. The learned masks are then used by LDF with a local motion-aware constraint to facilitate more accurate depth predictions in these regions. Concurrently, the two modules enable our EventDC to address both global misalignment and local depth inaccuracies for handling complex scenarios involving motion.

Additionally, depth completion based on event cameras remains an underexplored area with no existing event-based depth completion datasets to date. To address this gap, we introduce the first event-based depth completion benchmark, which includes a real-world dataset *EventDC-Real*, a semi-synthetic dataset *EventDC-SemiSyn*, and a fully synthetic dataset *EventDC-FullSyn*.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to introduce EventDC, a novel event-driven depth completion framework designed to address the challenges of dynamic environments.
- We present two event-driven modules: EMA and LDF which are designed to mitigate the global misalignment caused by ego-motion and local depth inaccuracies due to object motion. Additionally, these two modules are jointly supported by two dedicated loss constraints.
- To foster further research, we build the first event-based benchmark for depth completion. Extensive experiments across these datasets demonstrate the superiority of our approach, with up to 12.8% improvement on the best-performing dataset over suboptimal methods.

2 Related Work

Depth Completion. Early depth completion methods [48, 33, 25, 7, 49, 32] focus on predicting dense depth maps directly from sparse inputs. For example, IP-Basic [25] uses traditional image processing techniques to densify sparse depth without deep learning. In contrast, Uhrig et al. [48] introduce Sparsity Invariant CNNs, which adapt convolutional operations to varying input densities to ensure consistent performance. S2D [33] employs an encoder-decoder architecture to progressively densify sparse depth input. FusionNet [49] integrates global context and local structures with a confidence-driven refinement mechanism. Eldesokey et al. [7] present a confidence propagation method within CNNs to improve sparse depth regression by modeling uncertainty. Guided depth completion using color images has gained significant traction [46, 70, 65, 20, 69, 73, 47, 63, 64]. Dynamic filtering techniques [46, 61, 62] generate adaptive filtering kernels from color images for effective extraction of depth features. Methods such as FuseNet [1], PointDC [67], BEVDC [73], and TPVD [65] further enhance depth completion by incorporating raw point clouds. Moreover, priors of the depth foundation models are used to improve generalization [38, 37, 51, 16]. Recently, SigNet [66] redefines depth completion as enhancement, densifying sparse depth with non-CNN methods, and then refines

it through a degradation-aware framework. In addition, SPN techniques [3, 4, 36, 59, 20, 28, 65], which serve as effective refinement modules, can further enhance performance.

Event-Based Depth Estimation. Depth estimation with event cameras [9, 15, 10, 8, 13, 17, 30, 58] attracts growing interest due to the high temporal resolution, dynamic range, and low latency of asynchronous vision sensors. Early methods reconstruct depth solely from event streams such as the end-to-end framework by Hidalgo-Carrió et al. [17], the multi-view stereo pipeline EMVS [39], and unsupervised learning approaches for depth and egomotion [74]. DERD-Net [18] further exploits 3D convolutions and recurrence on event-based disparity space images, and Zhu et al. [77] propose a self-supervised framework for joint depth and optical flow estimation. Recent works leverage additional modalities to enhance event-based depth estimation. EMoDepth [75] temporally aligns events and intensity frames for self-supervised monocular depth learning. Muglikar et al. [34] propose event-guided illumination control for active depth sensors. SRFNet [35] fuses frame and event features for fine-grained depth prediction with improved structure in both daytime and nighttime scenes. SDT [68] combines spiking neural networks and transformers for efficient depth estimation. Furthermore, contrast maximization that emerges as a fundamental principle for event-based motion, depth, and optical flow estimation [9, 41] has inspired many subsequent works.

Dynamic Convolution. Dynamic convolution is a method that adjusts the convolution operation based on input features, and it gains significant attention in computer vision tasks. Techniques such as graph convolution and deformable convolution serve as specific manifestations of dynamic convolution. For example, ACMNet and GraphCSPN build graph structures to enable effective multi-modal fusion and refinement. STN [21] introduces the concept of spatially transforming features within a network, although training such a mechanism is a challenging task. Following this, DFN [23] proposes an approach that adapts filter parameters based on input features despite maintaining fixed kernel sizes. Deformable Convolution [5, 76] takes a different approach with the focus on dynamically adjusting sampling locations by generating offsets based on the geometric properties of objects. Similarly, Active Convolution [22] improves sampling by adjusting the locations while keeping the kernel shape fixed. More recently, GuideNet [46] develops a guided convolution block specifically designed for multi-modal data. Despite these innovations, dynamic mechanisms often add considerable complexity. To address this issue, RigNet [61] simplifies the dynamic guidance process by employing convolution factorization combined with attention [19].

3 Our Method

3.1 Background

The core of dynamic convolution lies in the adaptive determination of sampling positions and weights. Graph Convolutional Networks (GCNs) [24, 50] and Deformable Convolutional Networks (DCNs) [5, 76] serve as representative implementations of this concept. GCNs define sampling locations as neighboring nodes within the graph structure and compute adaptive weights during the aggregation stage. On the other hand, DCNs determine sampling locations through learned offsets and obtain adaptive weights by modulating predefined kernel weights with learned scalars. Both GCNs and DCNs can be viewed as extensions of standard convolutional operations, where the sampling locations and weights are made learnable and structure-aware. We use DCNs as an example to illustrate this dynamic learning process.

Specifically, DCNv1 [5] introduces learnable offsets for each sampling location to shift adaptively. Subsequently, DCNv2 [76] further incorporates a learnable modulation scalar for each sampling position that enables the assignment of varying importance to different locations. Given an input feature map $\mathbf{x}$ and a convolutional kernel with K sampling positions, let $\mathbf{w}_k$ and $\mathbf{p}_k$ denote the weight and the pre-defined offset of the k-th position, respectively. DCNv2 can be formulated as:

$$\hat{\mathbf{x}}(\mathbf{p}_0) = \sum_{k=1}^{K} \mathbf{w}_k \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k) \cdot \Delta\mathbf{m}_k, \tag{1}$$

where $\mathbf{p}_0$ denotes the reference location, and $\Delta\mathbf{p}_k$ and $\Delta\mathbf{m}_k$ are the learnable offset and modulation scalar, respectively. Note that $\mathbf{w}_k$ and $\Delta\mathbf{m}_k$ can be jointly interpreted as a unified learnable term. As a result, the adaptive adjustment of offset and weight in DCNv2 provides a foundation for leveraging event data to tackle the challenges posed by fast motion in depth completion.

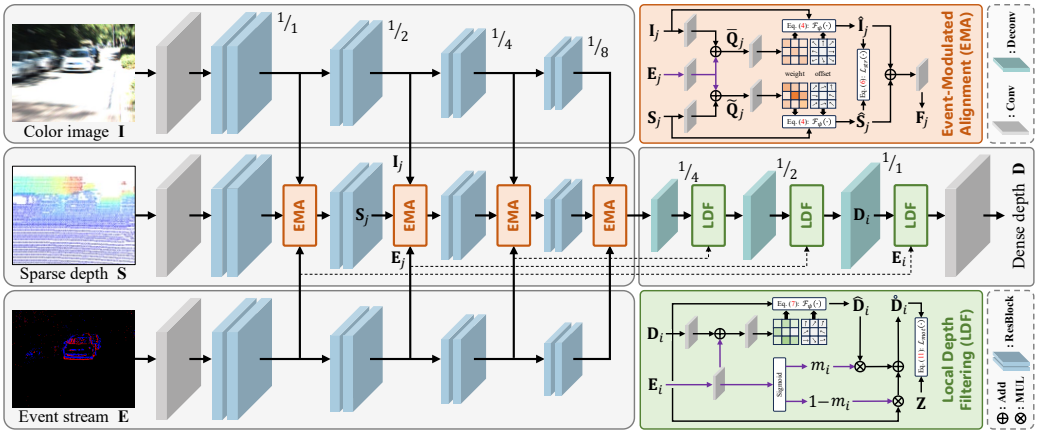


Figure 2: Pipeline of our EventDC. The color image $\mathbf{I}$, sparse depth $\mathbf{S}$, and event stream $\mathbf{E}$ are first processed by three structurally identical encoders. At each stage, the Event-Modulated Alignment (EMA) block leverages event features to align and fuse RGB-D representations. In the decoder, the Local Depth Filtering (LDF) unit further enhances depth estimation around moving objects, guided by the inherent sensitivity of events to motion and reinforced by local motion-aware constraints.

3.2 EventDC Architecture

Overview. In highly dynamic environments, the proposed approach is designed to alleviate the adverse effects of fast motion that include global misalignment and local depth inaccuracies caused by ego-motion and object motion, respectively. Fig. 2 illustrates the pipeline of our EventDC, which begins by employing three structurally consistent encoders to extract features from the color image $\mathbf{I}$, sparse depth $\mathbf{S}$, and event data $\mathbf{E}$. This yields multi-scale representations $\{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\}$, $\{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4\}$, and $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \mathbf{E}_4\}$ at the $\{1/1, 1/2, 1/4, 1/8\}$ stages, respectively. In the decoder, three deconvolution layers are applied to progressively generate $\{\mathbf{D}_3, \mathbf{D}_2, \mathbf{D}_1\}$ at the $\{1/4, 1/2, 1/1\}$ stages, respectively. Furthermore, EventDC incorporates two key components: Event-Modulated Alignment (EMA) and Local Depth Filtering (LDF). At each encoder stage, EMA predicts spatial offsets from event features and uses them to adjust the pixel distributions of RGB and depth features. This enables more precise multi-modal alignment and fusion. In addition, a structure-aware loss is introduced to further enhance the consistency. At the decoder stage, LDF leverages event features to estimate motion masks that identify moving objects. It then refines the depth values within these regions using dynamic convolutions and a local motion-aware loss, ultimately enhancing depth accuracy around the moving objects.

Event-Modulated Alignment. As depicted in Fig. 2, at the j-th ($j \in \{1, 2, 3, 4\}$) stage of the three encoders, the EMA module takes as input the color image feature $\mathbf{I}_j$, sparse depth feature $\mathbf{S}_j$, and event feature $\mathbf{E}_j$, each with dimensions $\mathbb{R}^{C\times H\times W}$, where C, H, and W denote the channel, height, and width, respectively. These inputs are first individually processed by three separate 3×3 convolutional layers $\mathcal{F}_{\tau_{j1}}(\cdot)$, $\mathcal{F}_{\tau_{j2}}(\cdot)$ and $\mathcal{F}_{\tau_{j3}}(\cdot)$, with a stride of 1 and output channels of C, $2C$, and C, respectively. The transformed event feature is then fused with the transformed RGB and depth features, respectively, resulting in the intermediate features:

$$\bar{\mathbf{Q}}_j = \mathcal{F}_{\tau_{j1}}(\mathbf{I}_j) + \alpha \cdot \mathcal{F}_s\left(\mathcal{F}_{\tau_{j2}}(\mathbf{E}_j)\right), \tag{2a}$$

$$\tilde{\mathbf{Q}}_j = \mathcal{F}_{\tau_{j3}}(\mathbf{S}_j) + \beta \cdot \mathcal{F}_s\left(\mathcal{F}_{\tau_{j2}}(\mathbf{E}_j)\right), \tag{2b}$$

where $\mathcal{F}_s(\cdot)$ denotes the operation that splits the $2C$-channel feature into two C-channel parts. α and β are learnable terms [1] that control the contribution of the event term.

Subsequently, these two intermediate features are used to predict the offsets via two additional 3×3 convolutions, $\mathcal{F}_{\tau_{j4}}(\cdot)$ and $\mathcal{F}_{\tau_{j5}}(\cdot)$, producing $2K \times H \times W$ offsets and $K \times H \times W$ weights:

$$\Delta\bar{\mathbf{p}}_j, \bar{\mathbf{w}}_j = \mathcal{F}_{\tau_{j4}}(\bar{\mathbf{Q}}_j), \tag{3a}$$

$$\Delta\tilde{\mathbf{p}}_j, \tilde{\mathbf{w}}_j = \mathcal{F}_{\tau_{j5}}(\tilde{\mathbf{Q}}_j). \tag{3b}$$

[1] Implemented using `torch.nn.Parameter(zeros(1))` with the zero initialization designed to facilitate the progressive learning of event priors during training.

This step enables the model to adaptively determine the sampling locations by learning a prior from event data that is sensitive to fast motion. Consequently, the dynamic convolution in Eq. (1) can be used to perform pixel-wise adjustment of image and depth features formulated as follows:

$$\hat{\mathbf{I}}_j = \mathcal{F}_\psi(\mathbf{I}_j;\ \Delta\bar{\mathbf{p}}_j, \bar{\mathbf{w}}_j), \tag{4a}$$

$$\hat{\mathbf{S}}_j = \mathcal{F}_\psi(\mathbf{S}_j;\ \Delta\tilde{\mathbf{p}}_j, \tilde{\mathbf{w}}_j), \tag{4b}$$

where $\mathcal{F}_\psi(\cdot)$ denotes the generalized form of the operation defined in Eq. (1). Note that Eq.(4) emphasizes RGB-D pixel redistribution under highly dynamic conditions with offsets instead of adaptive weights. Consequently, in contrast to Eq.(1), both $\bar{\mathbf{w}}_j$ and $\tilde{\mathbf{w}}_j$ correspond to the predefined weights $\mathbf{w}$ with $\Delta\mathbf{m}$ being the identity matrix. Subsequently, the redistributed RGB-D features are further processed by a 3×3 convolution $\mathcal{F}_{\tau_{j6}}(\cdot)$ to obtain the fused feature $\mathbf{F}_j \in \mathbb{R}^{C\times H\times W}$, which is formulated as:

$$\mathbf{F}_j = \mathcal{F}_{\tau_{j6}}(\hat{\mathbf{I}}_j + \hat{\mathbf{S}}_j). \tag{5}$$

Additionally, a structure-aware loss $\mathcal{L}_{str}$ is introduced to enhance the consistency. Let $\mathcal{G}(\cdot)$ denote a sequence of single-channel convolution, Min-Max normalization, and gradient computation:

$$\mathcal{L}_{str} = \sum_{j=1}^{4} \frac{1}{n} \|\mathcal{G}(\hat{\mathbf{I}}_j) - \mathcal{G}(\hat{\mathbf{S}}_j)\|_2^2. \tag{6}$$

Local Depth Filtering. As shown in Fig.2, the LDF module takes the depth feature $\mathbf{D}_i$ and event feature $\mathbf{E}_i$ as inputs to adaptively generate offsets and weights at the i-th stage of the decoder ($i \in \{1, 2, 3\}$). Following the strategy used in Eqs.(2)–(4), this results in the updated depth feature:

$$\hat{\mathbf{D}}_i = \mathcal{F}_\psi(\mathbf{D}_i;\ \Delta\tilde{\mathbf{p}}_i, \tilde{\mathbf{w}}_i). \tag{7}$$

In contrast, the modulation scalar Δm within $\tilde{\mathbf{w}}_i$ is learned jointly from the depth and event inputs. Furthermore, to explicitly model regions of dynamic objects, LDF predicts a motion mask m_i based on $\mathbf{E}_i$ using a sigmoid activation $\sigma(\cdot)$ after a single-channel 3×3 convolution $\mathcal{F}_{\tau_{i6}}(\cdot)$:

$$m_i = \sigma(\mathcal{F}_{\tau_{i6}}(\mathbf{E}_i)). \tag{8}$$

By combining Eqs. (7) and (8), LDF refines depth with a focus on dynamic regions to get:

$$\mathring{\mathbf{D}}_i = m \cdot \hat{\mathbf{D}}_i + (1 - m) \cdot \mathbf{D}_i. \tag{9}$$

Finally, the output $\mathring{\mathbf{D}}_1$ from the last LDF module is passed through a 3×3 convolutional tail $\mathcal{F}_{\tau_t}(\cdot)$ to generate the dense depth prediction:

$$\mathbf{D} = \mathcal{F}_{\tau_t}(\mathring{\mathbf{D}}_1). \tag{10}$$

Additionally, we introduce a motion-aware loss to enhance the depth recovery around motion areas:

$$\mathcal{L}_{mot} = \sum_{i=1}^{3} \frac{1}{n} \|b_i \cdot \mathcal{H}(\mathring{\mathbf{D}}_i) - b_i \cdot \mathcal{F}_d(\mathbf{Z})\|_2^2, \tag{11}$$

where $\mathbf{Z}$ is the GT depth, $\mathcal{H}(\cdot)$ applies ReLU and a single-channel convolution, and b_i is a binary mask with $b_i = 1$ if m_i exceeds its mean, and 0 otherwise. $\mathcal{F}_d(\cdot)$ denotes the downsampling operation.

Discussion. In summary, EMA and LDF differ from previous dynamic convolution methods in two key aspects: **(1)** Unlike traditional methods that typically rely on single-modal and single-path inputs, our approach adopts a multi-modal and multi-path input design, where key convolutional parameters are derived from different modalities. **(2)** Our method is data-driven where we use event-based adaptation to address global misalignment and local depth inaccuracies caused by fast motion.

3.3 Loss Function

Given the predicted depth $\mathbf{D}$ and GT depth $\mathbf{Z}$ with n valid pixels, we adopt a commonly used reconstruction loss [36, 27, 69, 56, 55] to formulate the training objective:

$$\mathcal{L}_{rec} = \frac{1}{n}\left(\left\|\mathbf{D} - \mathbf{Z}\right\|_2^2 + \|\mathbf{D} - \mathbf{Z}\|_1\right). \tag{12}$$

By combining the reconstruction loss with the structure-constrained loss $\mathcal{L}_{str}$ in Eq. 6 and motion-aware loss $\mathcal{L}_{rec}$ in Eq. 11, the overall loss function is formulated as:

$$\mathcal{L}_t = \mathcal{L}_{rec} + \lambda\mathcal{L}_{str} + \mu\mathcal{L}_{mot}, \tag{13}$$

where λ and μ are weighting hyper-parameters that we empirically set to 1 and 0.1, respectively.

Table 1: Basic statistics of the EventDC benchmark.

Dataset	Color Camera	Depth Sensor	Event Camera	Train	Test	Resolution
EventDC-Real	FLIR BFS-U3-31S4C	Ouster OS1-128 LiDAR	DAVIS346	14,845	1,000	320×256
EventDC-SemiSyn	PointGrey Flea2	Velodyne HDL-64E LiDAR	-	7,094	2,213	1216×256
EventDC-FullSyn	-	-	-	21,000	500	512×256

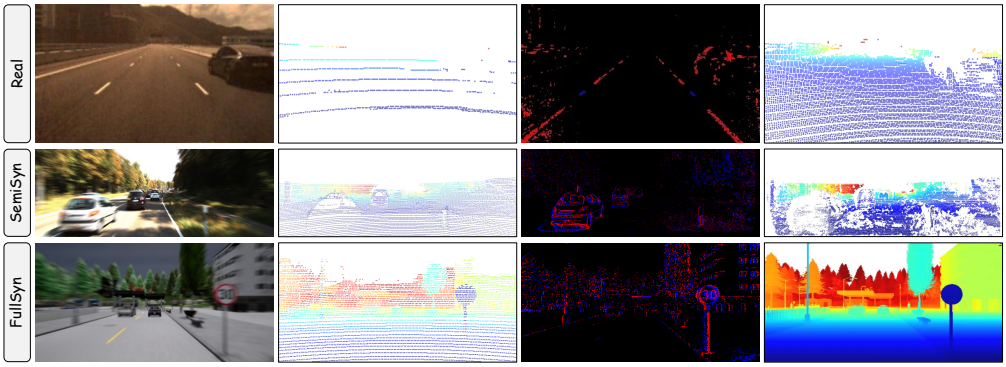


Figure 3: Visualizations of the proposed EventDC benchmark: EventDC-Real/SemiSyn/FullSyn.

4 EventDC Benchmark

Motivation. Traditional depth completion datasets [14, 42, 65, 44] rely on the fusion of color images and sparse depth maps to predict dense depth. However, this approach suffers in highly dynamic environments especially when dealing with fast ego-motion and object motion. This is due to unreliable low-frame-rate RGB images and sparse depth data from motion blur and sampling inconsistencies. Event cameras with the capability to capture high temporal resolution and sensitivity to rapid movements [10] provide an ideal solution to overcome these limitations. By asynchronously recording minute brightness variations, event cameras can offer accurate depth information in dynamic scenarios where conventional RGB-D sensors fail. In light of these characteristics, we propose an event-based depth completion benchmark that leverages the unique advantages of event data to address the challenges of depth completion in dynamic environments.

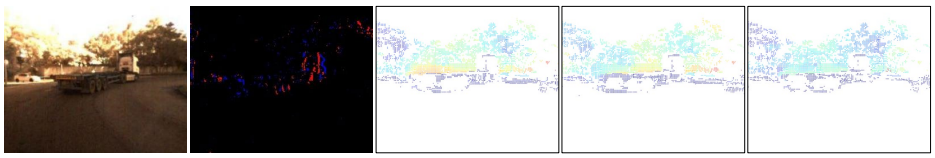
Data Collection. Tab. 1 provides an overview of the sensors used in the datasets with their respective specifications. *EventDC-Real* is a real-world dataset in which color images and event frames are captured using the FLIR BFS-U3-31S4C camera and the DAVIS346 sensor, respectively. The ground truth (GT) depth is acquired from a 128-line Ouster LiDAR, and the sparse depth is derived from its 16 sub-lines. *EventDC-SemiSyn* is a semi-synthetic dataset based on KITTI [14]. The sparse depth and GT depth come from the raw data of KITTI. For the color images, we apply radial motion blur by progressively scaling and transforming the image around its center to simulate a motion blur effect with adjustable strength and step count. Additionally, VID2E [12] is used to generate the event data with frames captured within 15 ms before and after the current timestamp. *EventDC-FullSyn* is a fully synthetic dataset generated using the CARLA simulator [6]. The color images are processed similarly with radial motion blur. Finally, to facilitate model training, the resolution of all datasets has been cropped to multiples of 32. Fig. 3 presents visual examples from these three datasets.

5 Experiment

Metric and Implementation Detail. Following previous depth completion methods [14, 46, 65, 26], we adopt RMSE (mm), MAE (mm), REL, and threshold accuracy δ (%) as evaluation metrics. Refer to the appendix for their full definitions. We implement EventDC using the PyTorch framework and conduct training on two NVIDIA RTX 4090 GPUs using the Distributed Data Parallel strategy for efficiency. Optimization is performed with the AdamW optimizer [31] in conjunction with the OneCycle learning rate policy [43]. The training process begins with a warm-up stage that linearly increases the learning rate from 0.00002 to 0.001 over the first 10% of iterations. Subsequently, a cosine annealing schedule gradually decays the learning rate to a final value of 0.0002. The batch size is set to 2 per GPU. In addition, to further enhance model performance, we employ a set of data augmentation strategies [46, 29], including random horizontal flip, rotation, cropping, and color jitter.

Table 2: Quantitative depth completion comparisons on the EventDC-Real dataset.

Method	RMSE ↓	MAE ↓	REL ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.15}$ ↑	Venue
CSPN [2]	858.5	284.6	0.0386	90.0	94.4	96.0	ECCV 2018
S2D [33]	984.1	410.8	0.0565	82.1	90.3	93.4	ICRA 2018
FusionNet [49]	658.1	262.4	0.0384	87.6	93.9	96.0	MVA 2019
RigNet [61]	685.4	234.6	0.0336	87.5	92.3	95.4	ECCV 2022
DySPN [27]	700.1	223.7	0.0285	91.3	95.1	96.6	AAAI 2022
Prompting [38]	670.7	205.1	0.0252	92.1	95.7	97.0	CVPR 2024
OGNI-DC [78]	709.7	231.1	0.0294	90.9	94.8	96.4	ECCV 2024
SigNet [66]	906.4	348.1	0.0345	83.5	90.6	93.2	CVPR 2025
LPNet [55]	911.2	389.0	0.0472	83.6	90.4	93.3	arXiv 2025
EventDC (our)	**574.0**	**179.0**	**0.0242**	**92.9**	**96.3**	**97.5**	-
Improvement ↑	*84.1*	*26.1*	*0.0010*	*0.8*	*0.6*	*0.5*	-



Color image Event Prompting OGNI-DC **EventDC**

Figure 4: Depth error comparisons on EventDC-Real. Warmer color indicates higher error.

5.1 Comparisons with State-of-the-arts

In this section, we compare our EventDC with well-known methods: CSPN [2], S2D [33], FusionNet [49], RigNet [61], DySPN [27], Prompting [38], OGNI-DC [78], SigNet [66], and LPNet [55]. For a fair comparison, we retrain all methods from scratch on the proposed benchmark. Note that BPNet [47], TPVD [65], and DMD^{3}C [26] are excluded from the comparison. This is because they require additional camera parameters during training which are not available in our settings.

EventDC-Real. We first evaluate the proposed EventDC on EventDC-Real, a real-world dataset collected using various devices such as handheld sensors and robotic platforms. The numerical results are summarized in Tab 2. Our EventDC achieves the overall lowest errors while maintaining the highest accuracy across the board. For example, it outperforms the second-best method by 84.1 mm in RMSE, 26.1 mm in MAE, 0.001 in REL, and 0.8 points in $\delta_{1.05}$. Compared to post-refinement methods such as CSPN [2] and DySPN [27], our EventDC without any post-processing consistently achieves better performance. Even when compared to the large-scale depth foundation model Prompting [38], our approach achieves superior results with significantly fewer model parameters. Fig. 4 presents the comparisons of depth error. It clearly shows that our EventDC produces more accurate depth results especially around moving objects.

EventDC-SemiSyn. To further validate the effectiveness of EventDC, we evaluate it on EventDC-SemiSyn, a semi-synthetic dataset comprising synthetically generated event frames and color images rendered under highly dynamic conditions. As reported in Tab. 3, our EventDC continues to deliver outstanding results across all evaluation metrics. On average, it outperforms recent methods: Prompting [38], OGNI-DC [78], and LPNet [55] by 18.9%, 30.8%, and 27.7% in RMSE, MAE, and REL, respectively, and by 3.9, 1.6, and 0.9 percentage points in $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.15}$, respectively. As illustrated in Fig. 5, our EventDC effectively reconstructs accurate depth details and structural consistency even under highly dynamic scenes.

EventDC-FullSyn. Apart from the real and semi-synthetic settings, we also validate EventDC on the fully synthetic dataset, EventDC-FullSyn, to further assess its generalization capability under diverse scenarios. As shown in Tab.4, our EventDC consistently outperforms all competing approaches by large margins. For example, it surpasses the second-best approach by 53.5 mm in RMSE and 19.4 mm in MAE. In addition, it achieves a 13.0% improvement in REL compared to the foundation model-based Prompting [38]. These results demonstrate the robustness of our EventDC in reducing both absolute and relative errors. Fig. 6 shows that our EventDC yields more refined details and sharper object boundaries than others, which highlight its effectiveness in fully synthetic scenarios.

Table 3: Quantitative comparisons on the EventDC-SemiSyn dataset.

Method	RMSE↓	MAE↓	REL↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.15}$ ↑
CSPN [2]	989.8	262.8	0.0189	94.6	97.2	98.1
S2D [33]	1097.3	366.4	0.0237	91.0	96.4	97.9
FusionNet [49]	877.6	333.1	0.0258	92.6	98.2	98.8
RigNet [61]	858.2	216.4	0.0156	95.1	97.8	98.1
DySPN [27]	897.7	207.5	0.0149	95.9	97.8	98.6
Prompting [38]	873.9	291.1	0.0198	92.6	97.1	98.4
OGNI-DC [78]	832.0	210.5	0.0143	95.7	98.0	98.7
SigNet [66]	1065.4	321.3	0.0226	91.1	97.0	98.1
LPNet [55]	1283.4	416.3	0.0242	90.0	95.3	97.2
EventDC (ours)	**778.8**	**196.2**	**0.0134**	**96.7**	**98.4**	**99.0**
Improvement ↑	*53.2*	*11.3*	*0.0009*	*0.8*	*0.2*	*0.2*

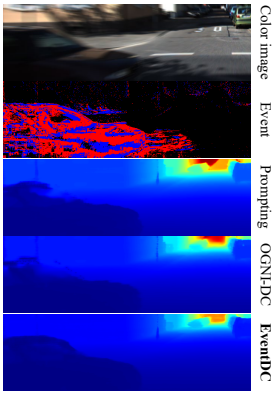


Figure 5: Visual results.

Table 4: Quantitative comparisons on the EventDC-FullSyn dataset.

Method	RMSE↓	MAE↓	REL↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.15}$ ↑
CSPN [2]	864.9	399.5	0.1193	62.7	80.9	87.6
S2D [33]	899.0	376.2	0.1243	69.8	83.8	89.1
FusionNet [49]	670.6	230.9	0.0931	77.3	86.6	90.4
RigNet [61]	723.4	166.3	0.0578	81.1	91.6	92.8
DySPN [27]	679.8	165.6	0.0646	87.2	92.6	94.6
Prompting [38]	709.7	180.9	0.0538	90.7	93.8	95.3
OGNI-DC [78]	673.7	162.5	0.0578	87.8	93.0	95.1
SigNet [66]	904.5	349.2	0.0902	76.3	84.1	90.3
LPNet [55]	920.2	357.3	0.0943	75.1	85.9	90.3
EventDC (ours)	**620.2**	**143.1**	**0.0468**	**92.1**	**95.5**	**96.8**
Improvement ↑	*53.5*	*19.4*	*0.0070*	*1.4*	*1.7*	*1.5*

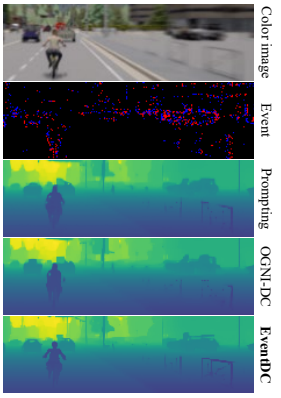


Figure 6: Visual results.

Complexity Analysis. Tab. 5 presents the complexity comparisons between our EventDC and other competing methods in terms of model parameters (Param.), memory consumption (Memo.), and inference time. Our EventDC not only achieves outstanding performance, but also maintains competitive efficiency. In particular, compared to the second-best method Prompting [38], our EventDC achieves a significantly lower RMSE by 96.7 mm with only about one-eighth the number of parameters and one-third the memory.

Table 5: Complexity on EventDC-Real.

Method	Param. (M) ↓	Memo. (GB) ↓	Time (ms) ↓	RMSE (mm) ↓
DySPN [27]	**26.3**	**0.9**	**9.8**	700.1
RigNet [61]	65.2	2.3	26.5	685.4
Prompting [38]	326.9	4.1	39.5	670.7
OGNI-DC [78]	84.4	3.7	314.1	709.7
LPNet [55]	29.6	1.1	18.4	911.2
EventDC (our)	43.2	1.5	41.5	**574.0**

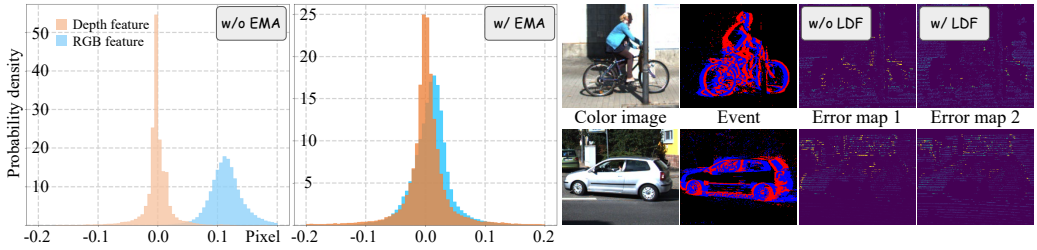
5.2 Ablation Studies

Tab. 6 summarizes the ablation results on EventDC-Real. EventDC-i serves as a UNet-style baseline that takes only sparse depth as input and employs additive skip connections.

(1) EventDC-ii further develops this approach by utilizing RGB images and integrating RGB-D features through additive fusion. Although depth input is sparse, RGB offers rich structural and semantic details. This leads to a significant decrease in error and substantial gains in accuracy. For example, the RMSE is reduced by 41.3 mm and the MAE by 32.1 mm. EventDC-iii enhances support for event streams, which are advantageous because of their fine temporal detail and motion sensitivity. Consequently, this makes them very effective in dynamic environments where they supplement depth data. EventDC-iv combines all three modalities to give consistent improvements across all evaluation metrics. Specifically, it surpasses the baseline by 12.2%, 20.7%, and 8.5% in RMSE, MAE, and REL, respectively. It concurrently improves $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.15}$ by 1.0, 0.3, and 0.2 percentage points. These results underscore the effectiveness of multi-modal fusion, where the integration of complementary modalities enables more accurate and complete depth reconstruction.

Table 6: Ablations on EventDC-Real. DConv/Enc/Dec: dynamic convolution/encoder/decoder.

EventDC	Modality			DConv		EMA	LDF	RMSE	MAE	REL	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.15}$
	Depth	RGB	Event	Enc	Dec	Enc	Dec						
i	✓							727.2	283.2	0.0341	89.3	94.4	96.2
ii	✓	✓						685.9	251.1	0.0328	89.8	94.5	96.2
iii	✓		✓					696.0	244.3	0.0314	90.0	94.5	96.3
iv	✓	✓	✓					638.3	224.7	0.0312	90.3	94.7	96.4
v	✓	✓	✓	✓				628.8	219.5	0.0292	91.0	95.0	96.7
vi	✓	✓	✓			✓		602.8	196.1	0.0276	91.7	95.6	97.1
vii	✓	✓	✓		✓			630.6	219.8	0.0295	91.0	94.8	96.5
viii	✓	✓	✓				✓	605.4	198.3	0.0279	91.5	95.6	97.1
ix	✓	✓	✓			✓	✓	574.0	179.0	0.0242	92.9	96.3	97.5



(a) Histogram comparison (b) Results around moving objects

Figure 7: Statistical and visual comparative analyses of the proposed EMA and LDF modules.

(2) EventDC-v to EventDC-ix conduct ablation studies to examine the impact of dynamic convolution (DConv), EMA, and LDF in the encoder (Enc) and decoder (Dec) stages. Specifically, the introduction of DConv in EventDC-v brings notable benefits. Furthermore, EventDC-vi with EMA further reduces RMSE by 26 mm. These results validate the efficacy of our event-based adaptive alignment strategy. Fig. 7(a) compares the distributions of RGB-D features with and without EMA. EMA works as intended in promoting better alignment between the two modalities with more consistent feature representations. Similarly, EventDC-viii with LDF outperforms EventDC-vii with DConv by 25.2 mm. This demonstrates its superior ability to recover fine-grained local depth which is further evident in Fig. 7(b). Finally, EventDC-ix which integrates both EMA and LDF modules achieves the best overall performance. It reduces RMSE by 16.0% (from 683.3 mm) and MAE by 20.3% (from 224.7 mm). In summary, each component contributes positively to the overall performance gains.

6 Conclusion

We propose EventDC in this work. Our EventDC is the first depth completion framework that tackles the challenges of dynamic scenes by harnessing the unique strengths of event data. To mitigate the adverse effects of fast ego-motion and object motion, our EventDC incorporates two event-driven modules: event-modulated alignment and local depth filtering. These modules, supported by two dedicated loss constraints, address global misalignment and local depth inaccuracies, respectively. To further support research in this area, we construct the first benchmark for event-based depth completion comprising one real-world and two synthetic datasets. Extensive experiments demonstrate the effectiveness of our EventDC and its superior performance in challenging dynamic environments.

Limitation and Broader Impact. Despite achieving promising results in dynamic scenes, our EventDC relies on high-quality event data and precise sensor alignment that may not be easily attainable in all real-world settings. The EMA and LDF modules introduce additional computational costs, potentially limiting deployment on resource-constrained devices. Moreover, the scale and diversity of our real-world dataset are limited, and future work is needed to evaluate generalization across more diverse environments and motion patterns. Despite these limitations, our EventDC offers a step forward in robust depth perception under motion blur and rapid dynamics with potential applications in autonomous driving, robotics, AR/VR, *etc*. By introducing a dedicated benchmark, we aim to promote research in event-based depth completion. As with all perceptual systems, responsible deployment requires attention to reliability, fairness, and safety in complex real-world conditions.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008) and the Tier 1 grant T1-251RES2305 from the Singapore Ministry of Education.

References

[1] Yun Chen, Bin Yang, Ming Liang, and Raquel Urtasun. Learning joint 2d-3d representations for depth completion. In *ICCV*, pages 10023–10032, 2019.

[2] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. In *ECCV*, pages 103–119, 2018.

[3] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2361–2379, 2019.

[4] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI*, pages 10615–10622, 2020.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, pages 1–16. PMLR, 2017.

[7] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2423–2436, 2020.

[8] Guillermo Gallego, Jon EA Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-dof camera tracking from photometric depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2402–2412, 2017.

[9] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *CVPR*, pages 3867–3876, 2018.

[10] 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.

[11] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.

[12] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *CVPR*, pages 3586–3595, 2020.

[13] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[15] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation from ego-motion using ray density fusion. *arXiv preprint arXiv:2210.08927*, 2022.

[16] Jakub Gregorek and Lazaros Nalpantidis. Steeredmarigold: Steering diffusion towards depth completion of largely incomplete depth maps. *arXiv preprint arXiv:2409.10202*, 2024.

[17] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *3DV*, pages 534–542. IEEE, 2020.

[18] Diego de Oliveira Hitzges, Suman Ghosh, and Guillermo Gallego. Derd-net: Learning depth from event-based ray densities. *arXiv preprint arXiv:2504.15863*, 2025.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[20] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *ICRA*, 2021.

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015.

[22] Yunho Jeon and Junmo Kim. Active convolution: Learning the shape of convolution for image classification. In *CVPR*, pages 4201–4209, 2017.

[23] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[25] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *CRV*, pages 16–22. IEEE, 2018.

[26] Yingping Liang, Yutao Hu, Wenqi Shao, and Ying Fu. Distilling monocular foundation model for fine-grained depth completion. *arXiv preprint arXiv:2503.16970*, 2025.

[27] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *AAAI*, pages 1638–1646, 2022.

[28] Yuankai Lin, Hua Yang, Tao Cheng, Wending Zhou, and Zhouping Yin. Dyspn: Learning dynamic affinity for image-guided depth completion. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

[29] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *AAAI*, pages 2136–2144, 2021.

[30] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Kaiyue Lu, Nick Barnes, Saeed Anwar, and Liang Zheng. From depth what can you see? depth completion via auxiliary image reconstruction. In *CVPR*, pages 11306–11315, 2020.

[33] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *ICRA*, pages 4796–4803. IEEE, 2018.

[34] Manasi Muglikar, Diederik Paul Moeys, and Davide Scaramuzza. Event guided depth sensing. In *3DV*, pages 385–393. IEEE, 2021.

[35] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *ICRA*, pages 10695–10702. IEEE, 2024.

[36] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*, 2020.

[37] Jin-Hwi Park and Hae-Gon Jeon. A simple yet universal framework for depth completion. In *NeurIPS*, 2024.

[38] Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *CVPR*, pages 9859–9869, 2024.

[39] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018.

[40] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *ICCV*, pages 7931–7940, 2023.

[41] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[43] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications*, pages 369–386. SPIE, 2019.

[44] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[45] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang. Channel attention based iterative residual learning for depth map super-resolution. In *CVPR*, pages 5631–5640, 2020.

[46] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020.

[47] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *CVPR*, pages 9763–9772, 2024.

[48] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, pages 11–20, 2017.

[49] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *MVA*, pages 1–6, 2019.

[50] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[51] Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. *arXiv preprint arXiv:2412.13389*, 2024.

[52] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *ICRA*, page 4976–4982. IEEE, 2024.

[53] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025.

[54] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025.

[55] Kun Wang, Zhiqiang Yan, Junkai Fan, Jun Li, and Jian Yang. Learning inverse laplacian pyramid for progressive depth completion. *arXiv preprint arXiv:2502.07289*, 2025.

[56] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *ICCV*, pages 9422–9432, 2023.

[57] Yufei Wang, Ge Zhang, Shaoqian Wang, Bo Li, Qi Liu, Le Hui, and Yuchao Dai. Improving depth completion via depth feature upsampling. In *CVPR*, pages 21104–21113, 2024.

[58] Hexiang Wei, Jianhao Jiao, Xiangcheng Hu, Jingwen Yu, Xupeng Xie, Jin Wu, Yilong Zhu, Yuxuan Liu, Lujia Wang, and Ming Liu. Fusionportablev2: A unified multi-sensor dataset for generalized slam across diverse platforms and scalable environments. *The International Journal of Robotics Research*, page 02783649241303525, 2024.

[59] Zheyuan Xu, Hongche Yin, and Jian Yao. Deformable spatial propagation networks for depth completion. In *ICIP*, pages 913–917. IEEE, 2020.

[60] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Guangyu Li, Jun Li, and Jian Yang. Learning complementary correlations for depth super-resolution with incomplete data in real world. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[61] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *ECCV*, pages 214–230, 2022.

[62] Zhiqiang Yan, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet++: Efficient repetitive image guided network for depth completion. *arXiv preprint arXiv:2309.00655*, 2023.

[63] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *AAAI*, pages 3109–3117, 2023.

[64] Zhiqiang Yan, Yupeng Zheng, Kun Wang, Xiang Li, Zhenyu Zhang, Shuo Chen, Jun Li, and Jian Yang. Learnable differencing center for nighttime depth perception. *arXiv preprint arXiv:2306.14538*, 2023.

[65] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *CVPR*, pages 4874–4884, 2024.

[66] Zhiqiang Yan, Zhengxue Wang, Kun Wang, Jun Li, and Jian Yang. Completion as enhancement: A degradation-aware selective image guided network for depth completion. *arXiv preprint arXiv:2412.19225*, 2024.

[67] Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *ICCV*, pages 8732–8743, 2023.

[68] Xin Zhang, Liangxiu Han, Tam Sobeih, Lianghao Han, and Darren Dancey. A novel spike transformer network for depth estimation from event cameras via cross-modality knowledge distillation. *arXiv preprint arXiv:2404.17335*, 2024.

[69] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *CVPR*, pages 18527–18536, 2023.

[70] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing*, 2021.

[71] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, 2022.

[72] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *ICCV*, pages 12547–12558, 2023.

[73] Wending Zhou, Xu Yan, Yinghong Liao, Yuankai Lin, Jin Huang, Gangming Zhao, Shuguang Cui, and Zhen Li. Bev@ dc: Bird's-eye view assisted training for depth completion. In *CVPR*, pages 9233–9242, 2023.

[74] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019.

[75] Junyu Zhu, Lina Liu, Bofeng Jiang, Feng Wen, Hongbo Zhang, Wanlong Li, and Yong Liu. Self-supervised event-based monocular depth estimation using cross-modal consistency. In *IROS*, pages 7704–7710. IEEE, 2023.

[76] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019.

[77] Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based visual inertial odometry. In *CVPR*, pages 5391–5399, 2017.

[78] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *ECCV*, pages 78–95. Springer, 2024.

A Metrics.

We adopt the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Relative Error (REL), and threshold accuracy δ_θ as evaluation metrics, where θ is set to 1.05, 1.10, and 1.15. The definitions of these metrics are shown in Tab. 7.

Table 7: Definition of evaluation metrics.

Given the predicted depth $\mathbf{D}$ and GT depth $\mathbf{Z}$ with n valid pixels:			
– RMSE:	$\sqrt{\frac{1}{n}\sum(\mathbf{D}-\mathbf{Z})^2}$	– MAE:	$\frac{1}{n}\sum\lvert\mathbf{D}-\mathbf{Z}\rvert$
– REL:	$\frac{1}{n}\sum\lvert\mathbf{D}-\mathbf{Z}\rvert/\mathbf{Z}$	– δ_θ:	$\frac{q}{n},\ q:\max\left(\frac{\mathbf{D}}{\mathbf{Z}},\frac{\mathbf{Z}}{\mathbf{D}}\right)<\theta$

Figure 8: More visual examples of the proposed event-based depth completion benchmark.

B More Visualizations

Fig. 8 presents some RGB-D-Event examples from our event-based depth completion benchmark, showcasing its high quality and strong cross-modal consistency, as well as the close correlation among the RGB, depth, and event modalities. Figs. 9, 10 and 11 show visual comparisons on the EventDC-Real, EventDC-SemiSyn, and EventDC-FullSyn datasets. These results further validate that our approach effectively improves depth predictions through the event-driven module designs.

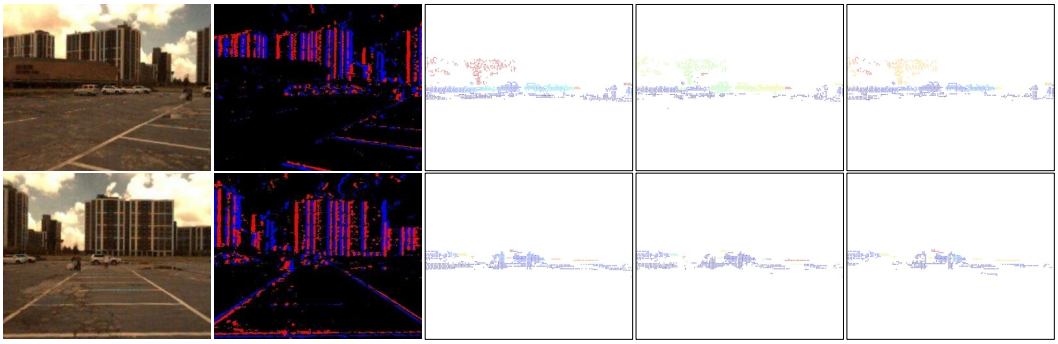


Figure 9: More depth error comparisons on EventDC-Real. Warmer colors indicate higher errors.

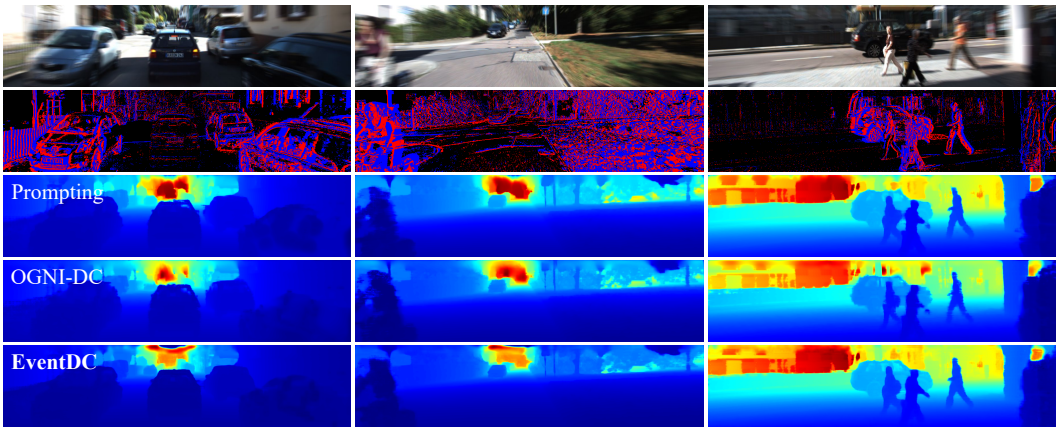


Figure 10: More depth visualization comparisons on the proposed EventDC-SemiSyn dataset.

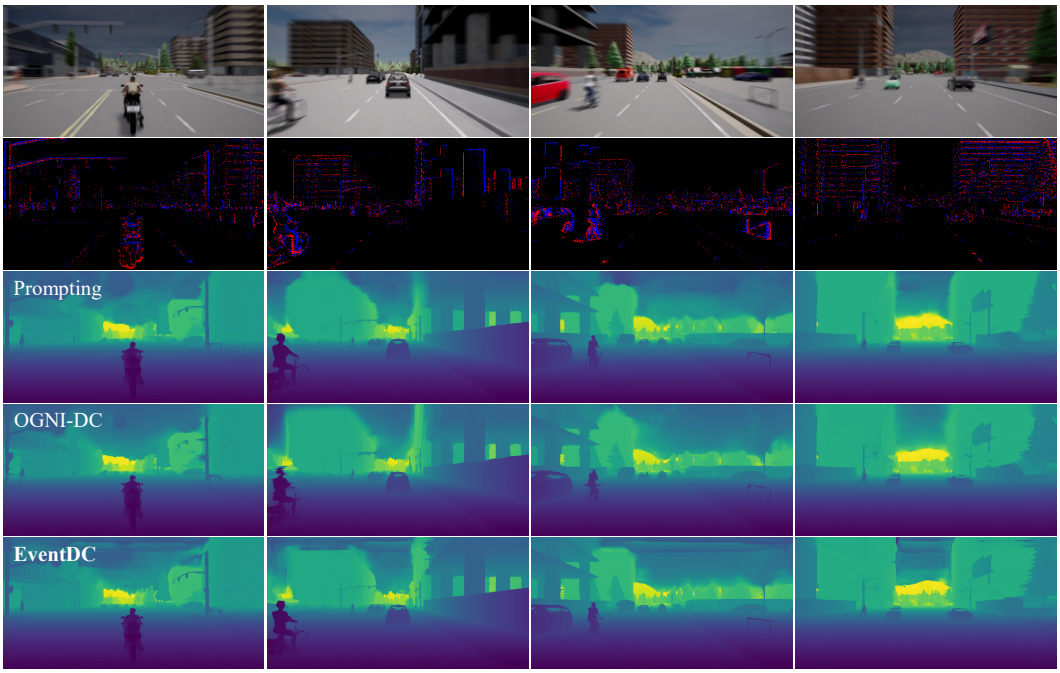


Figure 11: More depth visualization comparisons on the proposed EventDC-FullSyn dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction have accurately reflected the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the section of 'Limitation and Broader Impact'.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the steps taken to make our results reproducible or verifiable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data and source codes will be publicly available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have made sure to preserve anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the section of 'Limitation and Broader Impact'.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CC-BY 4.0.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have communicated the details of the dataset/code/model as part of our submissions via structured templates.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.