Depth Hypothesis Guided Iterative Refinement for Event–Image Monocular Depth Estimation

Supplementary Material

6. Network Details

HypoDepth. Our model performs depth refinement progressively from 1/16 to 1/4 resolution. The detailed network configurations are summarized in Tab. 8, including the backbone, channel dimensions of core modules, the number of iterative updates, and the parameters related to cost-volume lookup. **Ours** denotes the event–image fusion model, while **Ours-E** represents the event-only variant, in which the image encoder and EIFusion module are removed. Notably, we adopt resolution-dependent lookup scales (l in Fig. 4 (a)) across different stages in our base model. At lower resolutions (such as 1/16 resolution), the lookup radius r already covers most or even all spatial regions of the feature map, making the resulting correlation effectively global. Further downsampling would cause the lookup region to span nearly the entire feature map again, leading to redundant correlations and reducing the effectiveness of the multi-scale lookup strategy. For the scaled-down model in the ablation studies, we train it with a batch size of 3 for 40k iterations.

PCDepth-Ours. Since PCDepth [22] performs iterative feature-level fusion refinement at 1/8 resolution for depth optimization, we also conduct refinement at the same scale for fair comparison. The contextual features are obtained from PCDepth's fused encoder output, while the parameters of our iterative refinement decoder are listed in Tab. 8.

7. More Results

More Comparison on MVSEC and DSEC. In Tab. 5, we report the results of our event-only (Ours-E-B) and image-only (Ours-I-B) variants on DSEC and MVSEC to verify the benefit of event–image fusion. In addition, we simulate image motion blur on DSEC to mimic high-speed scenarios, and directly evaluate the model trained on the original DSEC under zero-shot transfer to the blurred setting. Tab. 5 compares our image-only variant (w/o events) with PCDepth in this challenging scenario. The results show that incorporating events consistently improves performance, with larger gains in challenging conditions such as nighttime (MVSEC_on1) and high-speed motion (blurred DSEC). In these cases, frame appearance is severely degraded by low illumination or motion blur, whereas events still provide rich structural and texture cues, leading to more reliable correspondence and estimation; in contrast, frame-only methods degrade notably. Moreover, our method generalizes better than PCDepth under high-speed blur. The qualitative comparisons in Fig. 9 further confirm our advantage in the presence of image blur.

Table 5. More results on MVSEC [40], DSEC [9], and blurred DSEC. The best are in bold.

Dataset	Input	Method	$\delta 1 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$
DSEC	I	Ours-I-B	0.859	0.116	4.183
	E	Ours-E-B	0.839	0.129	4.442
	E+I	**Ours-B**	**0.901**	**0.099**	**3.583**
MVSEC_od1 / on1	I	Ours-I-B	0.692 / 0.534	0.233 / 0.330	7.020/8.537
	E	Ours-E-B	0.690 / 0.600	0.243 / 0.289	6.732 / 7.201
	E+I	Ours-T	0.682 / 0.573	0.260 / 0.312	7.157 / 7.597
	E+I	**Ours-B**	**0.698 / 0.614**	**0.212 / 0.268**	**6.556 / 6.834**
Blurred DSEC	I	Ours-I-B	0.121	0.5207	15.9315
	E+I	**Ours-B**	**0.665**	**0.1828**	**6.3642**
	E+I	PCDepth	0.209	0.454	14.627

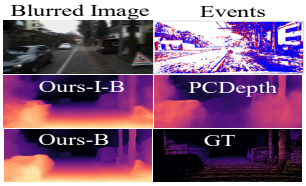

Figure 9. Qualitive results on blurred DSEC dataset.

Duration of Event Slices. We investigate the impact of different event-stream window lengths, including 20 ms, 50 ms, and 100 ms. Results in Tab. 6 show that 20 ms and 50 ms achieve comparable performance, indicating robustness to the temporal window. In contrast, longer windows cause excessive event accumulation under large motion, increasing misalignment and motion artifacts, blurring the matching distribution, and degrading performance. We therefore adopt a 20 ms window to reduce latency, better leverage the strengths of event cameras, and ensure a fair performance comparison.

Table 6. Ablation studies on the duration of event slices.

Duration (ms)	$\delta 1 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$
<u>20</u>	<u>0.885</u>	**0.105**	<u>3.857</u>
50	**0.887**	<u>0.106</u>	**3.705**
100	0.876	0.112	3.869

Depth-range of DHV. We investigate the effect of the depth-range on accuracy. DHV defines a coarse search space, while the subsequent continuous residual updates refine depth and alleviate discretization errors. Experiments show that an overly narrow depth range degrades accuracy, whereas expanding the range further has a marginal impact.

Table 7. Ablation studies on the depth-range of DHV.

Depth range	$\delta 1 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$
$[0.05 * d_{min}, 2 * d_{max}]$	0.884	0.107	3.833
$[d_{min}, d_{max}]$	0.885	0.105	3.857
$[2 * d_{min}, 0.7 * d_{max}]$	0.793	0.141	4.533

8. Visualizations

Qualitative Results on DSEC. In Fig. 11, we present additional qualitative comparisons on the DSEC [9] among **Ours-B**, PCDepth [22], and **PCDepth-Ours**. The results demonstrate that on this complex-structured dataset, our approach consistently produces more geometrically plausible and spatially accurate depth maps under varying illumination conditions compared to PCDepth. Moreover, when equipped with our iterative refinement module, **PCDepth-Ours** also benefits from sharper and more accurate object boundaries. Thanks to the additional 1/4-resolution refinement stage in our model, our depth predictions exhibit finer structural details than those of **PCDepth-Ours**.

Qualitative Results on MVSEC. We present a qualitative comparison between our method and PCDepth [22] on both daytime and nighttime sequences of the MVSEC [40] in Fig. 12. The visual results demonstrate that our approach accurately estimates both distant (rows 2 and 3) and nearby (rows 1 and 4) structures under challenging lighting conditions and sparse event inputs. In particular, PCDepth fails to recover distant regions, whereas our model produces stable and continuous estimates, which are crucial for enabling autonomous systems to perceive distant scenes in advance.

More visualizations of refinement on DSEC. We visualize additional multi-resolution depth refinement processes and corresponding error in Fig. 13. At the 1/16 resolution, the model performs globally consistent regression, producing coarse estimations. When refined at the 1/8 level, the scene geometry becomes more pronounced. Further refinement at the 1/4 resolution sharpens both distant and nearby structures, demonstrating the effectiveness of our method in achieving coarse-to-fine depth reconstruction.

9. Limitations and Future Work

We develop a depth-hypothesis–guided iterative refinement framework for event–image monocular depth estimation, achieving high-performance results. To further improve the method, we analyze several remaining limitations.

First, the superior performance of **PCDepth-Ours** over **Ours-B** indicates that our multimodal fusion strategy still has space for improvement. A more adaptive fusion mechanism could better leverage the complementary advantages of the two modalities. For instance, emphasizing image features in static regions where events fail, and focusing on event-based spatiotemporal cues when image appearance becomes unreliable.

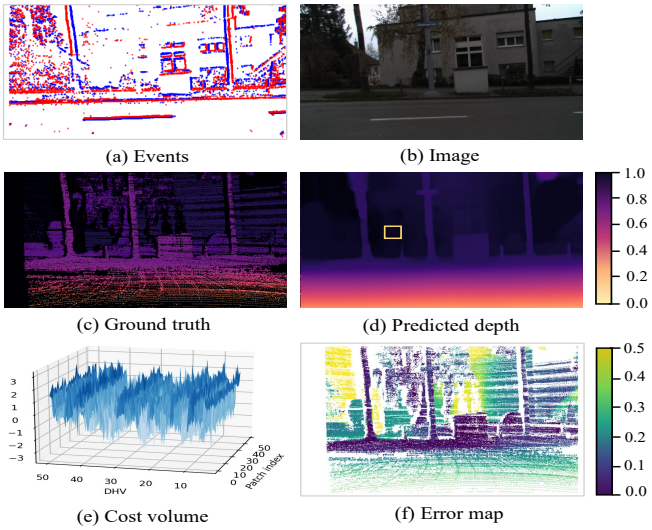

Figure 10. Failure case analysis. (e) shows the cost-volume distribution of the yellow patch in (d). (f) shows the log-normalized error. Large errors appear between the predicted depth and the ground truth. In particular, low-texture or homogeneous regions lead to ambiguous matching patterns in the cost volume, resulting in unreliable depth refinement.

Second, as illustrated in Fig. 10, our method fails under certain challenging conditions. The camera motion in this scenario is not forward-facing, resulting in weak parallax between foreground and background objects. Consequently, the spatiotemporal structure captured by events is limited, and the geometric boundaries become ambiguous, $e.g.$, between the window and the transformer box beneath it, or between the tree and the house behind it. Additionally, the scene contains textureless, homogeneous regions such as road surfaces and building facades. These areas exhibit little brightness change and therefore generate few events, while also lacking discriminative texture in the image modality. As a result, the cost volume becomes ambiguous (as shown in Fig. 10 (e)), offering limited guidance for reliable matching. In such cases, the iterative optimization tends to rely more on contextual priors and the depth distribution rather than meaningful correlation evidence obtained from lookup operations.

In future work, we will focus on exploring more effective fusion strategies for images and events to leverage their complementary characteristics fully. To address the limitations of depth estimation under non-forward motion, we will develop multi-view depth estimation models with either multiple temporal continuous frames or multi-perspective inputs. Additionally, we will investigate incorporating additional information (such as semantic segmentation, active structured light [17], or LiDAR data) to

Table 8. Detailed network parameter settings. **PCDepth-Ours** only performs iterative refinement at 1/8 resolution. The values in '()' correspond to the parameters for stages 1/16, 1/8, and 1/4, respectively.

	Ours-B	Ours-E-B	Ours-E-S	Ours-E-T	PCDepth-Ours	Ours scaled-down
Backbone	Swin-T	Swin-T	Swin-T	ResNet	Swin-T	Swin-T
Backbone out dims	(128, 96, 64)	(128, 96, 64)	(96, 64, 32)	(64, 64, 32)	-	(128, 64, 48)
EIFusion out dims	(128, 96, 64)	-	-	-	-	(128, 64, 48)
DHV embedding dims	(128, 96, 64)	(128, 96, 64)	(96, 64, 32)	(64, 64, 32)	128	(128, 64, 48)
Iterations	(8, 8, 8)	(8, 8, 8)	(3, 3, 3)	(2, 2, 1)	8	(8, 8, 8)
D values of DHV	(32, 64, 96)	(32, 64, 96)	(32, 64, 96)	(16, 32, 64)	64	(32, 64, 96)
Lookup radius r	(8, 8, 8)	(8, 8, 8)	(8, 8, 8)	(4, 4, 4)	8	(8, 8, 8)
Multi-scale (l) in Lookup	(1, 2, 3)	(1, 2, 3)	(1, 1, 1)	(1, 1, 1)	1	(1, 2, 3)
Geometry encoder out dims	(128, 128, 64)	(128, 128, 64)	(96, 96, 48)	(32, 32, 16)	64	(96, 64, 32)
Hidden dims of GRU	(128, 96, 64)	(128, 96, 64)	(64, 32, 24)	(32, 32, 16)	128	(64, 32, 24)
Mid-channel of head	(192, 128, 96)	(192, 128, 96)	(96, 64, 64)	(32, 32, 16)	256	(128, 96, 64)

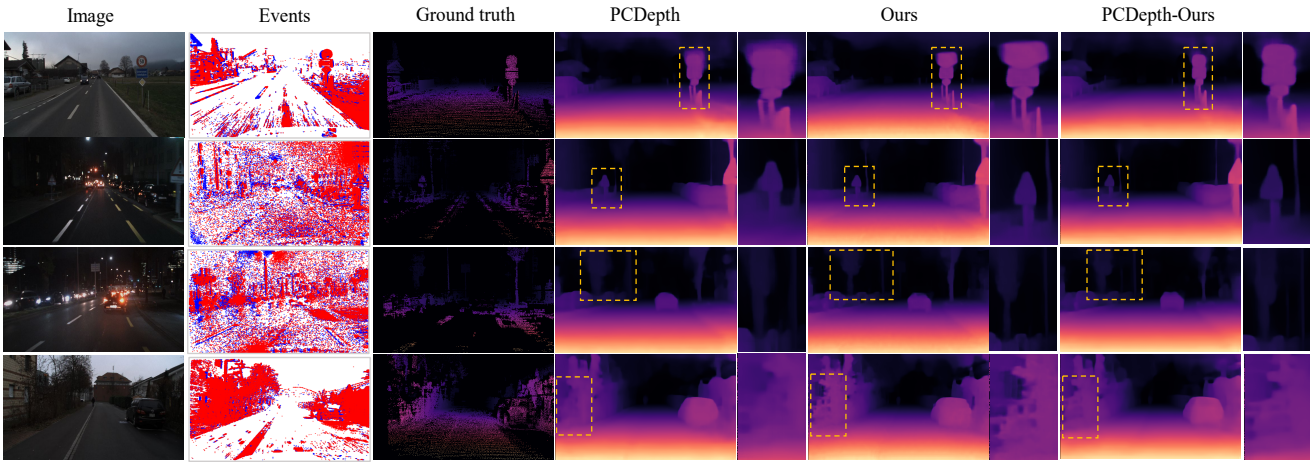

Figure 11. More qualitative monocular depth estimation results on DSEC [9]. The first and fourth rows correspond to daytime scenes, while the second and third rows represent night scenes.

enhance contextual features and alleviate the ambiguity in cost volume matching, or employ attention mechanisms to enhance matching accuracy. Moreover, cross-frame-rate frameworks [24] naturally align with the asynchronous, high-temporal-resolution of events. We plan to extend HypoDepth to asynchronous multimodal fusion, preserving the strength of spatial iterative refinement while effectively aggregating temporal event cues to achieve temporally consistent, low-latency, and high-frequency depth estimation.

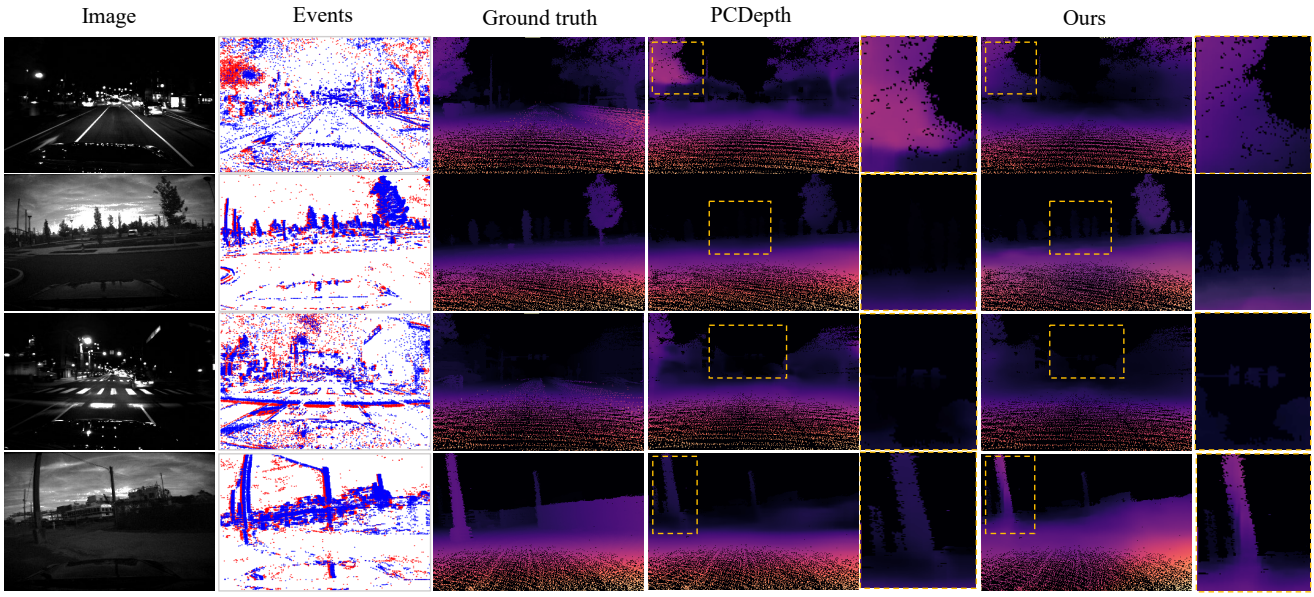

Figure 12. Qualitative monocular depth estimation results on MVSEC [40]. The first and third rows correspond to night scenes, while the second and fourth rows represent daytime scenes.

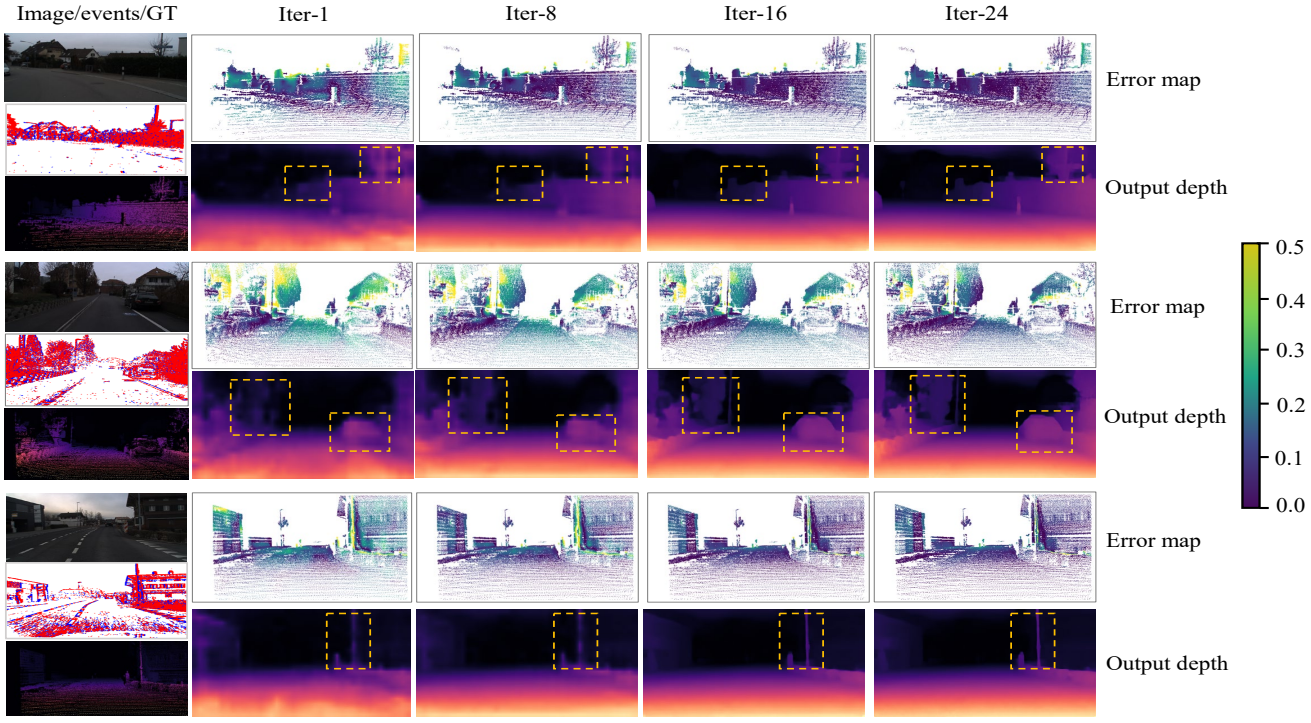

Figure 13. Qualitative results on the outdoor_day1 sequence on MVSEC [40].