BLMT-Stereo: Breaking the Local Minima Trap of Iterative Stereo Matching

Supplementary Material

A. Implementation Details

In this section, we provide detailed specifications for the BLMT-Stereo framework, which were briefly summarized in Section 3 of the main paper.

A.1. Backbone of Feature Extraction and Feature Pyramid

A.1.1 Feature Extraction

Our framework leverages the pre-trained DepthAnythingV2 [48] ViT-Large model as a monocular encoder. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the encoder produces a hierarchy of features $\{\Phi^{s\times}\}$, where $s \in \{4, 8, 14, 28\}$ denotes the downsampling factor relative to the input resolution.

To construct multi-scale context features for the recurrent update module, we employ a lightweight, trainable top-down path. High-level semantic features are progressively upsampled and fused with lower-level features via standard 1×1 convolutions and addition. This process yields three context tensors:

- $\text{inp}_{16} \in \mathbb{R}^{H/16\times W/16\times C_{dim}}$
- $\text{inp}_{8} \in \mathbb{R}^{H/8\times W/8\times C_{dim}}$
- $\text{inp}_{4} \in \mathbb{R}^{H/4\times W/4\times C_{dim}}$

where we typically set $C_{dim} = 128$. These context features provide the necessary context guidance for both the GRU update (Sec. 3.2) and the semantic-aware propagation (Sec. 3.3).

A.1.2 Matching Embeddings Projection

For constructing the cost volume, we utilize the $1/4$-scale backbone features $\Phi^{4\times}$. To enhance their discriminativeness for stereo matching, we project them using a specifically designed head, $\mathcal{P}_{\text{match}}$.

Based on the provided implementation, the head $\mathcal{P}_{\text{match}}$ is designed as a residual block consisting of a three-layer convolutional sequence:

1. A 3×3 convolution that expands the feature channels (typically from 96 to a hidden dimension of 128), followed by Group Normalization (GN) and a ReLU activation.
2. A second 3×3 convolution maintaining the hidden dimension, also followed by GN and ReLU.
3. A final 1×1 convolution that projects the features to the desired output dimension (typically 96).

A residual skip connection is added to this sequence. If the input and output channels differ, a 1×1 convolution is applied to the skip connection to align dimensions.

Crucially, the output features are pixel-wise ℓ_2 normalized to produce the final matching embeddings $\mathbf{M} \in \mathbb{R}^{H/4\times W/4\times D}$:

$$\mathbf{M}(u, v) = \frac{\mathbf{F}(u, v)}{\|\mathbf{F}(u, v)\|_2 + \epsilon}, \tag{14}$$

where $\mathbf{F} = \mathcal{P}_{\text{match}}(\Phi^{4\times})$ and $\epsilon = 10^{-6}$. This normalization ensures that the subsequent dot-product cost volume computation effectively calculates the **cosine similarity** between left and right features. This is vital for our dynamic window selection (Sec. 3.1) as it bounds the cost values within $[-1, 1]$, providing stable statistics across different lighting conditions and scenes.

A.2. Inference Strategy on High-Resolution Benchmarks

While our model is trained on cropped patches (e.g., 512×960), real-world benchmark like Middlebury [30] often contain images with significantly higher resolutions (up to 3000×2000). Directly feeding such high-resolution images into the ViT-based backbone introduces two critical issues: (1) GPU Memory Exhaustion, and more importantly, (2) Receptive Field Shift, where the changed aspect ratio and global attention scope deviate from the distribution learned during training, leading to suboptimal feature extraction.

To bridge this domain gap and ensure consistent performance, we employ a **Tile-Based Inference with Overlap-Fusion** strategy specifically for the Middlebury datasets.

A.2.1 Consistent-Receptive-Field Tiling

Instead of resizing the full image or simple non-overlapping cropping, we simulate the training crop distribution during inference. Given a high-resolution input pair $\mathbf{I}_L, \mathbf{I}_R \in \mathbb{R}^{H\times W\times 3}$, we decompose them into a set of N tiles using a sliding window of size $H_t \times W_t$.

Crucially, to prevent boundary artifacts and maintain context continuity, we enforce a spatial overlap (O_h, O_w) between adjacent tiles. The stride for the sliding window is defined as $S_h = H_t - O_h$ and $S_w = W_t - O_w$. The resulting tiles $\{\mathbf{T}_i\}_{i=1}^{N}$ form a batch input tensor:

$$\mathcal{B}_{tiles} = \text{Stack}(\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N\}) \in \mathbb{R}^{B_{tiles}\times3\times H_t\times W_t} \tag{15}$$

where H_t, W_t are chosen to be close to the training crop size (e.g., 512×960 or similar aspect ratios), ensuring that the backbone encoder Φ operates within its learned receptive field conditions.

A.2.2 Batch Processing and Weighted Fusion

The batch of tiles is processed in parallel by the shared feature extractor (Backbone + FPN + Projection Head) to yield feature tiles $\{\mathbf{F}_i\}_{i=1}^{N}$. To reconstruct the full-resolution feature maps $\mathbf{F}_{full} \in \mathbb{R}^{C\times H/4\times W/4}$, we employ an accumulation-averaging mechanism.

We initialize zero-filled tensors for feature accumulation $\mathbf{A}_{feat}$ and weight accumulation $\mathbf{A}_{weight}$. For each processed tile $\mathbf{F}_i$ corresponding to the spatial location (y_i, x_i) in the original image:

$$\mathbf{A}_{feat}[y_i : y_i + H_t, x_i : x_i + W_t] \leftarrow \mathbf{A}_{feat}[\dots] + \mathbf{F}_i \tag{16}$$

$$\mathbf{A}_{weight}[y_i : y_i + H_t, x_i : x_i + W_t] \leftarrow \mathbf{A}_{weight}[\dots] + \mathbf{1} \tag{17}$$

where $\mathbf{1}$ represents a tensor of ones with the same spatial dimensions as the feature tile.

Finally, the fused full-resolution feature map is obtained by normalizing the accumulated features by the visit counts:

$$\mathbf{F}_{full} = \frac{\mathbf{A}_{feat}}{\max(\mathbf{A}_{weight}, \epsilon)} \tag{18}$$

This averaging operation effectively suppresses noise in the overlapping regions and ensures seamless transitions between tiles. By strictly aligning the inference crop size with the training distribution, this strategy allows BLMT-Stereo to generalize robustly to the large-scale Middlebury and ETH3D benchmarks without requiring re-training on high-resolution data.

A.3. Memory-Efficient Implementation of Semantic-Aware Propagation

The Semantic-Aware Propagation (SAP) module aggregates information from a large receptive field using a dilated neighborhood ($K = 7$, dilation $d = 3$). A naive implementation utilizing standard convolution or full-map unfolding would result in significant memory consumption, particularly for high-resolution inputs, as it requires storing a tensor of shape $B \times (K^2 \cdot C) \times H \times W$.

To mitigate this memory bottleneck and enable training on limited GPU resources, we implement a memory-efficient chunking strategy. The inference and training process within the SAP module is decomposed into sequential blocks along the spatial dimension. The detailed implementation pipeline is described as follows:

A.3.1 Patch Unfolding and Flattening

Given the input disparity $D \in \mathbb{R}^{B\times 1\times H\times W}$ and confidence map $C \in \mathbb{R}^{B\times 1\times H\times W}$, we first extract local patches using a sliding window with dilation. Instead of processing the entire spatial map simultaneously, we flatten the spatial dimensions:

$$\mathcal{P}_{disp} = \text{Unfold}(D, K, d) \xrightarrow{\text{view}} \mathbb{R}^{B\times K^2\times N} \tag{19}$$

where $N = H \times W$ is the total number of pixels, $K = 7$ is the kernel size, and $d = 3$ is the dilation rate. The confidence patches $\mathcal{P}_{conf}$, similarity map S, and semantic mask M_{sem} are similarly flattened to $\mathbb{R}^{B\times K^2\times N}$ or $\mathbb{R}^{B\times 1\times N}$.

A.3.2 Chunk-based Processing

We split the flattened pixel index range $[0, N)$ into M chunks to lower the peak memory footprint. For each chunk i covering indices $[start_i, end_i)$, we load only the corresponding subset of tensors. The operations within a chunk are implemented as follows:

- **Candidate Selection & Weight Computation:** We execute the candidate filtering and weight normalization steps following the exact mathematical formulations detailed in **Sec. 3.3** of the main paper. By restricting these dense computations (e.g., pair-wise similarity and Softmax) to the current chunk, we avoid materializing the full $B \times K^2 \times H \times W$ tensor.
- **Residual Aggregation with Fallback:** The update is computed as the weighted sum of neighbor disparities using the learned weights. *Implementation Detail:* To ensure numerical stability, we introduce a specific fallback mechanism not detailed in the main text: if the sum of validity masks for a pixel is negligible (indicating no reliable neighbors), the aggregated value is explicitly reset to the center disparity D_p to prevent invalid updates.

A.3.3 Proximal Clipped Update

Finally, after aggregating the chunks and reshaping them back to the spatial resolution $B \times 1 \times H \times W$, we apply a hard constraint to the residual. This step ensures training stability by preventing the refined disparity from drifting too far from the initial estimate in a single iteration. This implementation strategy effectively reduces the memory complexity of the SAP module from $\mathcal{O}(HWK^2)$ to $\mathcal{O}(\frac{HW}{M}K^2)$, allowing us to efficiently train with large batch sizes or high-resolution inputs (e.g., Middlebury) on standard GPUs.

B. Additional experiments

B.1. Qualitative comparison across real-world weather conditions

Figure 9 confirms the robustness of our proposed model in diverse conditions: confidence-guided updates stabilize refinement when photometric cues are unreliable,

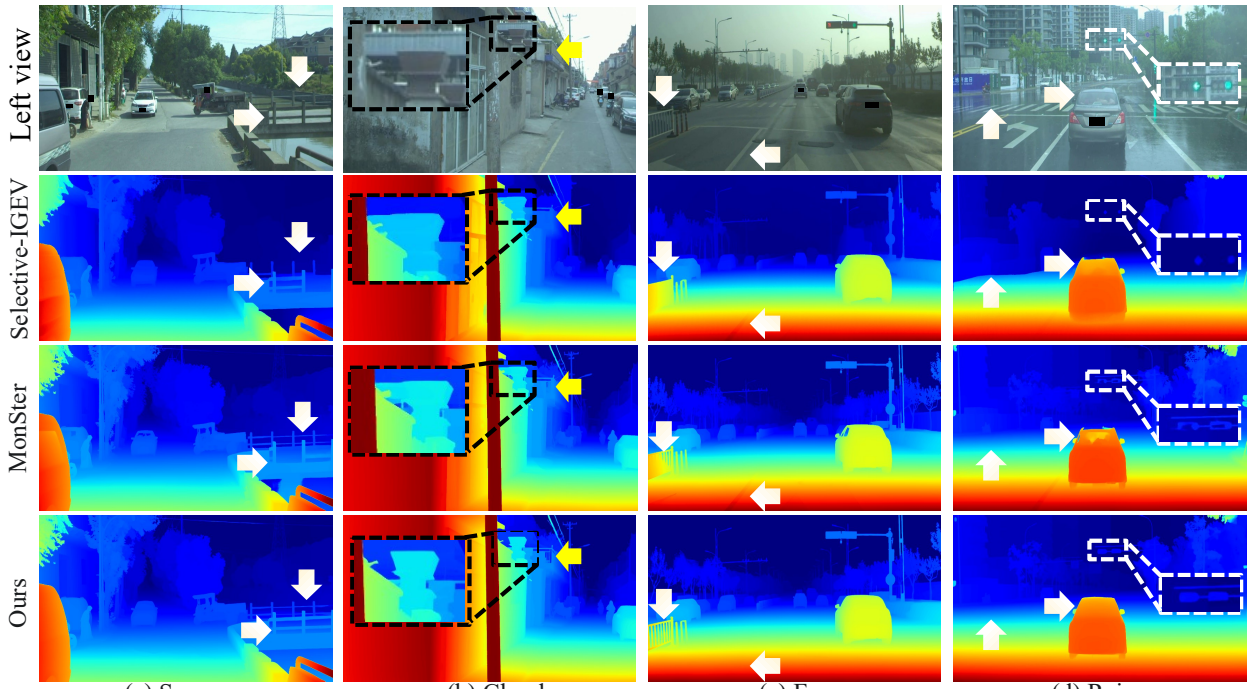


Figure 9. **Qualitative comparison across real-world weather conditions on DrivingStereo [47].** Columns: (a) *Sunny*, (b) *Cloudy*, (c) *Foggy*, (d) *Rainy*. Rows show the left image, predictions from Selective-IGEV and MonSter, and *Ours* (BLMT-Stereo). Across all scenarios, *Ours* produces more accurate and geometrically coherent disparities, with fewer breaks on thin structures and more uniform estimates on weakly textured surfaces. **Specifically**: in (c), our method preserves continuity on *thin railings* and suppresses *spurious ground artifacts*; in (d), it better handles *car-body specularities* and *wet-road reflections* while retaining *rain-blurred traffic-light details*.

while semantic-aware propagation preserves boundaries and propagates reliable hypotheses into low-confidence regions, producing coherent disparity maps across adverse weather.

B.2. Ablation Study Visualization

B.2.1 Confidence-Guided Refinement

Figure 10 visualizes the confidence-conditioned modulation features alongside evolving disparities. The encoded modulation features progressively concentrate on reliable structures and de-emphasize uncertain zones as iterations proceed (t=1, 2, 8, 32), suppressing the spread of erroneous disparities and stabilizing updates—consistent with the gains observed after adding CMG and $\mathcal{L}_{\text{conf}}$ in Table 5.

B.2.2 Hierarchical Initialization Effect

Figure 11 illustrates the effect of hierarchical initialization (HI) on disparity probability distributions. The refined stage suppresses spurious modes and sharpens the peak around the correct disparity, creating a steeper selection landscape for subsequent refinement.

B.3. 3D Geometric Consistency Analysis

To intuitively demonstrate the superior geometric accuracy of BLMT-Stereo, we visualize the reconstructed 3D point clouds in Figure 12. By back-projecting the estimated disparity maps into 3D space, we observe distinct advantages in our method compared to state-of-the-art baselines (RAFT-Stereo, Selective-IGEV, and MonSter).

First, in **weak-textured regions** such as the background wall, BLMT-Stereo successfully recovers a smooth **planar structure**, effectively suppressing the surface noise and depth distortions observed in competing methods. Second, in complex **occluded areas** like the staircase railings, our model preserves sharp geometric boundaries without the structural breakage or over-smoothing present in other approaches. Finally, BLMT-Stereo significantly reduces **flying pixels** (floating artifacts) at depth discontinuities, resulting in a much cleaner 3D representation. This qualitative superiority is corroborated by the quantitative error maps, where our method achieves a substantially lower Bad-2.0 error rate of 5.40%, compared to over 15% for the baselines. For a comprehensive dynamic view of the 3D geometry, please refer to the animated GIF provided in the supplementary material.

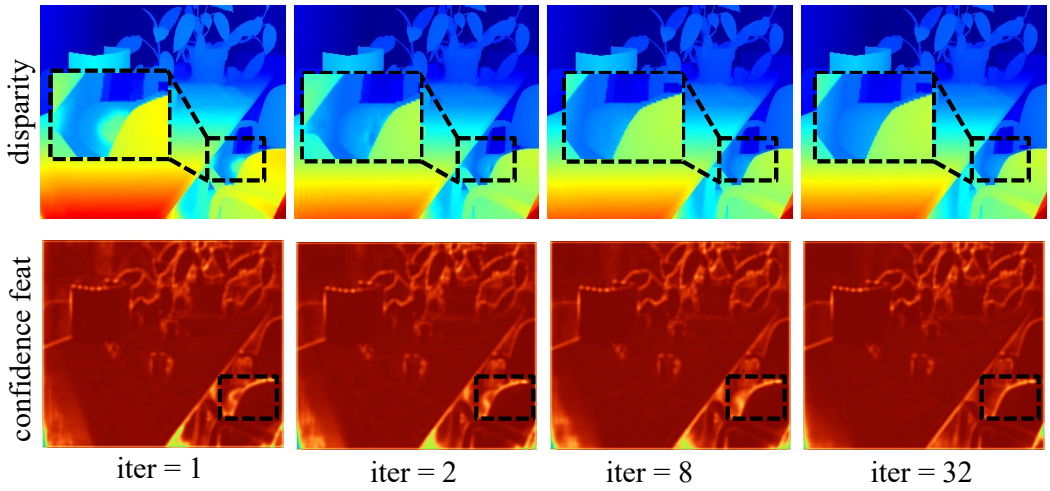


Figure 10. **Confidence-conditioned modulation over iterations.** *Top:* disparityat iterations $t \in \{1, 2, 8, 32\}$. *Bottom:* confidence-modulation features obtained by encoding the confidence map, used to modulate disparity features in GRU updates. The modulation adapts with the evolving estimate, concentrates on high-confidence structures, and down-weights uncertain regions, suppressing the diffusion of erroneous disparities (dashed boxes).

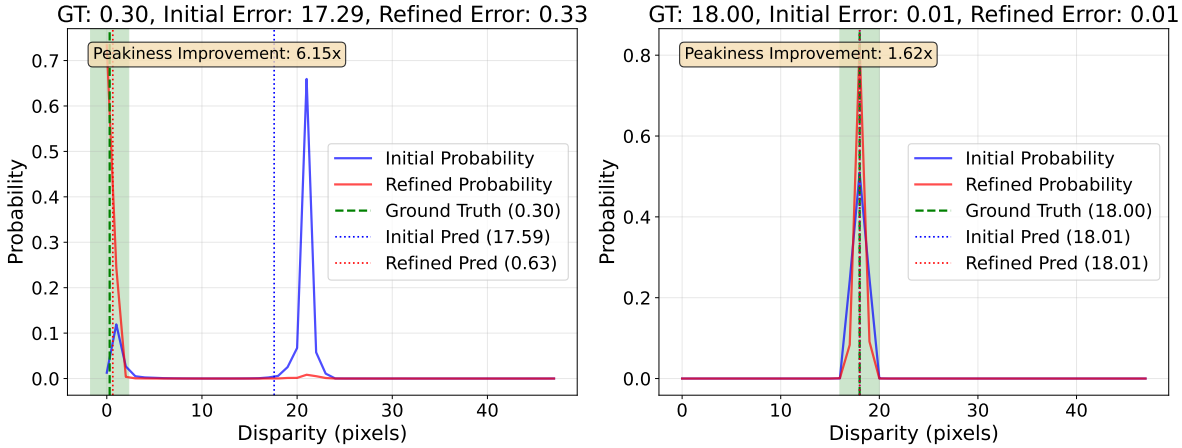


Figure 11. **Effect of hierarchical initialization on disparity probability distributions.** Visualized on SceneFlow, comparing initial (blue) and refined (red) probability distributions for two representative pixels against ground truth (green). *Left:* High-ambiguity case where a multi-modal initial distribution causes significant error. Refinement resolves this by suppressing spurious modes and forming a sharp, uni-modal peak at the correct disparity. *Right:* For an already accurate estimate, refinement further sharpens the distribution, increasing probability concentration and confidence.

B.4. Performance in complex scenes

To further validate the robustness of our proposed BLMT-Stereo, we present additional qualitative results on the challenging Flickr1024 [37] dataset. This dataset is renowned for its diverse, in-the-wild images, which contain many of the pathological cases discussed in the main paper, such as complex geometry, non-ideal illumination, and textureless regions. In this section, we compare our BLMT-Stereo against Selective-IGEV [36] and MonSter [8] across a gauntlet of these challenging conditions.

Figure 13 presents an evaluation of performance on diverse architectural and urban scenes from Flickr1024 [37]. These environments introduce a mixture of challenging elements, such as fine-grained railings (top row), thin structures against textureless backgrounds (middle row), and complex urban clutter (bottom row). As detailed in the caption, BLMT-Stereo is the only method that consistently preserves these fine structures while maintaining a clean, artifact-free result, demonstrating a superior balance of precision and robustness.

Non-ideal illumination, a notorious failure case for stereo matching, is examined in Figure 14 and Figure 15. In the low-light scenario (Figure 14), the proposed method produces a clean, geometrically coherent result, whereas

competing methods suffer from significant matching artifacts on poorly-lit surfaces. A similar trend is observed in Figure 15, which features a combination of dense spotlights and highly reflective surfaces. Here, baseline methods are either confounded by artifacts or fail to resolve the true scene geometry. In sharp contrast, BLMT-Stereo remains robust to these effects and accurately captures the complex, recessed geometry of the ceiling.

Performance analysis is further extended to two classic pathological cases: large textureless regions and extremely fine-scale natural geometry. In Figure 16, the model yields a perfectly smooth and consistent estimation for the untextured sky while simultaneously preserving the entire fine-scale antenna mast—a structure completely lost by competitors. Figure 17 highlights this capability on intricate organic structures. BLMT-Stereo demonstrates the unique ability to resolve the semi-transparent, filamentous structure of the dandelion (top row), a scene where MonSter fails catastrophically. It also excels on the dense network of tree branches (bottom row), cleanly separating the fine foreground geometry from the textureless background, a task where competitors compromise on one or the other.

B.5. Performance on Public Leaderboards

To validate our state-of-the-art performance, we evaluate BLMT-Stereo on the competitive ETH3D [31] and Middlebury v3 [30] public benchmarks, achieving the #1 position on a significant number of evaluation metrics. At the time of submission (November 2025), our method achieves the #1 rank on both leaderboards, underscoring its superior accuracy and robustness on challenging real-world data. Official snapshots from both leaderboards are presented in Figure 18 and Figure 19.

C. Structural Properties and Theoretical Analysis

In this section, we analyze the structural properties of the proposed Adaptive Range Disparity Initialization (ARDI). Rather than seeking rigorous guarantees for the complex non-linear dynamics, we employ stylized theoretical frameworks to provide intuitive insights into the mechanisms by which these modules promote robustness and variance reduction.

C.1. ARDI: Variance Reduction Trade-off

We analyze the effect of cost volume truncation on estimator variance using a mixture model.

Model (Assumption 2). Disparity distribution $P(x) = (1-\alpha)P_S(x) + \alpha P_N(x)$ on $[-R, R]$.

- Signal P_S: Symmetric, zero-mean, variance σ_S^2, support $[-\Delta, \Delta]$.
- Noise P_N: Uniform on $[-R, R]$, variance $\sigma_N^2 = R^2/3$.
- Truncation: Interval $\mathcal{S} = [-\Delta, \Delta]$, with $\rho = \Delta/R < 1$.

Proposition 2 (Variance Reduction Threshold). *The truncated distribution Q satisfies $Var_Q < Var_P$ if and only if:*

$$\sigma_N^2 > \sigma_S^2 \cdot \Phi(\alpha, \rho), \tag{20}$$

where the threshold function Φ is defined as:

$$\Phi(\alpha, \rho) = \frac{(1-\alpha)(1-\rho)}{(1-\alpha) + \alpha\rho - \rho^3}. \tag{21}$$

Proof. The normalization constant is $Z = (1-\alpha) + \alpha\rho$. The variance of the truncated distribution is $\text{Var}_Q = \frac{1}{Z}[(1-\alpha)\sigma_S^2 + \alpha\rho^3\sigma_N^2]$. The condition for strict variance reduction, $\text{Var}_Q < \text{Var}_P$, is equivalent to $\text{Var}_P - \text{Var}_Q > 0$. Expanding terms:

$$\alpha\sigma_N^2\left(1 - \frac{\rho^3}{Z}\right) > (1-\alpha)\sigma_S^2\left(\frac{1}{Z} - 1\right). \tag{22}$$

Multiplying both sides by Z (noting $Z > 0$):

$$\alpha\sigma_N^2(Z - \rho^3) > (1-\alpha)\sigma_S^2(1 - Z). \tag{23}$$

Substitute $Z = (1-\alpha) + \alpha\rho$ into the terms:

- RHS term: $1 - Z = 1 - [(1-\alpha) + \alpha\rho] = \alpha(1-\rho)$.
- LHS term: $Z - \rho^3 = (1-\alpha) + \alpha\rho - \rho^3$.

The inequality becomes:

$$\alpha\sigma_N^2\left[(1-\alpha) + \alpha\rho - \rho^3\right] > (1-\alpha)\sigma_S^2\left[\alpha(1-\rho)\right]. \tag{24}$$

Dividing both sides by α (since $\alpha > 0$) and isolating σ_N^2:

$$\sigma_N^2 > \sigma_S^2 \cdot \frac{(1-\alpha)(1-\rho)}{(1-\alpha) + \alpha\rho - \rho^3}. \tag{25}$$

This matches the definition of $\Phi(\alpha, \rho)$, completing the proof.

Trade-off Analysis via Monotonicity. To understand the implications, we analyze the derivative of Φ with respect to ρ:

$$\frac{\partial\Phi}{\partial\rho} = \frac{-(1-\alpha)\left[(1-\alpha) + \alpha(3\rho^2 - 2\rho^3)\right]}{\left((1-\alpha) + \alpha\rho - \rho^3\right)^2}. \tag{26}$$

Since $\rho \in (0, 1)$, the term in the brackets is positive, making the derivative strictly negative ($\frac{\partial\Phi}{\partial\rho} < 0$). This reveals a **Stability-Precision Trade-off**:

1. **Aggressive Truncation ($\rho \to 0$):** Φ is large. The condition $\sigma_N^2 > \sigma_S^2 \cdot \Phi$ is harder to satisfy, requiring dominant ambiguity to justify the truncation.
2. **Conservative Truncation ($\rho \to 1$):** Φ decreases, making the condition easier to satisfy, but the variance reduction gain diminishes.

This suggests ARDI's dynamic window selection implicitly optimizes this trade-off by adapting ρ based on the local signal-to-noise ratio estimates.

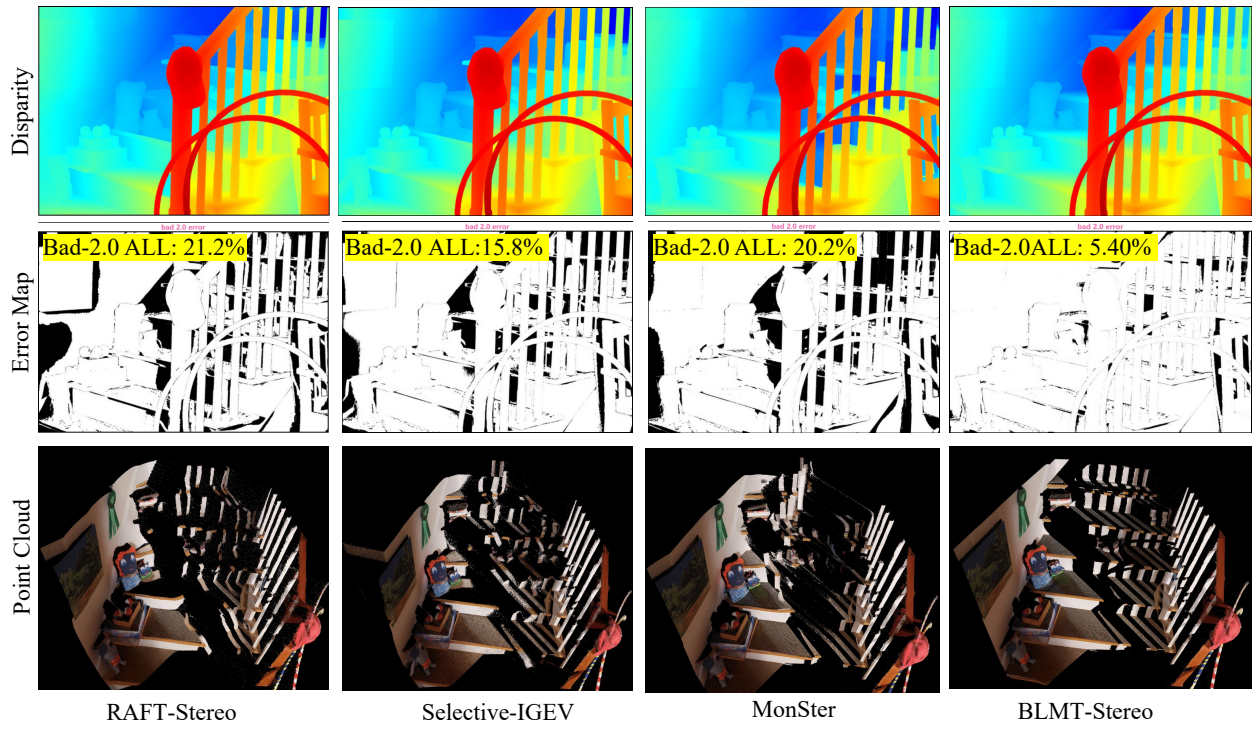


Figure 12. **Qualitative comparison of 3D point cloud reconstruction.** We back-project the estimated disparity maps into 3D point clouds to visualize geometric consistency. Compared with RAFT-Stereo, Selective-IGEV, and MonSter, our BLMT-Stereo yields significantly more accurate 3D structures. Specifically, it recovers smoother planar surfaces in textureless regions (e.g., the background wall) and preserves sharp geometric boundaries in complex occluded areas (e.g., the staircase railings). Notably, our method effectively suppresses flying pixels and outliers, resulting in a much cleaner point cloud representation with the lowest Bad-2.0 error rate (5.40%).

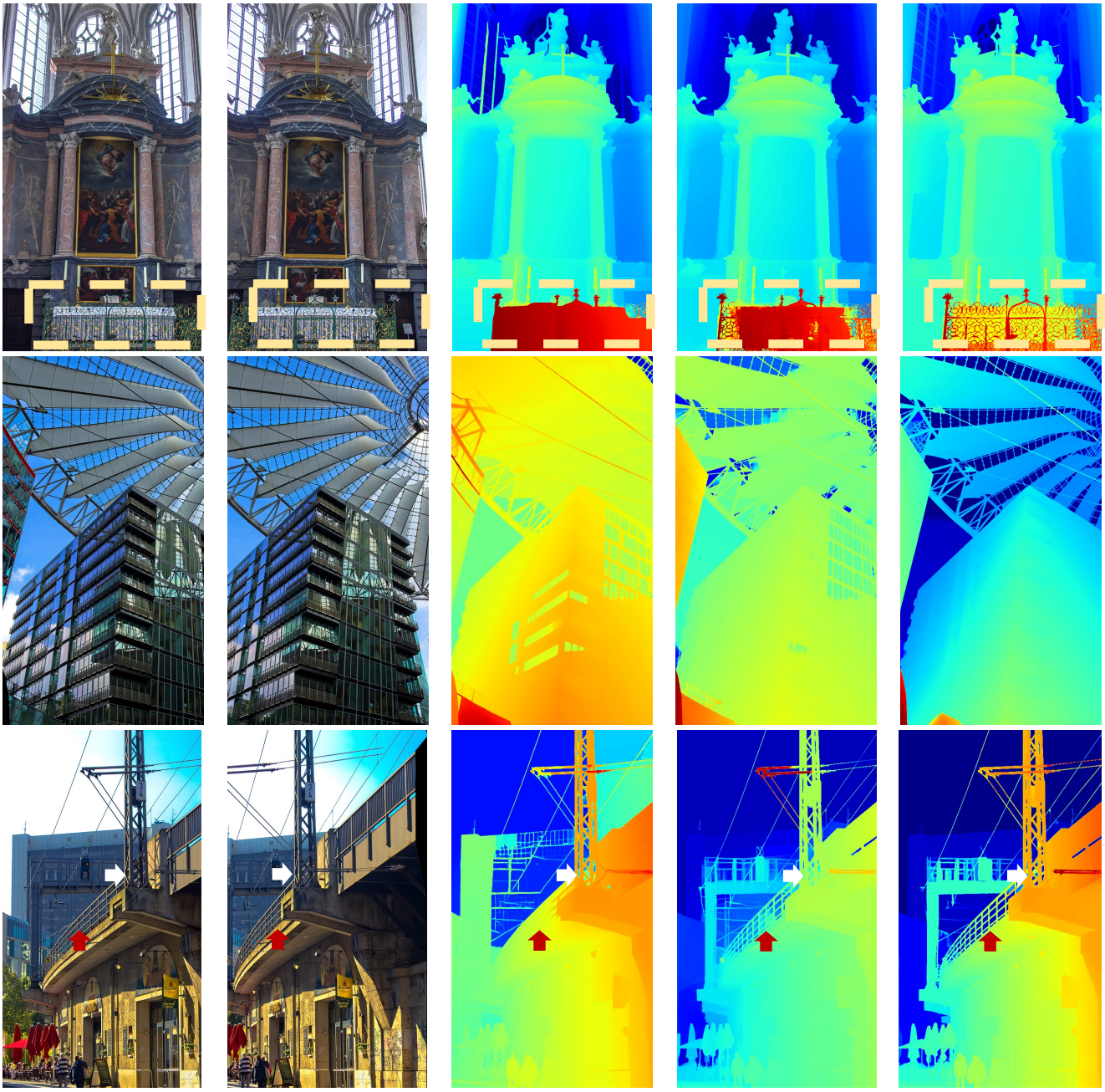
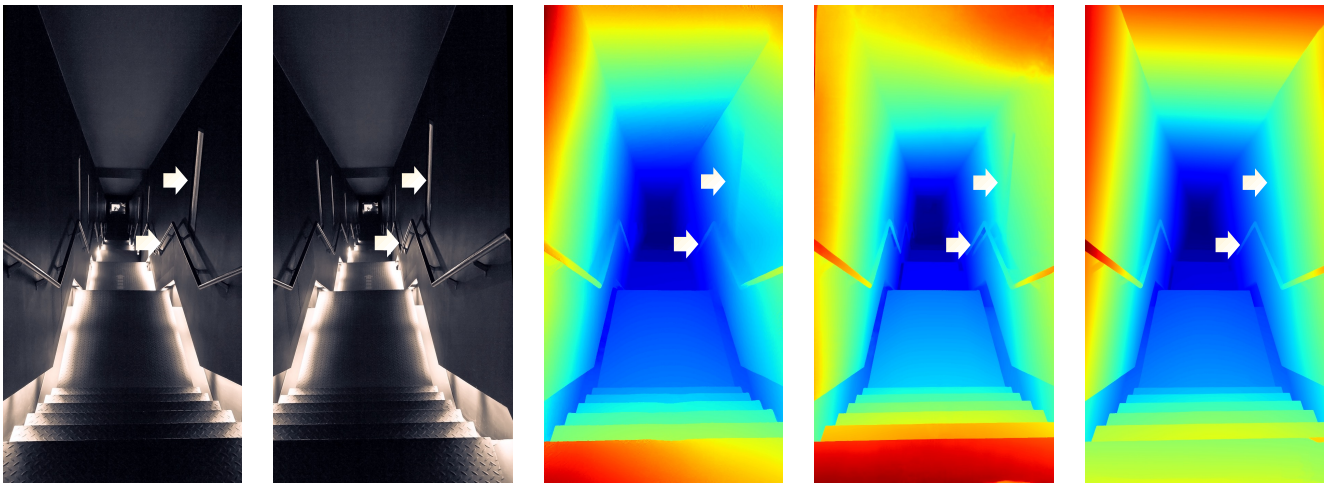
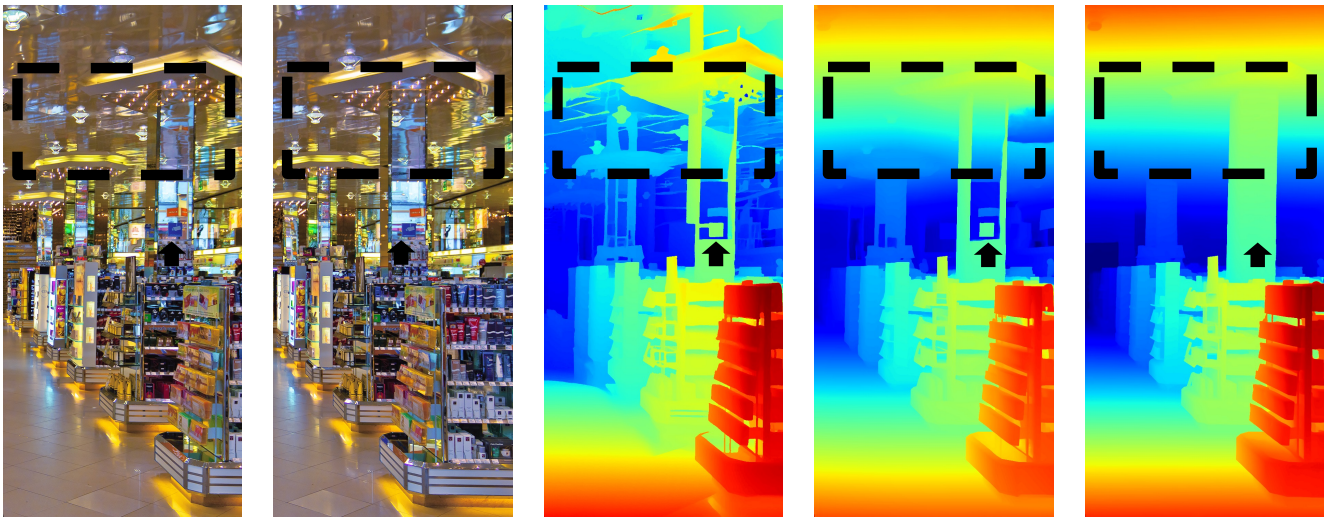


Figure 13. Qualitative evaluation of our BLMT-Stereo on complex architectural and urban scenes. We compare our method (rightmost) against Selective-IGEV and MonSter in three distinct challenging scenarios (one per row). (Top row) In a cluttered indoor scene, our method successfully distinguishes the intricate, fine-grained railings from the complex background, producing a robust and complete disparity map where others fail. (Middle row) For scenes with complex, thin structures against large, textureless areas (e.g., sky), our BLMT-Stereo demonstrates superior geometric precision, accurately capturing the fine roof geometry. Note its accurate estimation on the building facade, which features challenging illumination. (Bottom row) In an outdoor urban scene, our method clearly preserves extremely thin structures (e.g., poles and wires, indicated by arrows) and delivers a much cleaner, artifact-free result for the support structure. These results highlight the superior robustness and precision of our BLMT-Stereo in handling thin structures, textureless regions, and complex foreground-background separation.



Left view | Right view | Selective-IGEV | MonSter | BLMT-Stereo

Figure 14. Robustness evaluation under challenging low-light conditions. This scene presents a difficult matching scenario due to large dark regions. Both Selective-IGEV and MonSter suffer from significant matching artifacts and noise, particularly on the poorly-lit wall surfaces (indicated by arrows). In contrast, our BLMT-Stereo remains robust, producing a clean, artifact-free, and geometrically coherent disparity map that accurately captures the planar structures.



Left view | Right view | Selective-IGEV | MonSter | BLMT-Stereo

Figure 15. Performance evaluation in a challenging indoor scene with complex illumination and reflections. The scene features a difficult combination of dense spotlights and highly reflective surfaces (e.g., ceiling and floor). Selective-IGEV (middle-left) produces significant artifacts and fails to capture the scene structure. MonSter (middle-right) struggles with the ceiling geometry, incorrectly smoothing over the recessed area (indicated by the arrow). In sharp contrast, our BLMT-Stereo (rightmost) demonstrates superior robustness to these challenging conditions, producing a clean, artifact-free result. Notably, it is the only method that correctly resolves the complex, recessed geometry of the ceiling (see dashed box and arrow), proving its higher geometric accuracy..

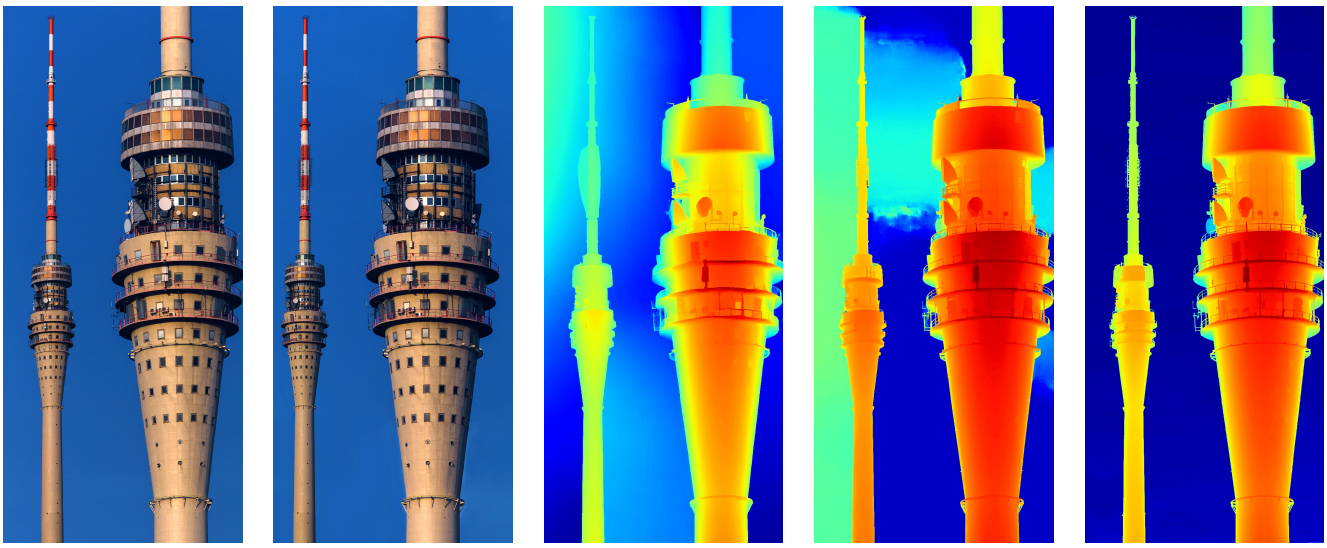


Figure 16. Qualitative evaluation on a scene with large textureless regions. The input views feature a tower against a large, untextured sky, a classic challenge for stereo matching. Competing methods like Selective-IGEV and MonSter produce severe streaking artifacts and noise in the sky region. Furthermore, they fail to preserve the fine-scale geometry of the antenna mast. Our BLMT-Stereo (rightmost) demonstrates superior robustness, yielding a clean, consistent, and artifact-free disparity map for the textureless sky, while simultaneously and accurately capturing the entire thin structure of the antenna.

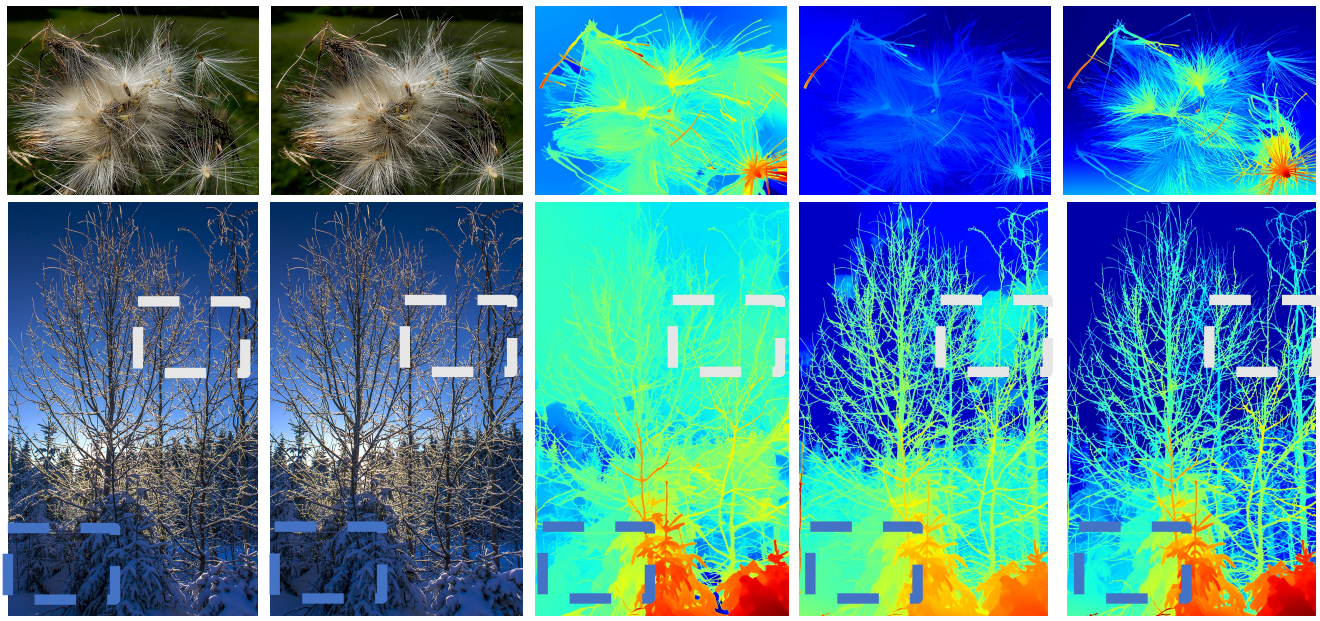
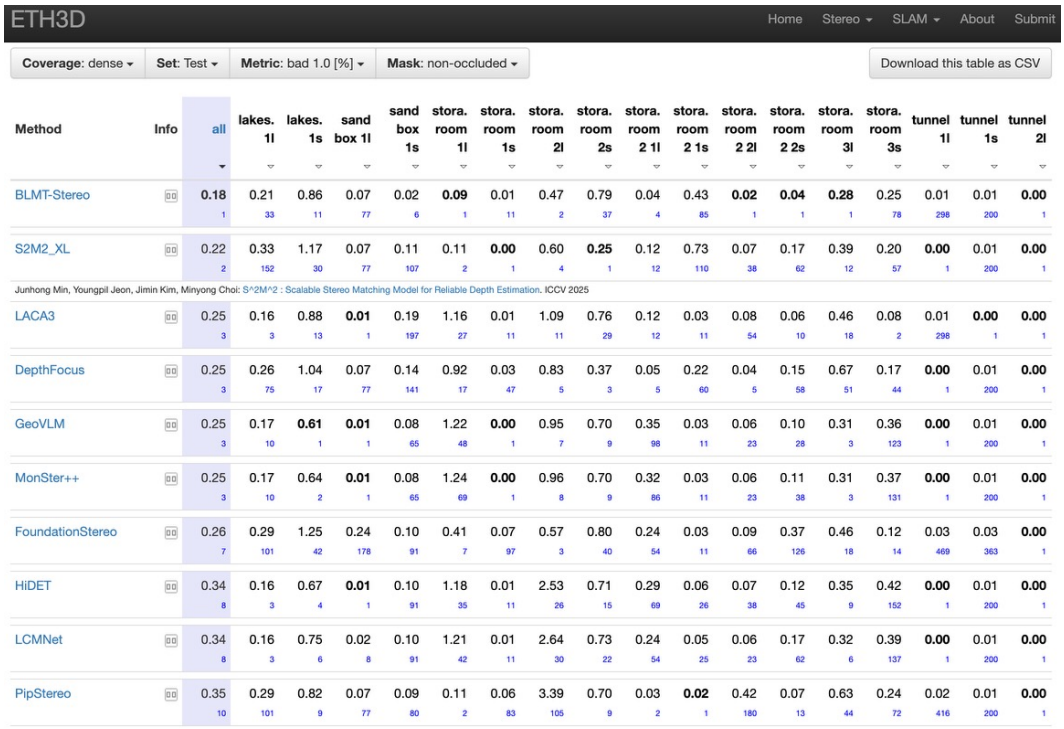


Figure 17. Qualitative comparison on challenging natural scenes with extremely fine-grained geometry. (Top row) We evaluate on a highly complex scene of dandelion seeds, which consist of intricate, semi-transparent filaments. Selective-IGEV produces a noisy, unusable result, while MonSter catastrophically fails, missing the entire foreground object. Our BLMT-Stereo is the only method to successfully resolve the fine-scale filamentous structure and capture its 3D form. (Bottom row) In a scene with a dense network of thin tree branches against a textureless sky, competing methods either produce significant artifacts (Selective-IGEV) or erase the fine geometry (MonSter, see dashed boxes). Our method again demonstrates superior precision, robustly preserving the intricate branch structures while maintaining a clean, consistent estimation for the textureless background.



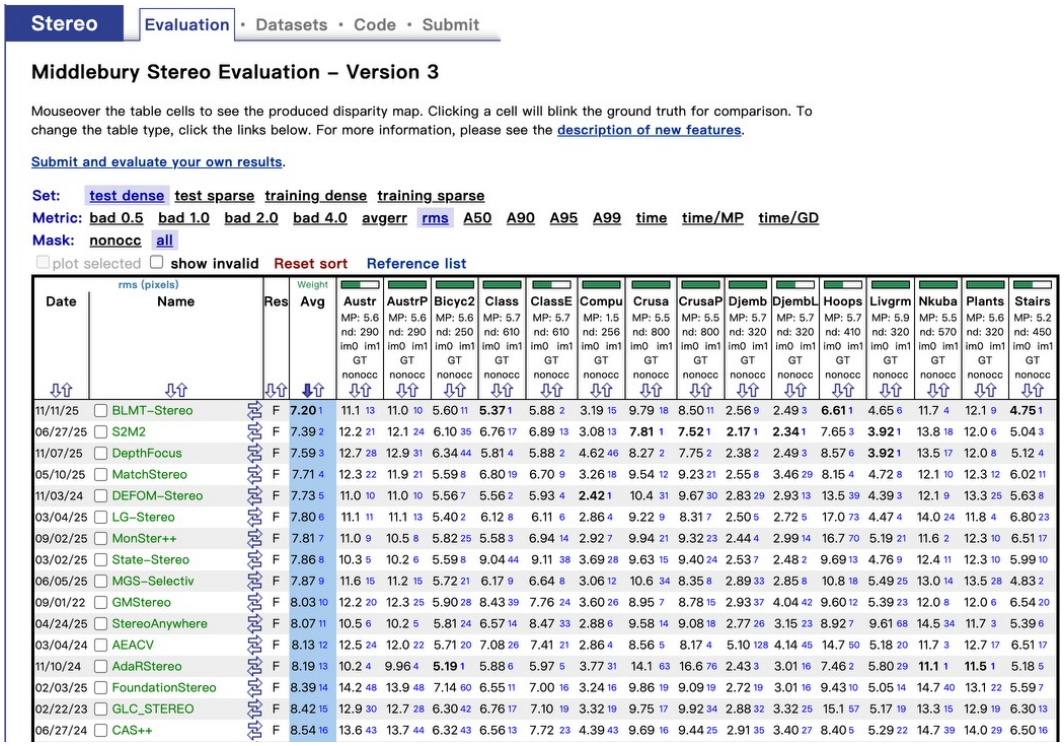
ETH3D

Home | Stereo | SLAM | About | Submit

Coverage: dense | Set: Test | Metric: bad 1.0 [%] | Mask: non-occluded | Download this table as CSV

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunnel 2l
BLMT-Stereo		**0.18** 1	0.21 33	0.86 11	0.07 77	0.02 6	**0.09** 1	0.01 11	0.47 2	0.79 37	0.04 4	0.43 85	**0.02** 1	**0.04** 1	**0.28** 1	0.25 78	0.01 298	0.01 200	**0.00** 1
S2M2_XL		0.22 2	0.33 152	1.17 30	0.07 77	0.11 107	0.11 2	**0.00** 1	0.60 4	**0.25** 1	0.12 12	0.73 110	0.07 38	0.17 62	0.39 12	0.20 57	**0.00** 1	0.01 200	**0.00** 1
Junhong Min, Youngpil Jeon, Jimin Kim, Minyong Choi: S^2M^2 : Scalable Stereo Matching Model for Reliable Depth Estimation. ICCV 2025																			
LACA3		0.25 3	0.16 3	0.88 13	**0.01** 1	0.19 197	1.16 27	0.01 11	1.09 11	0.76 29	0.12 12	0.03 11	0.08 54	0.06 10	0.46 18	0.08 2	0.01 298	**0.00** 1	**0.00** 1
DepthFocus		0.25 3	0.26 75	1.04 17	0.07 77	0.14 141	0.92 17	0.03 47	0.83 5	0.37 3	0.05 5	0.22 60	0.04 5	0.15 58	0.67 51	0.17 44	**0.00** 1	0.01 200	**0.00** 1
GeoVLM		0.25 3	0.17 10	**0.61** 1	**0.01** 1	0.08 65	1.22 48	**0.00** 1	0.95 7	0.70 9	0.35 98	0.03 11	0.06 23	0.10 28	0.31 3	0.36 123	**0.00** 1	0.01 200	**0.00** 1
MonSter++		0.25 3	0.17 10	0.64 2	**0.01** 1	0.08 65	1.24 69	**0.00** 1	0.96 8	0.70 9	0.32 86	0.03 11	0.06 23	0.11 38	0.31 3	0.37 131	**0.00** 1	0.01 200	**0.00** 1
FoundationStereo		0.26 7	0.29 101	1.25 42	0.24 178	0.10 91	0.41 7	0.07 97	0.57 3	0.80 40	0.24 54	0.03 11	0.09 66	0.37 126	0.46 18	0.12 14	0.03 469	0.03 363	**0.00** 1
HiDET		0.34 8	0.16 3	0.67 4	**0.01** 1	0.10 91	1.18 35	0.01 11	2.53 26	0.71 15	0.29 69	0.06 26	0.07 38	0.12 45	0.35 9	0.42 152	**0.00** 1	0.01 200	**0.00** 1
LCMNet		0.34 8	0.16 3	0.75 6	0.02 8	0.10 91	1.21 42	0.01 11	2.64 30	0.73 22	0.24 54	0.05 25	0.06 23	0.17 62	0.32 6	0.39 137	**0.00** 1	0.01 200	**0.00** 1
PipStereo		0.35 10	0.29 101	0.82 9	0.07 77	0.09 80	0.11 2	0.06 83	3.39 105	0.70 9	0.03 2	**0.02** 1	0.42 180	0.07 13	0.63 44	0.24 72	0.02 416	0.01 200	**0.00** 1

Figure 18. The official ETH3D low-resolution stereo benchmark leaderboard (accessed November 2025). Our model, BLMT-Stereo, is ranked first, surpassing all previous state-of-the-art methods.



Stereo | Evaluation · Datasets · Code · Submit

Middlebury Stereo Evaluation – Version 3

Mouseover the table cells to see the produced disparity map. Clicking a cell will blink the ground truth for comparison. To change the table type, click the links below. For more information, please see the description of new features.

Submit and evaluate your own results.

Set: test dense | test sparse | training dense | training sparse

Metric: bad 0.5 | bad 1.0 | bad 2.0 | bad 4.0 | avgerr | rms | A50 | A90 | A95 | A99 | time | time/MP | time/GD

Mask: nonocc | all

plot selected | show invalid | Reset sort | Reference list

Date	Name	Res	Weight Avg	Austr MP: 5.6 nd: 290 im0 im1 GT nonocc	AustrP MP: 5.6 nd: 290 im0 im1 GT nonocc	Bicyc2 MP: 5.6 nd: 250 im0 im1 GT nonocc	Class MP: 5.7 nd: 610 im0 im1 GT nonocc	ClassE MP: 5.7 nd: 610 im0 im1 GT nonocc	Compu MP: 1.5 nd: 256 im0 im1 GT nonocc	Crusa MP: 5.5 nd: 800 im0 im1 GT nonocc	CrusaP MP: 5.5 nd: 800 im0 im1 GT nonocc	Djemb MP: 5.7 nd: 320 im0 im1 GT nonocc	DjembL MP: 5.7 nd: 320 im0 im1 GT nonocc	Hoops MP: 5.7 nd: 410 im0 im1 GT nonocc	Livgrm MP: 5.9 nd: 320 im0 im1 GT nonocc	Nkuba MP: 5.5 nd: 570 im0 im1 GT nonocc	Plants MP: 5.6 nd: 320 im0 im1 GT nonocc	Stairs MP: 5.2 nd: 450 im0 im1 GT nonocc
11/11/25	BLMT-Stereo	F	**7.20** 1	11.1 13	11.0 10	5.60 11	**5.37** 1	5.88 2	3.19 15	9.79 18	8.50 11	2.56 9	2.49 3	**6.61** 1	4.65 6	11.7 4	12.1 9	**4.75** 1
06/27/25	S2M2	F	7.39 2	12.2 21	12.1 24	6.10 35	6.76 17	6.89 13	3.08 13	**7.81** 1	**7.52** 1	**2.17** 1	**2.34** 1	7.65 3	**3.92** 1	13.8 18	12.0 6	5.04 3
11/07/25	DepthFocus	F	7.59 3	12.7 28	12.9 31	6.34 44	5.81 4	5.88 2	4.62 46	8.27 2	7.75 2	2.38 2	2.49 3	8.57 6	**3.92** 1	13.5 17	12.0 8	5.12 4
05/10/25	MatchStereo	F	7.71 4	12.3 22	11.9 21	5.59 8	6.80 19	6.70 9	3.26 18	9.54 12	9.23 21	2.55 8	3.46 29	8.15 4	4.72 8	12.1 10	12.3 12	6.02 11
11/03/24	DEFOM-Stereo	F	7.73 5	11.0 10	11.0 10	5.56 7	5.56 2	5.93 4	**2.42** 1	10.4 31	9.67 30	2.83 29	2.93 13	13.5 39	4.39 3	12.1 9	13.3 25	5.63 8
03/04/25	LG-Stereo	F	7.80 6	11.1 11	11.1 13	5.40 2	6.12 8	6.11 6	2.86 4	9.22 9	8.31 7	2.50 5	2.72 5	17.0 73	4.47 4	14.0 24	11.8 4	6.80 23
09/02/25	MonSter++	F	7.81 7	11.0 9	10.5 8	5.82 25	5.58 3	6.94 14	2.92 7	9.94 21	9.32 23	2.44 4	2.99 14	16.7 70	5.19 21	11.6 2	12.3 10	6.51 17
03/02/25	State-Stereo	F	7.86 8	10.3 5	10.2 6	5.59 8	9.04 44	9.11 38	3.69 28	9.63 15	9.40 24	2.53 7	2.48 2	9.69 13	4.76 9	12.4 11	12.3 10	5.99 10
06/05/25	MGS-Selectiv	F	7.87 9	11.6 15	11.2 15	5.72 21	6.17 9	6.64 8	3.06 12	10.6 34	8.35 8	2.89 33	2.85 8	10.8 18	5.49 25	13.0 14	13.5 28	4.83 2
09/01/22	GMStereo	F	8.03 10	12.2 20	12.3 25	5.90 28	8.43 39	7.76 24	3.60 26	8.95 7	8.78 15	2.93 37	4.04 42	9.60 12	5.39 23	12.0 8	12.0 6	6.54 20
04/24/25	StereoAnywhere	F	8.07 11	10.5 6	10.2 5	5.81 24	6.57 14	8.47 33	2.88 6	9.58 14	9.08 18	2.77 26	3.15 23	8.92 7	9.61 68	14.5 34	11.7 3	5.39 6
03/04/24	AEACV	F	8.13 12	12.5 24	12.0 22	5.71 20	7.08 26	7.41 21	2.86 4	8.56 5	8.17 4	5.10 128	4.14 45	14.7 50	5.18 20	11.7 3	12.7 17	6.51 17
11/10/24	AdaRStereo	F	8.19 13	10.2 4	9.96 4	**5.19** 1	5.88 6	5.97 5	3.77 31	14.1 63	16.6 76	2.43 3	3.01 16	7.46 2	5.80 29	**11.1** 1	**11.5** 1	5.18 5
02/03/25	FoundationStereo	F	8.39 14	14.2 48	13.9 48	7.14 60	6.55 11	7.00 16	3.24 16	9.86 19	9.09 19	2.72 19	3.01 16	9.43 10	5.05 14	14.7 40	13.1 22	5.59 7
02/22/23	GLC_STEREO	F	8.42 15	12.9 30	12.7 28	6.30 42	6.76 17	7.10 19	3.32 19	9.75 17	9.92 34	2.88 32	3.32 25	15.1 57	5.17 19	13.3 15	12.9 19	6.30 13
06/27/24	CAS++	F	8.54 16	13.6 43	13.7 44	6.32 43	6.56 13	7.72 23	4.39 43	9.69 16	9.44 25	2.91 35	3.40 27	8.40 5	5.29 22	14.7 39	14.0 29	6.50 16

Figure 19. The official Middlebury v3 benchmark leaderboard (accessed November 2025). Our method secures the top rank, confirming its high accuracy and robustness for challenging real-world stereo estimation.