BridgeDepth: Bridging Monocular and Stereo Reasoning with Latent Alignment

Supplementary Material

This supplementary material provides additional insights and evaluations to support the findings presented in our main paper. In Sec. A, we elaborate on the details of the stereo feature network of BridgeDepth-L. Sec. B presents additional ablation studies, including comparing the performance of our framework when using different monocular foundation backbones, and a comparison with a post-hoc monocular-stereo combination strategy. In Sec. C, we report results from the official ETH3D benchmark, accompanied by a quantitative comparison between our refined monocular depth estimates and well-known monocular foundation models.

A. BridgeDepth-L variant

To assess the scalability of our approach, we introduce a variant named **BridgeDepth-L**, which replaces the standard `BasicEncoder` with an enlarged stereo feature network. This enhanced network is built upon a pre-trained ConvNext-Tiny model [31], a cutting-edge convolutional neural network recognized for its efficiency and robust feature extraction capabilities across various vision tasks.

In the BridgeDepth-L architecture, we utilize the first three "blocks" of the ConvNext-Tiny model. These blocks produce hierarchical feature maps at multiple spatial resolutions: 1/4, 1/8, and 1/16 of the input image size. This multi-scale feature extraction enables the model to capture both fine-grained local details (at higher resolutions) and broader contextual information (at lower resolutions), which are critical for high-quality correspondence estimation. To integrate these multi-resolution features effectively, we employ a DPT fusion module. The DPT fuses the features from different scales into a cohesive representation at 1/4 resolution with a channel dimension of 128. Subsequently, this fused feature map is average-pooled to yield a feature representation at 1/8 resolution, which serves as the input for coarse-level inference within our bidirectional alignment pipeline.

This variant demonstrates the flexibility of our framework, as it can seamlessly scale to incorporate more powerful backbones. By replacing `BasicEncoder` with enlarged stereo feature network, BridgeDepth-L serves as a proof-of-concept for adapting our approach to more sophisticated architectures, potentially yielding superior performance in demanding scenarios (*e.g.*, in Tab. 2).

B. Additional ablations

In this section, we present additional ablation studies to further investigate the robustness and versatility of the proposed framework. These experiments focus on two key aspects: (1) evaluating the performance of our framework when integrated with different monocular foundation backbones, and (2) comparing our iterative bidirectional alignment mechanism against a post-hoc monocular-stereo combination strategy. The results of these studies validate the generalization of our method and highlight the critical role of our alignment module in achieving superior disparity estimation performance.

Table 7. **In-domain accuracy on Scene Flow test set and zero-shot generalization on well-known real-world datasets.**

Backbone	SceneFlow EPE ↓	KITTI-12 D1 ↓	KITTI-15 D1 ↓	ETH3D BP-1 ↓	Middlebury (Q) BP-2 ↓
baseline	0.47	4.5	5.2	3.7	8.0
DAv2-B	0.37	3.8	4.7	1.4	4.7
DAv2-L	0.37	3.7	4.5	1.3	4.4
UDv2	0.36	3.7	4.8	1.8	5.1
MoGe	0.36	3.3	4.5	1.8	5.1
DAv2-B (post-hoc)	0.44	4.1	4.8	3.2	7.4

B.1. Generalization with monocular backbones

To demonstrate that our framework is not limited to a single monocular backbone (*i.e.*, DepthAnythingV2), we test its performance using multiple state-of-the-art monocular foundation models: DepthAnythingV2-L (DAv2-L) [54], DepthAnythingV2-B (DAv2-B) [54], UniDepthV2 (UDv2) [36], and MoGe [46]. Each backbone is integrated into our framework, and their performance is evaluated using in-domain and zero-shot experiments. All models are trained exclusively using Scene Flow data. The results, summarized in , indicate that the monocular backbone indeed affects generalization to some extent, our method exhibits significant improvements over baseline under all backbones. Thus, generality is validated.

B.2. Post-hoc fusion

We also compare our iterative bidirectional alignment mechanism with a post-hoc fusion strategy. In our proposed framework, monocular contextual features are iteratively aligned with stereo hypothesis volumes during the stereo aggregation process. In contrast, the post-hoc baseline directly fuses the monocular features with the aggregated stereo hypothesis volume using a lightweight ResNet, bypassing the iterative alignment step. For this experiment, we employ the ViT-B version of DAv2 as the monocular backbone. The performance of the post-hoc fusion strategy is reported in the last row of Table 7. While this baseline yields some improvement over purely monocular or stereo methods, it underperforms compared to our iterative alignment approach. This gap in performance suggests that the

Table 8. **Results on ETH3D leaderboard (test set)**

Method	BP-1	BP-2	AvgErr	RMS
LoS [27]	1.03	0.32	0.15	0.34
IGEV++ [52]	1.58	0.76	0.19	0.74
Selective-IGEV [47]	1.56	0.51	0.15	0.57
DEFOM-Stereo [21]	0.78	<u>0.19</u>	**0.11**	**0.26**
FoundationStereo [49]	**0.48**	0.26	0.13	0.61
BridgeDepth (Ours)	<u>0.50</u>	**0.16**	**0.11**	**0.26**

Table 9. **Monocular evaluation.**

Model	KITTI-12		KITTI-15		ETH3D		Middlebury	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DAv2-L	0.093	0.925	0.084	0.933	0.055	0.965	0.081	0.943
UDv2	0.054	0.967	0.067	0.949	0.052	0.972	0.056	0.982
Ours	**0.033**	**0.982**	**0.048**	**0.975**	**0.020**	**1.000**	**0.019**	**0.990**

primary gains in our framework arise from the dynamic, bidirectional alignment of monocular and stereo cues, rather than a static fusion of features. By enabling mutual refinement between modalities, our method produces more accurate and reliable correspondence estimates.

C. Additional results

In this section, we report the performance of our method on the official ETH3D benchmark, a widely adopted dataset designed for evaluating stereo matching methods. The benchmark is renowned for its challenging scenes, encompassing both indoor and outdoor environments with intricate geometries, textureless regions, and reflective surfaces. Our approach achieves state-of-the-art (SOTA) performance on this benchmark, underscoring its ability to deliver robust and accurate depth estimates across diverse conditions.

We also present a quantitative comparison between our refined monocular depth estimates and those generated by well-known monocular foundation models. This comparison, which will be detailed in Tab. 9, highlights the superior accuracy of our method. The key to this improvement lies in the integration of geometric precision from stereo data, which introduces additional constraints absent in purely monocular approaches. By leveraging stereo information, our refined monocular depth estimates gain the geometric consistency and precision inherent in stereo vision, resulting in significantly enhanced depth maps.