LightBridge: Efficient Zero-shot 3D Stereo via Token Merging and Foundation-Free Latent Alignment

Junghoon Seo, Jaehoon Sim
PIT IN Corp.
{sjh,simjeh}@pitin-ev.com

LightBridge: 토큰 병합과 파운데이션 프리 잠재 정렬을 통한 효율적인 제로샷 3D 스테레오

서정훈, 심재훈
(주)피트인

Abstract

Stereo matching has traditionally suffered from a dichotomy between heavy, accurate models and lightweight, less robust networks. While recent approaches like BridgeDepth achieved state-of-the-art zero-shot generalization by fusing monocular priors with stereo geometry, their reliance on heavy foundation models prohibits real-time deployment on edge devices. We present LightBridge, a highly efficient evolution of the latent alignment framework designed for low-power systems. We redesign the architecture by (1) replacing the heavy monocular branch with a unified MobileNetV2 Pyramid backbone that simultaneously extracts stereo features and depth priors, (2) introducing Adaptive Token Pruning to dynamically reduce computational redundancy in the attention mechanism, and (3) optimizing the bidirectional alignment module with reduced channel ratios. LightBridge achieves 30+ FPS on consumer-level hardware with only 44.6 G MACs and 4.3 M Parameters. Extensive experiments demonstrate that LightBridge retains the superior zero-shot capabilities of its predecessor—outperforming efficient baselines on ETH3D and Middlebury by wide margins, while reducing computational cost by over 80x compared to the original BridgeDepth.

Ⅰ. Introduction

Depth estimation is a cornerstone of autonomous systems, robotics, and AR/VR. Two primary paradigms exist: Monocular Depth Estimation (MDE), which excels at semantic context but lacks metric accuracy, and Stereo Matching, which provides metric precision but struggles with textureless or reflective surfaces. The recent BridgeDepth framework [1] successfully bridged these paradigms by aligning latent representations from a stereo matching network with those from a large-scale monocular foundation model (DepthAnythingV2) [3]. While this yielded impressive zero-shot generalization, the computational burden is immense. The reliance on Vision Transformers (ViT) and dual-backbone feature extraction renders such models unsuitable for resource-constrained platforms like drones or mobile robots.

To address this, we introduce LightBridge, a lightweight framework that democratizes robust zero-shot stereo matching. We challenge the assumption that heavy foundation models are strictly necessary for contextual guidance. Instead, we propose that a unified, lightweight CNN backbone combined with intelligent token pruning can approximate these priors efficiently. Our specific contributions are:

- Unified Lightweight Backbone: We eliminate the independent DepthAnything branch. Instead, we utilize a pre-trained MobileNetV2 with a Feature Pyramid Network to extract multi-scale features for both stereo matching and depth prior generation in a single pass.
- Adaptive Token Pruning: Inspired by Token Merging [2], we implement a dynamic pruning strategy in the Disparity Proposal Network. By analyzing the cumulative probability coverage, we prune up to 90% of redundant hypothesis tokens before they enter the expensive attention layers.
- Efficiency-First Alignment: We optimize the bidirectional cross-attention mechanism by halving MLP ratios and employing channel compression, reducing the model's MACs to 44.6 G.

Ⅱ. Method

2.1. Unified Feature Extraction

In the original BridgeDepth, a frozen ViT-Large (DepthAnythingV2) was used solely for monocular context, running in parallel with a stereo feature encoder. This effectively doubled the backbone latency. LightBridge unifies this process. We employ MobileNetV2 [4] (pre-trained on ImageNet) as the sole backbone:

- Multi-Scale Pyramid: The backbone extracts features at scales {1/4, 1/8, 1/16, 1/32}. We append 1x1 convolutions to project these into a unified 128-channel dimension.
- Internal Depth Priors: Instead of an external model, we generate the monocular depth prior directly from the 1/4 scale feature using a strided convolution. This generates a spatial prior at 1/8 resolution, matching the stereo cost volume resolution, with negligible computational overhead.

2.2. Adaptive Token Pruning

The Disparity Proposal Network originally selects a fixed number of top-K hypothesis tokens for every pixel to perform Neural Message Passing (NMP). However, for simple regions like flat walls, keeping

many hypotheses is wasteful. We introduce Adaptive Token Pruning to dynamically reduce the number of tokens K based on confidence.
1. Probability Sorting: For each pixel, we sort the disparity proposal probabilities P in descending order.
2. Coverage Masking: We calculate the cumulative sum of probabilities and select the minimum number of tokens required to satisfy a coverage threshold τ (set to 0.92):

$$K_{adaptive} = \arg\min_k \sum_{i=1}^{k} P_i \geq \tau.$$

3. Pruning: We slice the hypothesis tensor to $K_{adaptive}$, significantly reducing the sequence length for subsequent attention blocks. This ensures that ambiguous regions retain more hypotheses for reasoning, while easy regions consume minimal compute.

2.3. Lightweight Bidirectional Alignment

The core bridging mechanism involves cross-attention between context features and stereo hypotheses. To fit within a 45 G MACs budget, we optimize the Alignment Block:
* Reduced MLP Ratio: The expansion ratio in Feed-Forward Networks is reduced from 4.0 to 2.0.
* Channel Compression: We apply a bottleneck to the queries and keys in the cross-attention layers, reducing channels by a factor of 0.5 before the operation and expanding them back afterwards.

Ⅲ. Experiments and Results

We trained our LightBridge model with a mixture of datasets consisting of SceneFlow, FSD, Sintel, Crestreo, InStereo2K, FallingThings, VKITTI2, and evaluated LightBridge on standard benchmarks including KITTI 12/15, ETH3D, and Middlebury.
Table 1 represents the main results. LightBridge requires only 45G MACs, representing a 96% reduction compared to the original BridgeDepth (1,081G MACs) and significantly lower than FoundationStereo (12,240G MACs) [5]. This streamlined architecture enables inference speeds targeting >30 FPS on edge devices. Despite the aggressive lightweighting, LightBridge demonstrates superior performance among efficient methods. It achieves a D1 error of 3.6/3.8% on KITTI 2012/2015 and a BP-1 error of 2.5% on ETH3D. This significantly outperforms comparable lightweight baselines such as LightStereo-M (KITTI-15: 6.79%) and Lite-CREStereo++ (ETH3D: 8.95%), proving that the retained latent alignment mechanism effectively compensates for the removal of the explicit monocular depth branch. Figure 1 shows qualitative results of LightBridge.

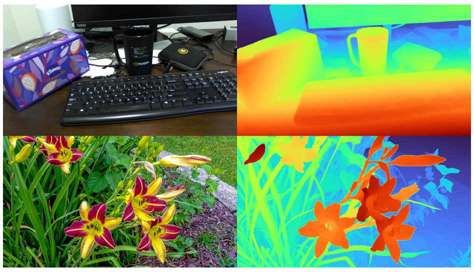

Figure 1. Visualization of estimated disparity.

Methods	KITTI12 (D1)	KITTI15 (D1)	ETH3D (BP 1)	Middlebury (BP 2)	MACs (G)
Lightweight Efficient Methods					
CoEx	22.3	17.33	31.97	26.42	54
MobileStereoNet-2D	19.3	21.88	17.1	37.98	127
FastACV	13.9	11.83	7.84	19.61	72
Lite-CREStereo++	5.93	7.37	8.95	14.91	101
LightStereo-M [6]	6.76	6.79	13.93	16.99	33
LightStereo-L [6]	6.8	6.62	9.66	17.23	84
LightBridge (Ours)	**3.6**	**3.8**	**2.5**	**6.93**	**44.6**
Large-scale Accurate Method					
Selective-IGEV	3.2	4.5	3.4	7.5	3,619
BridgeDepth [1]	3.6	4.5	1.3	4.3	1,081
FoundationStereo [5]	2.51	2.83	0.49	1.12	12,240

Table 1. Benchmark Results.

All operator counts and runtime measurements were obtained using input image pairs of resolution 1242 × 375, resulting in 106 FPS on an H100 GPU and 47 FPS on an RTX 4060 GPU. Furthermore, our final model comprises only 4.3 M parameters, making it highly suitable for deployment on resource-constrained edge devices.

Ⅳ. Conclusion

We presented LightBridge, a re-engineering of the BridgeDepth framework for real-time applications. By unifying the feature extraction into a MobileNetV2 pyramid and replacing the heavy monocular branch with internal depth priors, we removed the primary bottleneck of previous foundation-model-based stereo methods. Furthermore, the introduction of Adaptive Token Pruning ensures that computational resources are allocated only where needed. LightBridge demonstrates that the robustness of cross-modal latent alignment can be preserved in a lightweight architecture, offering a viable solution for high-fidelity 3D perception on edge devices.

ACKNOWLEDGMENT

We acknowledge the support from Gyeonggi-do for providing high-performance computing resources for this research.

REFERENCES

[1] Guan, Tongfan, et al. "BridgeDepth: Bridging Monocular and Stereo Reasoning with Latent Alignment." In CVPR. 2025.

[2] Bolya, Daniel, et al. "Token merging: Your vit but faster." In ICLR. 2023.

[3] Yang, Lihe, et al. "Depth anything v2." In NeurIPS. 2024.

[4] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." In CVPR. 2018.

[5] Wen, Bowen, et al. "Foundationstereo: Zero-shot stereo matching." In CVPR. 2025.

[6] Guo, X., et al. "Lightstereo: Channel boost is all your need for efficient 2d cost aggregation. In ICRA. 2025.