
Appendix for PPMStereo: Pick-and-Play Memory Construction for Consistent Dynamic Stereo Matching

Anonymous Author(s)

Affiliation

Address

email

1 Appendix for PPMStereo

2 Our supplementary material provides extensive additional analysis, implementation details, and
3 discussions, organized as follows: (A) Demonstration Video and More Visualization (Sec. 1). We
4 include a comprehensive demo video (included in demo_outputs.zip) showcasing: (1) Real-world
5 dynamic scene reconstructions, (2) Corresponding disparity maps, (3) Comparative results under
6 varying conditions. (B) Implementation Details (Sec. 2). We present complete technical specifications
7 for our PPMStereo_VDA framework, including: (1) Model architecture: Detailed network config-
8 uration. (2) Datasets: Descriptions of all benchmark datasets used for evaluation. (3) Algorithmic
9 details: Detailed pseudo-codes. (4) Computational analysis: Runtime and GPU memory comparisons.
10 (5) Memory buffer visualization: Evidence of long-range relationship modeling. (C) Additional
11 discussions on limitations and future work. We offer a more detailed discussion of the limitations and
12 potential future directions (Sec. 3).



Figure 1: Qualitative comparisons on the Dynamic Replica test set. They are rendered with a camera displaced by 15 degree angles. Our method exhibits smoother reconstruction results.

13 1 More Visualizations on Real-world Scenes

14 Figure 1 demonstrates the reconstruction performance of our method on the Dynamic Replica (DR)
15 test set. The results illustrate our approach’s ability to accurately recover fine-grained details while
16 preserving the global structural integrity of the scene, even under challenging dynamic conditions.
17 Figure 2 and Figure 3 showcase the performance of our method in outdoor real-world scenarios,
18 highlighting its robustness under varying lighting conditions and complex backgrounds. For indoor
19 environments, Figure 4 and Figure 5 provide a comprehensive comparison, demonstrating consistent
20 accuracy even in confined spaces with occlusions and dynamic objects. Additional qualitative
21 results (e.g., thin structures and reconstructed results) are available in the supplementary materials
22 (demo_outputs.zip).

23

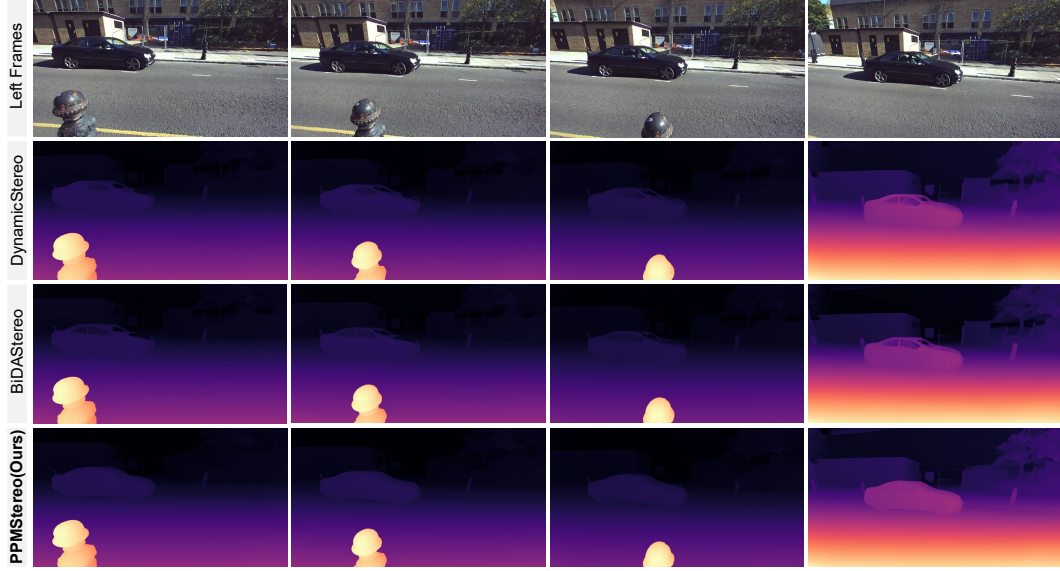


Figure 2: Qualitative comparison on a dynamic outdoor scenario from the South Kensington SV dataset [9].

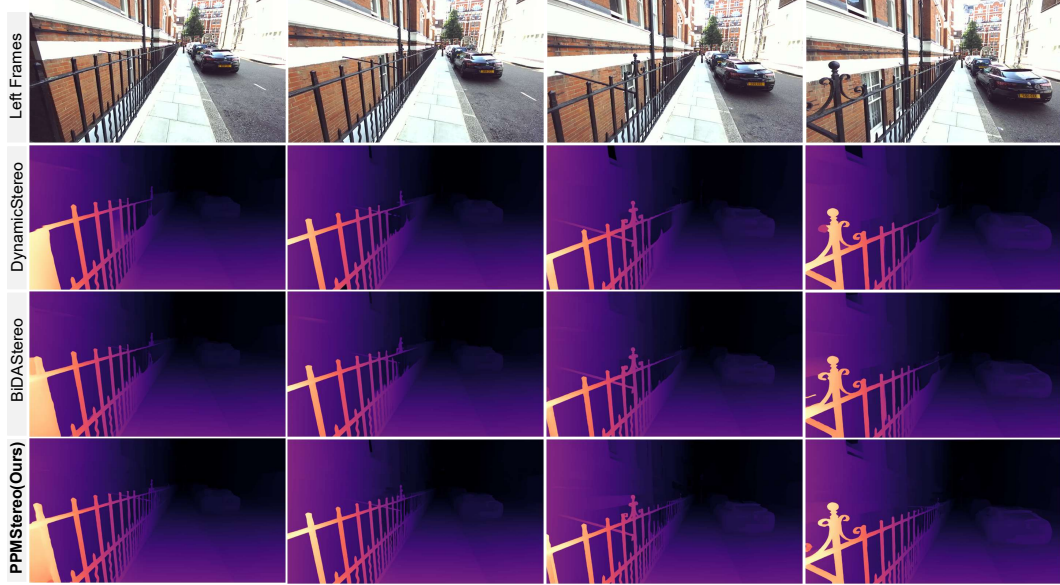


Figure 3: Qualitative comparison on a dynamic outdoor scenario from the South Kensington SV dataset [9].

24 2 Implementation Details

25

26 2.1 PPMStereo_VDA

27 For PPMStereo_VDA model, we use VideoDepthAnything [2] to replace our feature extractor.
 28 Specifically, in the feature extraction stage, when processing a video sequence with the monocular
 29 video depth model, we first resize it to ensure its dimensions are divisible by 14, maintaining

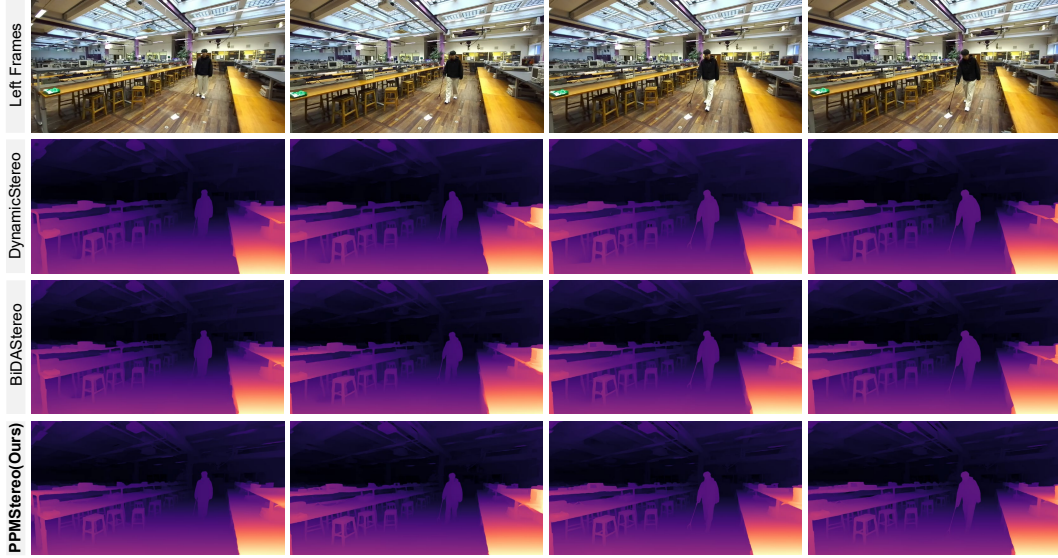


Figure 4: Qualitative comparison on a dynamic indoor scenario from the South Kensington SV dataset [9].

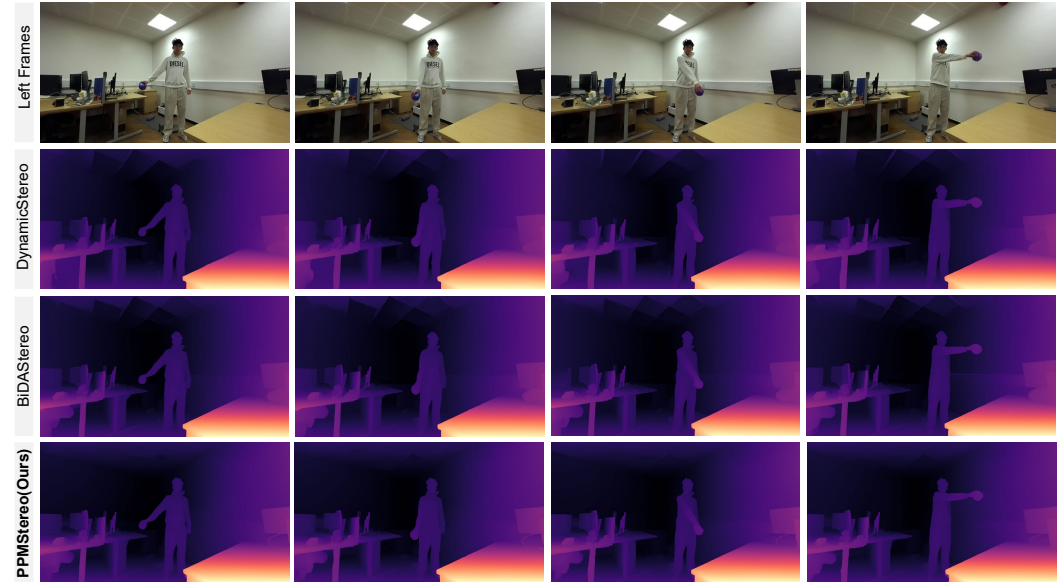


Figure 5: Qualitative comparison on a dynamic indoor scenario from the South Kensington SV dataset [9].

30 consistency with the model’s pretrained patch size. After obtaining the feature maps, we resize the
 31 image back to its original dimensions. The monocular depth model produces feature maps with 64
 32 channels, while the CNN encoders extract both image and context features with 128 channels each.
 33 These feature maps are concatenated to form a 192-channel representation, a decoder is used then to
 34 obtain a 128-channel representation, which serves as input to the subsequent correlation module.

35 2.2 Datasets.

36 **SceneFlow (SF)** SceneFlow [11] consists of three subsets: FlyingThings3D, Driving, and Monkaa.



Figure 6: For the target frame (11th frame), the occlusion point is highlighted by a yellow circle. Unlike conventional approaches that rely on adjacent frames, our PPMStereo method dynamically selects and aggregates features from the most informative and diverse frames across the entire sequence ($T=20$). By adaptively bypassing occluded or unreliable neighboring frames, PPMStereo ensures robust and occlusion-aware feature representation, enhancing both accuracy and generalization.

- FlyingThings3D is an abstract dataset featuring moving shapes against colorful backgrounds. It contains 2,250 sequences, each spanning 10 frames.
- Driving includes 16 sequences depicting driving scenarios, with each sequence containing between 300 and 800 frames.
- Monkaa comprises 48 sequences set in cartoon-like environments, with frame counts ranging from 91 to 501.

Sintel Sintel [1] is generated from computer-animated films. It consists of 23 sequences available in both clean and final rendering passes. Each sequence contains 20 to 50 frames. We use the full sequences of Sintel for evaluation.

Dynamic Replica Dynamic Replica [10] is designed for longer sequences and the presence of non-rigid objects such as animals and humans. The dataset includes:

- 484 training sequences, each with 300 frames.
- 20 validation sequences, each with 300 frames.
- 20 test sequences, each with 900 frames.

Following prior methods [10, 8], we use the entire training set for model training and evaluate on the first 150 frames of the test set.

South Kensington SV South Kensington SV [8] is a real-world stereo dataset capturing daily life scenarios for qualitative evaluation. It consists of 264 stereo videos, each lasting between 10 and 70 seconds, recorded at 1280×720 resolution and 30 fps. We conduct qualitative evaluations on this dataset.

2.3 Computational Costs

As illustrated in Fig. 7, we conduct a comprehensive comparison of the competing methods across three critical metrics: model size (parameters), training GPU memory consumption, and computational complexity (multiply-accumulate operations, MACs). Our proposed method achieves an optimal trade-off among these efficiency criteria while simultaneously delivering the lowest error rate. Notably, compared to the previous state-of-the-art approach, BiDAStereo [8], our method demonstrates a significant performance improvement while maintaining comparable computational costs. The advantage of enhanced accuracy and superior efficiency makes our approach particularly suitable for real-world applications.

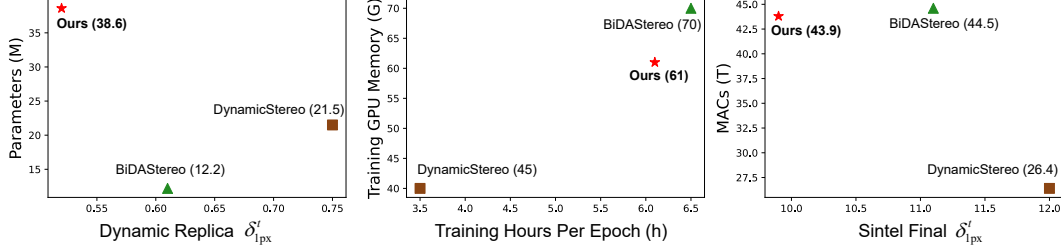


Figure 7: (a) δ^t_{1px} on DR vs. parameters. (b) Training GPU memory at 320×512 vs. Training hours per epoch. (c) δ^t_{1px} on Sintel vs. MACs (20 frames \times 768×1024).

2.4 Memory Reference

Here, we visualize the memory aggregation process (Section 3) by showing the candidate frames, some of the selected reference frames, and the corresponding aggregation weights. As illustrated in Figure 6 we observe semantically meaningful regions to be focused.

3 Limitations and Future

While our method advances the state of dynamic scene modeling, it shares a common limitation with existing approaches: the inability to proactively distinguish between dynamic and static regions, which is crucial for maintaining temporal consistency. Also, our method occasionally in textureless areas (e.g., blank walls) or transparent surfaces (e.g., glass), where current techniques, including ours, may produce inconsistencies. To address these limitations, we plan to pursue two key directions: (1) integrating high-quality memory cues to improve scene understanding and consistency, and (2) developing a lightweight variant of our model for resource-constrained applications [6, 7, 5, 3, 4, 12]. Looking forward, we aim to create a comprehensive model zoo featuring both full-capacity and efficient versions of our approach, facilitating adoption across different hardware scenarios.

References

- [1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 611–625. Springer, 2012.
- [2] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. 2025.
- [3] Hong Huang and Dapeng Wu. Quaff: Quantized parameter-efficient fine-tuning under outlier spatial stability hypothesis. *arXiv preprint arXiv:2505.14742*, 2025.
- [4] Hong Huang, Decheng Wu, Rui Cen, Guanghua Yu, Zonghang Li, Kai Liu, Jianchen Zhu, Peng Chen, Xue Liu, and Dapeng Wu. Tequila: Trapping-free ternary quantization for large language models. *arXiv preprint arXiv:2509.23809*, 2025.
- [5] Hong Huang, Hai Yang, Yuan Chen, Jiaxun Ye, and Dapeng Wu. Fedrts: Federated robust pruning via combinatorial thompson sampling. *arXiv preprint arXiv:2501.19122*, 2025.
- [6] Hong Huang, Lan Zhang, Chaoyue Sun, Ruogu Fang, Xiaoyong Yuan, and Dapeng Wu. Distributed pruning towards tiny neural networks in federated learning. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, pages 190–201. IEEE, 2023.
- [7] Hong Huang, Weiming Zhuang, Chen Chen, and Lingjuan Lyu. Fedmef: Towards memory-efficient federated dynamic pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27548–27557, 2024.

- 100 [8] Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment
101 for consistent dynamic stereo matching. In *European Conference on Computer Vision (ECCV)*,
102 pages 415–432. Springer, 2024.
- 103 [9] Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirec-
104 tional alignment. 2024.
- 105 [10] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and
106 Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In
107 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,
108 pages 13229–13239, 2023.
- 109 [11] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy,
110 and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow,
111 and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and*
112 *Pattern Recognition (CVPR)*, pages 4040–4048, 2016.
- 113 [12] Shuguang Wang, Qian Zhou, Kui Wu, Jinghuai Deng, Dapeng Wu, Wei-Bin Lee, and Jianping
114 Wang. Interventional root cause analysis of failures in multi-sensor fusion perception systems.
115 *perception*, 4:5, 2025.