StereoSurg: A Benchmark for Stereo Depth Estimation in Surgical Scenes

Jiaxin Guo[1,2], Wenzhen Dong[1,2], Tongfan Guan[1,2], Ziyi Wang[1,2], Qi Dou[1,2], and Yun-Hui Liu[1,2(✉)]

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Hong Kong Center for Logistics Robotics, Hong Kong SAR, China

Abstract. Stereo endoscopy provides metric depth through binocular geometry and is an important sensing modality for robot-assisted minimally invasive surgery. However, surgical stereo matching remains challenging due to narrow baselines, textureless tissue, specular reflections, non-Lambertian appearance, and instrument occlusion. Meanwhile, stereo matching has rapidly advanced with monocular depth priors, large-scale pretraining, and foundation stereo models, yet their effectiveness in surgical scenes has not been systematically compared under a unified protocol. In this work, we introduce StereoSurg, a benchmark of recent stereo depth estimation methods, including IGEV++, MonSter, MonSter++, S^2M^2, FoundationStereo, Fast-FoundationStereo, BridgeDepth, and StereoMamba. We use SCARED as the quantitative benchmark and evaluate disparity accuracy using EPE, D1, Bad-k metrics, and metric-depth error. We further compare computational cost in terms of model size, inference speed, and GPU memory, and qualitatively evaluate challenging in-vivo scenes from Hamlyn and StereoMIS. Under our unified protocol, FoundationStereo achieves the best EPE at 2.659 px and the lowest Bad1 and Bad3 rates, while BridgeDepth achieves the best D1 at 20.63%. StereoMamba reports a slightly lower EPE of 2.64 px and the best Bad2 at 41.49% under its SCARED fine-tuning protocol. Fast-FoundationStereo performs best on Dataset 8 with a mean depth error of 2.83 mm, whereas S^2M^2 performs best on Dataset 9 with 1.95 mm. Fast-FoundationStereo also achieves 28.02 FPS with 1.08 GB GPU memory. Qualitative comparisons reveal additional failure modes around specularities, weak texture, surgical instruments, and degraded regions. Overall, the benchmark highlights the trade-offs between accuracy, in-vivo robustness, and computational efficiency in surgical stereo estimation.

Keywords: Stereo depth estimation · Surgical scene reconstruction · Stereo matching · Minimally invasive surgery

1 Introduction

Accurate depth perception is fundamental to robot-assisted minimally invasive surgery, supporting surgical navigation [12], instrument localization [19, 28], 3D reconstruction [6, 7, 9, 22], tissue motion analysis [8, 10, 25], and increasingly

autonomous manipulation [11, 29]. Stereo endoscopy is particularly attractive because binocular correspondence provides metric depth from image observations and camera calibration, avoiding the scale ambiguity of monocular estimation. Early surgical vision systems therefore explored stereo reconstruction for intraoperative 3D perception and tissue tracking [15, 20]. With the rapid development of deep stereo matching, recent learning-based methods have achieved substantial progress in disparity estimation on general-domain benchmarks [1–4, 16, 23, 24, 27]. This raises an important question: *how well do these advances transfer to surgical scenes?*

Answering this question is challenging because surgical stereo images differ substantially from the natural-image domains on which most stereo networks are developed. Endoscopic images often contain textureless tissue regions, weak and spatially varying disparities due to short stereo baselines, specular reflections, non-uniform illumination, repetitive anatomical structures, and frequent instrument occlusions. Tissue deformation and instrument motion further introduce appearance changes across views and time, making reliable correspondence particularly difficult. Thus, strong performance on conventional stereo benchmarks does not necessarily translate into reliable depth in surgical environments.

Meanwhile, stereo matching has evolved rapidly from conventional cost-volume and recurrent architectures [26, 27] to monocular-guided [1, 2], scalable [16], and foundation stereo models [4, 23, 24]. These models introduce increasingly strong monocular priors and large-scale pretraining, offering substantially broader generalization than earlier stereo networks. However, their performance, efficiency, and robustness in surgical scenes have not been systematically characterized under a common evaluation protocol. Existing surgical stereo studies typically focus on individual architectures or limited benchmark settings [14, 21], leaving the behavior of recent general-purpose stereo models in surgery largely unclear.

In this work, we introduce **StereoSurg**, a systematic benchmark study of recent stereo depth estimation in surgical scenes. We evaluate seven recent stereo estimators, including IGEV++ [27], MonSter [2], MonSter++ [1], S^2M^2 [16], FoundationStereo [24], Fast-FoundationStereo [23], and BridgeDepth [4], under a unified evaluation protocol, and include StereoMamba [21] as an external surgical reference. We examine four aspects of surgical stereo perception: disparity accuracy, metric-depth accuracy, computational efficiency, and qualitative robustness under challenging in-vivo conditions. In addition, we analyze performance across individual SCARED test datasets and keyframes to examine whether aggregate metrics consistently reflect model behavior.

Our experiments reveal several findings. First, modern stereo models transfer effectively to surgical imagery: among the methods evaluated under our unified protocol, FoundationStereo achieves the best EPE of 2.659 px and the lowest Bad1 and Bad3 rates, while BridgeDepth obtains the best D1 of 20.63%. StereoMamba reports a slightly lower EPE of 2.64 px and the best Bad2 rate of 41.49% under its SCARED fine-tuning protocol. Second, metric-depth performance varies across the two SCARED test datasets: Fast-FoundationStereo performs best on Dataset 8 with a mean depth error of 2.83 mm, whereas S^2M^2

performs best on Dataset 9 with 1.95 mm. Fast-FoundationStereo also achieves 28.02 FPS with 1.08 GB GPU memory. These results reveal a gap between benchmark-level geometric accuracy and reliable behavior under challenging surgical conditions, motivating evaluation protocols that jointly consider accuracy, robustness, and computational efficiency.

Our contributions are summarized as follows:

- We establish a unified benchmark for evaluating seven recent stereo depth estimation methods in surgical scenes, covering recurrent, monocular-guided, scalable, and foundation stereo architectures.
- We provide an extensive and consistent evaluation using quantitative disparity and metric-depth metrics on SCARED, together with qualitative assessment of challenging in-vivo scenes from Hamlyn and StereoMIS.
- We systematically analyze the accuracy–efficiency trade-offs and sequence-dependent performance of stereo models.
- We identify representative failure modes around weak texture, specularities, surgical instruments, and degraded image regions, providing practical insights for surgical stereo perception.

2 Benchmark Setup

2.1 Surgical Stereo Datasets

We evaluate three complementary surgical stereo datasets, using SCARED for quantitative evaluation and Hamlyn and StereoMIS for qualitative assessment of in-vivo robustness.

SCARED. The SCARED dataset [13] provides calibrated stereo endoscopic sequences acquired with a da Vinci Xi system and dense metric depth from structured-light scanning. We use its standard test split for quantitative evaluation with EPE, D1, Bad-k, and metric-depth error.

Hamlyn. The Hamlyn Dataset [17, 18, 20] contains diverse in-vivo surgical videos with challenging conditions such as tissue deformation, specular reflections, smoke, and instrument occlusion. We use representative stereo frames for qualitative evaluation, as reliable dense depth ground truth is unavailable for the selected sequences.

StereoMIS. StereoMIS [14] provides in-vivo stereo videos from robotic porcine procedures, including breathing motion, tool motion, and non-rigid tissue deformation. Without dense metric depth ground truth, we use it for qualitative comparison of disparity structure and robustness in dynamic surgical scenes.

2.2 Compared Methods

We compare representative recent stereo matching methods covering iterative refinement, scalable stereo architectures, monocular-guided stereo estimation, and foundation-model-based stereo matching. IGEV++ [27] represents a strong

geometry-aware iterative refinement architecture, while S^2M^2 [16] adopts a scalable multi-scale design for reliable disparity estimation. MonSter [2] and MonSter++ [1] represent recent monocular-guided stereo models that leverage monocular depth priors to initialize stereo disparity and improve matching generalization. FoundationStereo [24] and Fast-FoundationStereo [23] represent recent foundation stereo models designed for zero-shot generalization, while Fast-FoundationStereo tailored for efficient inference. Finally, BridgeDepth [4] is included as a recent stereo framework that aligns monocular depth priors and neural MRF [5] stereo representation in latent space. All methods are evaluated using their released implementations and pretrained models under the same input resolution and evaluation protocol. For methods supporting different inference configurations, we use the recommended default configuration unless otherwise specified. We additionally report the SCARED results of StereoMamba [21] as an external reference. Its results are taken from the original paper and marked with †, since the reported model is fine-tuned on SCARED.

2.3 Evaluation Metrics

We quantitatively evaluate stereo depth estimation only on SCARED, where reliable dense depth ground truth is available. Given a predicted disparity map d and calibrated stereo parameters with focal length f and baseline B, metric depth is recovered as $Z = \frac{fB}{d}$. For disparity accuracy, we report end-point error (EPE), D1 error, and Bad-k rates. EPE measures the mean absolute disparity error in pixels. D1 reports the percentage of pixels whose absolute disparity error is larger than 3 pixels and exceeds 5% of the ground-truth disparity. Bad-k denotes the percentage of pixels with absolute disparity error larger than k pixels. Lower values indicate better performance.

For metric-depth accuracy, we report the mean absolute depth error (MAE) in millimeters following [13]: $\text{MAE} = \frac{1}{|\Omega|}\sum_{p\in\Omega}|Z(p) - Z^*(p)|$, where Z^* denotes the reference depth and Ω is the set of valid evaluation pixels.

For practical deployment, we additionally report the number of model parameters, inference speed in frames per second (FPS), and peak GPU memory consumption under a unified inference configuration. We further highlight the table cell to show red, orange, and yellow, indicating the best, second-best, and third-best results, respectively.

3 Experimental Results

3.1 Implementation Details

All experiments are conducted on NVIDIA GeForce RTX 4090 D. The input image resolutions are 1178×1014, 1280×1024, 640×480 for SCARED, StereoMIS, and Hamlyn, respectively. Stereo pairs are rectified before disparity estimation, and predicted disparities are converted to metric depth using the provided calibration parameters. Unless otherwise stated, the same validity masks and depth

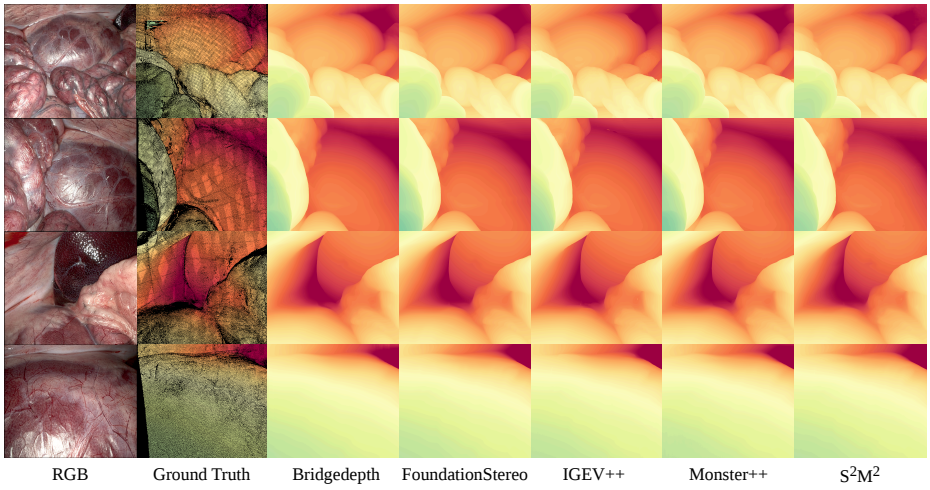

Fig. 1. Qualitative comparison of stereo depth on SCARED [13].

range are applied to all methods. Runtime is measured on the same GPU with batch size one.

3.2 Comparison on SCARED

Quantitative Results. Table 1 presents the quantitative comparison on SCARED. FoundationStereo achieves the best EPE of 2.659 pixels and the lowest Bad1 and Bad3 error rates, while BridgeDepth achieves the lowest D1 error of 20.63%. MonSter++ remains competitive with an EPE of 2.698 pixels and a D1 error of 20.77%, whereas Fast-FoundationStereo obtains 2.763 pixels and 21.33%, respectively. StereoMamba reports a slightly lower EPE of 2.64 pixels and the lowest Bad2 error of 41.49% among the reported results, but its result is obtained under an SCARED fine-tuning protocol. These results indicate that recent stereo models transfer effectively to surgical imagery, while their relative performance depends on the evaluation metric. FoundationStereo provides the strongest overall performance among the methods evaluated under our unified protocol, whereas BridgeDepth minimizes D1 error. The StereoMamba result further highlights the competitiveness of surgical-domain fine-tuning.

We further evaluate metric-depth accuracy on the two SCARED test datasets (Dataset 8 and Dataset 9) following [13]. Following the original SCARED challenge protocol [13], we report the mean absolute depth error (MAE) for five representative keyframes, with the final column giving their average. We also include the best challenge-period result for each dataset as the *Challenge Baseline*: J. C. Rosenthal for Dataset 8 and Trevor Zeffiro for Dataset 9 [13].

As shown in Table 2, recent models substantially improve over the challenge baselines. Fast-FoundationStereo achieves the lowest mean error on Dataset 8 (2.83 mm), while S^2M^2 performs best on Dataset 9 (1.95 mm). FoundationStereo

Table 1. Quantitative comparison on our unified surgical stereo benchmark. † denotes results reported from StereoMamba [21], which is fine-tuned on the SCARED surgical dataset. Metrics not reported in the original paper are denoted by '–'.

Method	EPE (px)↓	D1 (%)↓	Bad1 (%)↓	Bad2 (%)↓	Bad3 (%)↓
IGEV++ [27]	2.710	21.13	65.90	42.43	27.42
MonSter [2]	2.850	21.43	67.89	43.89	28.13
MonSter++ [1]	2.698	20.77	66.51	42.75	27.19
FoundationStereo [24]	2.659	20.72	65.46	41.97	26.84
Fast-FoundationStereo [23]	2.763	21.33	65.98	42.79	27.80
S^2M^2 [16]	2.919	22.74	67.04	43.96	28.94
BridgeDepth [4]	2.689	20.63	66.78	42.68	27.09
StereoMamba† [21]	2.640	–	–	41.49	26.99

Table 2. Mean absolute depth error (mm) on the SCARED test datasets. † denotes results reported from StereoMamba [21], which is fine-tuned on the SCARED surgical dataset.

Dataset	Method	Keyframe 0	Keyframe 1	Keyframe 2	Keyframe 3	Keyframe 4	Avg
Dataset 8	Challenge Baseline [13]	8.25	3.36	2.21	3.44	1.33	3.72
	IGEV++ [27]	8.28	2.47	1.53	1.70	0.50	2.90
	MonSter [2]	8.06	2.29	1.58	3.73	0.65	3.26
	MonSter++ [1]	7.83	2.27	1.58	3.66	0.50	3.17
	FoundationStereo [24]	7.98	2.35	1.55	3.69	0.44	3.20
	Fast-FoundationStereo [23]	7.98	2.35	1.57	1.78	0.46	2.83
	S^2M^2 [16]	8.02	2.30	1.85	3.49	0.47	3.23
	BridgeDepth [4]	7.85	2.21	1.54	3.66	0.57	3.17
	StereoMamba† [21]	8.57	2.61	1.46	2.03	0.77	3.09
Dataset 9	Challenge Baseline [13]	5.39	1.67	4.34	3.18	2.79	3.47
	IGEV++ [27]	3.95	0.97	3.08	1.61	0.35	1.99
	MonSter [2]	3.99	0.99	2.97	1.63	0.56	2.03
	MonSter++ [1]	4.08	1.01	2.88	1.60	0.45	2.00
	FoundationStereo [24]	4.00	0.95	3.04	1.57	0.33	1.98
	Fast-FoundationStereo [23]	4.04	0.99	3.07	1.61	0.38	2.02
	S^2M^2 [16]	3.99	0.94	2.79	1.58	0.44	1.95
	BridgeDepth [4]	3.93	0.95	2.95	1.59	0.52	1.99
	StereoMamba† [21]	4.19	0.92	3.07	1.46	0.41	2.01

and BridgeDepth follow closely on Dataset 9 with 1.98 mm and 1.99 mm, respectively. StereoMamba reports 3.09 mm and 2.01 mm on Dataset 8 and Dataset 9, which shows competitive performance but does not outperform the best directly evaluated model on either test dataset. The change in ranking between the two datasets indicates that metric-depth performance remains sequence-dependent.

Qualitative Results. Figure 1 provides a qualitative comparison on SCARED, showing that these methods recover the dominant anatomical surfaces and major depth transitions, consistent with their strong quantitative performance. The predictions are largely similar on the broad tissue regions, while smaller differences appear around depth boundaries and fine anatomical structures. FoundationStereo and BridgeDepth produce visually smooth surfaces that closely follow the reference geometry, whereas IGEV++, MonSter++, and S^2M^2 show slightly more local variation in some regions. Overall, the qualitative comparison confirms that the numerical differences in Table 1 correspond primarily to subtle local errors rather than large failures on the SCARED scenes.

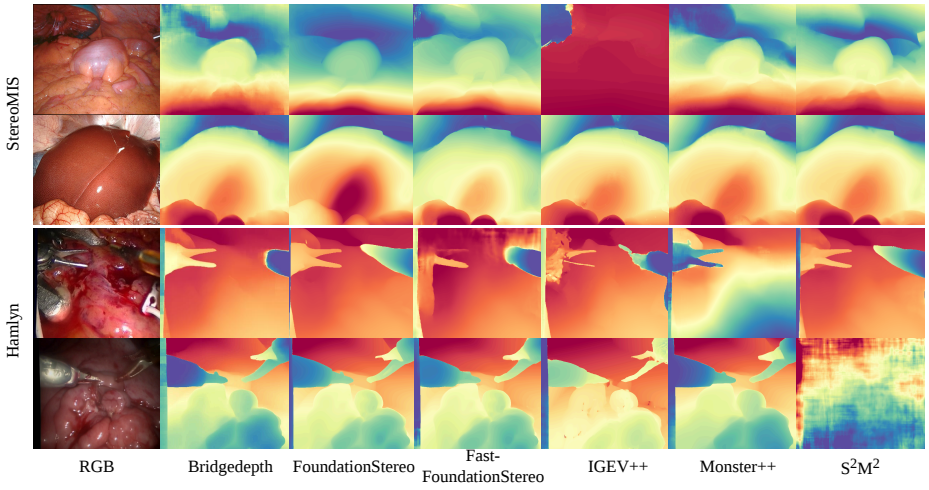

Fig. 2. Qualitative comparison of stereo depth on StereoMIS and Hamlyn.

3.3 Qualitative Generalization to In-Vivo Scenes

We further evaluate the models on challenging in-vivo scenes from StereoMIS and Hamlyn, where reliable dense metric-depth ground truth is unavailable. Figure 2 shows that most models recover the dominant anatomical surfaces, but their predictions differ substantially in challenging regions.

On StereoMIS, BridgeDepth, FoundationStereo, and Fast-FoundationStereo generally produce smooth and spatially coherent depth on large tissue surfaces, while IGEV++ and S^2M^2 exhibit more local variation in some regions. Around tissue boundaries and foreground structures, the models also differ in boundary sharpness and the amount of depth bleeding. These differences are particularly visible around the curved organ surfaces and instrument regions.

The Hamlyn examples reveal additional challenges caused by strong specularities, blood, weak texture, and surgical instruments. Most methods recover the coarse tissue geometry, but some predictions contain local artifacts or excessive smoothing near reflective and thin structures. S^2M^2 in particular exhibits visibly noisy responses in some degraded regions, whereas the foundation and monocular-guided models tend to produce smoother depth surfaces. These observations indicate that strong quantitative performance on SCARED does not necessarily translate into reliable geometry across diverse in-vivo conditions.

3.4 Efficiency Comparison

Table 3 compares computational efficiency at 640×480 resolution, reporting number of parameters, inference speed, and peak GPU memory consumption. Fast-FoundationStereo achieves the highest throughput (28.02 FPS) and lowest

Table 3. Efficiency comparison on Hamlyn in 640×480 resolution. We report the number of parameters, inference speed, and peak GPU memory consumption.

Method	Params (M)↓	Speed (FPS)↑	GPU (GB)↓
IGEV++ [27]	14.56	4.79	4.64
MonSter [2]	388.7	3.52	9.52
MonSter++ [1]	388.7	3.58	7.55
S^2M^2 [16]	406.5	7.96	6.23
FoundationStereo [24]	374.5	2.78	6.14
Fast-FoundationStereo [23]	16.87	28.02	1.08
BridgeDepth [4]	359.9	14.29	5.68

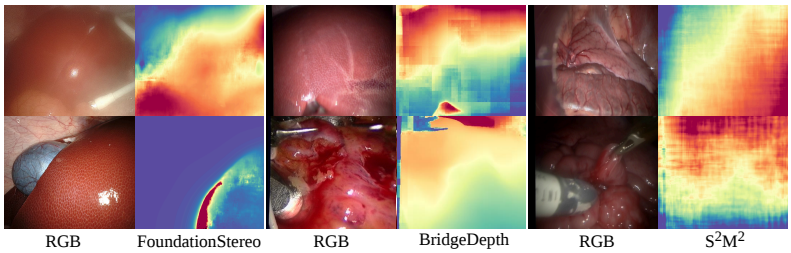

Fig. 3. Failure case visualization for FoundationStereo, BridgeDepth, and S^2M^2.

GPU memory consumption (1.08 GB), while IGEV++ has the fewest parameters (14.56M). BridgeDepth reaches 14.29 FPS with 5.68 GB memory, providing substantially higher throughput than the other high-capacity models. Overall, the results reveal a clear accuracy–efficiency trade-off: Fast-FoundationStereo provides the strongest efficiency profile, while BridgeDepth offers a favorable balance between accuracy and computational cost.

3.5 Failure Analysis and Discussion

As shown in Fig. 3, we visualize the representative failure cases of the latest models. Specular reflections on tissue surfaces can introduce local disparity artifacts because view-dependent reflections weaken stereo correspondence. Thin instruments and sharp tissue boundaries are also difficult because small localization errors can cause substantial depth deviations. In the example with blood and instrument, the predicted depth becomes less spatially consistent around fine structures. The S^2M^2 example additionally shows pronounced high-frequency artifacts in a degraded surgical region, illustrating a failure mode that is less evident from aggregate metrics. Overall, the qualitative results indicate that quantitative stereo accuracy alone does not fully characterize the reliability of local surgical geometry. Joint consideration of disparity accuracy, metric-depth accuracy, in-vivo behavior, and computational efficiency is therefore important for surgical stereo perception.

4 Conclusion

We introduce StereoSurg, a systematic benchmark of recent stereo depth estimation methods for surgical scenes, covering accuracy, efficiency, and in-vivo robustness. Among the methods evaluated under our unified protocol, FoundationStereo achieves the best EPE of 2.659 px and the lowest Bad1 and Bad3 rates, while BridgeDepth obtains the best D1 of 20.63%. StereoMamba reports a slightly lower EPE of 2.64 px and the best Bad2 rate of 41.49% under its SCARED fine-tuning protocol. Metric-depth performance is also sequence-dependent, with Fast-FoundationStereo best on Dataset 8 and S^2M^2 best on Dataset 9. Fast-FoundationStereo further provides the strongest efficiency at 28.02 FPS with 1.08 GB GPU memory. Qualitative results reveal failures under weak texture, specularities, instruments, and degraded conditions, highlighting a gap between strong benchmark accuracy and reliable geometry in real surgical scenes.

Acknowledgments. This work is supported in part by the HK RGC AoE under AoE/E-407/24-N, Grant 14218322, in part by RGC GRF Ref. 14207324, in part by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government via the Hong Kong Centre for Logistics Robotics, in part by the Multi-Scale Medical Robotics Centre, and in part by the CUHK T Stone Robotics Institute.

References

1. Cheng, J., Liao, W., Cai, Z., Liu, L., Xu, G., Wang, X., Wang, Y., Yuan, Z., Deng, Y., Zang, J., et al.: Monster++: Unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
2. Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X.: Monster: Marry monodepth to stereo unleashes power. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6273–6282. IEEE (2025)
3. Guan, T., Guo, J., Huang, T., Dong, J., Wang, C., Liu, Y.H.: Dispvit: Direct stereo disparity regression with a single-stream vision transformer. In: The Fourteenth International Conference on Learning Representations (2026)
4. Guan, T., Guo, J., Wang, C., Liu, Y.H.: Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 27681–27691. IEEE (2025)
5. Guan, T., Wang, C., Liu, Y.H.: Neural markov random field for stereo matching. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5459–5469. IEEE (2024)
6. Guo, J., Dong, W., Huang, T., Ding, H., Wang, Z., Kuang, H., Dou, Q., Liu, Y.H.: Endo3r: Unified online reconstruction from dynamic monocular endoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 170–180. Springer (2025)
7. Guo, J., Wang, J., Kang, D., Dong, W., Wang, W., Liu, Y.h.: Free-surgs: Sfm-free 3d gaussian splatting for surgical scene reconstruction. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 350–360. Springer (2024)
8. Guo, J., Wang, J., Li, Z., Jia, T., Dou, Q., Liu, Y.H.: Ada-tracker: Soft tissue tracking via inter-frame and adaptive-template matching. In: 2024 IEEE international conference on robotics and automation (ICRA). pp. 15463–15470. IEEE (2024)
9. J. Guo *et al.*: Uc-nerf: Uncertainty-aware conditional neural radiance fields from endoscopic sparse views. IEEE Transactions on Medical Imaging (2024)
10. Karaoglu, M.A., Ji, W., Abbas, A., Navab, N., Busam, B., Ladikos, A.: Litetracker: Leveraging temporal causality for accurate low-latency tissue tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 308–317. Springer (2025)
11. Kim, J.W., Chen, J.T., Hansen, P., Shi, L.X., Goldenberg, A., Schmidgall, S., Scheikl, P.M., Deguet, A., White, B.M., Tsai, D.R., et al.: Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. Science robotics **10**(104), eadt5254 (2025)
12. Lu, Y., Wei, R., Li, B., Chen, W., Zhou, J., Dou, Q., Sun, D., Liu, Y.h.: Autonomous intelligent navigation for flexible endoscopy using monocular depth guidance and 3-d shape planning. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 1–7. IEEE (2023)
13. M. Allan *et al.*: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
14. M. Hayoz *et al.*: Learning how to robustly estimate camera pose in endoscopic videos. International journal of computer assisted radiology and surgery **18**(7), 1185–1192 (2023)
15. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al.: Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Medical image analysis **17**(8), 974–996 (2013)
16. Min, J., Jeon, Y., Kim, J., Choi, M.: S^2M^2: Scalable stereo matching model for reliable depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
17. P. Mountney *et al.*: Three-dimensional tissue deformation recovery and tracking. IEEE Signal Process. Mag. **27**(4), 14–24 (2010)
18. P. Pratt *et al.*: Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: MICCAI. pp. 77–85 (2010)
19. Shen, D., Yang, S., Low, C.H., Li, Q., Xu, M., Dou, Q., Jin, Y.: Surfsurg6d: Geometry consistent dense correspondence for textureless surgical instrument pose estimation. IEEE Robotics and Automation Letters (2026)
20. Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 275–282. Springer (2010)
21. Wang, X., Xu, J., Zhang, S., Huang, B., Stoyanov, D., Mazomenos, E.B.: Stereomamba: Real-time and robust intraoperative stereo disparity estimation via long-range spatial dependencies. IEEE Robotics and Automation Letters **10**(10), 10682–10689 (2025)
22. Wei, R., Guo, J., Lu, Y., Zhong, F., Liu, Y., Sun, D., Dou, Q.: Scale-aware monocular reconstruction via robot kinematics and visual data in neural radiance fields. Artificial Intelligence Surgery **4**(3), 187–198 (2024)

23. Wen, B., Dewan, S., Birchfield, S.: Fast-foundationstereo: Real-time zero-shot stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7513–7524 (2026)
24. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5249–5260. IEEE (2025)
25. Wu, T., Miao, Y., Guo, J., Chen, Z., Zhao, S., Li, Z., Tang, Z., Huang, B., Yu, L.: Endowave: Rational-wavelet 4d gaussian splatting for endoscopic reconstruction. arXiv preprint arXiv:2510.23087 (2025)
26. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21919–21928 (2023)
27. Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., Yang, X.: Igev++: Iterative multi-range geometry encoding volumes for stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(8), 7108–7122 (2025)
28. Xu, H., Weld, A., Xu, C., Roddan, A., Cartucho, J., Karaoglu, M.A., Ladikos, A., Li, Y., Li, Y., Shen, D., et al.: Surgripe challenge: Benchmark of surgical robot instrument pose estimation. Medical Image Analysis **105**, 103674 (2025)
29. Yang, B., Sui, C., Zhong, F., Liu, Y.H.: Modal-graph 3d shape servoing of deformable objects with raw point clouds. The International Journal of Robotics Research **42**(14), 1213–1244 (2023)