

Occlusion-aware Disparity Estimation via Depth Prior Embedded Cost Volume for Stereo Endoscopic Images

Ziteng Liu¹, Dongdong He², Chenghong Zhang², Weirong Wang³, Wenpeng Gao^{1,2}*, and Yili Fu¹**

¹ State Key Laboratory of Robotics and System, Harbin Institute of Technology, 2 Yikuang Str., Nangang District, Harbin, 150080, China

² School of Life Science and Technology, Harbin Institute of Technology, 2 Yikuang Str., Nangang District, Harbin 150080, China

³ School of Intelligent Science and Technology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

Abstract. Accurate disparity estimation from stereo laparoscopic images is essential for 3D surgical navigation but remains difficult in occlusion regions, where conventional cost volumes produce unreliable matching costs. To address this, we propose Depth Prior Embedded Disparity Estimation Network that leverages monocular depth cues—inherently less affected by occlusion—to refine disparity predictions. Our framework consists of two key components: a Coarse Disparity Estimation (CDE) module that generates an initial disparity prior from an affine-transformed depth map, and a novel Depth Prior Embedded Cost Volume (DPCV). The DPCV encodes this prior into a probability distribution over disparity levels and iteratively fuses it with the traditional correlation volume via a cross-attention mechanism, thereby improving matching robustness in occluded/unmatched regions. Evaluations on the SCARED dataset show that DEDENet achieves improved performance over several state-of-the-art methods, with clear gains observed particularly within challenging occlusion areas. Our work offers an effective approach for reliable 3D reconstruction in minimally invasive surgery. The code is made available: <https://homepage.hit.edu.cn/wpgao?lang=en>.

Keywords: Disparity estimation · Disparity refinement · Cost Volume.

1 Introduction

Laparoscopic surgery remains a clinical gold standard, yet its conventional 2D visualization limits depth perception and anatomical comprehension [4, 24]. Stereo laparoscopes address this limitation by enabling 3D reconstruction via disparity maps derived from stereo matching [20], which significantly enhances intra-

* Corresponding author: Wenpeng Gao (wpgao@hit.edu.cn)

** Corresponding author: Yili Fu (meylfu@hit.edu.cn)

operative spatial awareness [19]. Modern deep learning-based disparity estimators leverage end-to-end feature learning [10], typically relying on the construction and aggregation of cost volumes [13]. However, a fundamental challenge (Fig. 1(a)) persists in such frameworks: in stereo imaging, object occlusion and baseline separation between cameras create occluded/unmatched regions where pixel correspondences are absent [25]. This results in invalid or ambiguous entries within the cost volume at these locations (Fig. 1(b)), critically undermining both training stability and final estimation accuracy [5]. Therefore, developing methods to accurately infer disparity in these occluded/unmatched regions is crucial for advancing the robustness and performance of disparity estimation in surgical vision [23].

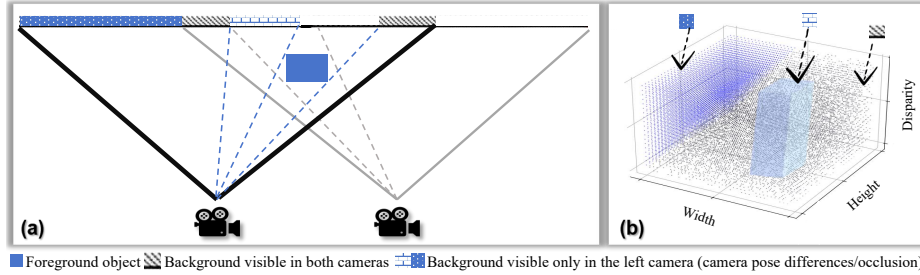


Fig. 1: Conceptual and structural impact of occlusion in stereo matching. (a) The formation of occluded/unmatched regions in a stereo pair due to object occlusion and camera pose differences. (b) 3D visualization of a corresponding cost volume.

As a fundamental component in stereo matching [15], the cost volume remains susceptible to such ambiguities [8] in occluded regions [21, 26]. To enhance feature expressiveness, advanced cost volume designs have been introduced, such as the group-wise correlation in GwcNet [6] and the attention-based concatenation volume in ACVNet [22]. Beyond correlation-based volumes, methods like IGEV-Stereo [21] employ a lightweight 3D regularization network to convert a standard cost volume into a geometry encoding volume, thereby enriching structural context. More recent approaches further explore multi-volume fusion strategies: DCVSMNet [17] couples two same-scale volumes built differently to extract complementary geometric information, while Miao et al. [12] design a dynamic cost volume using residual optical flow for moving objects, fusing it with a static volume to address occlusion and noise. In a different vein, MVSplat [3] adopts a plane-swept cost volume for multi-view feature matching to support 3D Gaussian localization, simplifying optimization by replacing regression with similarity measurement [27].

Nevertheless, a key limitation persists: the structure of the cost volume itself can adversely affect the cost aggregation process during training [2]. In occluded regions, for example, cost values may be incorrectly inferred or leak from neigh-

boring pixels [7], compromising the integrity of matching [14]. Consequently, issues such as information leakage during aggregation and the ghosting effect in the warping process [18] remain unresolved, negatively impacting the accuracy of subsequent disparity estimation. This indicates that further improvements in volume design, regularization, and fusion mechanisms are still needed.

Since occlusion is an inherent challenge in disparity estimation tasks, we propose a **Depth Prior Embedded Disparity Estimation Network (DEDENet)** to refine occluded regions in disparity maps using depth information, as monocular depth estimation is less affected by occlusion. The key contributions of our work are threefold: First, we introduce the DEDENet framework, which systematically integrates monocular depth cues into stereo matching to address occlusion ambiguities. Second, we design a novel Depth Prior Embedded Cost Volume (DPCV) that encodes coarse inverse depth priors into a probability distribution and iteratively fuses them with traditional correlation volumes through cross-attention, improving matching reliability in occluded/unmatched regions. Third, we develop complementary modules including a Coarse Disparity Estimation (CDE) module for robust initial guidance.

2 Method

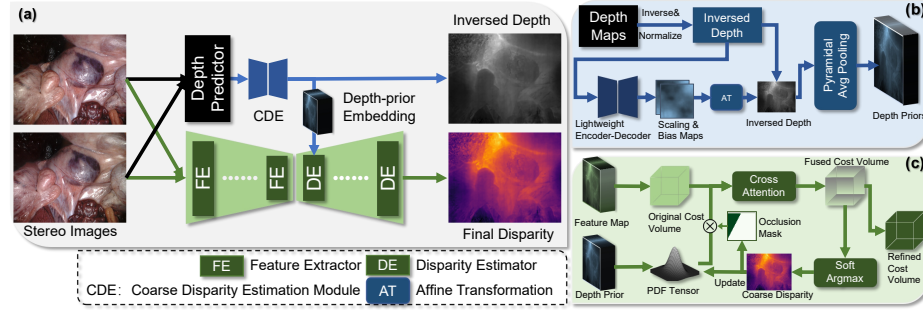


Fig. 2: Overview of the DEDENet framework. (a) Main pipeline. (b) Coarse Disparity Estimation Module. (c) Depth Prior Embedded Cost Volume in the disparity estimator.

As shown in Fig. 2(a), the proposed framework builds upon a lightweight stereo matching backbone [13], which includes pyramid feature extractors and disparity estimators. The key modifications are twofold: first, a Coarse Disparity Estimation (CDE) module (Fig. 2(b)) is integrated to provide complementary guidance for the disparity estimation; second, the standard cost volume is replaced by our proposed Depth Prior Embedded Cost Volume (DPCV) module, depicted in Fig. 2(c).

2.1 Coarse Disparity Estimation

The CDE module generates an initial disparity estimate from the stereo pair. First, a pre-trained depth network produces a depth map d , which is then normalized via inverse depth transformation and percentile clamping:

$$\hat{d}_{inv} = \frac{\min\left(\frac{1}{d+\epsilon}, Q_{0.999}\left(\frac{1}{d+\epsilon}\right)\right) - \min\left(\frac{1}{d+\epsilon}\right)}{\max\left(\frac{1}{d+\epsilon}\right) - \min\left(\frac{1}{d+\epsilon}\right) + \epsilon} \quad (1)$$

where $Q_{0.999}(\cdot)$ denotes the 99.9th percentile operator, and ϵ is a small constant for numerical stability. Subsequently, a lightweight encoder-decoder network predicts pixel-wise scaling and bias maps, which are used to refine $\hat{d}_{inv}$ via an affine transformation, yielding the final inverse depth prior $\hat{D}_{inv} \in \mathbb{R}^{B \times 1 \times H \times W}$. Finally, a pyramidal average pooling operation is applied to $\hat{D}_{inv}$ to extract multi-scale contextual features, thereby providing hierarchical depth priors for the subsequent stereo matching stages.

2.2 Depth Prior Embedded Cost Volume

To enhance the robustness of stereo matching in occluded regions, we propose a DPCV module that incorporates geometric priors into the cost volume construction process. The proposed DPCV module enhances conventional stereo matching by integrating explicit depth priors with traditional feature correlation volumes. Given left and right feature maps $\mathbf{F}_L, \mathbf{F}_R \in \mathbb{R}^{B \times C \times H \times W}$ and the inverse depth $\hat{D}_{inv}$, the module operates through three key stages: probability distribution tensor generation, cross-attention fusion, and iterative refinement.

Probability Distribution Tensor Generation We convert the coarse inverse depth prior into a probability distribution function (PDF) tensor that encodes uncertainty across the disparity range. Given an inverse depth prior, we generate a PDF tensor with shape $\mathbb{R}^{B \times D \times H \times W}$ (where D denotes the number of preset maximum disparity levels) by employing a discretized normal distribution. For each pixel location (i, j) , we sample a normal distribution centered at the coarse disparity estimate $\mu_{i,j}$ with variance σ^2 :

$$\mathbf{V}_{\text{pdf}}^{(i,j)}(d) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d - \mu_{i,j})^2}{2\sigma^2}\right) \quad (2)$$

where $d \in \{d_0, d_1, \dots, d_{D-1}\}$ represents discrete disparity levels uniformly spaced in the range $[0, D_{\max}]$, $\mu_{i,j}$ corresponds to the mean values derived from the inverse depth prior, and $\sigma = 0.2$ controls the spread of the distribution. The resulting 4D tensor $\mathbf{V}_{\text{pdf}} \in \mathbb{R}^{B \times D \times H \times W}$ provides a probabilistic matching prior, with higher values indicating more probable disparities. The PDF values are normalized such that their sum equals one across all disparity levels.

Cross-Attention Fusion The original cost volume $\mathbf{V}_{\text{og}}$ is iteratively refined with the probability distribution function (PDF) tensor $\mathbf{V}_{\text{pdf}}^{(t)}$ via a cross-attention fusion mechanism [9]. In this process, $\mathbf{V}_{\text{og}}$ serves as the Query, while $\mathbf{V}_{\text{pdf}}^{(t)}$ provides the Key and Value. These are obtained through learnable linear projections parameterized by weight matrices W_q , W_k , and W_v , respectively. The attention weights are computed as the scaled dot-product between the Query and the Key, followed by a softmax normalization. With C denotes the channel dimension, the fused volume $\mathbf{V}_{\text{fused}}^{(t)}$ is then generated by applying these attention weights to the Value and projecting the result with an output weight matrix W_o :

$$\mathbf{V}_{\text{fused}}^{(t)} = W_o \left(\text{softmax} \left(\frac{(W_q \mathbf{V}_{\text{og}})(W_k \mathbf{V}_{\text{pdf}}^{(t)})^\top}{\sqrt{C}} \right) (W_v \mathbf{V}_{\text{pdf}}^{(t)}) \right) \quad (3)$$

Iterative Refinement Process The DPCV incorporates an iterative refinement mechanism to progressively enhance depth estimation. As illustrated in Fig. 2(c) and detailed in Algorithm 1, this process sequentially updates the PDF tensor and refines the cost volume fusion. Updates are selectively applied only to occluded regions, which are identified by a binary occlusion mask $\mathbf{M}_{\text{occ}}^{(t)}$ generated via a left-right consistency (LRC) check [29]. The number of refinement iterations is set to $T = 1$ by default.

Algorithm 1 Iterative Refinement in DPCV

Require: Original cost volume $\mathbf{V}_{\text{og}}$, inverse depth prior $\hat{D}_{\text{inv}}$, maximum iterations T

Ensure: Refined cost volume $\mathbf{V}_{\text{final}}$

- 1: Initialize $\mathbf{V}_{\text{pdf}}^{(0)} \leftarrow \text{GeneratePDF}(\hat{D}_{\text{inv}})$
 - 2: Initialize $\mathbf{V}_{\text{fused}}^{(0)} \leftarrow \text{CrossAttentionFusion}(\mathbf{V}_{\text{og}}, \mathbf{V}_{\text{pdf}}^{(0)})$
 - 3: **for** $t = 1$ to T **do**
 - 4: Compute probability distribution: $\mathbf{P}^{(t-1)} \leftarrow \text{softmax}(\mathbf{V}^{(t-1)})$
 - 5: Estimate refined disparity using soft argmax: $\mathbf{D}^{(t)} \leftarrow \sum_{d=0}^{D-1} d \cdot \mathbf{P}^{(t-1)}(\cdot, d, \cdot, \cdot)$
 - 6: Update PDF tensor: $\mathbf{V}_{\text{pdf}}^{(t)} \leftarrow \text{GeneratePDF}(\mathbf{D}^{(t)})$
 - 7: Update $\mathbf{M}_{\text{occ}}^{(t)} \leftarrow \text{LRC}(\mathbf{D}^{(t)})$
 - 8: Fuse with cross-attention: $\mathbf{V}_{\text{fused}}^{(t)} \leftarrow \text{CrossAttentionFusion}(\mathbf{V}_{\text{og}}, \mathbf{V}_{\text{pdf}}^{(t)})$
 - 9: $\mathbf{V}^{(t)} \leftarrow \mathbf{V}^{(t-1)} + \mathbf{M}_{\text{occ}}^{(t)} \odot \mathbf{V}_{\text{fused}}^{(t)}$
 - 10: **end for**
 - 11: $\mathbf{V}_{\text{final}} \leftarrow \mathbf{V}^{(T)}$
 - 12: **return** $\mathbf{V}_{\text{final}}$
-

3 Experiments

3.1 Datasets and Metrics

We pre-train our model on the clean pass version of the FlyingThings3D dataset [11], which provides 21,818 synthetically generated stereo samples at a resolution of 960×540 . For domain adaptation and evaluation on surgical data, we use the SCARED dataset [1]. Following established protocols [4, 13], datasets 1, 2, 3, 6, and 7 are used for fine-tuning, while datasets 8 and 9 are reserved for evaluation (datasets 4 and 5 are excluded due to known calibration errors). To address temporal misalignment between the RGB video and interpolated ground truth, we construct two evaluation subsets: SCARED-evaluation-full (SEF) uses all video frames, and SCARED-evaluation-keyframe (SEK) uses only synchronized keyframes. The fine-tuning set contains 25 samples, SEF contains 5,915 samples, and SEK contains 10 samples.

For stereo matching, the evaluation metrics include the average end-point error (EPE), defined as the mean absolute error between the reference and predicted disparity values, and the percentage of pixels with a disparity error greater than 3 pixels (Bad3). To evaluate depth prediction performance, we employ the widely used root mean squared error (RMSE).

3.2 Loss Function

In this study, the loss function is defined as follows:

$$L_{total} = L_{gt}(\hat{D}, D_{gt}) + L_{CDE}(\hat{D}_{inv}, D_{gt}) \quad (4)$$

Here, L_{gt} is an L1 loss applied between the final refined disparity estimate $\hat{D}$ and the ground-truth disparity map D_{gt} . The term L_{CDE} is also an L1 loss, computed between the coarse inverse depth $\hat{D}_{inv}$ (estimated by the CDE module) and the ground truth D_{gt} .

3.3 Implementation Details

Our network is implemented in PyTorch and trained on two NVIDIA A10 GPUs. Following common practice [16], we utilize LeReS (ResNet50 version) [28] as the monocular depth predictor. The AdamW optimizer is employed with a weight decay of 1×10^{-4} . We initialize the stereo matching backbone with pre-trained weights from [13]. The model is first pre-trained on the FlyingThings3D dataset for 20 epochs with a batch size of 12, using a cosine annealing learning rate scheduler starting at 1×10^{-4} . Subsequently, it is fine-tuned on the SCARED dataset for 400 epochs with a batch size of 4 and a constant learning rate of 1×10^{-4} .

4 Results and Discussion

4.1 Quantitative and Qualitative Results

We evaluate the performance of our method by comparing it with six state-of-the-art (SOTA) disparity estimation approaches: ACVNet [22], HSM [25], DeepPruner [1], GMStereo [23], and MSDESIS [13]. The accuracy of the monocular depth predictor LeRes [28] is also reported for comparison.

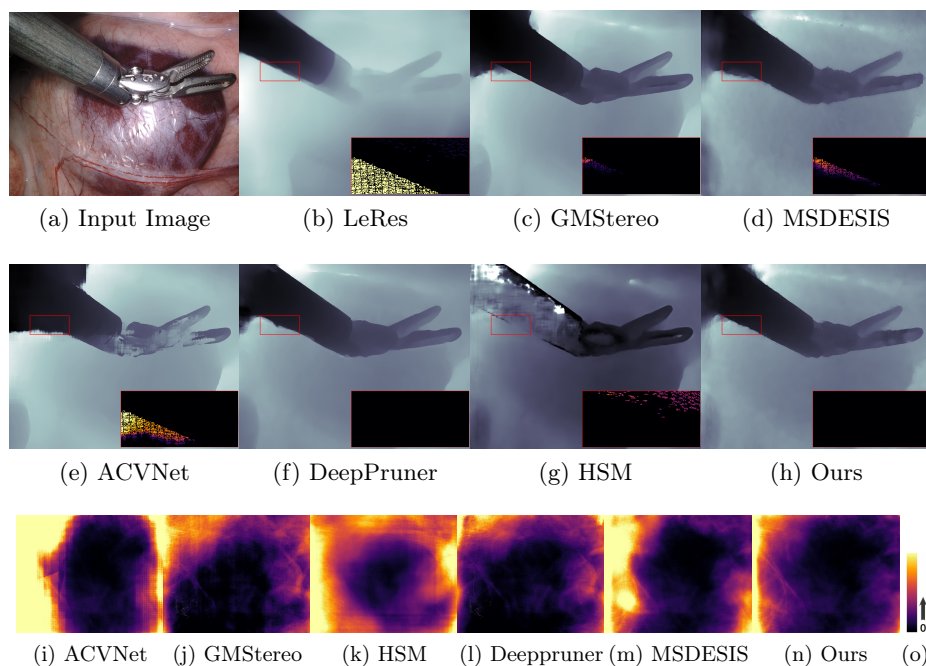


Fig. 3: Predictions from SOTA methods and the proposed approach. Each prediction includes an embedded zoomed-in view of the corresponding mean squared error (MSE) map in the bottom-right corner. (i)-(n) show the MSE distribution maps aggregated across all samples in the SEF dataset. (o) MSE color bar.

The quantitative comparison on the SEF and SEK datasets is presented in Table 1. Our method achieves competitive performance across all metrics, notably surpassing several recent approaches in key areas. On the SEK dataset, our model attains the best RMSE (3.42) and Bad3 (10.45%) among all compared methods, while also achieving a best EPE of 1.74. These results demonstrate the accuracy of our disparity estimates, particularly in complex surgical scenes.

Qualitative comparisons are presented in Fig. 3(b)–(h). Compared to the backbone method (Fig. 3(d)) and other state-of-the-art methods, our approach

Table 1: Quantitative evaluation results on the SEF and SEK datasets. The runtime (milliseconds) corresponds to the average inference time for 1280×1024 resolution inputs.

Method	SEF			SEK			Runtime
	EPE	Bad3	RMSE	EPE	Bad3	RMSE	
LeRes	N/A	N/A	9.33	N/A	N/A	7.31	19.58
ACVNet	6.28	48.35	15.87	2.59	15.71	3.72	53.23
HSM	6.05	61.85	6.85	4.92	57.00	6.14	40.71
DeepPruner	4.08	46.45	<u>5.06</u>	<u>1.89</u>	<u>11.67</u>	3.69	168.36
GMStereo	4.04	43.95	5.03	1.95	13.93	3.33	183.65
MSDESIS	4.45	46.04	5.40	2.30	13.50	3.99	17.01
Ours	<u>4.05</u>	<u>44.68</u>	5.16	1.74	10.45	<u>3.42</u>	45.23

Table 2: Ablation study on SEK.

Method	EPE	Bad3	RMSE
Proposed	1.74	<u>10.45</u>	3.42
w/o DPCV	2.30	13.50	3.99
w/o $\mathbf{M}_{occ}$	2.48	13.70	4.26
w/o L_{CDE}	1.78	10.86	3.52
T=0	1.88	10.97	3.63
T=3	<u>1.77</u>	10.30	<u>3.43</u>

visibly reduces errors in occluded regions. The mean squared error (MSE) distribution maps in Fig. 3(i)–(n) further demonstrate that our method consistently yields lower errors, particularly in comparison to the backbone (Fig. 3(m)). This improvement is evident not only on the left side—where stereo occlusions arise from viewpoint differences—but also across the image as a whole. These visual results confirm the effectiveness of the proposed modules, indicating that embedding reliable geometric priors into the cost volume enhances overall disparity estimation accuracy.

In terms of efficiency, the proposed DEDENet module runs at 25.65 ms for 1280×1024 images, and the total inference time is 45.23 ms (19.58 ms from LeRes). This is still faster than computationally expensive models such as DeepPruner (168.36 ms) and GMStereo (183.65 ms). Although the lightweight MSDESIS model achieves a slightly lower runtime, our method achieves better accuracy across all metrics on both datasets. This indicates that our design offers an excellent trade-off between computational efficiency and estimation precision.

4.2 Ablation Study

- **Depth prior embedding cost volume.** As shown in Table 2, removing the DPCV module results in a performance drop across all metrics. Specifically, EPE increases by 32.2% (from 1.74 to 2.30) and Bad3 increases by 29.2% (from 10.45% to 13.50%), which validates the module’s critical role in providing geometric guidance. Furthermore, applying the depth prior globally without $\mathbf{M}_{occ}$ leads to even worse performance, with EPE increasing by 42.5% (to 2.48), demonstrating that selectively enhancing only unreliable regions is more effective than global application.
- **Coarse Disparity Estimation Loss.** Training without L_{CDE} (Eq. 4) results in a consistent degradation across all metrics, though relatively minor. Bad3 increases by 3.9%, confirming that this auxiliary supervision provides valuable regularization during training, leading to more stable convergence and a better initialization for the subsequent refinement stage.
- **Iterative refinement process.** Disabling the iterative process (T=0) causes a noticeable performance decline, with EPE increasing by 8.0% and Bad3

increasing by 5.0%, underscoring the necessity of the iterative update mechanism for robust estimation. Increasing iterations to $T=3$ yields marginal gains in Bad3 (10.30%, a -1.4% change relative to our method), indicating that a single refinement cycle ($T=1$) offers a favorable trade-off between accuracy and computational efficiency.

5 Conclusion

This paper introduces DEDENet, a framework that addresses stereo occlusion in endoscopic 3D reconstruction by integrating monocular depth priors. Our core contribution is the DPCV, which encodes depth estimates as a probability distribution and fuses them with correlation features via cross-attention. Experiments show SOTA accuracy and real-time efficiency, demonstrating that the proposed method can enhance the robustness of stereo matching in occluded/unmatched regions. Future work will extend the proposed method to other endoscopic imaging scenarios and explore temporal fusion for dynamic scene reconstruction.

Acknowledgments. This work was supported by National Key R&D Program of China (Grant No.2023YFB4705700), National Natural Science Foundation of China (92048202), Zhejiang Province Key Research and Development Program: development of core components and system of surgical robot (NO. 2023C03010), Self-Planned Task (NO. SKLRS202010B, SKLRS202209B), China Postdoctoral Science Foundation (2025M784391), State Key Laboratory of Robotics and Systems (HIT, SKLRS-2026-KF-08), the "Open Bidding for selecting the best candidates" for Artificial Intelligence Medical Device Innovation Tasks of the Ministry of Industry and Information Technology (No.2025-AIFZZL-02005).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., McLeod, A.J., Wang, C., Rosenthal, J., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogyios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S.: Stereo correspondence and reconstruction of endoscopic data challenge. CoRR **abs/2101.01133** (2021)
2. Chao, W., Duan, F., Wang, X., Wang, Y., Lu, K., Wang, G.: Occcasnet: Occlusion-aware cascade cost volume for light field depth estimation. IEEE Transactions on Computational Imaging **10**, 1680–1691 (2024)
3. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 370–386. Springer Nature Switzerland, Cham (2025)

4. Cheng, X., Zhong, Y., Harandi, M., Drummond, T., Wang, Z., Ge, Z.: Deep laparoscopic stereo matching with transformers. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. pp. 464–474. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_44
5. Dong, L., Han, Y., Hu, M., Luo, H., Wang, Y.: Stereo matching method based on high-precision occlusion-recovering and discontinuity-preserving. *Displays* **80**, 102573 (2023)
6. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3268–3277 (2019). <https://doi.org/10.1109/CVPR.2019.00339>
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 15979–15988 (2022)
8. Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(2), 504–511 (2013)
9. Lin, H., Cheng, X., Wu, X., Shen, D.: Cat: Cross attention in vision transformer. In: *2022 IEEE International Conference on Multimedia and Expo (ICME)*. pp. 1–6 (2022). <https://doi.org/10.1109/ICME52920.2022.9859720>
10. Luo, H., Wang, C., Duan, X., Liu, H., Wang, P., Hu, Q., Jia, F.: Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images. *Computers in Biology and Medicine* **140**, 105109 (2022)
11. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4040–4048 (2016). <https://doi.org/10.1109/CVPR.2016.438>
12. Miao, X., Bai, Y., Duan, H., Huang, Y., Wan, F., Xu, X., Long, Y., Zheng, Y.: Ds-depth: Dynamic and static depth estimation via a fusion cost volume. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(4), 2564–2576 (2024)
13. Psychogios, D., Mazomenos, E., Vasconcelos, F., Stoyanov, D.: Msdesis: Multitask stereo disparity estimation and surgical instrument segmentation. *IEEE transactions on medical imaging* **41**(11), 3218–3230 (2022)
14. Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1599–1610 (2023)
15. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8934–8943 (2018)
16. Sun, L., Bian, J.W., Zhan, H., Yin, W., Reid, I., Shen, C.: Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(1), 497–508 (2024)
17. Tahmasebi, M., Huq, S., Meehan, K., McAfee, M.: Dcvsmnet: Double cost volume stereo matching network. *Neurocomputing* **618**, 129002 (2025)
18. Wang, W., Ye, S., Wang, X., Zhao, Y.: Omnet: Real-time stereo matching with unsupervised occlusion mask. In: *2022 IEEE International Conference on Image Processing (ICIP)*. pp. 1241–1245 (2022)
19. Wei, R., Li, B., Mo, H., Lu, B., Long, Y., Yang, B., Dou, Q., Liu, Y., Sun, D.: Stereo dense scene reconstruction and accurate localization for learning-based navigation

- of laparoscope in minimally invasive surgery. *IEEE Transactions on Biomedical Engineering* **70**(2), 488–500 (2023). <https://doi.org/10.1109/TBME.2022.3195027>
20. Xia, W., Chen, E.C.S., Pautler, S., Peters, T.M.: A robust edge-preserving stereo matching method for laparoscopic images. *IEEE Transactions on Medical Imaging* **41**(7), 1651–1664 (2022). <https://doi.org/10.1109/TMI.2022.3147414>
 21. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21919–21928 (2023)
 22. Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X.: Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
 23. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(11), 13941–13958 (2023)
 24. Yang, B., Xu, S., Chen, H., Zheng, W., Liu, C.: Reconstruct dynamic soft-tissue with stereo endoscope based on a single-layer network. *IEEE Transactions on Image Processing* **31**, 5828–5840 (2022). <https://doi.org/10.1109/TIP.2022.3202367>
 25. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5515–5524 (2019)
 26. Yang, X., Feng, Z., Zhao, Y., Zhang, G., He, L.: Edge supervision and multi-scale cost volume for stereo matching. *Image and Vision Computing* **117**, 104336 (2022)
 27. Yang, Y., Cao, J., Zhao, H., Wang, W.: Generalizable 3d gaussian splatting via multi-view stereo and consistency constraints. *Neurocomputing* **658**, 131696 (2025). <https://doi.org/https://doi.org/10.1016/j.neucom.2025.131696>
 28. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 204–213 (2021)
 29. Zitnick, C., Kanade, T.: A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7), 675–684 (2000). <https://doi.org/10.1109/34.865184>