



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

ESRT: Endoscopic Stereo Reconstruction Transformer

Jiun-Feng Chang¹, Po-Wei Tsai¹, Chih-Tsung Shen², and Chun-Rong Huang^{1*}

¹ Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

² Industrial Technology Research Institute, Hsinchu, Taiwan
crhuang@cs.nycu.edu.tw

Abstract. Accurate endoscopic stereo reconstruction is essential for disease diagnosis and 3D visualizations in surgery. To achieve endoscopic stereo reconstruction from endoscopic stereo images under specular reflections, illumination variations, and smooth texture conditions, we propose the endoscopic stereo reconstruction transformer (ESRT). ESRT captures fine-grained textures by introducing a row-wise cross-attention mechanism that explicitly constrains correspondence search along the horizontal direction and prevents attention from being computed over geometrically invalid regions. Furthermore, we present a distillation-based training strategy that transfers geometric knowledge from a pretrained model without requiring endoscopic 3D annotations. The proposed method achieves the best performance on the SCARED dataset while maintaining real-time efficiency.

Keywords: Depth and Pose Estimation · Endoscopy · Stereo Reconstruction.

1 Introduction

Endoscopy is the clinical gold standard for the diagnosis of diseases and is essential to detect life-threatening conditions such as cancer and precancerous lesions [2, 8, 11]. Accurate assessment of lesion morphology is critical for early diagnosis and treatment planning. However, conventional endoscopy provides only 2D images without explicit depth information, making it difficult for physicians to reliably estimate the height, depth, and true physical size of the lesion [7, 9, 16].

Due to the lack of labeled training data, recent endoscopic 3D reconstruction methods predominantly rely on self-supervised monocular frameworks, where scene geometry and camera motion are jointly optimized using photometric consistency between consecutive frames [13, 19, 21]. Early approaches such as Fang et al. [4] and Monodepth2 [5] introduced self-supervised depth estimation from monocular sequences. Endo-SfM [12] and AF-SfMLearner [14] extended structure from motion (SfM)-based learning to endoscopic reconstruction. To further improve 3D reconstruction performance, DARES [20], EndoDAC [3], and

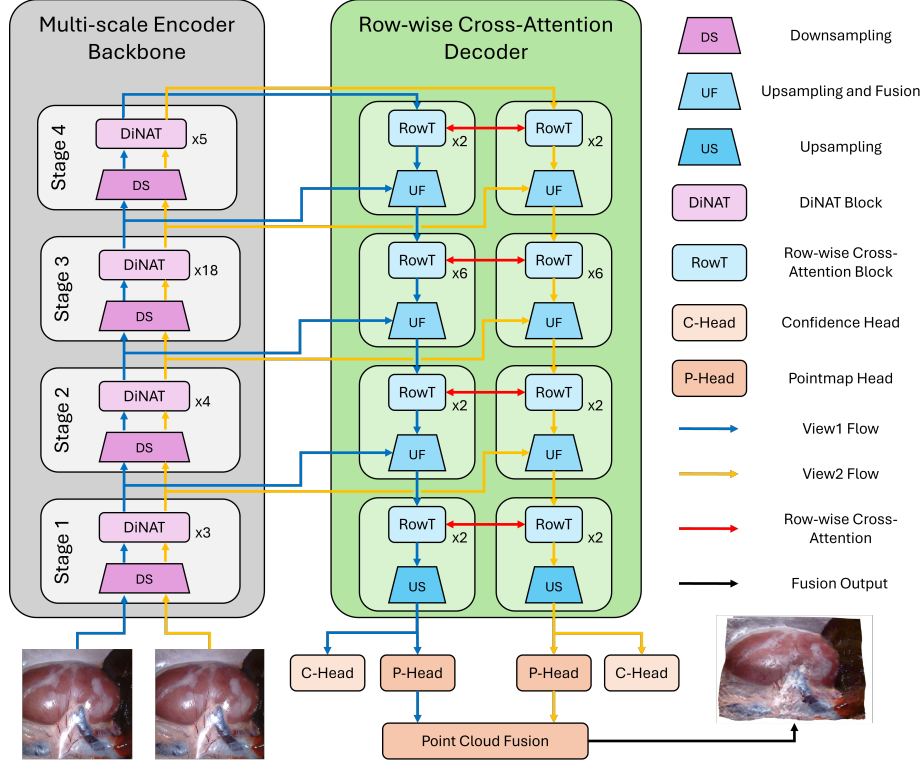


Fig. 1. The architecture of the proposed transformer, which consists of a multi-scale encoder backbone to capture multi-scale features for surface representation, a row-wise cross-attention decoder to capture corresponding regions from stereo views, and an aggregation module to reconstruct 3D point clouds and camera poses.

Endo-FAST3r [15] used pretrained foundation models for depth estimation. In EndoSfM3D [21], intrinsic parameters are jointly estimated with depths and camera motions to achieve more accurate depth estimation results with foundation models. In endoscopic scenes, specular reflections, illumination variations, and smooth textures of organs lead to unstable depth estimation for SfM-based methods. Moreover, these methods remain sensitive to correspondence errors and inaccuracies in camera parameter estimation. These limitations also lead to unstable geometry in reconstructing endoscopic scenes.

To overcome these limitations, transformer-based 3D reconstruction methods [17, 18] perform dense cross-view matching and joint correspondence reasoning. They achieve strong performance on general image datasets. Despite these advances, applying transformer-based 3D reconstruction methods to endoscopic images remains challenging. In particular, endoscopic scenes often exhibit limited texture and strong specular reflections, which degrade correspondence estimation. Moreover, stereo endoscopes produce disparity mainly along

the horizontal direction, where unconstrained correspondence search introduces geometric ambiguity. Furthermore, transformer-based methods typically require 3D ground-truth supervision, which is difficult to obtain for in vivo endoscopic data. Finally, transformer-based reconstruction frameworks also incur high computational costs, limiting their use in real-time clinical settings.

To address these challenges, we propose the Endoscopic Stereo Reconstruction Transformer (ESRT), a multi-scale encoder–decoder framework specifically designed for endoscopic stereo reconstruction as shown in Fig. 1. ESRT captures fine-grained textures through hierarchical feature extraction while improving robustness to specular reflections. To explicitly model the horizontal disparity characteristic of stereo endoscopy, we introduce a row-wise cross-attention mechanism that constrains the correspondence search to the horizontal direction and prevents attention from being computed over geometrically invalid regions. Furthermore, we propose a distillation-based training strategy that effectively transfers geometric knowledge from a pretrained model without endoscopic 3D annotations. By integrating the efficient multi-scale encoder–decoder framework, ESRT significantly reduces computational cost while maintaining stable and accurate reconstruction, enabling real-time deployment in clinical environments. Our contributions can be summarized as follows:

1. ESRT is proposed to learn robust representations for endoscopic stereo reconstruction under visually similar texture and specular reflections.
2. A row-wise cross-attention mechanism is introduced to explicitly enforce horizontal epipolar constraints, reducing geometric ambiguity, and preventing correspondence search over invalid regions.
3. A distillation-based training strategy is developed to transfer geometric priors from a teacher network, reducing the fine-tuning efforts for the downstream task.
4. The proposed ESRT achieves efficient and stable reconstruction with substantially reduced computational cost, enabling real-time clinical applicability compared with transformer-based methods.

2 Method

2.1 Overview

As illustrated in Fig. 1, the proposed transformer consists of three components: a multi-scale encoder backbone [6] to capture multi-scale features for surface representation, a row-wise cross-attention decoder to capture corresponding regions from stereo views, and an aggregation module to reconstruct 3D point clouds and camera poses. Given a pair of stereo endoscopic images $\mathbf{I}_1$ and $\mathbf{I}_2 \in \mathbb{R}^{H \times W \times 3}$, the two views are first processed by the shared-weight multi-scale encoder backbone to extract multi-scale features. For each view, the backbone produces four-stage hierarchical feature representations, which provide multi-scale contextual cues for stereo correspondence estimation. To establish geometrically consistent matching, we design a row-wise cross-attention decoder. The decoder contains

two symmetric branches, each enhancing features of one view. At each stage, row-wise cross-attention is performed between features of identical resolution across views to accelerate correspondence search along the horizontal epipolar line. A hierarchical multi-scale fusion strategy with upsampling is further applied to progressively recover high-resolution spatial details for the following decoder stages. Finally, each branch is connected to an aggregation module composed of a confidence head and a pointmap head, predicting a confidence map and a dense 3D point cloud in the coordinate system of the first view.

2.2 Multi-scale Encoder Backbone

To extract hierarchical representations for stereo correspondence learning, we adopt DiNAT [6] as the shared encoder backbone. The backbone consists of four stages, progressively downsampling the feature maps to spatial resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input resolution. Stage 1, Stage 2, Stage 3, and Stage 4 contain 3, 4, 18, and 5 DiNAT blocks, respectively. Given $\mathbf{I}_1$ and $\mathbf{I}_2$, the backbone extracts multi-scale features for each view, denoted as $\{\mathbf{f}_s^{(v)}\}_{s=1}^4$, where $v \in \{1, 2\}$ indicates the index of the view and s is the index of the stage. The hierarchical multi-scale design enhances feature robustness under smooth texture and specular reflections, providing stable representations that facilitate reliable correspondence learning in the subsequent decoder.

2.3 Row-wise Cross-Attention Decoder

The row-wise cross-attention decoder progressively establishes dense stereo correspondences from larger field-of-views to smaller field-of-views with hierarchical multi-scale fusion of encoder features at different scales. By imposing these encoder features on different scales, the decoder can better represent tissue structures during correspondence matching. The decoder adopts a symmetric dual-branch structure, where each branch refines features for one view. Given encoder features $\{\mathbf{f}_s^{(v)}\}_{s=1}^4$, the decoder operates from larger field-of-views (Stage 4) to smaller field-of-views (Stage 1).

Row-wise Cross-Attention. Let $\mathbf{d}_s^{(v)} \in \mathbb{R}^{H_s \times W_s \times C_s}$ denote the decoder feature at stage s for view v . Because of static stereo settings, for each row i in view 1, we compute row-wise cross-attention with features of the same row i in view 2. Specifically, the query $\mathbf{q}_{i,j}^{(1)}$ of column j for view 1, and key $\mathbf{k}_{i,k}^{(2)}$ and value $\mathbf{v}_{i,k}^{(2)}$ of column k for view 2 are obtained via learnable linear projections as follows:

$$\mathbf{q}_{i,j}^{(1)} = \mathbf{W}_Q \mathbf{d}_s^{(1)}(i, j), \quad \mathbf{k}_{i,k}^{(2)} = \mathbf{W}_K \mathbf{d}_s^{(2)}(i, k), \quad \mathbf{v}_{i,k}^{(2)} = \mathbf{W}_V \mathbf{d}_s^{(2)}(i, k),$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C_s \times C_s}$ are learnable projection matrices. The row-wise cross attention $\alpha_{i,j,k}$ between column j and column k of $\mathbf{d}_s^{(1)}$ and $\mathbf{d}_s^{(2)}$ with respect to row i is defined as follows:

$$\alpha_{i,j,k} = \text{Softmax} \left(\frac{\mathbf{q}_{i,j}^{(1)} \cdot \mathbf{k}_{i,k}^{(2)\top}}{\sqrt{C_s}} \right) \mathbf{v}_{i,k}^{(2)},$$

where the $\text{Softmax}(\cdot)$ is the softmax function. The refined feature $\tilde{\mathbf{d}}_s^{(1)}(i, j)$ of $\tilde{\mathbf{d}}_s^{(1)}$ for row i and column j is obtained as follows:

$$\tilde{\mathbf{d}}_s^{(1)}(i, j) = \sum_{k=1}^{W_s} \alpha_{i,j,k}.$$

Thus, the decoder only handles features within the same image row of the opposite view. The same operation is symmetrically applied to view 2 to obtain $\tilde{\mathbf{d}}_s^{(2)}$ for two-sided matching, which can improve correspondence matching performance on tissues with smooth texture and specular reflections.

Hierarchical Multi-scale Fusion. After row-wise cross-attention, features are progressively upsampled and fused with encoder features via feed-forward connections as follows:

$$\mathbf{d}_{s-1}^{(v)} = \text{UF}(\tilde{\mathbf{d}}_s^{(v)}, \mathbf{f}_{s-1}^{(v)}), \quad v \in \{1, 2\}.$$

where $\text{UF}(\cdot)$ is the upsampling and fusion function. At the finest scale, the upsampling function restores the feature maps to the original spatial resolution. With row-wise cross-attention and hierarchical multi-scale fusion, the decoder refines correspondence-aware representations from larger field-of-views to smaller field-of-views, providing geometry-consistent features for subsequent 3D point cloud aggregation.

2.4 Aggregation Module

The aggregation module transforms the final decoder features into 3D point clouds. For each view v , the finest-scale decoder feature $\mathbf{d}_1^{(v)} \in \mathbb{R}^{H \times W \times C_1}$ is processed by two lightweight heads: a pointmap head (P-Head) and a confidence head (C-Head). The P-Head predicts a dense 3D point map $\mathbf{P}^{(v)} \in \mathbb{R}^{H \times W \times 3}$ in the coordinate system of view 1, while the C-Head outputs a confidence map $\mathbf{C}^{(v)} \in \mathbb{R}^{H \times W}$ indicating the reliability of each prediction. These predictions constitute the final outputs of the network and are subsequently incorporated into the distillation-based training strategy. Finally, point cloud fusion produces 3D results based on the textures of $\mathbf{I}_1$ and $\mathbf{I}_2$.

2.5 Distillation-Based Training Strategy

Due to the absence of ground-truth 3D annotations in real endoscopic data, we adopt a distillation-based training strategy. As illustrated in Fig. 2, a pretrained model on large-scale general stereo datasets is employed as the teacher model, while the proposed ESRT serves as the student model. During training, $\mathbf{I}_1$ and $\mathbf{I}_2$ serve as the input for both the teacher model and the student model. For each view v , the teacher predicts a dense 3D point map $\mathbf{P}_T^{(v)} \in \mathbb{R}^{H \times W \times 3}$ and a confidence map $\mathbf{C}_T^{(v)} \in \mathbb{R}^{H \times W}$. The student model also predicts its own dense 3D point maps $\mathbf{P}_S^{(v)}$ and confidence maps $\mathbf{C}_S^{(v)}$. During training, the student

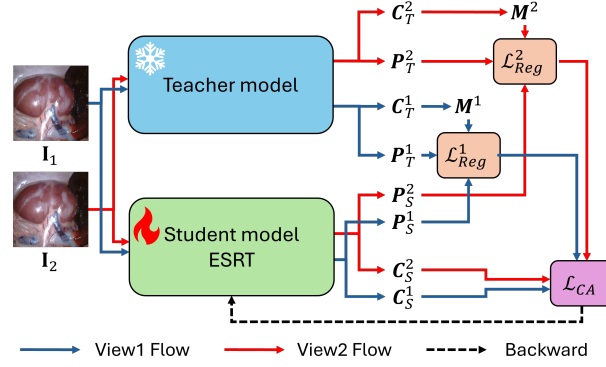


Fig. 2. The distillation-based training strategy with the teacher and student models.

model is optimized to match the geometric predictions of the teacher model, which have higher confidence values. The teacher model remains fixed during training and is not required at inference time. To update the student model, the regression loss $\mathcal{L}_{\text{Reg}}^{(v)}(\mathbf{x})$ for each pixel $\mathbf{x}$ in view v and confidence-aware loss $\mathcal{L}_{\text{CA}}$ for optimizing errors of both views are proposed.

Regression Loss. $\mathcal{L}_{\text{Reg}}^{(v)}(\mathbf{x})$ is defined as follows:

$$\mathcal{L}_{\text{Reg}}^{(v)}(\mathbf{x}) = \left\| \mathbf{P}_S^{(v)}(\mathbf{x}) - \mathbf{P}_T^{(v)}(\mathbf{x}) \right\|_2, \quad \mathbf{x} \in \mathbf{M}^{(v)},$$

where $\mathbf{P}_S^{(v)}(\mathbf{x})$ and $\mathbf{P}_T^{(v)}(\mathbf{x})$ are the 3D predictions of pixel $\mathbf{x}$ of the student model and teacher model with respect to view v , and $\mathbf{M}^{(v)}$ is a binary valid mask to select pixels whose teacher model's confidence exceeds the spatial mean $\overline{\mathbf{C}}_T^{(v)}$ of $\mathbf{C}_T^{(v)}$ as follows:

$$\mathbf{M}^{(v)}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{C}_T^{(v)}(\mathbf{x}) > \overline{\mathbf{C}}_T^{(v)} \\ 0, & \text{otherwise} \end{cases}.$$

Confidence-Aware Loss. $\mathcal{L}_{\text{CA}}$ incorporates the student-predicted confidence to weight the regression loss and is defined as follows:

$$\mathcal{L}_{\text{CA}} = \sum_{v=1}^2 \frac{1}{|\mathbf{M}^{(v)}|} \sum_{\mathbf{x} \in \mathbf{M}^{(v)}} \left(\mathbf{C}_S^{(v)}(\mathbf{x}) \mathcal{L}_{\text{reg}}^{(v)}(\mathbf{x}) - \log \mathbf{C}_S^{(v)}(\mathbf{x}) \right),$$

where $\mathbf{C}_S^{(v)}(\mathbf{x}) > 0$, and $|\mathbf{M}^{(v)}|$ is the number of valid pixels in $\mathbf{M}^{(v)}$. This formulation encourages the student model to learn geometrically accurate predictions with higher confidence, while preventing trivial solutions. Based on the training strategy, our model can effectively learn geometric knowledge from a pretrained teacher network without endoscopic 3D annotations.

Table 1. Quantitative comparisons on the SCARED dataset.

Method	Venue	AbsRel↓	SqRel↓	Rmse↓	RmseLog↓
Fang et al. [4]	WACV’20	0.078	0.794	6.794	0.109
Monodepth2 [5]	ICCV’19	0.069	0.577	5.546	0.094
Endo-SfM [12]	MIA’21	0.062	0.606	5.726	0.093
AF-SfMLearner [14]	MIA’22	0.059	0.435	4.925	0.082
DARES [20]	ECCVW’24	0.052	0.356	4.483	0.073
EndoDAC [3]	MICCAI’24	0.052	0.362	4.464	0.073
Endo-FAST3r [15]	MICCAI’25	0.051	0.354	4.480	-
EndoSfM3D [21]	MICCAIW’25	0.050	0.389	4.749	0.070
DUST3R [18]	CVPR’24	0.059	0.516	4.947	0.080
ESRT-Full (Ours)	-	0.053	0.454	4.791	0.076
ESRT-Row (Ours)	-	0.048	0.372	4.242	0.069

3 Experimental Results

Training. We trained and evaluated our method on the SCARED dataset [1], which contains 35 endoscopic videos with point clouds, depth maps, camera intrinsics, and pose estimation results, all acquired from fresh porcine cadaver abdominal anatomy. We followed the training and testing settings in [14, 15, 21] for fair comparisons. The rectified stereo pairs from the SCARED dataset were used for training and evaluation, satisfying the horizontal epipolar constraint adopted by the proposed row-wise cross-attention. During distillation-based training, DUST3R [18], pretrained on the natural 3D stereo datasets, is used as the teacher model. The student model is optimized by using AdamW [10] with a peak learning rate of 5×10^{-4} and a weight decay of 1×10^{-6} . The training is performed on a computer with two 4090 GPUs for 10 epochs with a batch size 12. For efficiency evaluation, the inference speed is measured on a single NVIDIA RTX 4090 GPU with a batch size of 1 and an input resolution of 320×256 .

Depth Estimation. We compared our method with several state-of-the-art depth estimation approaches [3–5, 12, 14, 15, 18, 20, 21] designed for endoscopic reconstruction, where all competing results are reported from their original publications. Among previous methods, EndoSfM3D [21] achieves good performance by jointly optimizing depth, camera motions, and intrinsic parameters, thereby improving geometric consistency across frames. Although DUST3R [18], relying on dense correspondence estimation, demonstrates competitive performance, its reconstruction accuracy remains affected by smooth texture and strong specular reflections in endoscopic images.

Our ESRT with full attention (ESRT-Full) employs a full cross-attention mechanism in the decoder to aggregate stereo features. Compared with DUST3R, ESRT-Full benefits from a multi-scale feature aggregation mechanism, allowing the model to capture both global structure and fine geometric details and leading to improved robustness and better depth estimation accuracy. By further introducing row-wise cross-attention, our method (ESRT-Row) can better exploit stereo constraints by restricting attention to epipolar-consistent directions,

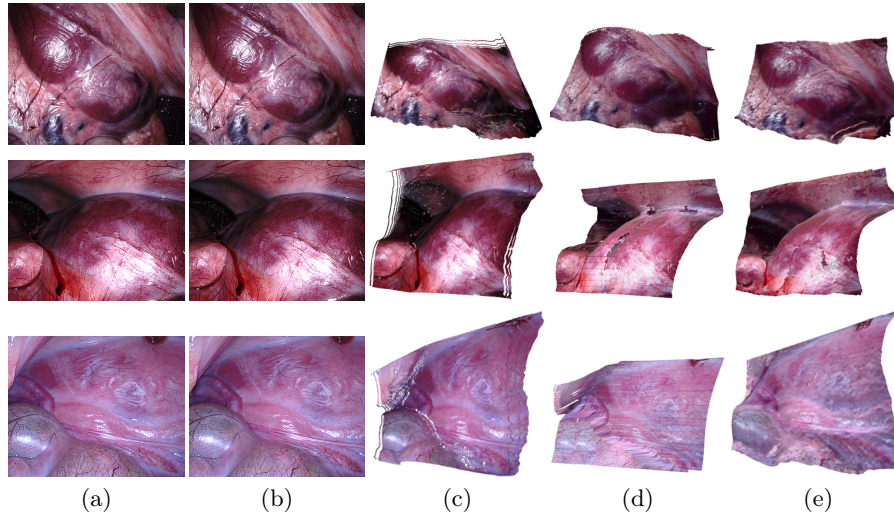


Fig. 3. Visualizations of 3D reconstruction. (a) Left image, (b) Right image, (c) Results of EndoSfM3D [21], (d) Results of DUST3R [18], and (e) Results of the proposed ESRT.

enabling more reliable feature matching and more effective stereo correspondence learning. As a result, ESRT-Row achieves the best performance among all compared methods for endoscopic stereo reconstruction. In addition, compared with the number of parameters 532.34M in DUST3R [18], the number of parameters of ESRT is only 154.45M, and the GFLOPs of ESRT are around 20% of those of DUST3R [18]. The inference FPS of ESRT is around 44. Such results show the efficiency of our method compared with SOTA transformer-based methods.

3D Visualization. Figs. 3(a) and (b) show the left and right stereo endoscopic images, which contain smooth textures and specular reflections. Figs. 3(c), (d), and (e) show 3D reconstruction results for the second-best SfM-based method EndoSfM3D [21], the transformer-based method DUST3R [18], and the proposed ESRT, respectively. The SfM-based method is unable to provide dense point clouds, resulting in fragmented scenes. Although DUST3R [18] provides dense point clouds, it suffers from geometric discontinuities due to smooth textures and specular reflections in the endoscopic environment. In contrast, our method successfully reconstructs organs with correct shapes and dense appearances, effectively addressing the challenge in the medical scenario.

4 Conclusions

In this paper, we propose a novel endoscopic stereo reconstruction transformer that uses a row-wise cross-attention mechanism and a distillation-based training strategy. Our method can effectively and efficiently achieve endoscopic stereo

reconstruction compared with state-of-the-art methods. In the future, we plan to apply the method for gastric cancer and precancerous lesion diagnosis.

Acknowledgments. This research was supported in part by the Industrial Technology Research Institute, Taiwan, under the ITRI Advanced Research Program Grant R301AR3630, and in part by the National Science and Technology Council (NSTC), Taiwan, under Grants NSTC 114-2221-E-A49-034-MY3, NSTC 114-2634-F-006-002, and NSTC 114-2640-E-006-010.

Disclosure of Interests. The authors have no competing interests to declare in the paper.

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S.: Stereo correspondence and reconstruction of endoscopic data challenge. ArXiv (2021)
2. Chen, Y.T., Yang, E.H., Chang, W.L., Lin, J., Cheng, H.C., Huang, C.R.: Mask focal modulation network for gastric intestinal metaplasia segmentation. In: Proc. Int’l Joint Conf. Neural Networks. pp. 1–8 (2024)
3. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: EndoDAC: Efficient Adapting Foundation Model for Self-Supervised Depth Estimation from Any Endoscopic Camera . In: Proc. Int’l Conf. Medical Image Computing and Computer-Assisted Intervention. pp. 208–218 (2024)
4. Fang, Z., Chen, X., Chen, Y., Gool, L.V.: Towards good practice for cnn-based monocular depth estimation. In: Proc. IEEE Winter Conf. Applications of Computer Vision. pp. 1091–1100 (2020)
5. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proc. Int’l Conf. Computer Vision. pp. 3828–3838 (2019)
6. Hassani, A., Shi, H.: Dilated neighborhood attention transformer. ArXiv (2022)
7. Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., Sznitman, R.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery **18**, 1185–1192 (2023)
8. Jhang, J.Y., Tsai, Y.C., Hsu, T.C., Huang, C.R., Cheng, H.C., Sheu, B.S.: Gastric section correlation network for gastric precancerous lesion diagnosis. IEEE Open Journal of Engineering in Medicine and Biology **5**, 434–442 (2024)
9. Li, B., Liu, B., Zhu, M., Luo, X., Zhou, F.: Image intrinsic-based unsupervised monocular depth estimation in endoscopy. IEEE Journal of Biomedical and Health Informatics **29**(12), 9084–9094 (2025)
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. Int’l Conf. Learning Representations. pp. 1–18 (2019)

11. Nien, C.M., Yang, E.H., Chang, W.L., Cheng, H.C., Huang, C.R.: Criss-cross attention based multi-level fusion network for gastric intestinal metaplasia segmentation. In: Proc. MICCAI Workshop on Imaging Systems for GI Endoscopy. pp. 13–23 (2022)
12. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., Sahin, H., Araujo, H., Alexandrino, H., Durr, N.J., Gilbert, H.B., Turan, M.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis* **71**, 102058 (2021)
13. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J.M.M., Civera, J.: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters* **6**(4), 7225–7232 (2021)
14. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical Image Analysis* **77**, 102338 (2022)
15. Sheikh Zeinoddin, M., Hoque, M.I., Tandogdu, Z., Shaw, G.L., Clarkson, M.J., Mazomenos, E.B., Stoyanov, D.: Endo-FASt3r: Endoscopic Foundation model Adaptation for Structure from motion . In: Proc. Int’l Conf. Medical Image Computing and Computer-Assisted Intervention. pp. 117–126 (2025)
16. Stoyanov, D.: Surgical vision. *Annals of biomedical engineering* **40**(2), 332–345 (2012)
17. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 5294–5306 (2025)
18. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
19. Yang, Z., Pan, J., Dai, J., Sun, Z., Xiao, Y.: Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. *IEEE Trans. on Medical Imaging* **43**(5), 1934–1944 (2024)
20. Zeinoddin, M.S., Lena, C., Qu, J., Carlini, L., Magro, M., Kim, S., Momi, E.D., Bano, S., Grech-Sollars, M., Mazomenos, E., Alexander, D.C., Stoyanov, D., Clarkson, M.J., Islam, M.: Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model. In: ECCV Workshops. pp. 1–11 (2024)
21. Zhang, C., Clarkson, M.J., Hoque, M.I.: Endosfm3d: Learning to 3d reconstruct any endoscopic surgery scene using self-supervised foundation model. In: Proc. MICCAI Workshop on Efficient Medical Artificial Intelligence. pp. 299–309 (2025)