EndoDAC: Efficient Adapting Foundation Model for Self-Supervised Depth Estimation from Any Endoscopic Camera

Beilei Cui[*,1,4][0009−0009−7900−8032], Mobarakol Islam[*,2][0000−0002−7162−2822], Long Bai[*,1,4][0000−0002−9762−6821], An Wang[1,4][0000−0001−5515−0653], and Hongliang Ren[†,1,3,4][0000−0002−6488−1551]

[1] Dept. of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Dept. of Medical Physics and Biomedical Engineering, University College London, UK
[3] Dept. of Biomedical Engineering, National University of Singapore, Singapore
[4] Shenzhen Research Institute, CUHK, Shenzhen, China
beileicui@link.cuhk.edu.hk, mobarakol.islam@ucl.ac.uk,
b.long@link.cuhk.edu.hk, wa09@link.cuhk.edu.hk, hlren@ee.cuhk.edu.hk

Abstract. Depth estimation plays a crucial role in various tasks within endoscopic surgery, including navigation, surface reconstruction, and augmented reality visualization. Despite the significant achievements of foundation models in vision tasks, including depth estimation, their direct application to the medical domain often results in suboptimal performance. This highlights the need for efficient adaptation methods to adapt these models to endoscopic depth estimation. We propose Endoscopic Depth Any Camera (EndoDAC) which is an efficient self-supervised depth estimation framework that adapts foundation models to endoscopic scenes. Specifically, we develop the Dynamic Vector-Based Low-Rank Adaptation (DV-LoRA) and employ Convolutional Neck blocks to tailor the foundational model to the surgical domain, utilizing remarkably few trainable parameters. Given that camera information is not always accessible, we also introduce a self-supervised adaptation strategy that estimates camera intrinsics using the pose encoder. Our framework is capable of being trained solely on monocular surgical videos from any camera, ensuring minimal training costs. Experiments demonstrate that our approach obtains superior performance even with fewer training epochs and unaware of the ground truth camera intrinsics. Code is available at https://github.com/BeileiCui/EndoDAC.

Keywords: Foundation models · Monocular depth estimation · Self-supervised learning.

---

[*] Authors contributed equally to this work.
[†] Corresponding Author.

1   Introduction

Depth estimation holds immense value in minimally invasive endoscopic surgery, enabling enhanced navigation, accurate surface reconstruction, and immersive augmented reality experiences [5, 32]. However, due to complex internal environments, low lighting conditions, and sparse feature textures, accurate surgical depth estimation remains a challenging task [22]. Traditional multi-view geometry-based methods such as structure from motion (SfM) [20] and simultaneous localization and mapping (SLAM) [10] perform poorly in surgical scenarios with low lighting and lack of texture. Deep learning method has been widely proposed for depth estimation in natural environments [2, 24]. Due to security, privacy, and professionalism issues, obtaining large-scale, precise surgical ground truth depth information for supervised training is challenging. Researchers therefore have been focusing on the self-supervised learning (SSL) method where the depths are constrained by the geometric relationship between video frames [1,17]. Shao *et al.* [22] utilize appearance flow to resolve inconsistent lighting problems in endoscopic depth estimation. Yang *et al.* [29] designed a lightweight framework combining CNN and Transformers to compress the model parameters effectively.

Recently, foundation models have attracted extremely increasing attention for their amazing performances in various tasks [13,14,16]. Leveraging the large amount of parameters and training data with integrated training methods, foundation models can learn highly generalizable information achieving state-of-the-art performance in multiple downstream tasks involving vision, text, and multi-modal inputs [6]. However, foundation models may experience significant performance degradation when applied to specific domains such as endoscopic scenes [25]. Training a medical-specific foundational model from scratch presents numerous challenges due to the scarcity of annotated data in the medical domain and the inadequate availability of computational resources. Hence, there has been considerable discourse on the adaptation of existing foundational models to various sub-domains, optimizing the utilization of pre-trained model parameters, and fine-tuning foundation models for specific application scenarios [4, 27].

The majority of the current adaptation of foundation models to the medical domain focuses on medical image segmentation and detection instead of regression tasks like depth estimation and annotated prompts are still required for the fine-tuning process [31]. To this end, we make the effort to explore efficiently adapting foundation models to surgical self-supervised depth estimation where only surgical videos are required for fine-tuning. To be specific, we design Dynamic Vector-Based Low-Rank Adaptation (DV-LoRA) which requires only a small number of parameters to be fine-tuned specifically for medical scenarios. We also design a Convolution Neck block to enhance the model's ability to capture high-frequency information thus facilitating more accurate depth estimation. Meanwhile, to improve the universality of adaptation based on the fact that endoscopic camera intrinsic parameters are not always given, we add a decode head to estimate the camera's intrinsic parameters simultaneously leading our method can be applied with only surgical videos from any unknown cameras.

Our key contributions can be summarized as follows:
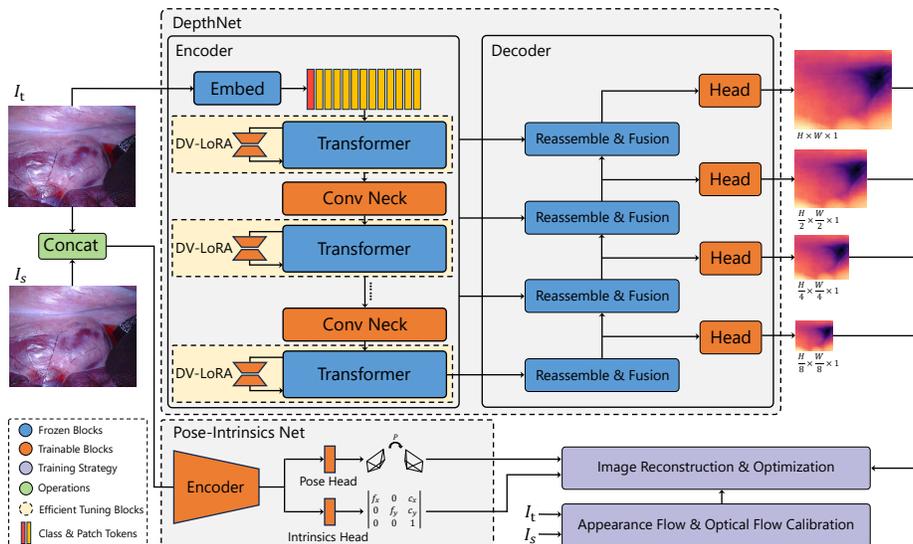
Fig. 1. Illustration of the proposed Endoscopic Depth Any Camera (EndoDAC) SSL depth estimation framework. ViT-based encoder and DPT-liked decoder pre-trained from Depth Anything [28] are employed for DepthNet. We utilize a small amount of trainable parameters (1.6M) including Dynamic Vector-Based LoRA (DV-LoRA), Convolutional Neck blocks and Multi-Scale Decoders to fine-tune the model. In Pose-Intrinsics Net, ego-motion and camera intrinsic parameters are predicted with the same encoder and separate decoders.

– We design DV-LoRA and a Convolution Neck block to efficiently adapt foundation models to surgical scene depth estimation with an exceptionally small amount of trainable parameters resulting in low computational resources and short training time.
– We present a self-supervised adaptation strategy where the depth, ego-motion, and camera's intrinsic parameters estimations are trained in parallel. Our method can be adapted to only surgical videos from any unknown camera which is broadly applicable to most surgical video datasets.
– Extensive experiments on two publicly available datasets have demonstrated the superior performance of our proposed method over other state-of-the-art SSL depth estimation methods with significantly fewer trainable parameters. It also reveals the important prospects of our proposed model in the endoscopic domain.

2 Method

2.1 Preliminaries

Foundation Models for Depth Foundation Models generally refer to powerful pre-trained models trained on extensive amounts of data which enable them

to exhibit strong generalization capabilities across multiple tasks and scenarios. Dense Prediction Transformer (DPT) [19] is a depth estimation foundation model based on Vision Transformer (ViTs). DINOv2 [16] is a semantic foundation model suitable for many vision tasks including depth estimation with separate decode decoders. In this work, we aim to adapt Depth Anything (DA) [28], which is a depth estimation foundation model trained on large-scale labeled and unlabeled data, to endoscopic scenes.

Low-Rank Adaptation (LoRA) [12] LoRA was proposed to fine-tune large-scale foundation models to downstream tasks. It was motivated by the observation that the pre-trained large model's learning ability remains unaffected even when randomly projected onto a smaller subspace. LoRA achieves a significant reduction in the number of trainable parameters for downstream tasks by incorporating trainable rank decomposition matrices into each layer of the Transformer architecture while keeping the pre-trained model weights frozen. To be specific, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA modifies the update to:

$$h = W_0 x + \Delta W x = W_0 x + BA x. \tag{1}$$

where $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$ with the rank $r \ll min(d, k)$; W_0 is frozen during training and only A and B receive gradient updates.

2.2   Proposed Framework: EndoDAC

As illustrated in Fig. 1, The architecture of our proposed Endoscopic Depth Any Camera (EndoDAC) framework aims to adapt the depth estimation foundation model - Depth Anything [28] - to the endoscopic domain in a self-supervised manner with minimal training cost. The framework mainly contains two sections: DepthNet and Pose-Intrinsics Net. The DepthNet estimates the multi-scale depth map of a single endoscopic image, while Pose-Intrinsics Net estimates the motion variation between adjacent images and the camera's intrinsic parameters. The DepthNet consists of a ViT-based encoder and a DPT-liked decoder with pre-trained weights. We implement the trainable DV-LoRA layers and Convolutional Neck blocks with the frozen transformer blocks to efficiently fine-tune the model. The Pose-Intrinsics Net estimates the camera ego-motion and intrinsic parameters with the same encoder but separate decoders. The estimated depth is reprojected back to the 2-D plane with the ego-motion information to generate the reconstructed image. The model can therefore be optimized by minimizing the loss between the reconstructed image and the target image. Finally, our EndoDAC only requires endoscopic videos for training and can be applied to any surgical videos without giving the camera intrinsic information.

DepthNet Different from fine-tuning the whole model, EndoDAC freezes the model and adds trainable DV-LoRA layers, Convolutional Neck blocks, and Multi-Scale Decoders, which largely reduces the required memory and computation resources for training and also benefits from convenient deployment.
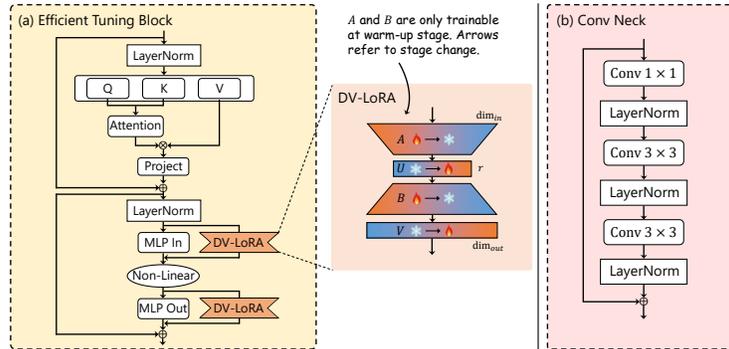
Fig. 2. Illustration of (a) Transformer Efficient Tuning Block with DV-LoRA and (b) Convolution Neck block. In DV-LoRA, we use the gradient color and arrows to represent the dynamic variation between training and frozen states.

A Vision Transformer (ViT), which DA is based on, is used as the backbone of the encoder. Different from conventional LoRA design, we innovatively introduce Dynamic Vector-Based Low-Rank Adaptation (DV-LoRA) to fine-tune the model more efficiently. Fig. 2(a) presents our fine-tuning architecture where DV-LoRA is only applied in two MLP layers for more comprehensive adaptation. In contrast to LoRA, our DV-LoRA is expressed as:

$$x_{out} = \hat{W}x_{in} = Wx_{in} + \Lambda_v B \Lambda_u A x_{in}, \tag{2}$$

where x_{in}, x_{out} are inputs and outputs of MLP layers; W is the frozen projection layer; A and B are trainable LoRA layers; Λ_v and Λ_u are trainable vectors U and V in diagonal matrices form. At the beginning of training, trainable vectors U and V are frozen while only LoRA layers A and B are trainable. After a warm-up phase, the state of DV-LoRA changes dynamically where A and B are frozen and U and V become trainable. We only train LoRA layers with a good initialization and utilize trainable vectors to fine-tune the proposed model with fewer parameters.

Furthermore, as demonstrated by [18], ViTs tend to weaken high-frequency signals, which could have negative effects on depth estimation. Therefore, inspired by [30], we employ Convolution Neck blocks to enhance our method. We incorporate a Convolutional Neck block after each $3^{rd}, 6^{th}, 9^{th}$ and 12th transformer efficient tuning block. The details are presented in Fig. 2(b) where three convolutional layers with LayerNorm and a residual connection are utilized to feed forward to the results of transformer blocks.

We utilize the DPT-liked decoder to estimate the depth maps. Different from the previous approach where only one depth decoder is applied at the highest resolution, we propose to exploit Multi-Scale Decoders at different fusion levels to generate multi-scale depth maps. Therefore, the feature representations are progressively reassembled and fused at various resolutions. The reassemble and fusion blocks are frozen and only depth decoders are trainable.

Pose-Intrinsics Net Simultaneous learning of depth and motion depends on the relationship between adjacent frames, related depth map and camera matrix:

$$z^{'}p^{'} = KRK^{-1}zp + Kt, \tag{3}$$

where K refers to the intrinsic matrix given by $K = \begin{vmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{vmatrix}$, p and $p^{'}$ are pixel coordinates before and after the transformation of rotation matrix R and translation vector t; z and $z^{'}$ are corresponding depths. Previous work [9] has demonstrated that given Equation. (3), no $\tilde{K}$ and $\tilde{R}$ exists such that $\tilde{K}\tilde{R}K^{-1} = KRK^{-1}$ leading the estimation of K, R and t converges simultaneously. Therefore, we utilize a shared ResNet [11] encoder which takes two adjacent color images as input and outputs the 6-DoF relative pose and camera intrinsics with two separate decoders.

Self-supervised Depth and Ego-motion Estimation With depth z, camera intrinsics K, rotation matrix R and translation vector t predicted by the networks and given source image I_s, the reconstructed Image $I_{s \to t}$ is obtained with the re-projection(π) defined by Equation. (3) described as:

$$I_{s \to t} = \pi\left(z, K, R, t, I_s\right). \tag{4}$$

We utilize an appearance flow network and an optical flow network proposed in [22] to compensate for the inconsistent lighting problem. A photometric loss combing $\mathcal{L}_1$ loss and structural similarities (SSIM) [26] is used to assess the image difference defined by:

$$\mathcal{L}_p = \alpha\frac{1 - \text{SSIM}\left(I_t, I_{s \to t}\right)}{2} + (1 - \alpha)\left|I_t - I_{s \to t}\right|. \tag{5}$$

An edge-aware loss [8] is also used to maintain the edges defined by:

$$\mathcal{L}_e = \left|\partial_x d\right| e^{-\left|\partial_x \mathbf{I}\right|} + \left|\partial_y d\right| e^{-\left|\partial_y \mathbf{I}\right|}, \tag{6}$$

where d represents the mean-normalized inverse depth of I.

3   Experiments and Results

SCARED Dataset. SCARED was first proposed for a challenge in MICCAI 2019 containing 35 endoscopic videos with 22950 frames of fresh porcine cadaver abdominal anatomy collected with a da Vinci Xi endoscope. Each video is accompanied by ground truth depth maps collected by a projector and ground truth poses and camera intrinsic. We followed the split scheme in [22] where the SCARED dataset is split into 15351, 1705, and 551 frames for the training, validation and test sets, respectively.

Hamlyn Dataset. Hamlyn is a laparoscopic and endoscopic video dataset taken from various surgical procedures with challenging in vivo scenes. We followed the selection in [21] with 21 videos for validation.

Implementation Details. The framework is implemented with PyTorch on NVIDIA RTX 3090 GPU. We utilize two AdamW [15] optimizers separately for DepthNet and Pose-Intrinsics Net with initial learning rates of 1×10^{-4}. The rank for DV-LoRA is set to 4, warm up step is set to 5000 and $\alpha = 0.85$ for $\mathcal{L}_p$. We propose the same training augmentations followed [22, 29] and batch size is set to 8 with 20 epochs in total.

Evaluation Settings. Following [6, 22, 28, 29], we compute the 5 standard metrics: Abs Rel, Sq Rel, RMSE, RMSE log and δ for evaluation. We re-scale the predicted depth map by a median scaling method [6, 22, 33] during evaluation. We also perform a 5-frame pose evaluation following [22] and adopt the metric of absolute trajectory error (ATE).

3.1 Results

Table 1. Quantitative depth comparison on SCARED dataset of SOTA self-supervised learning depth estimation methods. The best results are in bold and the second-best results are underlined. "G.I." refers to given camera intrinsic parameters. "Total." and "Train." refer to the total and trainable parameters utilized in DepthNet.

	Method	Year	G.I.	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	Total.(M)	Train.(M)	Speed (ms)
SCARED	Fang et al. [7]	2020	✓	0.078	0.794	6.794	0.109	0.946	136.8	136.8	-
	Defeat-Net [23]	2020	✓	0.077	0.792	6.688	0.108	0.941	14.8	14.8	-
	SC-SfMLearner [3]	2019	✓	0.068	0.645	5.988	0.097	0.957	14.8	14.8	-
	Monodepth2 [8]	2019	✓	0.069	0.577	5.546	0.094	0.948	14.8	14.8	-
	Endo-SfM [17]	2021	✓	0.062	0.606	5.726	0.093	0.957	14.8	14.8	-
	AF-SfMLearner [22]	2022	✓	0.059	0.435	4.925	0.082	0.974	14.8	14.8	8.0
	Yang et al. [29]	2024	✓	0.062	0.558	5.585	0.090	0.962	2.0	2.0	-
	DA (zero-shot) [28]	2024	✓	0.084	0.847	6.711	0.110	0.930	97.5	-	13.8
	DA (fine-tuned) [28]	2024	✓	0.058	0.451	5.058	0.081	0.974	97.5	11.2	13.8
	EndoDAC (Ours)	-	✓	**0.051**	**0.341**	**4.347**	**0.072**	**0.981**	99.0	1.6	17.7
	EndoDAC (Ours)	-	✗	<u>0.052</u>	<u>0.362</u>	<u>4.464</u>	<u>0.073</u>	<u>0.979</u>	99.0	1.6	17.7
Hamlyn	Endo Depth & Motion [21]	2021	✓	0.185	5.424	16.100	0.225	0.732	-	-	-
	AF-SfMLearner [22]	2022	✓	0.168	4.440	13.870	0.204	0.770	14.8	14.8	7.7
	DA (fine-tuned) [28]	2024	✓	0.170	4.413	13.920	0.205	0.765	97.5	-	12.5
	EndoDAC (Ours)	-	✓	**0.138**	**2.796**	**11.491**	**0.171**	**0.813**	99.0	1.6	15.7
	EndoDAC (Ours)	-	✗	<u>0.156</u>	<u>3.854</u>	<u>12.936</u>	<u>0.193</u>	<u>0.791</u>	99.0	1.6	15.7

Depth Estimation. The proposed method is compared with several SOTA self-supervised methods [3, 7, 8, 17, 21–23, 29], Depth Anything [28] model with and
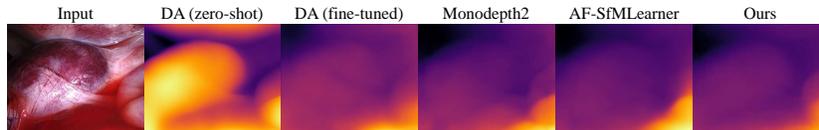
Fig. 3. Qualitative depth comparison on the SCARED dataset.

Table 2. Ablation study on the modules of EndoDAC with estimated intrinsics. Specifically, we (i) use the original LoRA [12] to replace DV-LoRA; (ii) remove the Convolution Neck blocks; (iii) use a single depth decoder to replace the multi-scale decoders.

DV-LoRA	Conv Neck	M.S. Decoders	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑
✓	✗	✗	0.054	0.397	4.718	0.076	0.978
✗	✓	✗	0.053	0.408	4.768	0.076	0.977
✗	✗	✓	0.051	0.379	4.621	0.074	0.977
✓	✓	✗	0.052	0.380	4.598	0.075	0.977
✓	✗	✓	0.052	0.378	4.614	0.075	0.978
✗	✓	✓	0.053	0.366	4.516	0.074	0.978
✓	✓	✓	0.052	0.362	4.464	0.073	0.979

without full parameter fine-tuning on two datasets. Note that we zero-shot evaluated on Hamlyn with the models trained on SCARED. All the other compared methods involving training have a total epoch greater or equal to 40 while our approach only trains for 20 epochs. Table 1 presents the comparison of quantitative results of the aforementioned methods. Our method exceeds all of the compared methods by a significant margin regardless of the awareness of camera intrinsics. It is also notable that only 1.6 million parameters are trainable for our model which accounts for 1.6% of the total parameters. Foundation models generally utilize large models leading our framework slower in inference speed compared to other methods, but a speed of 17.7 ms is still capable of real-time implementation which makes our framework applicable for a variety of real-time surgical applications. Fig. 3 shows several qualitative results visualization. We can observe that our method generates a more accurate geometry relation within the depth map while also preserving the global smoothness of tissues. More visualization of depth and 3D reconstruction are presented in Supplementary.

Ablation Studies. To further demonstrate the validity of our proposed model, we conduct ablation studies on the different modules of EndoDAC. As presented in Table 2, the ablation studies demonstrate the effectiveness of each module.

Pose and Intrinsics Estimation. Two sequences are selected followed [22,28] for evaluation of pose and intrinsic estimation. Pose estimation is evaluated on two sequences separately while intrinsic estimation is evaluated with a weighted average percentage error on two sequences. The results are presented in Table 3 and Table 4. Table 3 shows our proposed method obtains satisfactory performances on pose estimation with or without given intrinsic parameters. Our proposed method can also estimate accurate camera intrinsic parameters with a maximum percentage error of 4.02%. Visualization of pose estimation trajectories of two sequences are presented in Supplementary.

4   Conclusion

In this paper, we propose an efficient framework to adapt foundation models to SSL endoscopic depth estimation. Different from previous methods, we introduce DV-LoRA and Convolutional Neck to fine-tune depth foundation models

Table 3. Quantitative pose estimation comparison on two selected sequences. "G.I." refers to given camera intrinsic parameters.

Table 4. Quantitative intrinsic estimation results.

Method	G.I.	ATE ↓ (Seq.1)	ATE ↓ (Seq.2)
Monodepth2 [8]	✓	0.0769	0.0554
Endo-SfM [17]	✓	0.0759	0.0500
AF-SfMLearner [22]	✓	0.0742	0.0478
EndoDAC (Ours)	✓	0.0741	0.0512
EndoDAC (Ours)	✗	0.0762	0.0487

Intrinsics	Ground Truth	Percentage Error ↓(%)
fx	0.82	3.17
fy	1.02	4.02
cx	0.5	0.70
cy	0.5	1.70

to the endoscopic domain with limited trainable parameters and training costs. We also utilize a separate decoder to estimate camera intrinsic and ego-motion simultaneously. Our method is widely applicable to surgical datasets where only surgical videos are required for training. Extensive experiments demonstrate the superior performance of our proposed method even without knowledge of the camera and intrinsic parameters information.

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

References

1. Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N.: Monocular depth estimation: A thorough review. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
2. Bhat, S.F., Alhashim, I., Wonka, P.: Localbins: Improving depth estimation by learning local distributions. In: European Conference on Computer Vision. pp. 480–496. Springer (2022)
3. Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems **32** (2019)
4. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)
5. Collins, T., Pizarro, D., Gasparini, S., Bourdel, N., Chauvet, P., Canis, M., Calvet, L., Bartoli, A.: Augmented reality guided laparoscopic surgery of the uterus. IEEE Transactions on Medical Imaging **40**(1), 371–380 (2020)
6. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-dino: Adapter learning of foundation model for depth estimation in endoscopic surgery. arXiv preprint arXiv:2401.06013 (2024)
7. Fang, Z., Chen, X., Chen, Y., Gool, L.V.: Towards good practice for cnn-based monocular depth estimation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 1091–1100 (2020)

8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)

9. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8977–8986 (2019)

10. Grasa, O.G., Bernal, E., Casado, S., Gil, I., Montiel, J.: Visual slam for hand-held monocular endoscope. IEEE transactions on medical imaging **33**(1), 135–146 (2013)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

12. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)

13. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Ren, H.: Endo-4dgs: Distilling depth ranking for endoscopic monocular scene reconstruction with 4d gaussian splatting. arXiv preprint arXiv:2401.16416 (2024)

14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)

16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

17. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical image analysis **71**, 102058 (2021)

18. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2021)

19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)

20. Rattanalappaiboon, S., Bhongmakapat, T., Ritthipravat, P.: Fuzzy zoning for feature matching technique in 3d reconstruction of nasal endoscopic images. Computers in Biology and Medicine **67**, 83–94 (2015)

21. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., 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)

22. 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)

23. Spencer, J., Bowden, R., Hadfield, S.: Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14402–14413 (2020)

24. 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 (2023)
25. Wang, A., Islam, M., Xu, M., Zhang, Y., Ren, H.: Sam meets robotic surgery: An empirical study in robustness perspective. arXiv preprint arXiv:2304.14674 (2023)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
27. Wu, Q., Zhang, Y., Elbatel, M.: Self-prompting large vision models for few-shot medical image segmentation. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. pp. 156–167. Springer (2023)
28. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
29. Yang, Z., Pan, J., Dai, J., Sun, Z., Xiao, Y.: Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. IEEE Transactions on Medical Imaging (2024)
30. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: Boosting image matting with pre-trained plain vision transformers. Information Fusion **103**, 102091 (2024)
31. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
32. Zhang, P., Luo, H., Zhu, W., Yang, J., Zeng, N., Fan, Y., Wen, S., Xiang, N., Jia, F., Fang, C.: Real-time navigation for laparoscopic hepatectomy using image fusion of preoperative 3d surgical plan and intraoperative indocyanine green fluorescence imaging. Surgical endoscopy **34**, 3449–3459 (2020)
33. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)