EndoDAV: Depth Any Video in Endoscopy with Spatiotemporal Accuracy

Zanwei Zhou[1 [0000−0003−2222−4016]], Chen Yang[1 [0000−0003−4496−7849]], Piao Yang[2], Xiaokang Yang[1 [0000−0003−4029−3322]], and Wei Shen[†,1 [0000−0002−1235−598X]]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, Shanghai Jiao Tong University

[2] Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine

SJTU19zzw@sjtu.edu.cn, ycyangchen@sjtu.edu.cn, piaoyangyp@163.com, xkyang@sjtu.edu.cn, wei.shen@sjtu.edu.cn

Abstract. Video depth estimation has been applied to various endoscopy tasks, such as reconstruction, navigation, and surgery. Recently, many methods focus on directly applying or adapting depth estimation foundation models to endoscopy scenes. However, these methods do not consider temporal information, leading to an inconsistent prediction. We propose Endoscopic Depth Any Video (EndoDAV) to estimate spatially accurate and temporally consistent endoscopic video depth, which significantly expands the usability of depth estimation in downstream tasks. Specifically, we parameter-efficiently finetune a video depth estimation foundation model to endoscopy scenes, utilizing a self-supervised depth estimation framework which simultaneously learns depth and camera pose. Considering the distinct characteristics of endoscopic videos compared to common videos, we further design a novel loss function and a depth alignment inference strategy to enhance the temporal consistency. Experiments on two public endoscopy datasets demonstrate that our method presents superior performance in both spatial accuracy and temporal consistency. Code is available at https://github.com/Zanue/EndoDAV.

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

1 Introduction

In the realm of minimally invasive procedures, endoscopy serves as a cornerstone for diagnosis, treatment, and monitoring of gastrointestinal disorders. However, the 2D nature of endoscopic video limits the ability to perceive spatial relationships and depth, which are essential for precise navigation, lesion characterization, and therapeutic interventions. Therefore, accurate depth estimation in

† Corresponding Author.

endoscopic videos has emerged as a critical area of research. Recent advancements [2] [5] [12] [14] [16] [17] in computer vision and deep learning have enabled the development of depth estimation techniques tailored to endoscopic imaging, offering the potential to enhance clinical decision-making and procedural outcomes. By reconstructing 3D spatial information from endoscopic videos, these methods facilitate improved lesion size measurement, polyp detection, and tissue deformation analysis. Furthermore, depth estimation plays a pivotal role in robotic-assisted endoscopy, where accurate spatial awareness is crucial for autonomous navigation and instrument manipulation.

Recently, foundation models in depth estimation have garnered significant attention. Among these, Depth anything [19] [20], as a popular foundation model, represents a novel paradigm that leverages large-scale datasets and advanced neural network architectures to achieve robust and generalizable depth prediction from single images. Some methods [2] [11] [15] focus on directly applying or adapting this foundation models to endoscopy scenes. The work most closely related to ours is EndoDAC [2]. It adapts a self-supervised pipeline to efficiently finetune Depth Anything Model to endoscopic images. However, this method lacks an effective usage of temporal information, thus being unusable when attempting to acquire a consistent video depth prediction.

To address this issue, we parameter-efficiently finetune Video Depth Anything [1] to endoscopy scenes. Video Depth Anything is based on Depth Anything v2 [20] model and additionally trains a lightweight spatiotemporal head to extend its temporal awareness capability. We apply a self-supervised framework like EndoDAC, which simultaneously learns the depth and camera pose from the input video. In endoscopic videos, the soft tissues usually present slow motion and deformation. Many videos are captured by a camera navigating through the intestinal tract. Distinct from the common videos typically with a single fixed subject, endoscopic videos usually pose a greater challenge to the consistency of video depth estimation. Considering the distinct characteristics of endoscopic videos compared to common videos, we propose a projection loss and a depth alignment inference strategy to enhance the temporal consistency. The projection loss utilizes the predicted camera pose to project the predicted depth map from a source view to its adjacent target view, thus enhancing the depth consistency of two frames. At the inference stage, we carefully select previous key frames and current frames to feed into the depth estimation model, and align adjacent depth batches using the scale and shift over overlapped depth maps. Experiments on two public endoscopy datasets demonstrate that our method presents superior performance in both spatial accuracy and temporal consistency.

Our contributions are summarized as following:

1. We propose to estimate endoscopic video depth by parameter-efficiently fine-tuning a powerful video depth estimation foundation model with a self-supervised framework.
2. we propose a projection loss and a depth aligned inference strategy according to the distinct characteristics of endoscopic videos to further enhance the temporal consistency.

3. Extensive experiments on two publicly available datasets demonstrate the spatial accuracy and temporal consistency of our methods.

2 Method

2.1 Preliminaries

Single Image Depth Estimation Single Image Depth Estimation aims to predict relative or absolute depth values from one image. Recently, deep learning based methods have revolutionized this area by effectively learning from large scale depth datasets. Among these methods, Depth anything [19] [20] serves as a strong foundation model and has been wildly adapted to various downstream tasks. It utilizes a Dense Prediction Transformer(DPT) [10] structure and applies a powerful pre-trained vision model DINOv2 [8] as the backbone encoder. With an effective teacher-student framework and large scale synthetic and real data, Depth anything demonstrates astonishing generalization capability across various types of images. Some methods [2] [15] have utilized Depth anything to help single image depth estimation in endoscopy scenes. However, due to their inherent flaws in model design and training strategies, methods based on Depth anything cannot produce a consistent prediction on video depth estimation.
Video Depth Estimation Video Depth Estimation needs not only the accuracy in each single image depth, but also the temporal consistency during the whole video sequence. Recent video depth estimation foundation methods mainly utilize the priors from diffusion model [4] [11] [18] or train from a single image depth estimation foundation model [1] [6]. Among these methods, Video Depth Anything is based on Depth Anything v2 model and additionally trains a lightweight spatiotemporal head to extend its temporal awareness capability. With a proper temporal information aware training strategy and an effective long video inference strategy, Video Depth Anything achieves a satisfying video depth estimation performance. Due to its superior performance in general images, we adapt this powerful model to endoscopy scenes and meticulously develop a framework to efficiently finetune it.

2.2 EndoDAV

Framework We parameter-efficiently finetune Video Depth Anything [1] to endoscopy scenes by a self-supervised framework [2] [12]. As shown in Fig. 1, the framework consists of two parts: Video Depth Network and Pose Network. The Video Depth Network uses the structure of Video Depth Anything. To maximize the retention of Video Depth Anything model's capability and reduce the training parameters as much as possible, we select an efficient Low-Rank Adaptation(LoRA) method named 'Scaling the Subspace of Both left and right singular vector(SSB)' Lora [13] and add it to the feedforwad layers in attention blocks of the depth model. This Parameter-Efficient Finetuning(PEFT) strategy enables the finetuning process to be fast while minimizing resource consumption as much

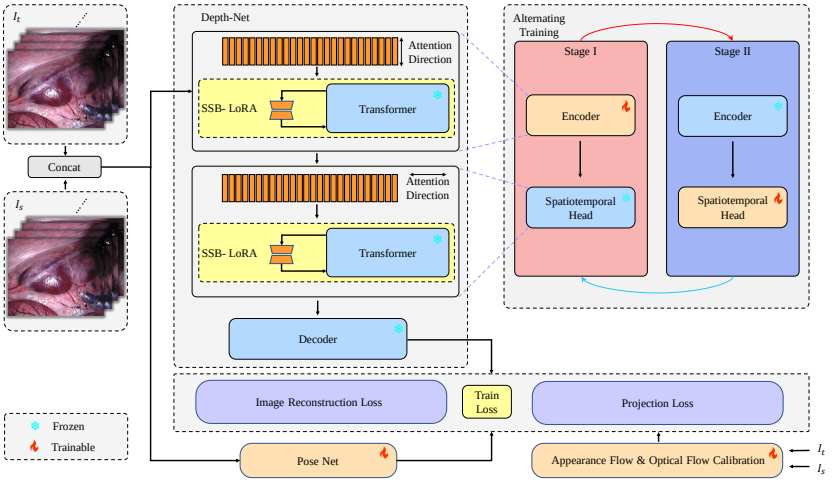

Fig. 1. Our framework consists of two parts: Video Depth Network and Pose Network. We use SSB LoRA [13] to alternatively finetune the spatial and temporal blocks. During training, a novel projection loss is introduced to enhance the temporal consistency.

as possible. The Pose Network is used to predict the intrinsic and extrinsic parameters of a camera. In this framework, pixels in a source image is projected to a target view using the predicted depth map, camera intrinsic parameters and relative extrinsic parameters between these two cameras. Then the same image reconstruction loss as EndoDAC [2] and AF-SfMLearner [12] is added to force the projected image to be as similar as possible with the target image. This process effectively learns depth and camera pose from unlabeled videos simultaneously.

PEFT Strategy In order to preserve the model's predictive capacity as much as possible while reducing the number of training parameters, we only add LoRA layers to the feedforwad layers in the model's attention blocks. To lower the training parameter requirements, we carefully select SSB LoRA [13] and apply it to our model. With no extra trainable parameters, our method only needs 0.17% of the total model parameters to be trainable. We find that with a meticulously devised training strategy, it is sufficient for the Video Depth Network to learn to predict an accurate results in endoscopy scenes.

Projection Loss In the self-supervised framework there are no extra constraints along the temporal dimension. Therefore, directly adapting Video Depth Anything Model to endoscopy scenes and finetuning it will result in the loss of temporal consistency. We thus propose a projection loss to overcome this issue. Given two adjacent frames I_s and I_t, our framework predicts their corresponding depth maps z_s, z_t, and the relative camera pose $T_{s\to t}$. Then the previous depth map z_s is projected to the t^{th} view by $T_{s\to t}$. This process produces a new depth map $z_{s\to t}$ and its corresponding pixel coordinate $u_{s\to t}$ at the t^{th} view:

$$u_{s\to t}, z_{s\to t} = \mathcal{R}(z_s; T_{s\to t}), \tag{1}$$

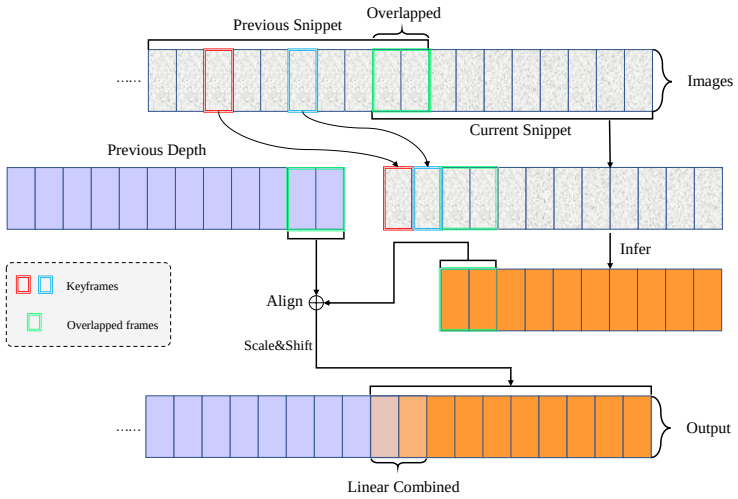

Fig. 2. Depth alignment strategy.

where $\mathcal{R}$ represents the projection operator. The pixel coordinate $u_{s\to t}$ is further utilized to sample the depth map z_t to get a resampled depth map $\hat{z}_t$:

$$\hat{z}_t = \mathcal{F}(z_t; u_{s\to t}), \tag{2}$$

where $\mathcal{F}$ represents bilinear interpolation operator. To filter the points derived from pixels out of bounds and invalid depth values, we also calculate a valid mask M. Our projection loss is then formulated as:

$$L_{proj} = M \cdot |z_{s\to t} - \hat{z}_t|, \tag{3}$$

Intuitively, this loss function acts as a consistency prior to force the depth maps from adjacent frames to align with each other. In the context of endoscopy, the movement of the camera is typically minimal, and the changes in the scene are relatively gradual. As a result, the depth maps of adjacent frames are often quite similar. This characteristic makes our loss function an effective constraint for temporal consistency.

Depth Alignment during Inference Due to the GPU memory constraints, in each step the depth model only receives a small video snippet as input at the training stage. However, during the inference stage it is necessary for the depth model to be equipped with the long video inference ability. General video depth estimation methods [1] [4] have explored different strategies, but they are not perfect in endoscopy scenes. In endoscopic videos, the soft tissues usually present slow motion and deformation. Many videos are captured by a camera navigating through the intestinal tract. Distinct from the common videos typically with a single fixed subject, endoscopic videos usually pose a greater challenge to the consistency of video depth estimation. Therefore, according to the endoscopy characteristics, we carefully design a Depth Alignment strategy during Inference.

We start from a simple but effective depth stitching strategy. Instead of directly taking the adjacent video snippets as input, we set the two snippets with L overlapped frames, which ensures the predictions to be more consistent. We then select T frames in the previous video snippet and concatenate them with the current snippet as input. Different from the selection strategy in Video Depth Anything [1] which actually keeps the first frame of the video to stay in each snippet input, we only select the frames from the previous snippet. Intuitively, this strategy extends the receptive field of the video depth estimation network along the temporal dimension, and also be flexible for a navigating video with a long trajectory. Then we calculate the shift and scale on the overlapped depth frames, and use them to align the next snippet. With enough overlapped frames, the adjacent depth snippets are aligned accurately. Then the current snippet are added to the result sequence, with the overlapped frames linearly combined with the previous depth snippet. The whole process is shown in Fig. 2.

3 Experiments

3.1 Evaluation

SCARED Dataset. The SCARED Dataset (Surgical Scene and Endoscopic Anatomy Recognition Dataset) is a pioneering resource in the field of surgical data science and medical imaging, providing a realistic and controlled environment for surgical simulation and analysis. The dataset comprises 35 endoscopic videos captured using a da Vinci Xi endoscope, totaling 22,950 frames of high-resolution imagery. Its annotations includes ground truth depth maps generated using a projector, as well as ground truth camera poses and intrinsic parameters. To evaluate video depth predictions, the SCARED dataset is split into 24, 3, and 8 video sequences for the training, validation and test sets, respectively.

Hamlyn Dataset. The Hamlyn Dataset is a widely recognized and valuable resource in the field of minimally invasive surgery, particularly for laparoscopic and endoscopic imaging research. It consists of a diverse collection of in vivo surgical videos captured during various real-world procedures. The dataset includes laparoscopic and endoscopic video sequences recorded from different anatomical regions, offering a realistic representation of the complexities encountered in clinical practice, such as tissue deformation, occlusions, and dynamic lighting conditions. We use the whole 21 video sequences for validation.

Implementation Details. Our framework is implemented with PyTorch [9] on NVIDIA RTX 3090 GPU. We apply AdamW [7] optimizer to train our framework with initial learning rates of $1e-4$. Training augmentations is the same as EndoDAC [2]. At the training stage we set the input sequence to be 16, and train for 20 epochs. For comparison fairness, we use the ViT-small backbone [3] for all the methods. For depth alignment, we use a slide window size 32 to process the input video. Frames in each window consist of $T = 2$ keyframes from the previous video snippet and 30 frames in the current video snippet, where the first $L = 8$ frames are the overlapped frames.

Table 1. Quantitative depth comparison on SCARED dataset. The best results are in bold. "Total." and "Train." refer to the total and trainable parameters utilized in Video Depth Network. Note that since Hamlyn dataset does not provide the camera pose annotations, we do not evaluate the TAE metric on it.

	Method	Year	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	TAE↓	Total.(M)	Train.(M)	Speed (ms)
SCARED	VDA [1]	2025	0.241	7.702	18.673	0.287	0.597	1.10	111.0	-	15.9
	EndoDAC [2]	2024	0.201	5.163	16.421	0.238	0.653	2.69	99.0	1.6	15.0
	EndoDAV(Ours)	-	**0.156**	**3.113**	**12.257**	**0.182**	**0.761**	**0.39**	111.3	**0.19**	16.0
Hamlyn	VDA [1]	2025	0.389	19.308	23.005	0.333	0.513	-	111.0	-	15.9
	EndoDAC [2]	2024	0.240	6.998	17.240	0.304	0.589	-	99.0	1.6	15.0
	EndoDAV(Ours)	-	**0.212**	**5.040**	**16.759**	**0.276**	**0.595**	-	111.3	**0.19**	16.0

Evaluation Protocols. We compute the 5 standard metrics: Abs Rel, Sq Rel, RMSE, RMSE log and *delta* for evaluation. We also use a TAE metric proposed in [18] to evaluate the temporal consistency. We evaluate affine-invariant depth predictions by previously aligning the scale and shift between the predicted depth and the ground truth. Note that, different from previous single image depth estimation methods [2], we follow the video depth estimation methods [1] [4] [6] to align depth using the median and scale across the entire depth sequence rather than each depth map. This is more challenging but necessary to ensure the temporal consistency to be fully evaluated. We average the result from each scene to obtain the final results.

3.2 Comparison

Quantitative results. Our proposed method is compared with a SOTA method EndoDAC [2] and a baseline method Video Depth Anything [1]. To ensure fairness, all the models use the ViT-S [3] encoder. Video Depth Anything is a pretrained model without retraining. EndoDAC is initialized with a pretrained Depth anything v2 [20] backbone, and finetuned under our experimental setup. Tabel 1 shows quantitative comparison results. On SCARED dataset, our method significantly surpasses the other two methods in both accuracy and temporal consistency. Finetuned from Video Depth Anything, our method significantly promotes it to predict more accurately in endoscopy scenes. Due to the lack of temporal information usage, EndoDAC fails to produce a consistent prediction and presents worst at TAE. Compared with EndoDAC, our method produces both spatial accurate and temporal consistent results. When achieving the best performance, our method also has the least trainable parameters. Since we do not add much parameters, the inference speed of our method is similar with the other two, being capable of dealing with real-time depth estimation.

Qualitative results. We show qualitative results on SCARED dataset in Fig. 3. Each model receives a video sequence as input. Video Depth Anything fails to produce accurate depth in endoscopy scenes. The depth in the green box is largely inconsistent with depth in other regions. EndoDAC produces a broken and inconsistent prediction. Depth of the soft tissue in the left top region and the

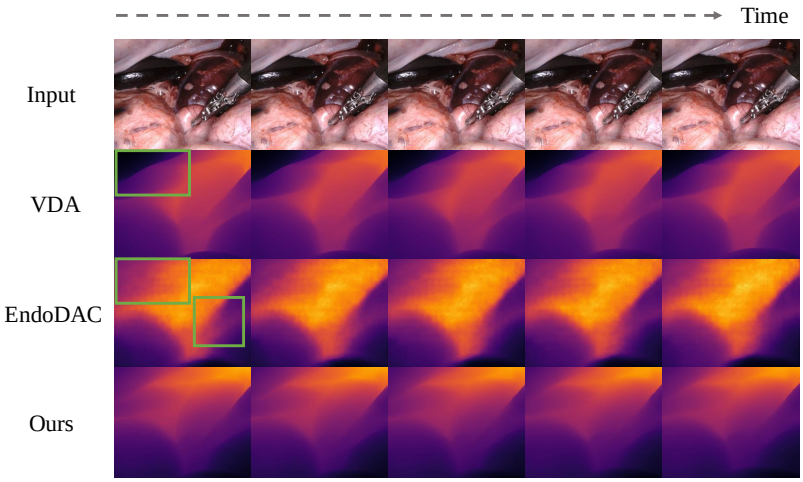

Fig. 3. Qualitative results with a video sequence as input. Video Depth Anything [1] cannot understand the endoscopy scene and predicts wrongly at the corner(in the green box). EndoDAC [2] presents a broken and inconsistent prediction(in the green box). Our method achieves both accurate and consistent depth results. For a more direct visualization, please refer to the video in our supplementary materials.

surgical forceps is unstable along the time dimension. Our method presents an spatial accurate and temporal consistent prediction, significantly enhancing the usability of the estimated video depth in downstream tasks. For a more direct visualization, please refer to the video in our supplementary materials.

3.3 Ablation Study

We conduct ablation study on our proposed methods. These methods are validated on SCARED Dataset. We set four experiments: with/without projection loss and with/without depth alignment inference strategy. All the models are finetuned by our framework with SSB Lora. Table 2 shows the experimental results. It can be seen that, both of the two strategies benefit the whole model to produce more accurate and consistent results.

4 Conclusion

To enable accurate and consistent video depth estimation in endoscopy scenes, we propose to efficiently adapt the video depth estimation foundation model utilizing a self-supervised framework. This framework predicts both depth and camera pose, which are learned simultaneously by projecting the depth at the source view to the target view using the predicted relative camera pose. Therefore, the depth network can effectively learn the modal information of the endoscopy scene. By

Table 2. Ablation study on SCARED dataset. The best results are in bold.

Projection Loss	Depth Alignment	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ ↑	TAE↓
×	×	0.195	4.273	15.768	0.224	0.639	1.03
✓	×	0.180	4.259	14.554	0.202	0.671	0.80
×	✓	0.162	3.957	13.215	0.208	0.665	0.40
✓	✓	**0.156**	**3.113**	**12.257**	**0.182**	**0.761**	**0.39**

utilizing a simple SSB Lora Layer, we only need 0.17% parameters to be trainable. To further enhance the model's ability of temporal consistency, we propose two simple but effective strategies. A projection loss is addressed to utilize the adjacent depths and relative camera pose to constrain the change of output depth stream. Considering the characteristics on endoscopic videos, an effective Depth Alignment Inference strategy is proposed to align the predicted depth snippet during inference. Our experiments on two public endoscopy datasets demonstrates the effectiveness on our methods.

Acknowledgments. This work was supported in part by the National Key R&D Program of China 2022YFF1202600, in part by the National Natural Science Foundation of China under Grant 62322604, 62176159 and in part by the Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

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

References

1. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. arXiv preprint arXiv:2501.12375 (2025)
2. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218. Springer (2024)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
4. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos. arXiv preprint arXiv:2409.02095 (2024)
5. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2024)
6. Ke, B., Narnhofer, D., Huang, S., Ke, L., Peters, T., Fragkiadaki, K., Obukhov, A., Schindler, K.: Video depth without video models. arXiv preprint arXiv:2411.19189 (2024)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

8. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
9. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
10. 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)
11. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. arXiv preprint arXiv:2406.01493 (2024)
12. 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)
13. Si, C., Yang, X., Shen, W.: See further for parameter efficient fine-tuning by standing on the shoulders of decomposition. CoRR (2024)
14. Wang, K., Yang, C., Wang, Y., Li, S., Wang, Y., Dou, Q., Yang, X., Shen, W.: Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 219–229. Springer (2024)
15. Wei, R., Li, B., Chen, K., Ma, Y., Liu, Y., Dou, Q.: Enhanced scale-aware depth estimation for monocular endoscopic scenes with geometric modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–273. Springer (2024)
16. Yang, C., Wang, K., Wang, Y., Dou, Q., Yang, X., Shen, W.: Efficient deformable tissue reconstruction via orthogonal neural plane. IEEE Transactions on Medical Imaging (2024)
17. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2023)
18. Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., He, T.: Depth any video with scalable synthetic data. arXiv preprint arXiv:2410.10815 (2024)
19. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
20. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2025)