B64 (Fri-AM)

MICCAI2025 Daejeon REPUBLIC OF KOREA

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

MICCAI

EndoDAV: Depth Any Video in Endoscopy with Spatiotemporal Accuracy

Zanwei Zhou[1], Chen Yang[1], Piao Yang[2], Xiaokang Yang[1], Wei Shen[1]

[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

Introduction

Background

- Video depth estimation has been applied to various endoscopy tasks, such as reconstruction, navigation, and surgery.
- Many methods focus on directly adapting depth estimation foundation models to endoscopy scenes, while do not consider temporal information, leading to an inconsistent prediction.

Contributions

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 public datasets demonstrate the spatial accuracy and temporal consistency of our methods.

Methodology

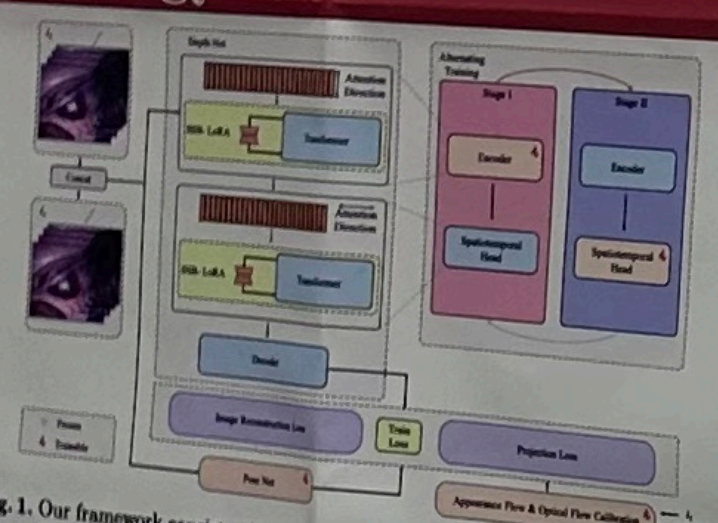

Fig. 1. Our framework consists of two parts: Video Depth Network and Pose Network. We use SSB LoRA [13] to alternatively finetune the spatial and temporal blocks. During training, a novel projection loss is introduced to enhance the temporal consistency.

PEFT Strategy

- We only add LoRA layers to the feed-forward layers.
- We carefully select SSB LoRA[1] to reduce the training parameters. Our method only needs **0.17%** of the model parameters to be trainable.

Projection Loss

Given two adjacent frames, our framework predicts their depth maps and the relative camera pose. Then the previous depth map can be projected:

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

The pixel coordinate is further utilized to sample the depth map to get a resampled depth map:

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

Our projection loss is then formulated as:

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

Depth Alignment during Inference

- Set the two snippets with L overlapped frames.
- Select T frames in the previous video snippet and concatenate them with the current snippet as input.
- Calculate the shift and scale on the overlapped depth frames, and use them to align the next snippet.

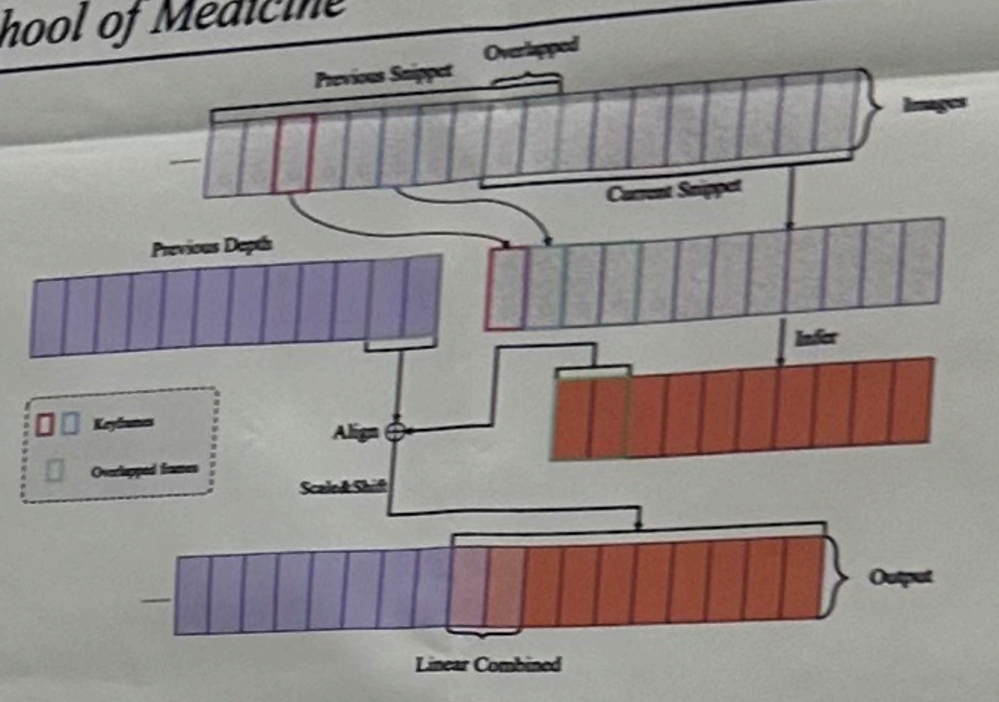

Fig. 2. Depth alignment strategy.

Experiments

Datasets

- SCARED Dataset: we split it into 24, 3, and 8 video sequences for the training, validation and test sets, respectively.
- Hamlyn Dataset: the whole 21 video sequences are for validation.

Quantitative Results

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

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**

Qualitative Results

Conclusion

To enable accurate and consistent video depth estimation in endoscopy scenes, we adapt the video depth estimation foundation model utilizing a self-supervised framework. By utilizing a simple SSB Lora Layer, we only set 0.17% parameters to be trainable. A projection loss is addressed to constrain the change of output depth stream. Depth Alignment Inference strategy is proposed to align the predicted depth snippet during inference. Experiments on two endoscopy datasets demonstrates the effectiveness.

References

[1] Si, C., Yang, X., Shen, W.: See further for parameter efficient fine-tuning by standing on the shoulders of decomposition. CoRR (2024)