Supplementary Material
Depth Any Endoscopy: Towards Self-Supervised Generalizable Depth Estimation in Monocular Endoscopy

Shuwei Shao[1] Kejin Zhu[2] Shixing Ma[1] Xinzhe Du[1] Baochang Zhang[3] Zhe Min[1,4*]
[1]School of Control Science and Engineering, Shandong University, Jinan, Shandong, China
[2]School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[3]School of Artificial Intelligence, Beihang University, Beijing, China
[4]Shenzhen Loop Area Institute, Shenzhen, China

1. Evaluation Metrics

Similar to [1, 2], we employ the standard evaluation metrics in our experiments, which are described in detail as follows:

- Relative Absolute Error (**Abs Rel**): $\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}|d-d^{gt}|/d^{gt}$; where d and d^{gt} stand for the predicted and ground-truth depth values, respectively;
- Relative Squared Error (**Sq Rel**): $\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}\left(d-d^{gt}\right)^2/d^{gt}$;
- Root Mean Squared Error (**RMSE**): $\sqrt{\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}\left(d-d^{gt}\right)^2}$;
- Root Mean Squared Logarithmic Error (**RMSE log**): $\sqrt{\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}\left(\log d-\log d^{gt}\right)^2}$;
- Threshold Accuracy (δ): $\%$ of d satisfies $\left(\max\left(\frac{d}{d^{gt}},\frac{d^{gt}}{d}\right)<1.25\right)$.

Table 1. **Comparison of model parameters and inference time**. Inf.: inference time; Total.: total parameters; Trainable.: trainable parameters.

Method	Backbone	Total. ↓	Trainable. ↓	Inf. ↓
AF-SfMLearner [2]	ResNet-18	14.3M	14.3M	0.012s
EndoDAC [1]	ViT-Base	99.1M	1.7M	0.029s
Depth Anythingv2 [3]	ViT-Base	97.5M	97.5M	0.024s
DAE (Ours)	ViT-Base	101.7M	4.3M	0.063s

2. Model Parameters and Inference Time

We compare DAE with EndoDAC, Depth Anythingv2, and AF-SfMLearner in terms of inference time and the number of model parameters, as reported in Table 1. The inference time is measured on the SCARED test set using a batch size of 1. It can be seen that the total number of parameters in the proposed DAE is higher than that of AF-SfMLearner and slightly higher than that of EndoDAC and Depth Anythingv2. However, its trainable parameters are fewer than those of Depth Anythingv2 and AF-SfMLearner. Although the MoE architecture increases inference time to some extent, the overall computational cost remains acceptable.

3. More Qualitative Depth and Point Cloud Results

To further compare the proposed DAE with previous state-of-the-art methods, we present more qualitative depth estimation results on C3VDv2, clinically collected arthroscopy data, and SCARED, as shown in Fig. 1. The results display that the proposed DAE produces more accurate depth predictions and maintains reliable performance across diverse scenarios. For better evaluation of depth predictions from the 3D shape, we convert depth predictions into point clouds and display more qualitative point cloud results for these datasets in Fig. 2. The results show that the proposed DAE effectively preserves prominent geometric features and is capable of recovering the underlying 3D anatomical structures reasonably.

References

[1] Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. 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*, pages 208–218. Springer, 2024. 1

[2] Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in en-

*Corresponding author

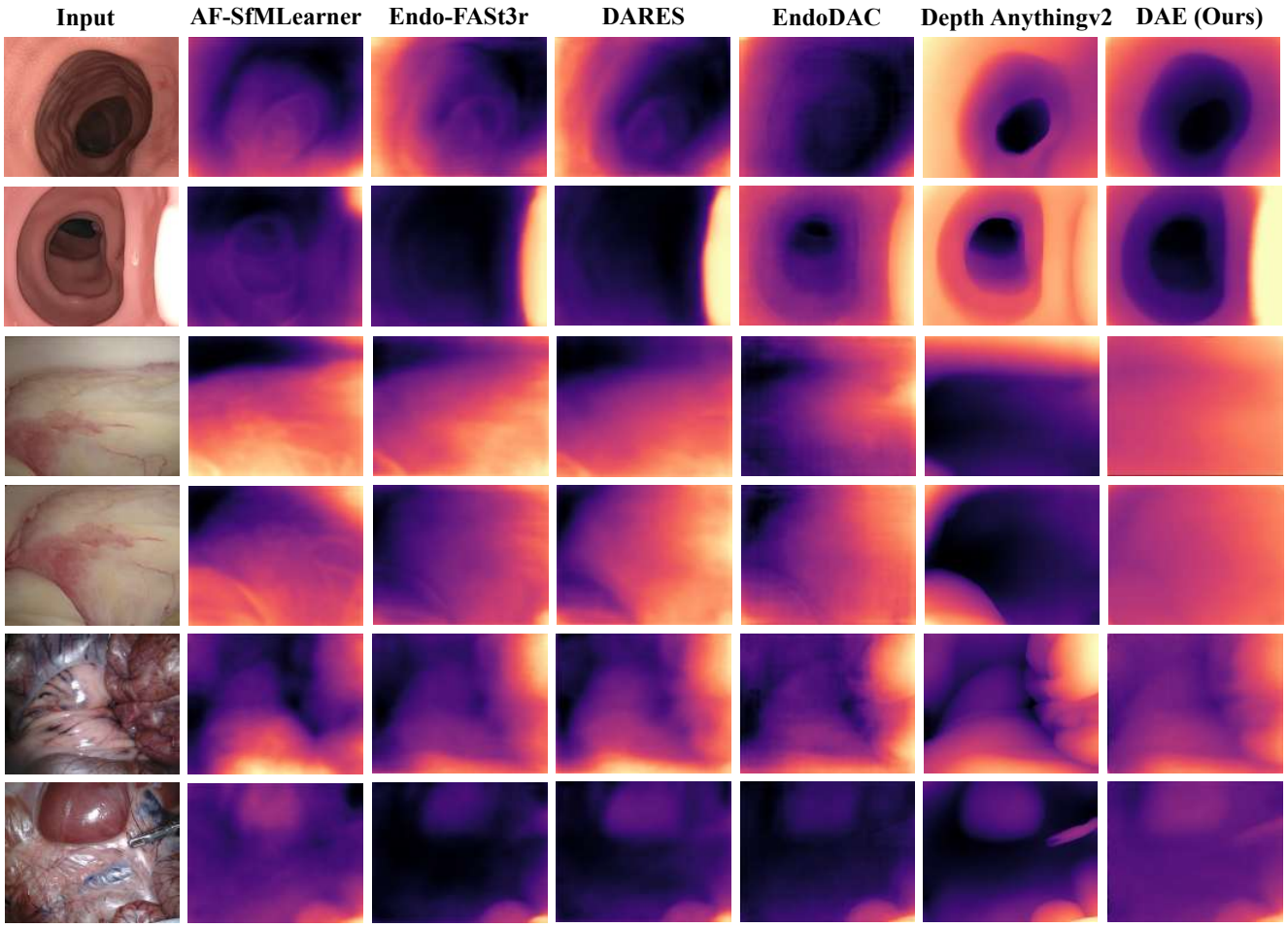

Figure 1. **More qualitative zero-shot and in-domain depth results across different datasets**.

doscopy: Appearance flow to the rescue. *Medical Image Analysis*, page 102338, 2022. 1

[3] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024. 1

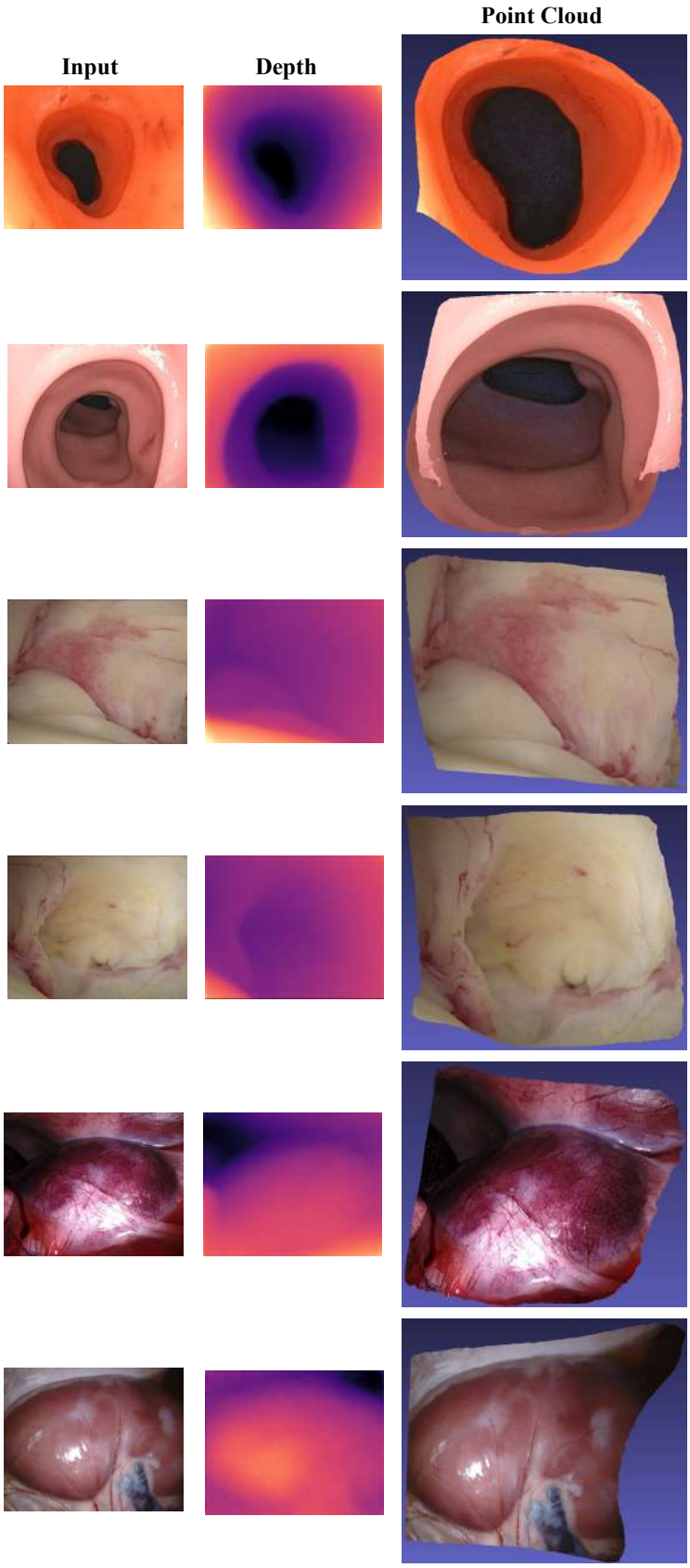

Figure 2. **Qualitative point cloud results of the proposed DAE**.