

EndoMamba: An Efficient Foundation Model for Endoscopic Videos via Hierarchical Pre-training

Qingyao Tian^{1,2}, Huai Liao³, Xinyan Huang³, Bingyu Yang^{1,2}, Dongdong Lei⁴,
Sebastien Ourselin⁵, and Hongbin Liu^{1,4,5}✉

- ¹ State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China liuhongbin@ia.ac.cn
² School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
³ The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
⁴ Centre for Artificial Intelligence and Robotics, Chinese Academy of Sciences, HK, China
⁵ School of Engineering and Imaging Sciences, King's College London, UK

Abstract. Endoscopic video-based tasks, such as visual navigation and surgical phase recognition, play a crucial role in minimally invasive surgeries by providing real-time assistance. While recent video foundation models have shown promise, their applications are hindered by (1) *computational inefficiencies* and (2) suboptimal performance caused by *limited data* for pre-training in endoscopy. To address these issues, we present **EndoMamba**, a foundation model designed for real-time inference while learning generalized spatiotemporal representations. *First*, to mitigate computational inefficiencies, we propose the EndoMamba backbone, optimized for real-time inference. Inspired by recent advancements in state space models, EndoMamba integrates Bidirectional Mamba blocks for spatial modeling within individual frames and vanilla Mamba blocks for past-to-present reasoning across the temporal domain. This design enables both strong spatiotemporal modeling and efficient inference in online video streams. *Second*, to improve data efficiency, we propose a self-supervised hierarchical pre-training diagram that enhances EndoMamba’s representation learning. Specifically, our approach combines masked reconstruction with auxiliary supervision, leveraging low-level reconstruction to capture spatial-temporal structures and high-level alignment to transfer broader knowledge from a pretrained general-video domain foundation model. Extensive experiments on four downstream tasks—classification, segmentation, surgical phase recognition, and localization—demonstrate that EndoMamba outperforms existing foundation models and task-specific methods while maintaining real-time inference speed. The source code is available at <https://github.com/TianCuteQY/EndoMamba>.

Keywords: Foundation model · Endoscopy video analysis · State space model.

1 Introduction

Tasks based on endoscopic videos, such as navigation [24] and surgical phase recognition [14], have emerged as a key area of research in medical image analysis. These tasks contribute significantly to minimally invasive surgeries by providing real-time, automated assistance. The core challenge in endoscopic video analysis lies in the efficient

estimation of spatiotemporal information [11]. To improve performance, most research has focused on designing specialized modules to learn task specific domain knowledge [14,13,24]. However, task-specific methods often struggle with limited generalization and adaptability to new data, restricting their applicability in real clinical settings [35].

Recently, foundation models have demonstrated remarkable performance in general video understanding [29,30,12]. These models have become the backbone of various video downstream applications, offering superior generalization and scalability compared to task-specific approaches [31]. Recognizing their potential in endoscopy, EndoFM [32] introduces a Transformer-based foundation model for endoscopic video analysis, achieving promising results on downstream tasks. However, two major challenges remain unaddressed: (1) *computational inefficiencies* in online inference and (2) *the limited quantity and deficient endoscopic data for effective pre-training* [32]. The limitation in inference speed primarily stems from the backbone architecture. Existing video foundation models, such as Transformer-based models [29,30] and bidirectional state space models (SSMs) [11], require recalculating all historical video frames repeatedly when new frames arrive during inference. This inefficiency significantly impacts real-time applications. Regarding data limitations, the main challenges are the relatively small dataset sizes and the lack of paired vision-language data in the endoscopic domain. These limitations restrict the potential for large-scale learning [30] and contrastive learning [22,12], which are essential for improving model performance.

In this work, we present EndoMamba, an efficient foundation model for endoscopy videos via hierarchical pre-training. To address the computational bottleneck, we propose the EndoMamba backbone. Motivated by recent advancements in SSMs [8,4] for long-term memory and fast inference, EndoMamba integrates Bidirectional Mamba (Bi-Mamba) for spatial modeling within individual frames and vanilla Mamba for causal reasoning across the temporal domain. This architecture enables both strong spatiotemporal reasoning and efficient inference in online video streams. To mitigate data limitations, we propose a hierarchical self-supervised pre-training diagram, enabling EndoMamba to capture spatiotemporal structures while leveraging broader knowledge from pretrained general-domain models. This diagram comprises two main components: low-level video reconstruction and high-level feature alignment. For the low-level reconstruction, inspired by VideoMAE [29], we adopt a masked autoencoder strategy for data-efficient pre-training, allowing EndoMamba to reconstruct video clips from heavily occluded visual cues. Additionally, we propose an auxiliary pre-training module for high-level feature alignment. This module leverages general-domain video foundation models trained on large-scale vision-text data to align features effectively. Extensive experiments are conducted on four downstream tasks. Results show that EndoMamba outperforms EndoFM with a +11.5% increase in segmentation Dice score and a +21.3% improvement in surgical phase recognition accuracy, while boosting inference speed from 9.2 frames per second (FPS) to 46.7 FPS with a memory length of 32 frames.

2 Methods

In this section, we first provide a brief overview of SSMs (§2.1). We then present a detailed description of the EndoMamba architecture (§2.2), which effectively captures

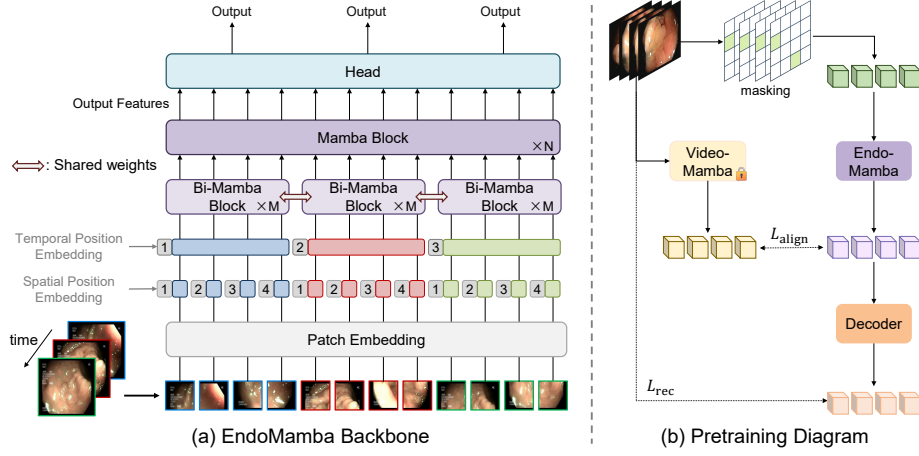


Fig. 1: Proposed EndoMamba model for endoscopic video analysis. (a) EndoMamba backbone structure with efficient recurrent inference ability, and (b) hierarchical pre-training diagram for enhanced representation learning.

generalized spatiotemporal representations while ensuring efficient inference. Finally, to mitigate the challenges associated with endoscopy data, we propose a pre-training strategy for EndoMamba that enhances its representation learning capability (§2.3).

2.1 Preliminaries

Deep State Space Models (SSMs), such as Mamba [8], have recently achieved success in natural language processing due to their ability to model long-range dependencies while maintaining computational efficiency. It is grounded by classical SSMs which map a one-dimensional function or sequence $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$. The system dynamics are modeled by the evolution matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, with projection parameters $\mathbf{B} \in \mathbb{R}^{N \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times N}$. The continuous system operates as follows:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t). \quad (1)$$

For efficient training with discrete data, SSMs discretizes the continuous system into a discrete form using the zero-order hold (ZOH) method, defined as:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, y_t = \mathbf{C}h_t, \quad (2)$$

where the discrete counterparts $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ are defined as:

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \quad (3)$$

Based on the SSMs, Mamba further proposes to selectively propagate or forget information along the sequence based on the current token, enabling efficient linear scalability in sequence length and strong long-range dependency modeling capability.

However, since vanilla Mamba processes input tokens in a past-to-present manner, it is not well-suited for vision tasks. To address this limitation, VisionMamba [36] uses the Bi-Mamba block, which applies bidirectional scanning to process the sequences from both forward and backward directions for image analysis.

2.2 EndoMamba Architecture

Motivated by recent advancements in SSMs, we propose EndoMamba, a Mamba-based backbone designed to enhance computational efficiency in video foundation models. An overview of the proposed EndoMamba is shown in Fig. 1 (a). As mamba blocks take input as 1-D sequence, we first transform the input videos $\mathbf{X} \in \mathbb{R}^{B \times 3 \times T \times H \times W}$ into flattened spatiotemporal patches $\mathbf{X}_p \in \mathbb{R}^{B \times N \times C}$ where $N = T \times h \times w$ ($h = \frac{H}{k}$, and $w = \frac{W}{k}$). Here, k denotes the kernel size, B is the batch size, T denotes the number of frames, and $H \times W$ represents frame spatial resolution.

To retain spatiotemporal position information, we add a learnable spatial position embedding \mathbf{p}_s , and a fixed sinusoid temporal embedding \mathbf{p}_t to capture temporal position information while allowing flexible input sequence length. The embedded input is then:

$$\mathbf{X}_{\text{embed}} = \mathbf{X}_p + \mathbf{p}_s + \mathbf{p}_t. \quad (4)$$

To perform spatial-temporal scanning while retaining recurrent inference, EndoMamba process $\mathbf{X}_{\text{embed}}$ by performing bidirectional scanning within each image patch and causal scanning along the time axis. Specifically, the embedded input $\mathbf{X}_{\text{embed}}$ are first rearranged into $\mathbf{X}_r \in \mathbb{R}^{J \times P \times C}$, where $J = B \times T$ and $P = h \times w$. Then, \mathbf{X}_r is passed through M stacked Bi-Mamba blocks for bidirectional scanning within each image frame in parallel. After processing, the tokens are rearranged back to $\mathbf{X}_b \in \mathbb{R}^{B \times N \times C}$ and passed through n stacked Mamba blocks, which perform causal scanning along the temporal axis. Finally, the processed tokens are fed into a multi-layer perceptron (MLP) head to produce the prediction. For frame-level tasks, the MLP head outputs $\mathbf{Y} \in \mathbb{R}^{B \times T}$.

For training, EndoMamba takes video clips as input and simultaneously outputs predictions for each frame, using information from the current frame as well as its past frames. For real-time applications, EndoMamba processes each frame of a live video stream individually, passing the past states as memory. The model then makes a final prediction for the current frame based on both the frame itself and its historical context.

2.3 Pretraining Diagram

To tackle the limitation of endoscopy data, we propose a hierarchical pre-training diagram for EndoMamba, as illustrated in Fig. 1(b). This pre-training combines low-level video reconstruction with high-level feature alignment to enhance representation learning. For the low-level reconstruction, we randomly mask a large portion of the input video, prompting EndoMamba to infer the missing regions. This process inherently encourages EndoMamba to capture contextual dependencies by leveraging spatiotemporal correlations [29]. Formally, the reconstruction loss is defined as follows:

$$L_{\text{rec}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\mathbf{X}(p) - \hat{\mathbf{X}}(p)\|^2, \quad (5)$$

Table 1: Overview of the pre-training dataset. Our full pre-training dataset MIX12 consists of 74,828 video clips, comprising a total of 11,303,853 frames.

	MIX7							Additional to MIX12				
Source	CS [16]	S&S [17,9]	LP [15]	HK [3]	KC [25]	CT [18]	EFM [32]	G [10]	hS [34]	EM [1]	R [23]	PB
Clips	210	1.0k	237	5.7k	1.0k	580	24.1k	104	236	1.7k	39.8k	179
Frames	36.5k	159.4k	40.2k	876k	159k	90.4k	3.66M	13.4k	35.6k	242k	5.96M	26.8k

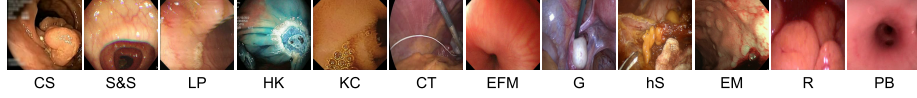


Fig. 2: Example frames of the pre-training data in MIX12.

where Ω is the set of masked video regions, p is the masked token index, and $\hat{\mathbf{X}}$ is the reconstructed video clip by EndoMamba.

To further enhance representation learning, we propose to align EndoMamba’s features with VideoMamba, a general-domain video model pretrained on large-scale video data as teacher model. The aligning is enforced by cosine similarity loss defined as:

$$L_{\text{align}} = 1 - \frac{1}{|\Omega|} \sum_{q \in \Omega} \cos(\mathbf{X}_f(q), \mathbf{X}_t(q)), \quad (6)$$

$$\cos(\mathbf{X}_f(q), \mathbf{X}_t(q)) = \frac{\mathbf{X}_f(q) \cdot \mathbf{X}_t(q)}{\|\mathbf{X}_f(q)\| \|\mathbf{X}_t(q)\|}, \quad (7)$$

where \mathbf{X}_f and \mathbf{X}_t are the output features of EndoMamba and the teacher model respectively; Ω is the set of unmasked video regions and q is the unmasked token index.

This feature alignment enables EndoMamba to inherit knowledge from a broader domain while introducing additional challenges in the pretraining process. Unlike the teacher model, EndoMamba is trained on significantly sparser inputs. Spatially, it operates on heavily masked video clips, while VideoMamba utilizes the full input. Temporally, VideoMamba leverages Bi-Mamba, enabling bidirectional scans across the temporal axis, whereas EndoMamba follows a causal scanning. Aligning these features thus necessitates a dual-axis completion of a robust general-domain representation.

Finally, the pre-training of EndoMamba uses the following loss function:

$$L = L_{\text{rec}} + \alpha L_{\text{align}}, \quad (8)$$

where α is the weight for feature alignment loss. Our pre-training diagram allows the model to capture contextual dependencies and inherit knowledge from a broader domain for enhanced generalization ability.

3 Experiments

3.1 Implementation Details

To increase data diversity and volume, EndoMamba is pretrained on MIX12, a combination of 12 endoscopic datasets. This includes MIX7 (shared with EndoFM), four curated from public datasets, and one from our porcine bronchoscopy dataset (PB), as detailed in Table 1 and Fig. 2. This increases the pre-training data from approximately 5M to 11M frames when moving from MIX7 to MIX12. The model is trained using the AdamW optimizer with a base learning rate of $1e-8$, a cosine learning rate schedule for 500 epochs, and a batch size of 48, with the first five epochs dedicated to linear warm-up. For EndoMamba architecture, we set $M=12$, $N=12$. Frames in each video clips are resized to spatial size of 224×224 and frame number of 16 as pre-training input. For feature alignment, the pretrained VideoMamba-S serves as the teacher model, with the feature alignment loss weight set to $\alpha = 0.25$ based on preliminary experiments.

To evaluate the generalization ability of EndoMamba, we transfer it to a range of downstream tasks, including 1) *PolypDiag for classification* [28] and 2) *CVC-12K for polyp segmentation* [2], using the same fine-tuning settings as EndoFM. Additionally, we extend its application to two other tasks: 3) *surgical phase recognition on the Autolaparo dataset* [33] and 4) *branch-level localization on patient bronchoscopy dataset*. Compared to classification and segmentation, surgical phase recognition and localization emphasis long-term memory and real-time inference. For the latter two tasks, we sample 32-frame input clips of spatial size 224×224 for fine-tuning, with data splits following existing methods [14,26]. The number of training epochs is set to 50 for surgical phase recognition and 20 for localization. No task-specific modules are used; instead, only an MLP head is added to the EndoMamba backbone for final output.

3.2 Comparison with State of the Art

We compare EndoMamba with recent state-of-the-art (SOTA) methods on the four downstream tasks. **1) Classification:** Compared to existing pre-training methods and foundation models [5,20,21,19,32,30,11], EndoMamba ranks first (Table 2), achieving a +4.3% improvement in F1 score over the best-performing SOTA method. **2) Segmentation:** EndoMamba maintains its leading performance in segmentation (Table 2) among compared pre-training methods and foundation models [5,20,21,19,32,30,11], achieving a +6.9% increase in Dice score over the best SOTA model. **3) Surgical Phase Recognition:** As shown in Table 3, EndoMamba outperforms all evaluated foundation models [30,11,32], achieving a +2.7% increase in video-level accuracy over VideoMamba. Additionally, it performs competitively with the top task-specific methods [6,7,13,14], such as SKiT, which employs a multi-stage framework, whereas EndoMamba achieves comparable results using only its backbone and an MLP head for end-to-end estimation. **4) Localization:** EndoMamba outperforms all evaluated foundation [26,30,11,32] and task-specific models [24,27] by a significant margin (Table 4). Bronchoscopy interventions often face severe visual occlusions, such as liquids and bubbles, making accurate localization reliant on strong spatiotemporal reasoning. EndoMamba effectively recovers positions after occlusions, as shown by reduced video-level accuracy variance, indicating more stable performance across cases. These results

Table 2: Results of classification on Polyp-Diag [28] and segmentation on CVC-12k [2]. **Best** and **second best** results are highlighted.

Methods	PolypDiag CVC-12K	
	F1 (%) \uparrow	Dice (%) \uparrow
FAME [5]	85.4 \pm 0.8	67.2 \pm 1.3
ProViCo [20]	86.9 \pm 0.5	69.0 \pm 1.5
VCL [21]	87.6 \pm 0.6	69.1 \pm 1.2
ST-Adapter [19]	84.8 \pm 0.7	64.3 \pm 1.9
EndoFM [32]	<u>90.7\pm0.4</u>	73.9 \pm 1.2
VideoMAEv2 [30]	87.5 \pm 1.6	72.1 \pm 0.9
VideoMamba [11]	75.6 \pm 1.9	<u>78.5\pm1.0</u>
Ours	95.0\pm1.3	85.4\pm0.2

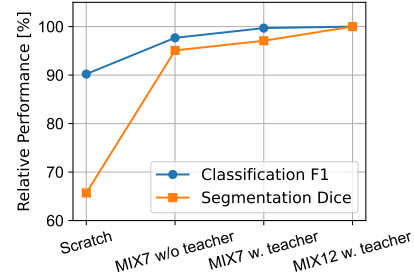


Fig. 3: Relative performance on classification F1 and segmentation Dice score, with gradual performance gains from pre-training data scaling and the addition of a teacher model.

Table 3: Surgical phase recognition results on the Autolaparo [33] dataset, comparing task-specific methods (top) and foundation models (bottom).

Method	Accuracy \uparrow	Jaccard \uparrow
Trans-SVNet [6]	78.3	50.7
AVT [7]	77.8	50.7
LoViT [13]	81.4 \pm 7.6	56
SKiT [14]	<u>82.9 \pm 6.8</u>	59.9
VideoMAEv2 [30]	77.0 \pm 5.7	53.1
VideoMamba [11]	80.3 \pm 8.2	54.0
EndoFM [32]	62.0 \pm 11.9	37.4
Ours	83.3 \pm 8.7	<u>57.6</u>

Table 4: Localization results on patient bronchoscopy dataset, comparing task-specific methods (top) and foundation models (bottom).

Methods	Accuracy \uparrow	F1 \uparrow
AirwayNet [24]	38.7 \pm 15.7	53.1
BronchoTrack [27]	57.0 \pm 24.6	61.3
EndoOmni [26]	78.6 \pm 16.2	<u>69.2</u>
VideoMAEv2 [30]	<u>78.9 \pm 9.7</u>	<u>69.2</u>
VideoMamba [11]	74.1 \pm 8.3	67.4
EndoFM [32]	62.2 \pm 19.1	56.7
Ours	83.0 \pm 5.5	73.5

demonstrate its ability to capture transferable representations robust enough to excel in complex downstream tasks without task-specific architectural modifications.

3.3 Ablation Studies

We explore various aspect of EndoMamba, regarding pre-training settings and real-time performance. **1) Hierarchical Pre-training:** We compare training from scratch with hierarchical pre-training, by evaluating downstream performance on classification and segmentation. As seen in Fig. 3, "MIX12 w. teacher", representing our full pre-training strategy, significantly outperforms "Scratch", which denotes random initialization. The improvement highlights the effectiveness of our pre-training, which includes low-level video reconstruction for learning spatiotemporal structure and high-level alignment for enhanced performance. **2) Teacher Guidance:** As seen in Fig. 3, "MIX7 w. teacher", which represents pre-training on MIX7 dataset with the teacher model guidance, brings

Table 5: Inference speed comparison of EndoMamba with existing video foundation models. Results are reported in frames per second (FPS) with network memory lengths set to $T=32, 64, 128$. Inference complexity [4] is measured by the number of image tokens (P), network memory lengths (T) and hidden state dimension (m).

Method	Backbone	Param Num.	Inference Complexity	FPS↑		
				T=32	T=64	T=128
EndoFM [32]	video ViT	121.26M	$O(P^2T^2m)$	9.2	4.8	2.4
VideoMAEv2 [30]	video ViT	22.26M	$O(P^2T^2m)$	27.8	9.7	2.7
VideoMamba [11]	VideoMamba	25.42M	$O(PTm^2)$	21.7	11.8	6.1
Ours	EndoMamba	24.46M	$O(Pm^2)$	46.7	47.1	46.6

an improvement of around 2% on the evaluated two tasks, comparing with "*MIX7 w/o teacher*", which uses the same dataset without teacher guidance. This improvement stems from high-level feature alignment, which transfers broader knowledge from a pretrained general-domain foundation model, thereby yielding more generalizable representations. **3) Dataset Scaling:** To explore the improvement brought by pre-training data scaling, we compare the performance of pre-training with MIX7, which aligns with existing endoscopy video foundation model EndoFM, and pre-training using our expanded MIX12. Results are shown in Fig. 3 "*MIX7 w. teacher*" and "*MIX12 w. teacher*" respectively. Dataset scaling up contribute to 3% improvement on the more challenging segmentation task, and a moderate improvement on classification. **4) Speed Analysis:** To evaluate the real-time applicability of EndoMamba, we compare its speed with existing foundation models, including the transformer-based EndoFM and VideoMAE v2, as well as the Mamba-based VideoMamba. Results for inference speed, computational complexity, and parameter counts of all models are presented in Table 5. For a fair comparison, inference speed is measured in frames per second (FPS) on an NVIDIA A800 GPU with memory lengths of 32, 64, and 128 frames. Results show that EndoMamba achieves the highest inference speed among all tested models with the evaluated memory length, significantly outperforming both transformer-based and Mamba-based architectures. Notably, its inference efficiency remains stable as memory size increases, benefiting from its ability to propagate past memory without redundant computation. In contrast, transformer-based models suffer from quadratic-complexity attention mechanisms, while Bi-Mamba blocks require recomputation, both of which hinder efficiency. By maintaining temporal continuity with minimal computational overhead, EndoMamba is particularly well-suited for real-time endoscopic applications.

4 Conclusion

In this work, we present EndoMamba, a foundation model designed for real-time endoscopic video analysis. The proposed EndoMamba backbone employs spatial bidirectional scanning and temporal causal scanning, enabling strong spatiotemporal modeling and efficient inference. To address the data limitation, we propose a hierarchical pre-training diagram that combines low-level video reconstruction for spatiotemporal rep-

representation learning, and high-level alignment with a pretrained general-domain foundation model, leveraging its broader visual knowledge to enhance representation learning. Experimental results on 4 diverse downstream tasks demonstrate that EndoMamba outperforms existing foundation models and SOTA task-specific methods while maintaining real-time inference efficiency for practical deployment.

Acknowledgments. This work was supported by the Centre of AI and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, sponsored by InnoHK Funding, HKSAR, and partially supported by Institute of Automation, Chinese Academy of Sciences.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al.: Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data* **10**(1), 671 (2023)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WMDOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics* **43**, 99–111 (2015)
3. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* **7**(1), 283 (2020)
4. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: *ICML* (2024)
5. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: *ICCV*. pp. 9716–9726 (2022)
6. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: *MICCAI*. pp. 593–603. Springer (2021)
7. Girdhar, R., Grauman, K.: Anticipative video transformer. In: *ICCV*. pp. 13505–13515 (2021)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: *First Conference on Language Modeling* (2024)
9. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research* **19**(6), 531–549 (2022)
10. Leibetseder, A., Kletz, S., Schoeffmann, K., Keckstein, S., Keckstein, J.: GLENDa: gynecologic laparoscopy endometriosis dataset. In: *MMM. Lecture Notes in Computer Science*, vol. 11962, pp. 439–450. Springer (2020)
11. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: *ECCV*. pp. 237–255. Springer (2024)
12. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. In: *ICCV*. pp. 19948–19960 (2023)
13. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: LoViT: Long video transformer for surgical phase recognition. *Medical Image Analysis* **99**, 103366 (2025)

14. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: SKiT: a fast key information video transformer for online surgical phase recognition. In: ICCV. pp. 21074–21084 (2023)
15. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: LDPolypVideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: MICCAI. pp. 387–396. Springer (2021)
16. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging* **35**(9), 2051–2063 (2016)
17. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy* **93**(4), 960–967 (2021)
18. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* **78**, 102433 (2022)
19. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: ST-Adapter: Parameter-efficient image-to-video transfer learning. *NeurIPS* **35**, 26462–26477 (2022)
20. Park, J., Lee, J., Kim, I.J., Sohn, K.: Probabilistic representations for video contrastive learning. In: ICCV. pp. 14711–14721 (2022)
21. Qian, R., Ding, S., Liu, X., Lin, D.: Static and dynamic concepts for self-supervised video representation learning. In: ECCV. pp. 145–164. Springer (2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
23. Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al.: Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge. *Medical image analysis* **70**, 101920 (2021)
24. Sganga, J., Eng, D., Graetzel, C., Camarillo, D.B.: Autonomous driving in the lung using deep learning for localization. *arXiv preprint arXiv:1907.08136* (2019)
25. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* **8**(1), 142 (2021)
26. Tian, Q., Chen, Z., Liao, H., Huang, X., Li, L., Ourselin, S., Liu, H.: EndoOmni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels. *arXiv preprint arXiv:2409.05442* (2024)
27. Tian, Q., Liao, H., Huang, X., Yang, B., Wu, J., Chen, J., Li, L., Liu, H.: BronchoTrack: Airway lumen tracking for branch-level bronchoscopic localization. *IEEE Transactions on Medical Imaging* (2024)
28. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: MICCAI. pp. 88–98. Springer (2022)
29. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS* **35**, 10078–10093 (2022)
30. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE v2: Scaling video masked autoencoders with dual masking. In: ICCV. pp. 14549–14560 (2023)
31. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: InternVideo2: Scaling foundation models for multimodal video understanding. In: ECCV. pp. 396–416. Springer (2024)

32. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: MICCAI. pp. 101–111 (2023)
33. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: MICCAI. pp. 486–496. Springer (2022)
34. Yoon, J., Lee, J., Heo, S., Yu, H., Lim, J., Song, C.H., Hong, S., Hong, S., Park, B., Park, S., et al.: hsd-b-instrument: instrument localization database for laparoscopic and robotic surgeries. In: MICCAI. pp. 393–402. Springer (2021)
35. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* pp. 1–13 (2024)
36. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: ICML (2024)