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

Qingyao Tian[1,2], Huai Liao[3], Xinyan Huang[3], Bingyu Yang[1,2], Dongdong Lei[4], Sebastien Ourselin[5], and Hongbin Liu[1,4,5]

[1] Institute of Automation, Chinese Academy of Sciences [2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] The First Affiliated Hospital, Sun Yat-sen University [4] Centre for Artificial Intelligence and Robotics, Chinese Academy of Sciences
[5] School of Engineering and Imaging Sciences, King's College London

Paper link: Code link:

BACKGROUND

- Endoscopic video analysis is emerging as a key research area.
- **Endoscopic imaging domain:** *Task-specific methods lack* generalization to new data.
- **General domain:** *Foundation models* achieve strong performance in video understanding.
- Question: How to build ***a foundation model for endoscopic videos***?

CHANLLENGES

- **Computational inefficiencies** when estimating video streams.
- **Data limitations**:
  - Data scarcity limits large-scale learning.
  - Lack of paired vision-language data restricts contrastive learning.

OVERVIEW

- We Propose **EndoMamba**, a **computational efficient** and **data efficient** endoscopic video foundation model.
- **Backbone** employs spatial bidirectional scanning and temporal causal scanning for:
  - Strong spatiotemporal modeling
  - Efficient inference
- Introduces a hierarchical **pre-training** strategy to boost representation learning.

MOTIVATION

- Why Mamba?
  1) Effectively captures **spatiotemporal representation**
  2) Enables **efficient inference** on live video stream
- Key Ideas
  - Hidden state evolution:

$$h'(t) = Ah(t) + Bx(t)$$

  - Output:

$$y(t) = Ch(t)$$

- Discrete Form

$$\boxed{ht = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = Ch_t}$$

➢ **efficient recurrent inference**

- Unroll over time:

$$y_1 = C\bar{B}x_1$$
$$y_2 = C\bar{A}\bar{B}x_1 + C\bar{B}x_2$$
$$y_3 = C\bar{A}^2\bar{B}x_1 + C\bar{A}\bar{B}x_2 + C\bar{B}x_3$$

- Recognizing Convolution
  - Each output y_t is a weighted sum of past inputs with coefficients:

$$\boxed{\begin{array}{c}\bar{K}_k = C\bar{A}^k\bar{B}, k = 0,1,2,\ldots \\ y = x * \bar{K}\end{array}}$$

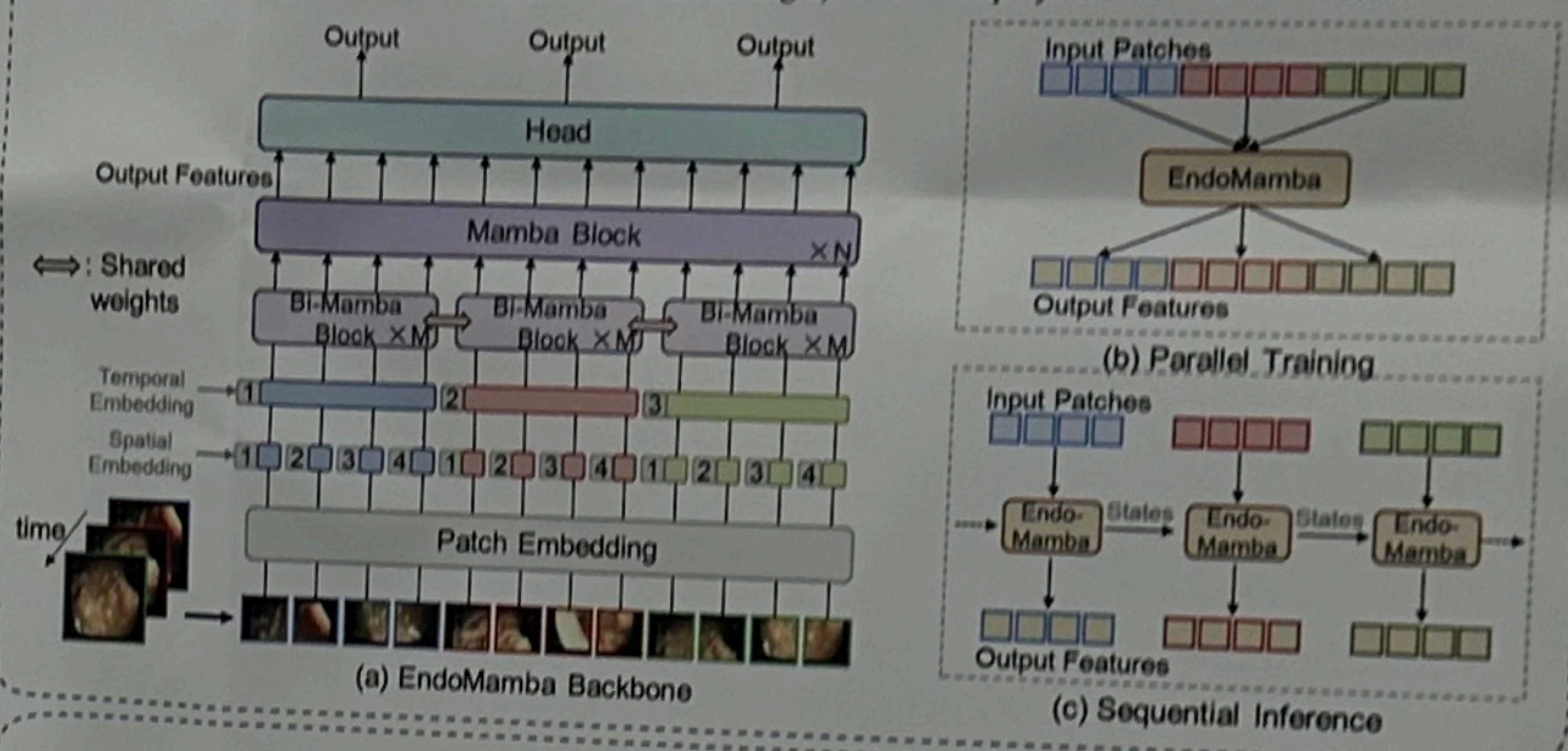
➢ **global receptive field**

➢ **efficient sequential training**

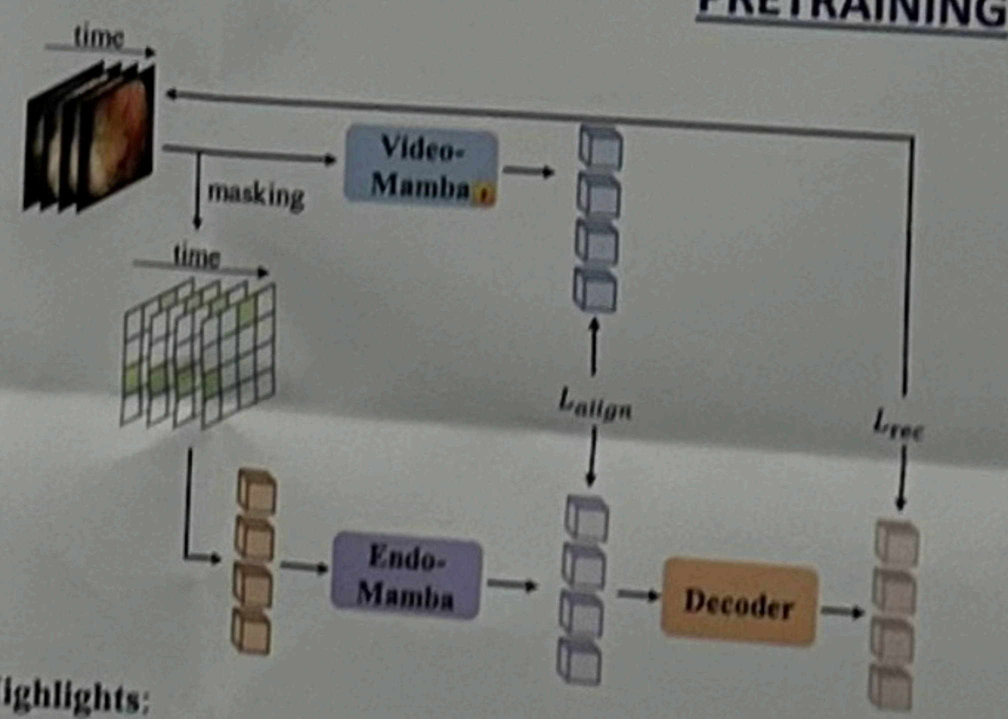
BACKBONE

Key features: 1) Bidirectional scanning within frames 2) Causal scanning across time

Highlights: 1) Strong spatiotemporal representation learning 2) Efficient deployment on live video stream



PRETRAINING



Key features:

1) Low-level video reconstruction

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

2) High-level feature alignment

$$L_{align} = 1 - \frac{1}{\Omega}\sum_{q\in\Omega} \cos\left(X_f(q), X_t(q)\right),$$

Pretraining loss

$$L = L_{rec} + \alpha L_{align}.$$

Highlights:

1) Encourages EndoMamba to capture contextual dependencies by leveraging spatiotemporal correlations.
2) Enables EndoMamba to inherit knowledge from a broader domain.

RESULTS

- Classification and segmentation

Methods	PolypDiag F1 (%)	CVC-12K Dice (%)
FAME	85.4±0.8	67.2±1.3
ProViCo	86.9±0.5	69.0±1.5
VCL	87.6±0.6	69.1±1.2
ST-Adapter	84.8±0.7	64.3±1.9
EndoFM	90.7±0.4	73.9±1.2
VideoMAEv2	87.5±1.6	72.1±0.9
VideoMamba	75.6±1.9	78.5±1.0
Ours	**95.0±1.3**	**85.4±0.2**

- Surgical phase recognition

Method	Paradigm	Video-level Accuracy↑	Jaccard↑
TeCNO	Two-stage	77.3	50.7
Trans-SVNet	Two-stage	78.3	50.7
AVT	Two-stage	77.8	50.7
LoViT	Two-stage	81.4 ± 7.6	56
SKiT	Two-stage	82.9 ± 6.8	**59.9**
VideoMAEv2	One-stage	77.0 ± 5.7	53.1
VideoMamba	One-stage	80.3 ± 8.2	54.0
EndoFM	One-stage	62.0 ± 11.9	37.4
Ours	One-stage	**83.0 ± 9.3**	56.5

- Airway anatomy detection and branch-level localization

Methods	Paradigm	Video Level Metric	Detection Metric		
		Accuracy↑	Precision↑	Recall↑	F1↑
AirwayNet	One-stage	38.7 ± 15.7	54.0	52.2	53.1
BronchoTrack	Multi-stage	57.0 ± 24.6	70.8	54.0	61.3
EndoOmni	One-stage	78.6 ± 16.2	66.4	72.2	69.2
VideoMAEv2	One-stage	78.9 ± 9.7	70.7	67.7	69.2
VideoMamba	One-stage	74.1 ± 8.3	64.7	70.3	67.4
EndoFM	One-stage	62.2 ± 19.1	59.5	54.1	56.7
Ours	One-stage	**83.0 ± 5.5**	**71.0**	**76.2**	**73.5**

- Speed analysis, measured by image tokens (P), network memory lengths (T), hidden state dimension (m).

Backbone	Param Num.	Complexity	FPS↑			
			T=16	T=32	T=64	T=128
video ViT	121.26M	$O(P^2T^2m)$	16.8	9.2	4.8	2.4
video ViT	22.26M	$O(P^2T^2m)$	70.3	27.8	9.7	2.7
VideoMamba	25.42M	$O(PTm^2)$	42.3	21.7	11.8	6.1
Ours	24.46M	$O(Pm^2)$	**47.3**	**46.7**	**47.1**	**46.6**

CONTACT

Email:
- tianqingyao2021@ia.ac.cn
- liuhongbin@ia.ac.cn

WeChat:

Join us for the oral presentation at:
Oral Session 5: Navigation and Surgical Workflows
Wednesday, September 24, 2025, 17:00 to 18:30 Main Hall