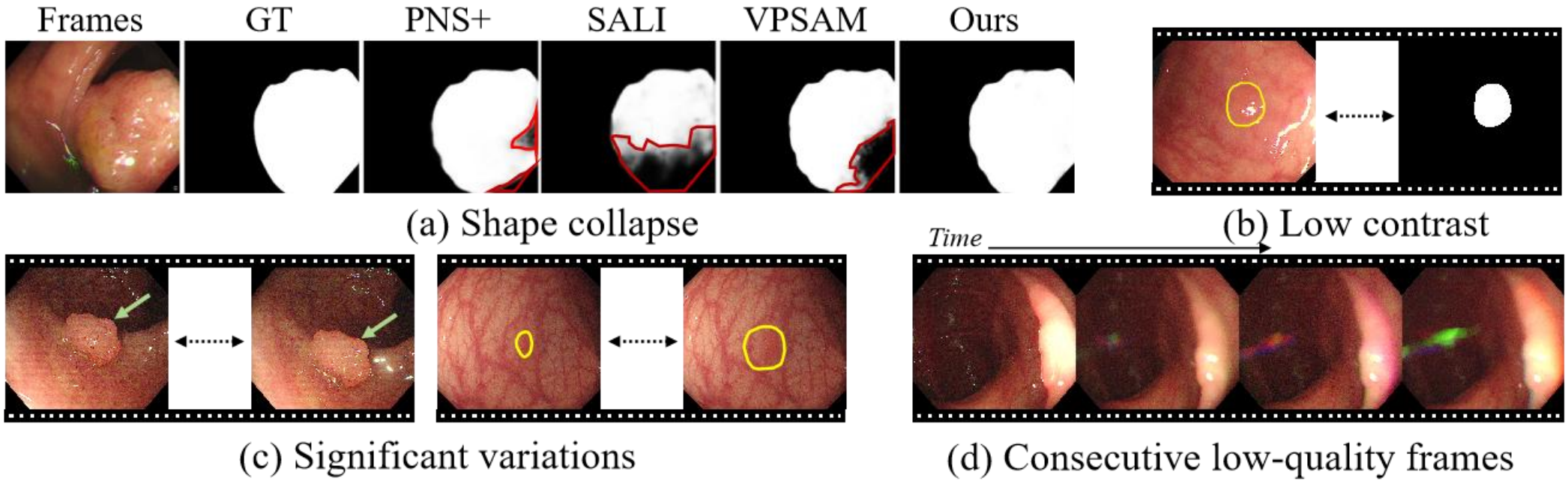
HFSTI-Net: Hierarchical Frequency-spatial-temporal Interactions for Video Polyp Segmentation

Yuanqin He[1], Guilian Chen[1], Yuhua Zhang[1], Huisi Wu[1]*, Jing Qin[2]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2400101064@mails.szu.edu.cn, hswu@szu.edu.cn

Background

- Colorectal cancer is the third most common cancer worldwide.
- Colonoscopy with polyp removal is the most effective prevention strategy.
- Diagnostic accuracy depends heavily on endoscopist experience and attention.
- Accurate, real-time video polyp segmentation is essential to assist detection, reduce missed cases, and improve consistency across clinicians.

Challanges



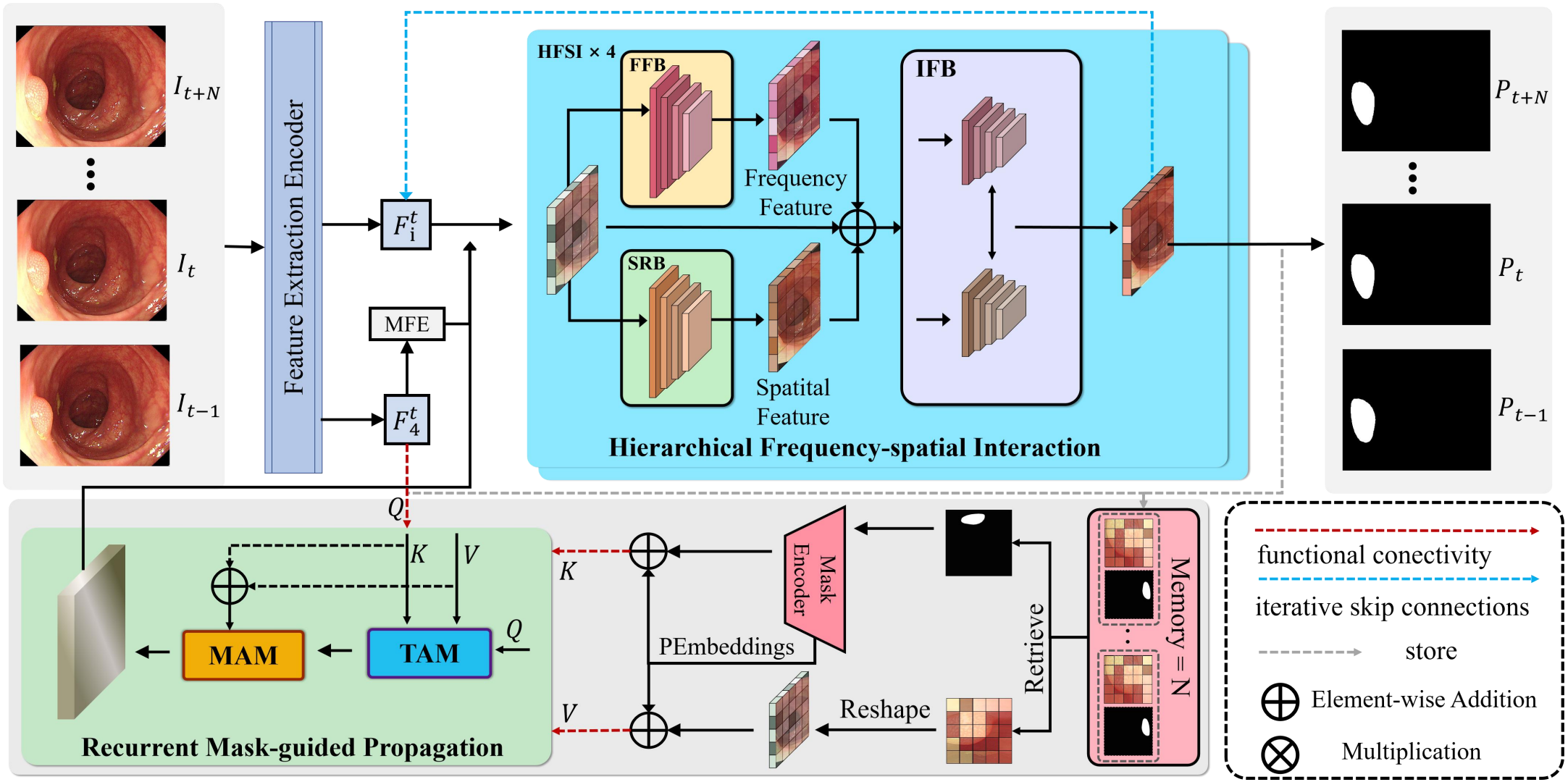
- Challenge 1: Shape Collapse Phenomenon

Low contrast causes poor boundary delineation and incomplete segmentation.

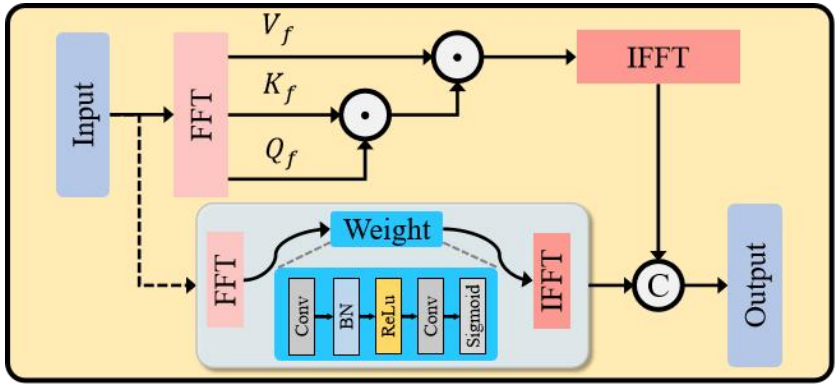
- Challenge 2: Episodic Amnesia in Videos

Motion and blur break temporal tracking, causing inconsistency and target loss.

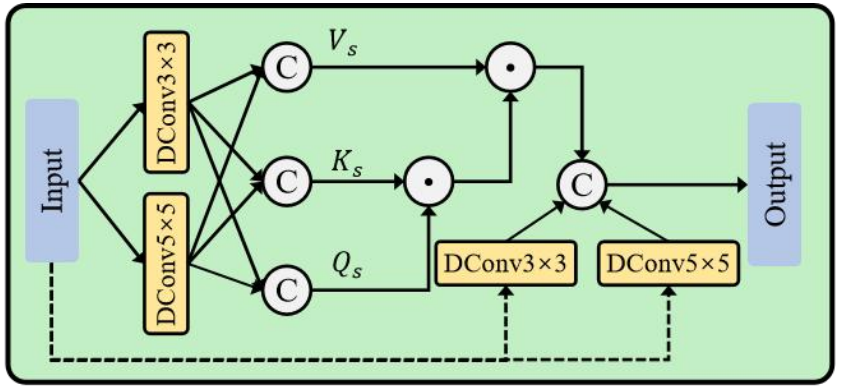
Method


I_{t+N}
I_t
I_{t-1}
Feature Extraction Encoder
F_i^t
MFE
F_4^t
HFSI × 4
FFB
SRB
Frequency Feature
Spatial Feature
IFB
Hierarchical Frequency-spatial Interaction
P_{t+N}
P_t
P_{t-1}
Q
K
V
MAM
TAM
Recurrent Mask-guided Propagation
Mask Encoder
PEmbeddings
Reshape
Retrieve
Memory = N
functional conectivity
iterative skip connections
store
Element-wise Addition
Multiplication

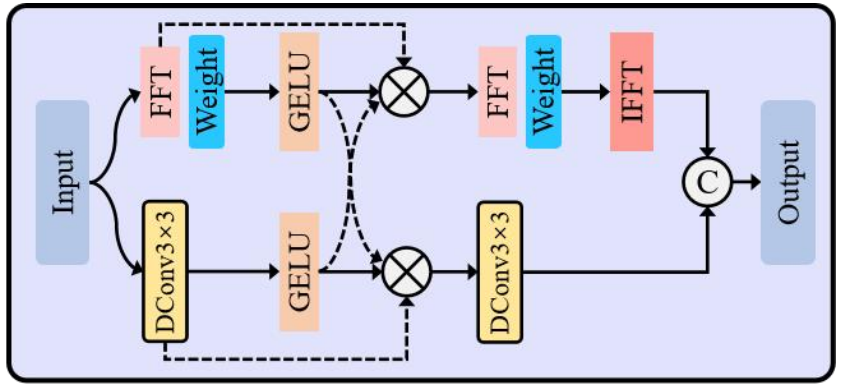
Method



(a) Frequency Filter Block (FFB)



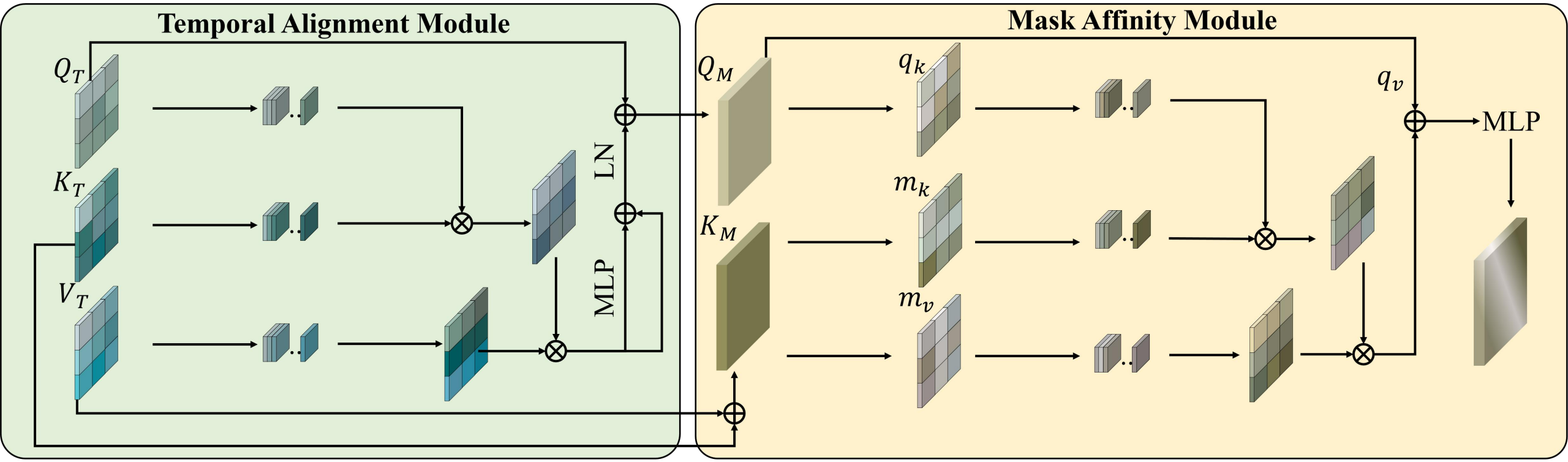
(b) Spatial Refinement Block (SRB)



(c) Interwoven Fusion Block (IFB)

- FFB applies Fast Fourier Transform to capture global semantic distributions.
- SRB preserves fine-grained edge details and local structures, enhancing boundary localization.
- IFB uses gated attention to align and fuse spectral–spatial features for accurate polyp boundary delineation.

Method



- TAM retrieves relevant context via cross-attention with memory bank.

- MAM aligns current predictions with past masks, maintaining polyp identity under blur, occlusion, or motion.

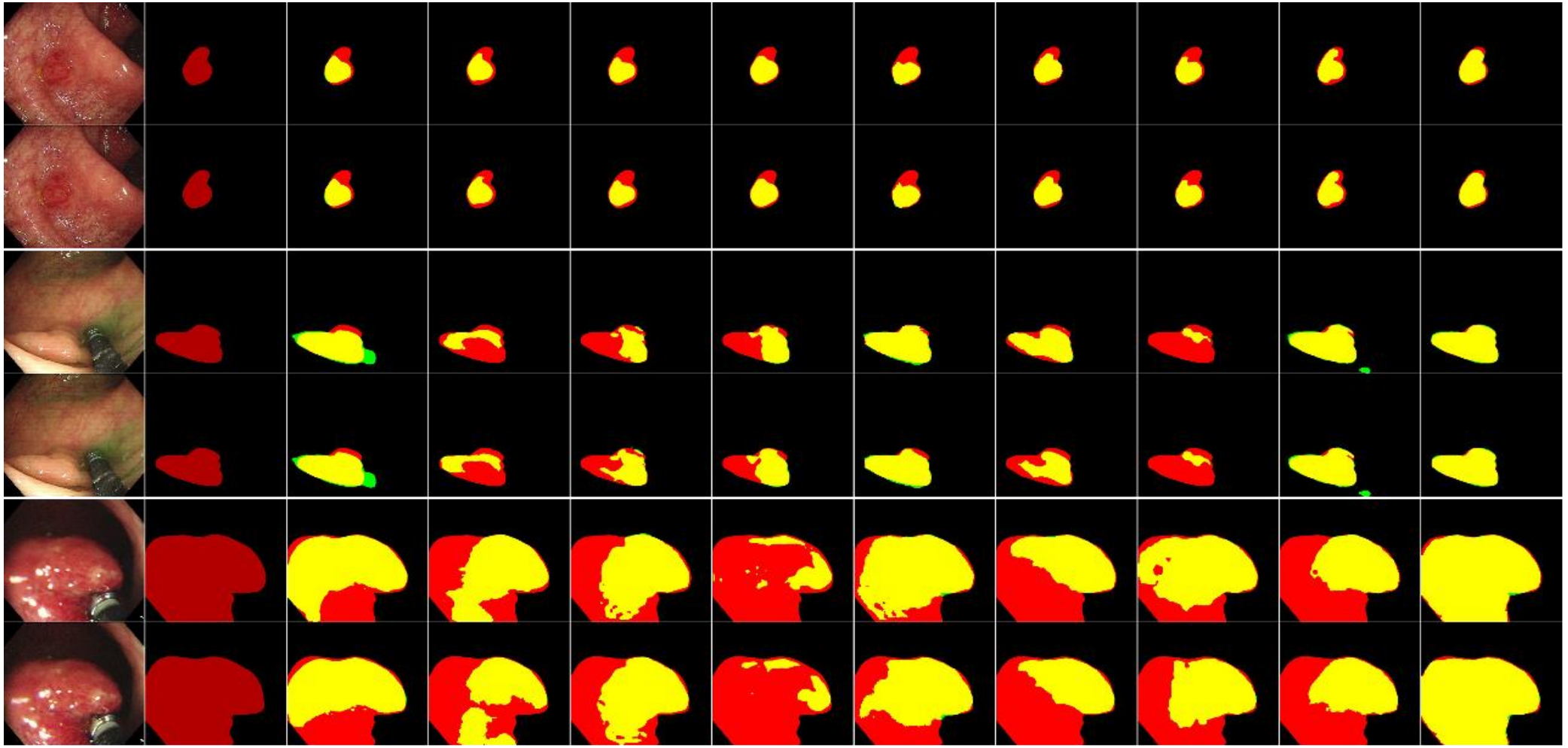
Experiment

- Evaluation spans SUN-SEG and CVC-612 datasets

- HFSTI-Net achieves 86.27% Dice on SUN-SEG-Hard, outperforming all methods and leading across Dice, IoU, and boundary metrics.

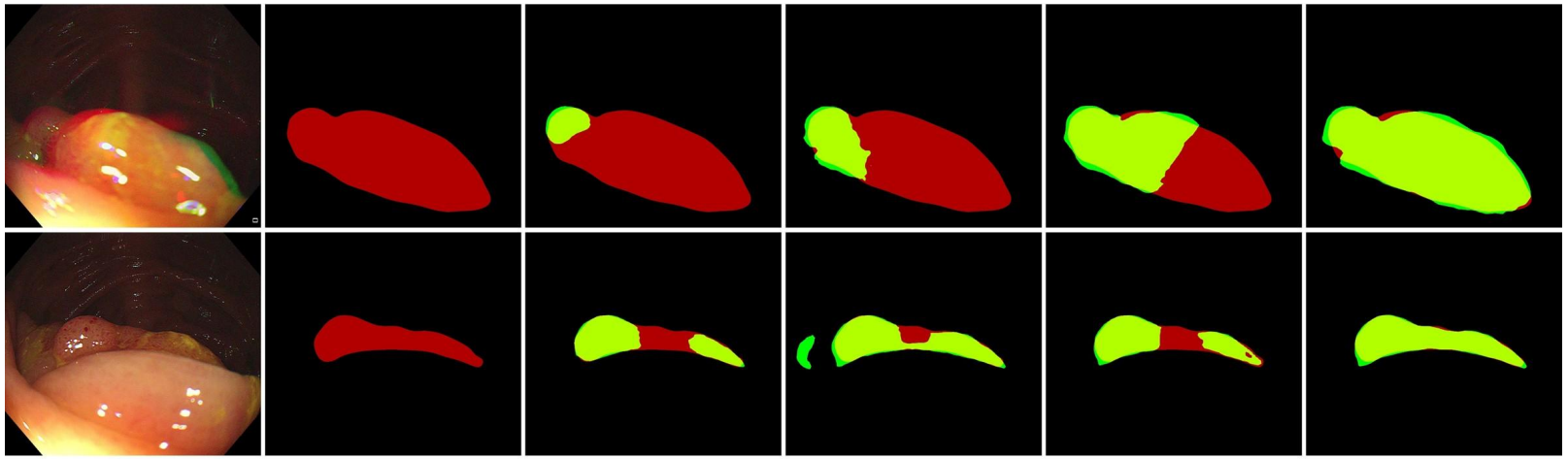
Model	Backbone	Class	SUN-SEG-Easy				SUN-SEG-Hard				CVC-612			
			S_α	E_ϕ^{mn}	F_β^w	Dice	S_α	E_ϕ^{mn}	F_β^w	Dice	S_α	E_ϕ^{mn}	F_β^w	Dice
ZoomNext	PVT-B2	NVS	88.33	90.48	80.66	85.49	87.64	90.84	80.25	83.51	94.66	97.83	92.45	93.17
SLTnet	PVT-B2	NVS	88.13	91.75	83.09	85.91	87.04	90.89	80.98	83.36	94.84	97.37	92.73	93.62
AutoSAM	VIT-B	IPS	86.28	91.67	78.36	81.28	83.57	89.93	73.59	77.37	91.52	95.38	87.47	88.73
WeakPolyp	PVT-B2	IPS	89.04	92.77	83.83	85.27	88.41	92.57	82.93	84.59	91.44	95.78	88.54	88.79
PNS+	Res-50	VPS	86.20	86.17	76.28	82.23	84.29	86.13	72.98	79.60	94.81	96.75	89.63	93.06
MAST	PVT-B2	VPS	84.53	89.81	77.04	78.43	86.17	91.42	77.76	80.32	92.03	95.38	87.47	90.84
VPSAM	VIT-B	VPS	89.31	92.34	82.86	85.62	88.93	92.13	82.98	85.28	93.26	95.75	89.63	92.33
SALI	PVT-B2	VPS	89.54	93.07	83.68	86.17	87.58	91.93	80.56	83.87	91.73	95.21	86.54	88.77
Ours	PVT-B2	VPS	**90.73**	**94.86**	**85.82**	**88.03**	**89.63**	**93.92**	**83.26**	**86.27**	**95.02**	**98.46**	**93.58**	**94.31**

Experiment


Frame
GT
ZoomNext
SLTnet
AutoSAM
WeakPolyp
PNS+
MAST
VPSAM
SALI
Ours

Ablation Study

HFSI	RMP	SUN-SEG-Easy				SUN-SEG-Hard			
		S_α	E_φ^{mn}	F_β^ω	Dice	S_α	E_φ^{mn}	F_β^ω	Dice
		89.51	92.93	83.72	86.20	88.03	92.29	80.62	83.67
	✓	89.71	93.68	84.15	87.04	88.59	92.91	81.24	84.03
✓		90.51	93.87	84.53	87.24	88.97	93.02	81.86	85.27
✓	✓	**90.73**	**94.86**	**85.82**	**88.03**	**89.63**	**93.92**	**83.26**	**86.27**



Frame GT w/o both w/ HFSI w/ RMP w/ both

Conclusion

- HFSTI-Net pioneers joint frequency-spatial-temporal modeling for VPS.
- HFSI eliminates shape collapse via spectral fusion;
- RMP cures episodic amnesia through explicit memory tracking, achieving breakthrough performance.
- State-of-the-art accuracy combined with 31.27 FPS real-time speed overcomes historical efficiency-precision trade-offs.

ICLR

THANK YOU