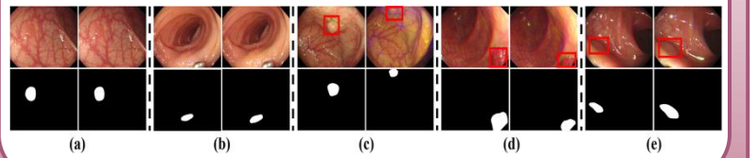
WavePolyp: Video Polyp Segmentation via Hierarchical Wavelet-Based Feature Aggregation and Inter-Frame Divergence Perception

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

ICLR

Motivation

- Automatic Video Polyp Segmentation (VPS) is crucial for early colorectal cancer diagnosis, but real-world colonoscopy videos present three major challenges:
- Camouflage. Polyps exhibit high visual similarity with surrounding mucosa, making precise boundary delineation difficult.
- Inter-frame Divergence. Significant variations in size, shape, and location across frames lead to temporal inconsistency and unstable predictions.
- Real-time Constraint. Clinical applications require efficient and fast inference for timely decision support.

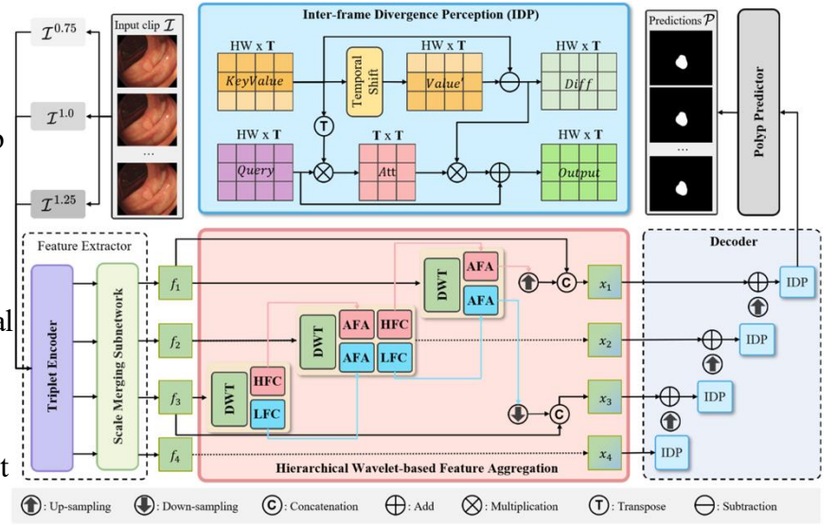


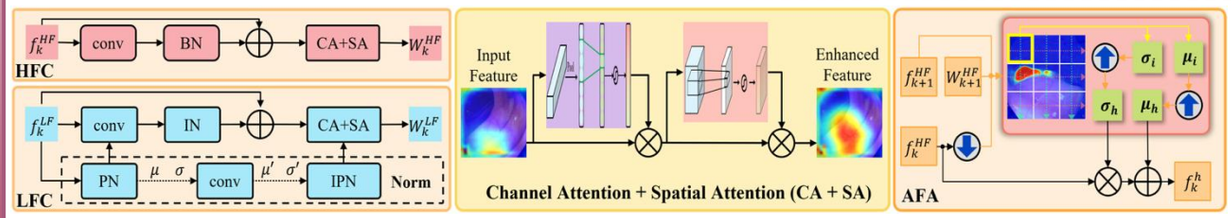
Contributions

- A novel VPS framework that jointly models intra-frame discrimination and inter-frame divergence for accurate and stable segmentation.
- Two key designs: HWFA leverages frequency-domain decomposition to enhance discriminative spatial features. IDP models temporal divergence to ensure consistent tracking across frames.
- Achieves state-of-the-art performance on SUN-SEG and CVC-612 while maintaining real-time inference speed (~23 FPS).

Method

- WavePolyp jointly models intra-frame discrimination and inter-frame divergence, which enables accurate polyp localization and stable video polyp segmentation.
- HWFA exploits frequency-domain decomposition to enhance discriminative spatial features.
- IDP captures temporal variations via divergence modeling, ensuring consistent tracking.

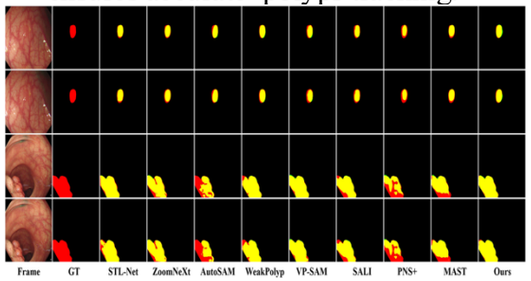




- Multi-level features are decomposed into high-frequency (HF) and low-frequency (LF) components via wavelet transform in the HWFA module, capturing complementary cues. HFC and LFC further refine texture details and global structures, respectively.
AFA hierarchically aggregates frequency-aware features across scales, enhancing intra-frame discriminative representations for camouflaged polyps.

Experiment

- Even under low contrast, rapid motion, and significant inter-frame variations, WavePolyp preserves fine details and ensures accurate polyps tracking.



- Achieves the best performance across all major metrics (Dice, Sα, Eϕ, Fβ) on both SUN-SEG and CVC-612 datasets.

Model	Backbone	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
SLT-Net	PVTv2-B5	90.39	93.75	84.35	87.15	88.06	92.05	80.31	83.55	94.61	97.70	92.06	92.96
ZoomNeXt	PVTv2-B5	89.81	92.25	84.64	87.55	88.51	91.34	82.21	85.22	94.54	97.53	90.91	92.73
AutoSAM	ViT-B	86.28	91.69	78.25	81.27	83.59	89.91	73.08	77.25	90.56	96.19	87.68	88.34
WeakPolyp	PVTv2-B5	90.51	93.72	84.89	87.57	90.19	93.77	83.74	86.73	91.51	95.18	88.74	89.07
PNS+	Res2Net-50	85.75	86.11	76.14	81.91	83.98	85.68	72.75	79.32	94.49	96.44	89.05	92.54
MAST	PVTv2-B2	87.91	92.87	81.40	84.44	87.44	92.82	80.27	83.79	93.54	96.07	89.93	90.12
SALI	PVTv2-B5	89.45	93.01	83.73	86.07	89.19	93.21	83.05	85.54	94.41	97.12	92.16	93.01
VP-SAM	ViT-B	90.49	93.68	85.64	88.19	90.03	93.74	83.30	86.94	94.68	97.83	92.51	93.45
Ours	PVTv2-B5	**90.93**	**94.15**	**86.74**	**88.96**	**90.28**	**93.96**	**85.17**	**87.55**	**95.24**	**98.61**	**93.57**	**94.36**

Conclusion

For more details, please read our paper.
Contact Email: *hswu@szu.edu.cn*
Codes Source:
https://github.com/FishballZhang/WavePolyp.