

# CMSA-Net: Causal Multi-scale Aggregation with Adaptive Multi-source Reference for Video Polyp Segmentation

Tong Wang<sup>1 2</sup>, Yaolei Qi<sup>1</sup>, Siwen Wang<sup>2</sup>, Imran Razzak<sup>2</sup>, Guanyu Yang<sup>1</sup>✉, Yutong Xie<sup>2</sup>✉

Code



<sup>1</sup> Southeast University, China

<sup>2</sup> Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE

Paper

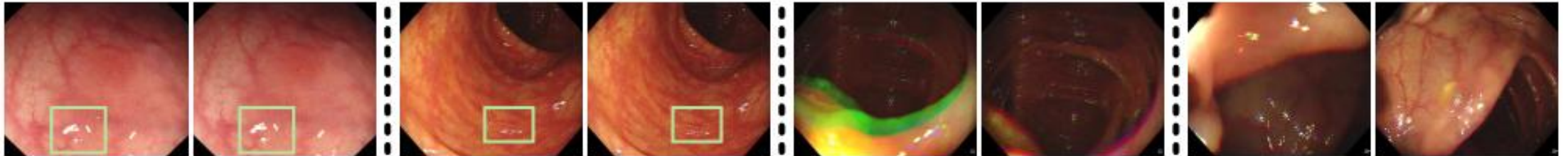


# Clinical Problem

---

**Challenge 1:** Weak semantic discrimination, caused by the low contrast between polyps and background.

**Challenge 2:** Large spatio-temporal variations across frames.

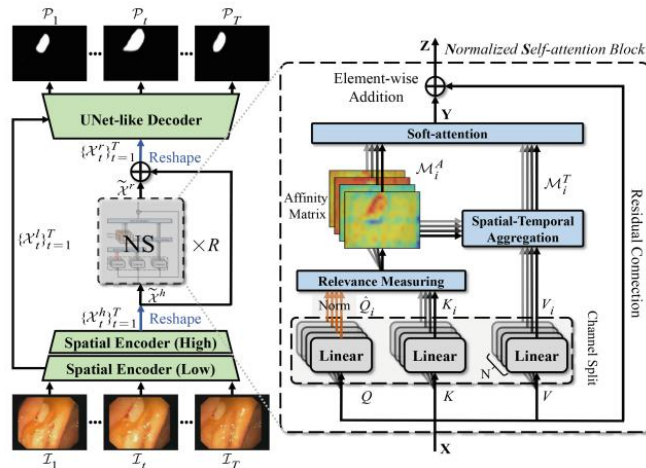


*Weak semantic discrimination*

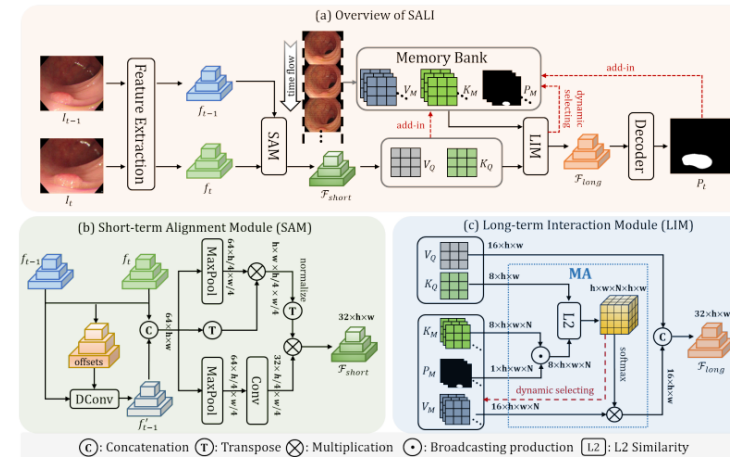
*Large spatio-temporal variation*



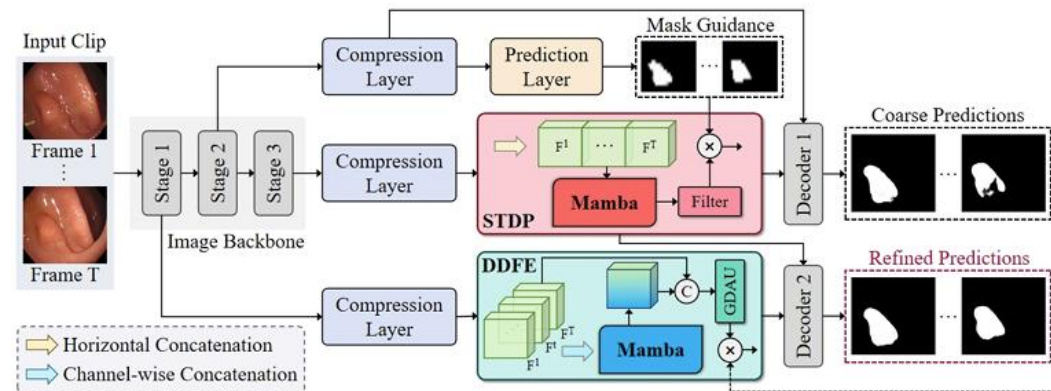
# Related Works



**PNS-Net (MICCAI 2021)**



**SALI (MICCAI 2024)**



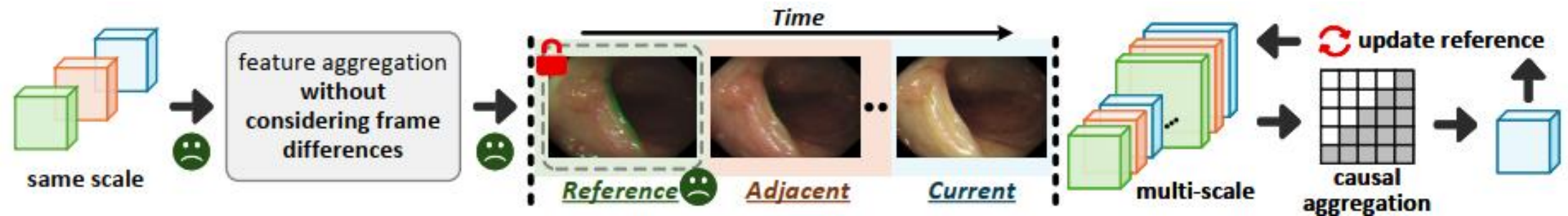
**STDD-Net (ICCV 2025)**



# Motivation

**Limitation 1:** Existing feature fusion methods for polyp semantic learning, which ignore multi-scale information and intrinsic identity relationships.

**Limitation 2:** Current approaches adopt a fixed single-source reference, failing to capture dynamic and diverse spatiotemporal cues.



*Current polyp semantic learning method*

*Current fixed single-source reference method*

*Our Method*



# Motivation

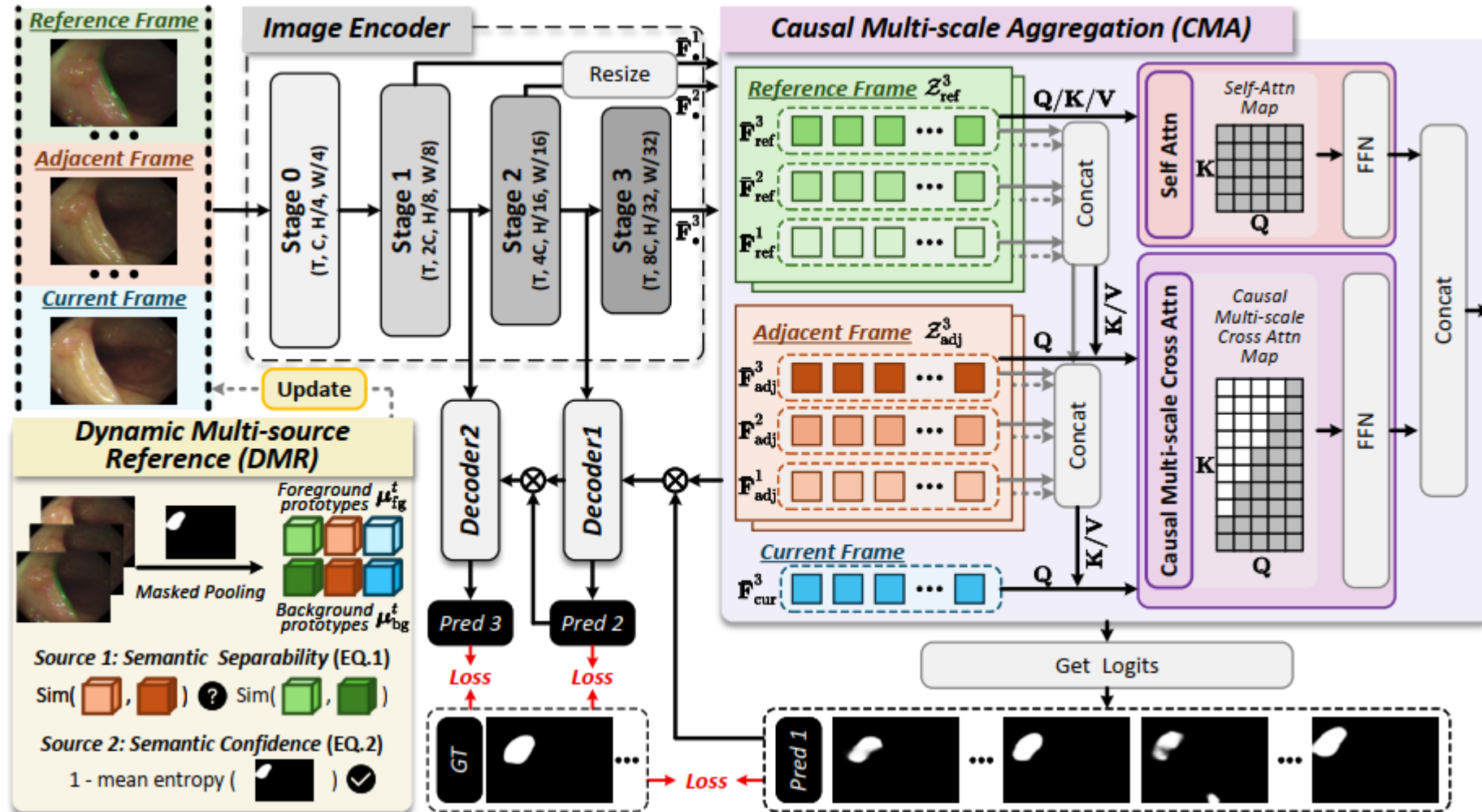
---

These observations motivate a more adaptive spatio-temporal modeling paradigm for video polyp segmentation:

1. **Multi-scale semantic modeling** is needed to distinguish low-contrast polyps from visually similar backgrounds.
2. **Frame-aware temporal propagation** is required to preserve useful historical information while suppressing noisy or inconsistent cues.
3. **Adaptive reference updating** is necessary to handle evolving target appearance and maintain reliable temporal guidance.



# Method Overview



# Method Overview

---

We introduce **CMSA-Net**, which integrates **Causal Multi-scale Aggregation (CMA)** and **Dynamic Multi-source Reference (DMR)**.

1. **CMA** aggregates multi-scale temporal cues from reference and adjacent frames under a causal constraint, enhancing current-frame polyp semantics while avoiding feature contamination.
2. **DMR** adaptively updates reliable multi-source references according to semantic separability and prediction confidence, providing robust and efficient guidance for real-time video polyp segmentation.



# Quantitative Comparisons

**Table 1.** Quantitative comparison on SUN-SEG-Easy dataset. The best is **bold**.

Method	Publication	Backbone	SUN-SEG-Easy-Seen (%)						SUN-SEG-Easy-Unseen (%)					
			$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice $\uparrow$	IoU $\uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice $\uparrow$	IoU $\uparrow$	MAE $\downarrow$
COSNet [16]	TPAMI'19	-	84.5	83.6	72.7	73.0	64.8	3.4	65.4	60.0	43.1	42.3	34.2	7.3
PCSA [11]	AAAI'20	-	85.2	83.5	68.1	70.9	60.4	3.9	68.0	66.0	45.1	45.0	35.3	7.8
2/3D [19]	MICCAI'20	-	89.5	90.9	81.9	82.9	75.6	2.1	78.6	77.7	65.2	65.6	57.0	4.4
PraNet [7]	MICCAI'20	Res2Net-50	91.8	94.2	87.7	88.3	82.5	2.0	78.1	78.8	66.3	66.5	58.2	5.2
ACSNet [27]	MICCAI'20	ResNet-34	92.0	94.2	87.4	88.2	82.8	1.7	77.2	76.6	63.0	63.8	56.4	4.6
SANet [22]	MICCAI'21	Res2Net-50	91.6	93.3	86.6	87.2	82.0	1.8	75.0	72.8	59.0	59.3	52.4	5.2
SEPNet [20]	TCSVT'24	PVTv2-B2	93.1	96.2	88.3	89.6	83.4	1.7	82.9	88.3	73.5	75.1	66.6	4.2
PNS [13]	MICCAI'21	Res2Net-50	90.6	91.0	83.6	84.1	78.3	2.0	76.7	74.4	61.6	61.8	54.5	4.8
PNS+ [14]	MIR'22	Res2Net-50	91.7	92.5	84.8	85.5	78.7	2.1	80.6	79.8	67.6	67.8	59.1	4.4
MAST [3]	Arxiv'24	PVTv2-B2	92.5	96.2	87.8	89.3	82.7	1.6	83.2	89.4	74.9	77.0	67.4	3.7
SALI [12]	MICCAI'24	PVTv2-B2	90.2	93.2	84.9	85.8	78.9	2.4	73.1	75.2	58.7	59.2	50.2	6.3
SALI [12]	MICCAI'24	PVTv2-B5	90.7	93.7	85.1	86.2	79.6	2.2	77.1	82.1	64.6	65.6	56.8	5.5
STDDNet [4]	ICCV'25	Res2Net-50	93.5	96.0	89.7	90.5	85.0	1.5	81.7	83.0	72.1	72.4	64.3	3.7
STDDNet [4]	ICCV'25	PVTv2-B2	94.1	96.9	90.5	91.5	86.1	1.4	86.0	<b>90.3</b>	78.6	80.1	72.4	3.4
Ours	-	Res2Net-50	94.5	97.3	90.5	91.9	86.5	1.2	84.4	90.1	75.3	77.5	69.2	3.5
Ours	-	PVTv2-B2	<b>95.1</b>	<b>97.5</b>	<b>91.6</b>	<b>92.6</b>	<b>87.6</b>	<b>1.1</b>	<b>86.7</b>	<b>90.3</b>	<b>79.3</b>	<b>80.3</b>	<b>72.6</b>	<b>2.9</b>

We evaluate CMSA-Net on four **SUN-SEG** subsets: Easy-Seen, Easy-Unseen, Hard-Seen, and Hard-Unseen.



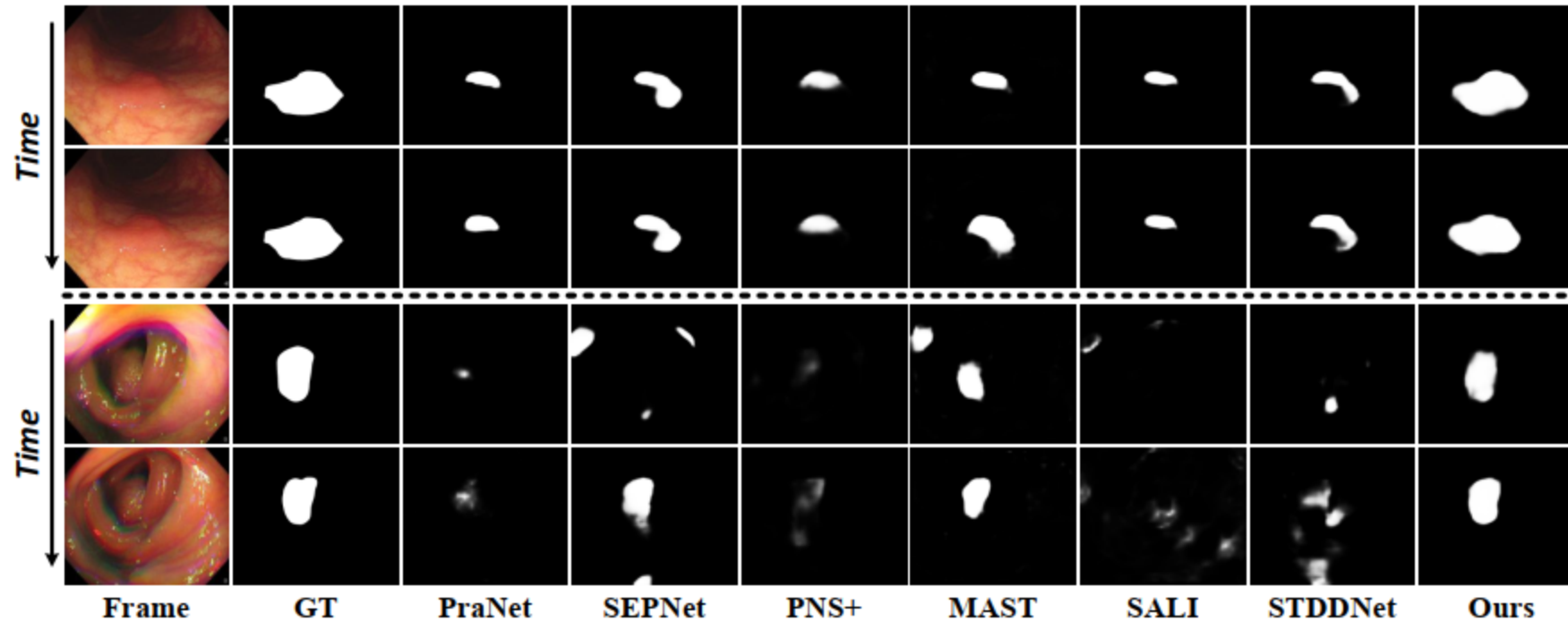
# Quantitative Comparisons

**Table 2.** Quantitative comparison on SUN-SEG-Hard dataset. The best is **bold**.

Method	Publication	Backbone	SUN-SEG-Hard-Seen (%)						SUN-SEG-Hard-Unseen (%)					
			$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice $\uparrow$	IoU $\uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice $\uparrow$	IoU $\uparrow$	MAE $\downarrow$
COSNet [16]	TPAMP'19	-	78.5	77.2	62.6	63.3	54.1	4.6	67.0	62.7	44.3	43.8	35.3	7.0
PCSA [11]	AAAI'20	-	77.2	75.9	56.6	58.5	47.9	5.7	68.2	66.0	44.2	45.0	35.1	8.0
2/3D [19]	MICCAI'20	-	84.9	86.9	75.3	76.4	67.1	3.5	78.6	77.5	63.4	64.4	55.8	4.4
PraNet [7]	MICCAI'20	Res2Net-50	88.4	91.9	83.1	83.9	76.6	3.1	78.7	80.2	66.7	67.5	58.7	5.3
ACSNet [27]	MICCAI'20	ResNet-34	87.2	91.0	80.6	82.0	74.8	3.6	76.2	77.6	61.0	62.4	54.7	5.3
SANet [22]	MICCAI'21	Res2Net-50	87.4	90.5	81.0	82.0	74.8	3.3	75.3	73.6	59.0	59.5	52.7	5.5
SEPNet [20]	TCSVT'24	PVTv2-B2	89.4	94.0	83.5	85.7	77.6	3.4	84.7	89.5	74.5	77.4	68.4	3.9
PNS [13]	MICCAI'21	Res2Net-50	87.0	89.2	78.7	79.6	72.1	3.3	76.7	75.5	60.9	61.5	53.9	5.0
PNS+ [14]	MIR'22	Res2Net-50	88.7	90.2	80.6	81.3	72.8	3.0	79.8	79.3	65.4	66.1	57.1	5.0
MAST [3]	Arxiv'24	PVTv2-B2	89.2	94.2	83.2	85.3	76.7	2.6	85.6	91.3	77.2	79.9	70.8	3.1
SALI [12]	MICCAI'24	PVTv2-B2	86.8	90.9	79.9	81.0	72.6	3.4	72.8	75.9	56.9	57.9	48.7	6.8
SALI [12]	MICCAI'24	PVTv2-B5	86.6	91.0	79.7	81.0	72.9	3.8	76.5	81.3	62.0	63.6	54.7	5.7
STDDNet [4]	ICCV'25	Res2Net-50	91.3	95.2	86.9	88.1	81.0	2.3	83.4	85.6	74.1	75.0	67.3	3.7
STDDNet [4]	ICCV'25	PVTv2-B2	91.1	95.0	86.0	87.8	80.6	2.8	86.3	90.2	78.1	80.2	72.2	3.5
Ours	-	Res2Net-50	<b>92.7</b>	<b>96.1</b>	<b>87.1</b>	<b>89.8</b>	<b>83.0</b>	<b>1.8</b>	85.1	89.8	75.0	78.0	69.5	3.6
Ours	-	PVTv2-B2	92.3	95.6	<b>87.1</b>	88.9	82.1	1.9	<b>87.3</b>	<b>91.0</b>	<b>79.6</b>	<b>81.3</b>	<b>73.7</b>	<b>2.9</b>



# Qualitative Comparisons

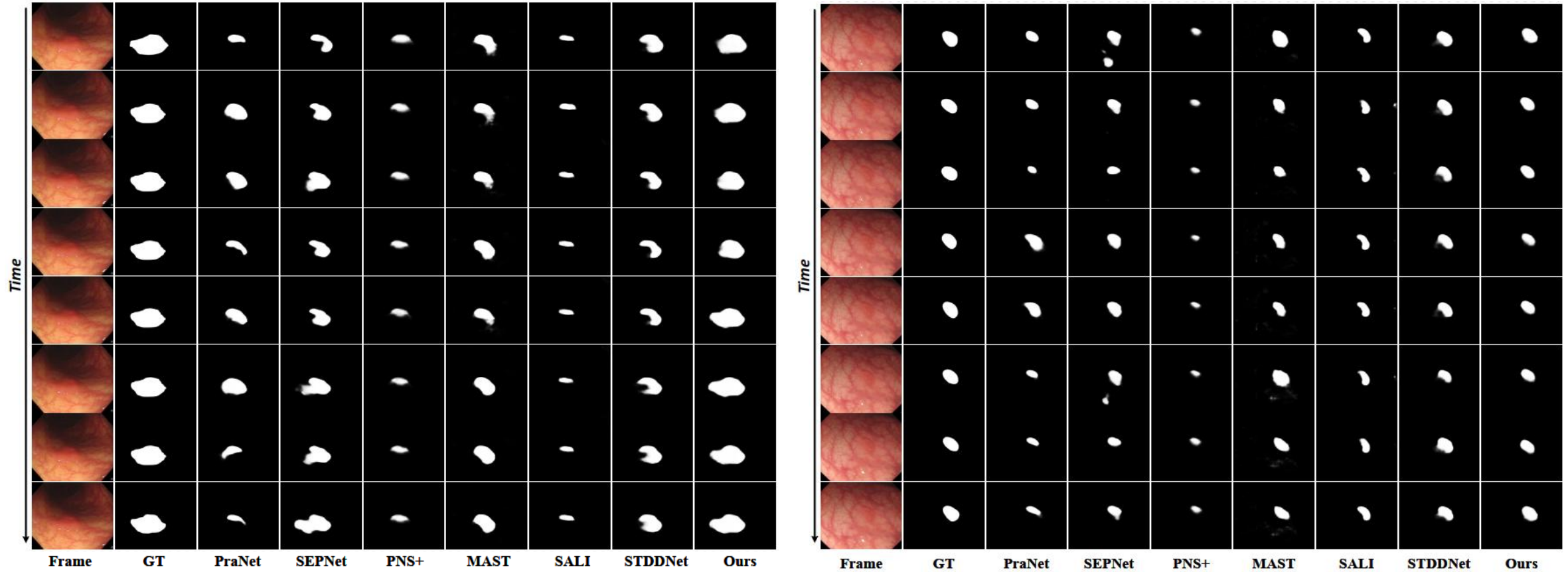


**Fig. 3.** Qualitative Comparisons. Upper (case91\_2): a sequence of consecutive low-contrast frames; Lower (case32\_4): significant variations between adjacent frames.

CMSA-Net produces more accurate and complete masks in challenging cases with low contrast and large inter-frame variations.



# Qualitative Comparisons



# Ablation Study

**Table 3.** Comparison of model parameters and inference time (clip = 6).

Efficiency	PraNet	ACSNet	SEPNet	PNS+	SALI	STDDNet (R)	STDDNet (P)	Ours (R)	Ours (P)
GFLOPs ↓	78.89	130.52	75.12	68.58	86.58	75.17	64.18	73.36	59.42
Param. (M) ↓	30.50	29.45	25.96	9.8	82.73	38.16	29.67	29.61	25.79
FPS ↑	42	21	33	64	10	45	36	47	38

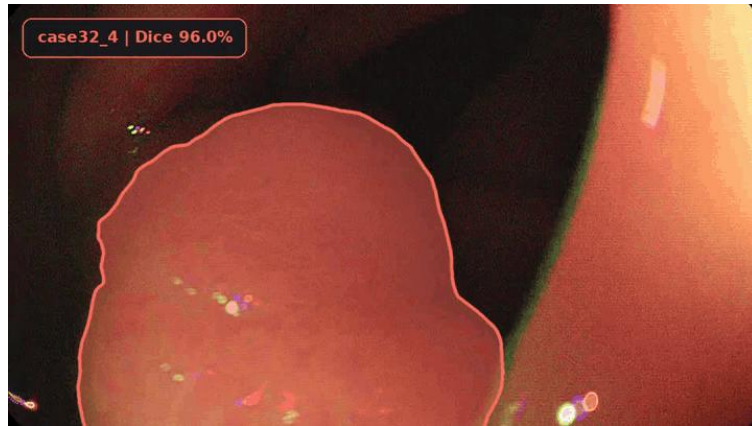
**Table 4.** Ablation study of key components. The best is **bold**.

Methods	SUN-SEG-Easy-Seen					SUN-SEG-Easy-Unseen					SUN-SEG-Hard-Seen					SUN-SEG-Hard-Unseen				
	$S_\alpha$	$E_\phi^{mn}$	$F_\beta^w$	Dice	IoU	$S_\alpha$	$E_\phi^{mn}$	$F_\beta^w$	Dice	IoU	$S_\alpha$	$E_\phi^{mn}$	$F_\beta^w$	Dice	IoU	$S_\alpha$	$E_\phi^{mn}$	$F_\beta^w$	Dice	IoU
CMSA-Net	<b>95.1</b>	<b>97.5</b>	<b>91.6</b>	<b>92.6</b>	<b>87.6</b>	<b>86.7</b>	<b>90.3</b>	<b>79.3</b>	<b>80.3</b>	<b>72.6</b>	<b>92.3</b>	<b>95.6</b>	<b>87.1</b>	<b>88.9</b>	<b>82.1</b>	<b>87.3</b>	<b>91.0</b>	<b>79.6</b>	<b>81.3</b>	<b>73.7</b>
w/o CMA	92.4	94.8	87.2	88.5	82.1	78.1	79.5	64.4	65.2	56.5	88.5	92.1	81.3	83.1	74.9	76.4	78.2	61.8	62.9	54.4
w/o DMR	93.8	96.2	89.1	90.3	84.7	80.2	81.6	67.5	68.8	60.3	90.4	94.2	84.8	86.6	79.1	79.1	80.2	65.3	67.0	58.4
w/o CMA+DMR	90.6	92.9	84.0	85.3	79.0	72.5	72.1	55.1	55.8	47.6	85.8	89.5	77.7	79.2	71.0	71.6	71.5	53.4	54.8	46.3
w/o Multi-scale	93.2	95.8	88.6	89.7	83.9	83.5	86.9	75.1	76.2	67.4	89.8	93.5	84.2	86.1	78.4	82.9	87.5	74.1	75.8	67.1
w/o Causal Attn	94.2	96.9	90.1	91.3	85.6	84.5	88.2	76.8	78.1	69.8	91.2	95.2	85.9	87.7	80.5	85.3	89.4	76.7	78.5	70.2
w/o Multi-source	94.0	96.7	89.9	91.1	85.5	86.7	90.7	79.7	80.9	73.0	91.0	95.0	85.6	87.5	80.3	86.9	89.9	78.4	80.4	72.6

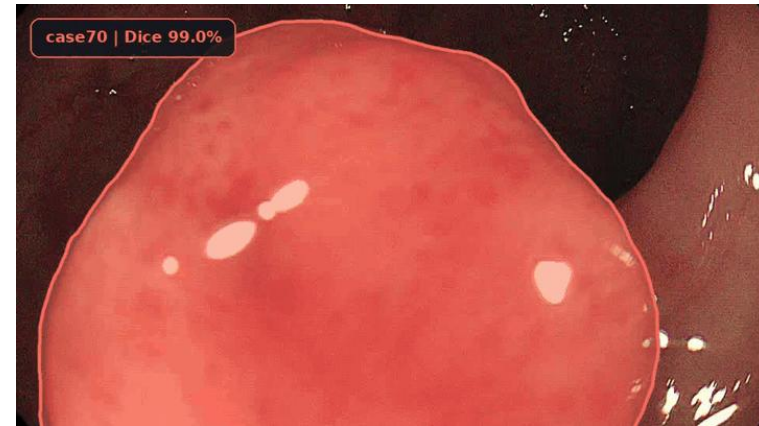
Ablation results show that removing CMA, DMR, or both significantly degrades performance, confirming their complementary contributions.



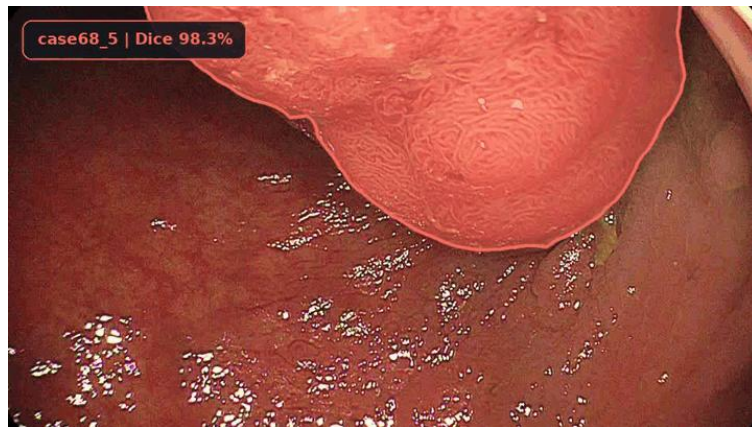
# Visualization (Video)



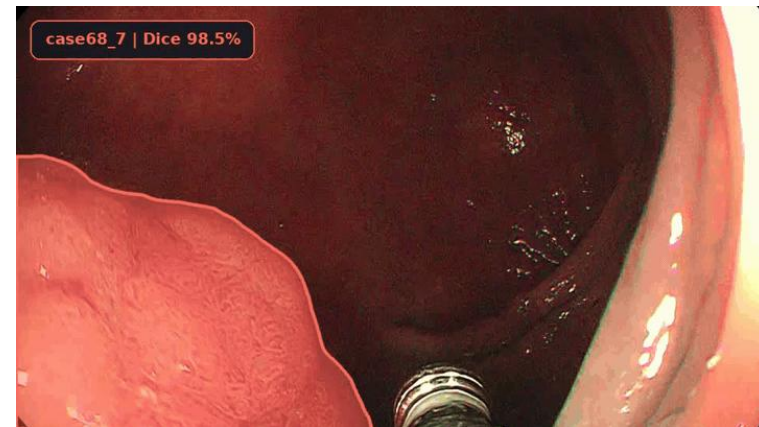
*case32\_4*



*case70*



*Case68\_5*



*Case68\_7*



# Resources

---

Home Page



Code



Paper



Checkpoint



Predicted Map



**Thank You !**

