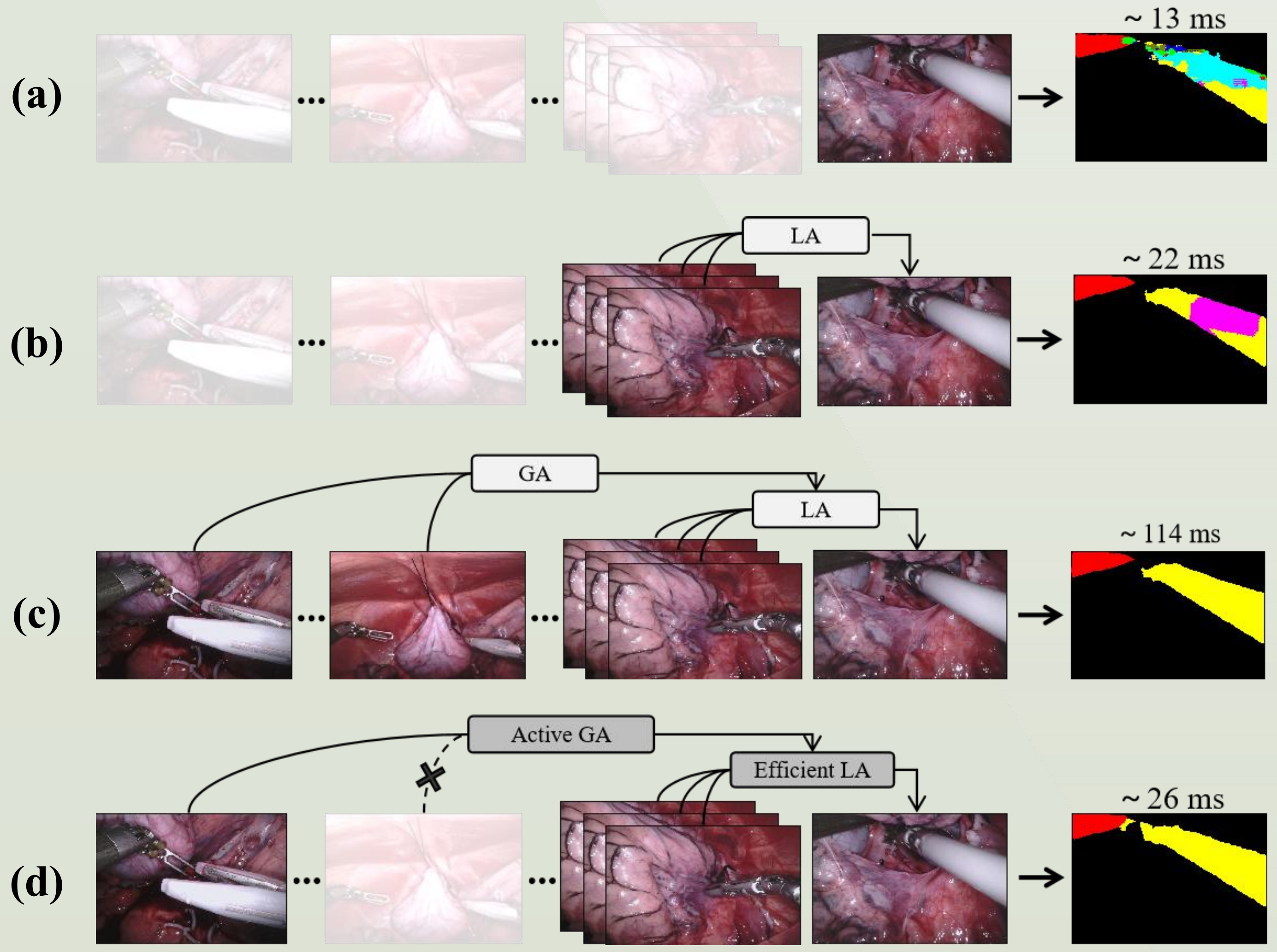
Efficient Global-Local Memory for Real-time Instrument Segmentation of Robotic Surgical Video

Jiacheng Wang [1], Yueming Jin [2], Liansheng Wang [1](☒), Shuntian Cai [3], Pheng-Ann Heng [2], Jing Qin [4]

[1] Xiamen University, [2] The Chinese University of Hong Kong, [3] Zhongshan Hospital affiliated to Xiamen University, [4] The Hong Kong Polytechnic University

INTRODUCTION

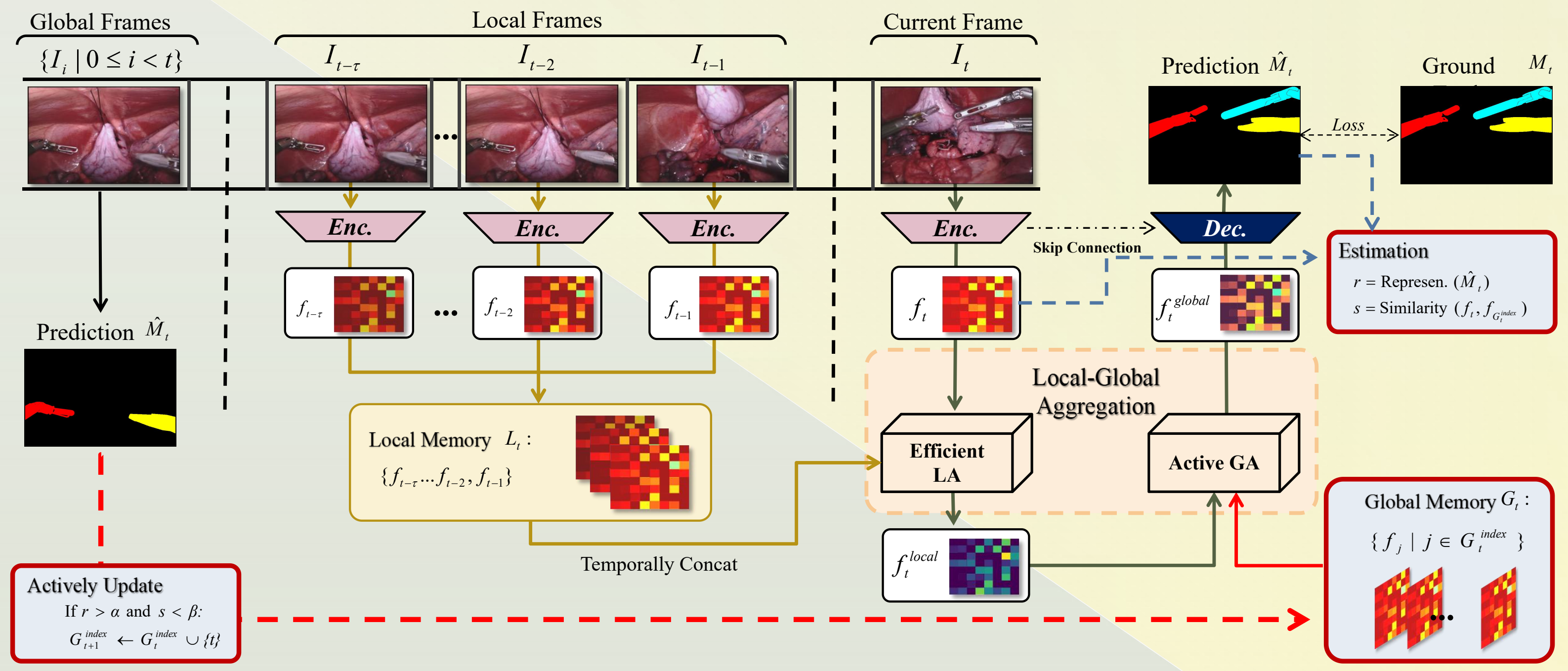
Robotic-assisted surgery has greatly improved the surgeon performance and patient safety. Semantic segmentation of instrument segmentation, aiming to separate instrument and identify its sub-type and parts, serves as an essential prerequisite in various applications in assisted surgery. Achieving high segmentation accuracy while with low latency for real-time prediction is vital in the real-world deployment. However, fast and accurate instrument segmentation from surgical video is very challenging, due to the complicated surgical scene, various lighting conditions, incomplete and distorted instrument structure caused by small FoV of endoscopic camera and inevitable visual occlusion by blood, smoke or instrument overlap.



(a) Most existing methods ignore the valuable clues in **temporal dimension**. It's hard to accurately recognize the correct type of instrument.
(b) Aggregating the local information (LA), with **unsatisfactory predictions** in a challenging case for instrument type segmentation.
(c) Putting feature maps of all the previous frames into the global memory and randomly select some samples when activating it has **significant memory-cost**.
(d) We propose a novel dual-memory network (DMNet) to wisely relate both **global and local spatio-temporal knowledge** to augment the current features, boosting the segmentation performance and retaining the real-time prediction capability.

METHODS

In this paper, we propose a novel Dual-Memory Network (DMNet) for achieving accurate and real-time instrument segmentation from surgical videos, by holistically and efficiently aggregating spatio-temporal knowledge. The dual-memory framework are based on two important intuitions for humans to perceive instruments in videos, i.e., local temporal dependence and global semantic information, therefore more temporal knowledge can be transferred to current semantic representation. More importantly, we are the first trials that enable such holistic aggregation in real-time setting by carefully considering the properties of these two-level horizons.

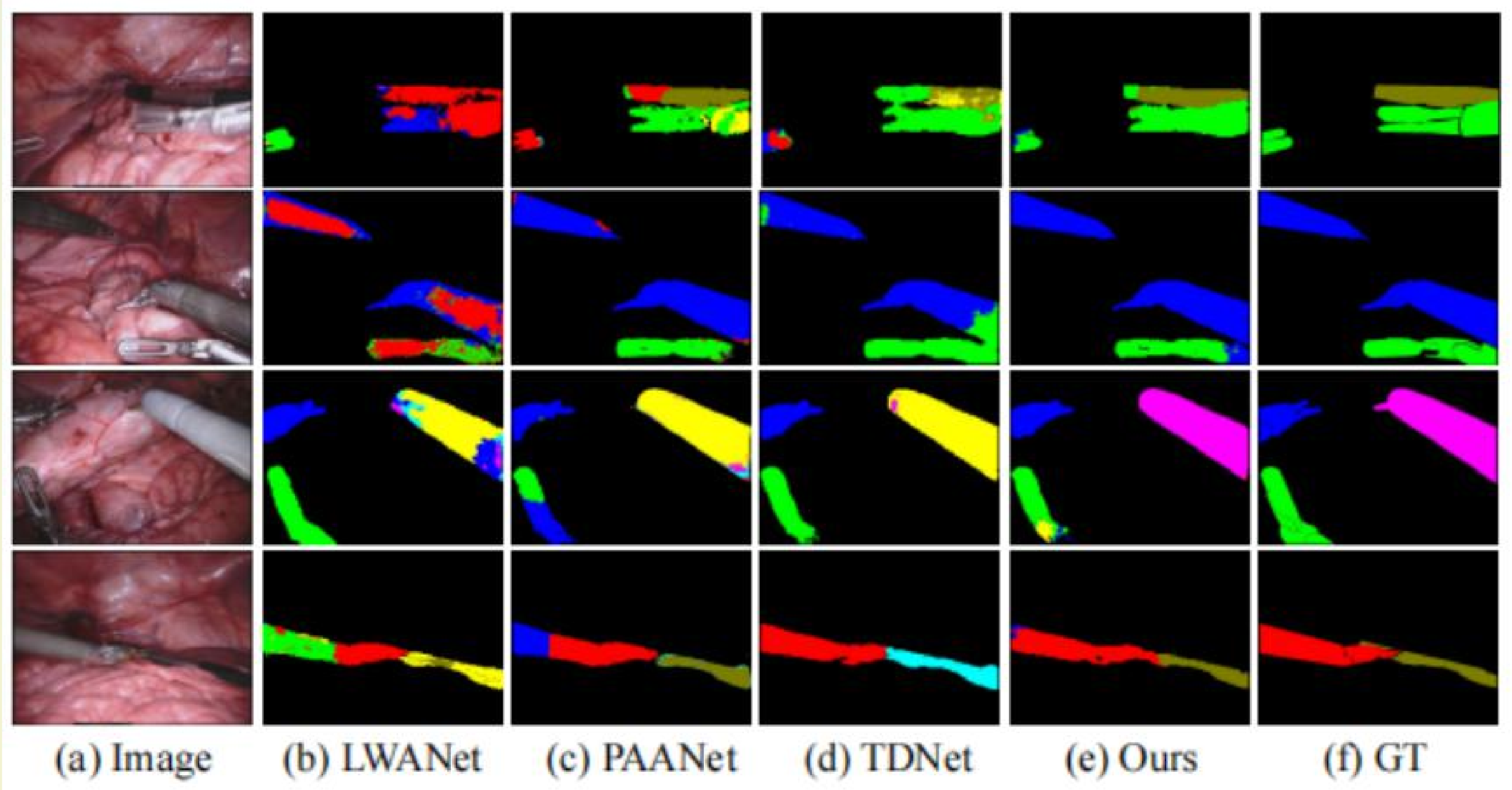


EXPERIMENTS

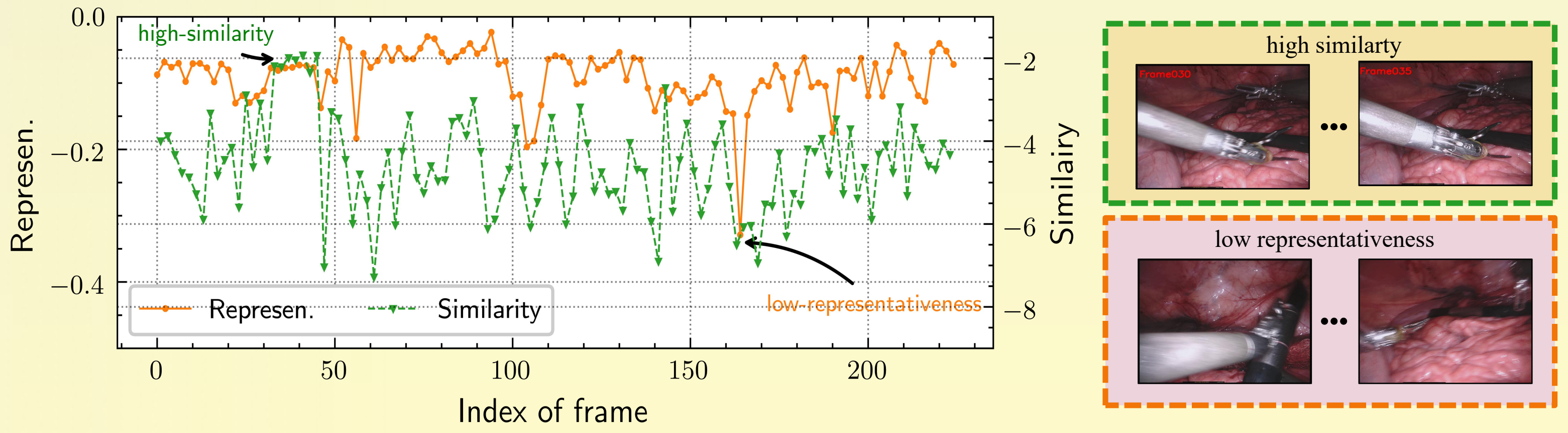
We extensively evaluate the proposed model on two popular public datasets: 2017 MICCAI EndoVis Instrument Challenge (EndoVis17) and 2018 MICCAI EndoVis Scene Segmentation Challenge (EndoVis18).

Methods	EndoVis17		EndoVis18		Param.	FLOPS	Time	FPS
	mDice(%)	mIOU(%)	mDice(%)	mIOU(%)	(M)	(G)	(ms)	
TernausNet [10]	44.95	33.78	61.78	50.25	36.92	275.45	58.32	17
MF-TAPNet [11]	48.01	36.62	-	-	-	-	-	-
PAANet [18]	56.43	49.64	75.01	64.88	21.86	60.34	38.20	26
LWANet [17]	49.79	43.23	71.73	61.06	**2.25**	**2.77**	13.21	76
TDNet [9]	54.64	49.24	76.22	66.30	21.23	47.60	22.23	45
DMNet	**61.03**	**53.89**	**77.53**	**67.50**	4.38	11.53	26.37	38

Type and part segmentation results of instrument on the EndoVis17 and EndoVis18 datasets, respectively. Note that underline denotes the methods in real-time.



Visual comparison of type segmentation on EndoVis17 produced by different methods.



Similarity and representativeness of a video sequence in global aggregation.

CONCLUSIONS

This paper presents a novel real-time surgical instruments segmentation model by efficiently and holistically considering the spatio-temporal knowledge in videos. We develop an efficient local cache for leveraging the most favorable region per frame for local-range aggregation, and an active global cache to select the most informative frames to cover the global cues using only few frames. Experimental results on two public datasets shows that our method outperforms state-of-the-arts by a large margin in accuracy while maintaining the fast prediction speed.

View Project Page