


# FlowRefineSeg: Lightweight Segmentation of Holistic Surgical Scenes with Spatial and Temporal Refinement

Muraam Abdel-Ghani<sup>1</sup> 

Muhammad Arsalan<sup>2</sup>

Fatmaelzahraa Ahmed<sup>1</sup>

Abdulaziz Al-Ali<sup>2</sup>

Khalid Al-Jalham<sup>1</sup>

Shidin Balakrishnan<sup>1</sup>

MURAAM.ABDELGHANI@OUTLOOK.COM

MUHAMMAD.ARSALAN@QU.EDU.QA

FATMAAHMED.HMC@GMAIL.COM

A.ALALI@QU.EDU.QA

KALJALHAM@HAMAD.QA

SBALAKRISHNAN1@HAMAD.QA

<sup>1</sup> *Department of Surgery, Hamad Medical Corporation*

<sup>2</sup> *Computer Science and Engineering Department, College of Engineering, Qatar University*

**Editors:** Under Review for MIDL 2026

## Abstract

Holistic surgical video segmentation is crucial for real-time applications such as proximity analysis of surgical components. For effective integration into clinical workflows, these models must deliver accurate and consistent outputs while being computationally efficient. However, current state-of-the-art (SOTA) architectures are complex, while lighter models fall short of baseline performance. Additionally, temporal consistency is often overlooked in existing surgical segmentation frameworks. To address these limitations, this work introduces FlowRefineSeg, a lightweight segmentation model that achieves SOTA performance with low computational costs. It features a Linear Self-Attention module for effective low-level feature processing, a Gaussian Refinement block to enhance spatial coherence, and a Temporal Refinery module to ensure consistency across video frames. Our experiments show that FlowRefineSeg achieves new benchmark performance on EndoVis18 (74% mIoU, 78% Dice) and SOTA performance on CholecSeg8k (75% mIoU, 80% Dice) with under 25M parameters, establishing a new standard for lightweight holistic surgical segmentation.

**Keywords:** Lightweight Segmentation Model, Holistic Surgical Scene Segmentation, Pyramidal Convolution, Lightweight Attention

## 1. Introduction

Surgical scene segmentation is crucial to accurately define the pixel-level boundaries of key surgical components (Ahmed et al., 2024), and is integral to various downstream applications, such as tool-anatomy interaction analysis during a surgical procedure (Sun and Chen, 2024). These tasks require precise localization of both instruments and anatomy to obtain a comprehensive understanding of their spatial relationships and interactions within the surgical context. However, prevailing research focuses mainly on tool segmentation (González et al., 2020; Matasyoh et al., 2024; Sun and Chen, 2024; Wei et al., 2024), with noticeably fewer studies addressing the complexities of holistic segmentation in these environments.

Current approaches to holistic segmentation range from lightweight models, such as work by Zhou et al. (2023); Ni et al. (2022), to more complex transformer architectures, including the works of Abdel-Ghani et al. (2025); Ahmed et al. (2025a); Liu et al. (2023b)

and video-centric segmentation frameworks such as [Grammatikopoulou et al. \(2024\)](#); [Jin et al. \(2022\)](#). While video-centric models achieve state-of-the-art (SOTA) performance, their high computational demands hinder clinical deployment ([Abdel-Ghani et al., 2025](#); [Ahmed et al., 2025b](#)). Minimally invasive surgical videos stream at 25-30 frames per second (FPS) ([Zi Ye, 2025](#)), and models must meet this throughput to minimize integration overhead in downstream applications. While lightweight models reduce processing times, they often fall short of performance benchmarks ([Zhang et al., 2026](#)). Thus, there is a clear need for high-precision, lightweight segmentation models for effective clinical integration.

Ensuring temporal consistency over consecutive frames poses another significant challenge for surgical video segmentation frameworks. These models often generate per-frame predictions independently, which may lead to discrepancies and contradictory predictions across subsequent frames, posing critical issues when integrated into surgical decision-making frameworks. Few surgical segmentation models address this challenge, despite its importance for clinical deployment ([Jin et al., 2022](#)). Temporal consistency solutions include compute time feature integration with concurrent predictions through MultiLayer Perceptrons ([Ayobi et al., 2023](#)), processing sequential frames for temporal information ([Grammatikopoulou et al., 2024](#); [Jin et al., 2022](#)) or video-specific augmentations to encourage temporal consistency in model finetuning ([Dhanakshirur et al., 2024](#)). These methods, however, either utilize sequential input to produce single refined frames or rely on heavy transformer backbones and high-end GPU resources, which limits reproducibility in less equipped environments.

To address the aforementioned challenges, we introduce FlowRefineSeg, a streamlined segmentation framework designed to deliver precise segmentation while optimizing computational efficiency. FlowRefineSeg utilizes a novel Linear Self-Attention module that enhances low-level features of surgical tools by applying transformer-inspired attention mechanisms, thereby improving holistic segmentation accuracy at reduced computational cost. The model also includes two refinement components. First, a Gaussian Refinement block enhances spatial coherence and reduces pixel-level inconsistencies in the segmentation results. Second, a Temporal Refinery module improves segmentation consistency across successive frames by effectively integrating optical flow information with predictions from previous frames. The integration of these two components serves to diminish extraneous noise and minimize inconsistencies in predictions over time. The final model architecture achieves high model performance while remaining under 25M parameters in size, supporting clinical deployability. Our code will be made publicly available upon acceptance of the paper.

## 2. Methodology

**Overview of FlowRefineSeg** As surgical video streams at 25-30 FPS, models must produce stable frame-by-frame segmentation of both instruments and anatomy without introducing latency. To this end, we propose a lightweight segmentation model, FlowRefineSeg, designed to balance high segmentation accuracy with computational constraints of real-time clinical deployment. FlowRefineSeg builds upon a Pyramidal Convolution ResNet-50 (PyConvResNet50) backbone ([Duta et al., 2020](#)), which excels in feature extraction by employing multiple kernel sizes within each convolution block for improved pixel approximation based on an effective receptive field ([Khan et al., 2022](#); [Duta et al., 2020](#)). A streaming

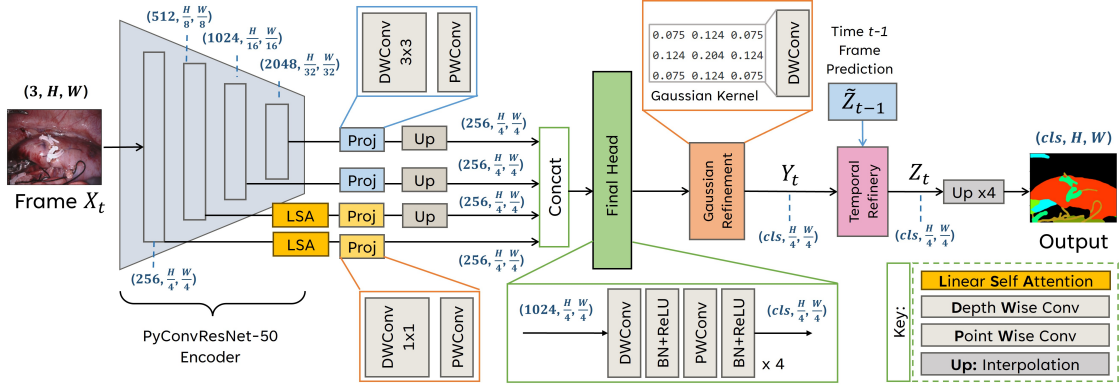


Figure 1: Proposed Architecture of FlowRefineSeg.  $X_t$ : Frame at Time  $t$ ;  $Y_t$ : Base Model Prediction for frame  $t$ ;  $\tilde{Z}_{t-1}$ : Warped Prediction of frame  $t-1$ ;  $Z_t$ : Refined Final Segmentation Output for frame  $t$

decoder processes multi-resolution features in parallel (Abdel-Ghani et al., 2025). The architecture is structured to address three key requirements: accurate delineation of fine structures; spatially coherent masks that avoid small gaps and jagged boundaries; and temporally consistent predictions to avoid distracting flicker or inconsistencies across successive frames. A Linear Self-Attention (LSA) block is incorporated in the decoder for low-level feature processing, improving edge and instrument representations at a low computational cost. The base model per-frame logits are processed by a Gaussian Spatial Refinement Block to smooth pixel-level inconsistencies within each class. Finally, a Temporal Refinery module uses optical flow and previous refined predictions to stabilize subsequent frame segmentation. Figure 1 presents the overall proposed architecture. The four encoder blocks output features into independent processing streams, with LSA applied only to the higher-resolution feature maps, following the approach of Abdel-Ghani et al. (2025). Each stream incorporates Depthwise (DWConv) and Pointwise (PWConv) Convolution layers, utilizing a DWConv kernel size of 1 in higher-resolution streams to preserve feature granularity. Stream outputs are resized via bilinear interpolation and concatenated. A segmentation head composed of alternating DWConv, PWConv, and Batch Normalization (BN) with ReLU activations facilitates gradual channel-wise compression to produce base logits  $Y_t$  for the frame at time  $t$ .  $Y_t$  is then processed through the Gaussian Spatial Refinement block followed by the Temporal Refinery module, yielding the final refined segmentation  $Z_t$ , which is upsampled to the original resolution through bilinear interpolation.

#### Linear Self-Attention (LSA) Block:

The use of a streaming decoder improves semantic segmentation of anatomical structures and surgical tools by leveraging multi-resolution features from the backbone (Abdel-Ghani et al., 2025). However, low-level features from early encoder blocks may contain substantial noise and irrelevant texture. Self-attention helps refine these features and reduce noise, but standard Multi-Head Self-Attention (MHSA) (Vaswani et al., 2017) incurs quadratic complexity with respect to spatial resolution, making it expensive for high-resolution feature maps common in surgical videos. Linear attention mechanisms have been proposed to reduce

the complexity of traditional transformer attention (Li et al., 2020; Liu et al., 2023a). This work introduces Linear Self-Attention (LSA), a variation that integrates transformer-style multi-head self-attention with linearized operations. Given a feature map  $X \in \mathbb{R}^{B \times C \times H \times W}$ . Let the projected embedding dimension be  $D$ , with  $h$  heads and  $d = D/h$  the per-head dimension. First, the query ( $Q$ ), key ( $K$ ) and value ( $V$ ) feature maps are computed by applying learnable PWConv onto the input feature map  $X$ , producing  $Q_{\text{full}}$ ,  $K_{\text{full}}$  and  $V_{\text{full}}$  such that  $Q_{\text{full}}, K_{\text{full}}, V_{\text{full}} \in \mathbb{R}^{B \times D \times H \times W}$ . These three feature maps are then flattened into linear vectors  $\{Q_{\text{linear}} = \text{reshape}(Q_{\text{full}}) \in \mathbb{R}^{B \times D \times N}, N = HW\}$  and divided into heads  $\{Q = [Q^{(1)}, \dots, Q^{(h)}], Q^{(r)} \in \mathbb{R}^{B \times d \times N}\}$ . To ensure non-negative entries in  $Q$  and  $K$ , a kernel activation  $\phi(x) = \text{ELU}(x) + 1$  is applied such that  $\tilde{Q}^{(r)} = \phi(Q^{(r)})$ ,  $\tilde{K}^{(r)} = \phi(K^{(r)})$ , and  $\tilde{Q}^{(r)}, \tilde{K}^{(r)} \in \mathbb{R}^{B \times d \times N}$ . Unlike previous studies, the vectors  $\tilde{Q}^{(r)}$ ,  $\tilde{K}^{(r)}$ , and  $V^{(r)}$  are utilized to implement self-attention for each head  $r$  and each batch  $b$ . This approach aims to replicate the multi-head self-attention mechanism of the transformer by computing the Einstein Summation separately for each head and each batch, using the following formulation in Eq. 1 and 2.

$$S^{(r)}(b) = \tilde{K}^{(r)}(b)(V^{(r)}(b))^{\top} \in \mathbb{R}^{d \times d} \quad (1)$$

$$O^{(r)}(b) = S^{(r)}(b)\tilde{Q}^{(r)}(b) \in \mathbb{R}^{d \times N} \quad (2)$$

The output for each head is concatenated and reshaped into  $O_{\text{full}} \in \mathbb{R}^{B \times D \times H \times W}$ . After applying Layer Normalization, the attention matrix passes through a final PWConv layer, and the output is added residual to the original feature map, followed by a BN layer to produce the final output,  $\tilde{X}$ , as shown in Eq. 3. LSA uses a compact  $d \times d$  per-head context shared across spatial positions, approximating multi-head attention with lower complexity, suitable for real-time processing of high-resolution features.

$$\tilde{X} = \text{BN} \left( X + \text{PWConv}(\text{LN}(\text{concat}_r O^{(r)})) \right) \quad (3)$$

### Gaussian Spatial Refinement Block:

Although the base architecture predicts class logits for each pixel, some pixels at object boundaries or thin structures may be inconsistently labeled, leading to small gaps or jagged edges in the final mask. To encourage spacial coherence without adding learnable parameters, we introduce a Gaussian Refinement block that applies classwise smoothing directly in logit space. Given a set of segmentation logits  $Y \in \mathbb{R}^{B \times C \times H \times W}$ ,  $C = \text{numClasses}$ . The Gaussian Refinement block utilizes a fixed  $3 \times 3$  Gaussian kernel in a DWConv operation, applying the refinement in channel-wise manner. The kernel values are designed using equations 4 through 6, where  $k$  is the size of the kernel,  $\sigma$  defines the distribution of weights within the kernel,  $g[i]$  is the Gaussian value in position  $i$  of the 1-D kernel, and  $K[i, j]$  is the Gaussian entry for the 2-D kernel. Eq. 7 presents the refinement operation, where  $Y_c$  is the logits in channel  $c$ . This class-wise Gaussian smoothing reduces small, isolated misclassifications within homogeneous regions while preserving inter-class boundaries, yielding

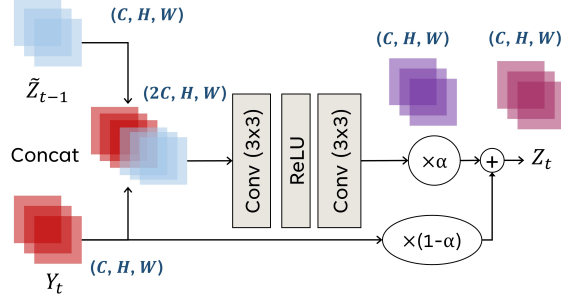


Figure 2: Architecture of Temporal Refinery module, where  $\tilde{Z}_{t-1}$  is the warped previous frame prediction,  $Y_t$  is the base model prediction, and  $\alpha$  is the tuning factor.

smoother and more clinically interpretable masks without additional training

$$g[i] = \exp\left(-\frac{(i-m)^2}{2\sigma^2}\right), \quad i = 0, \dots, k-1, \quad m = \frac{k-1}{2} \quad (4)$$

$$\bar{g}[i] = \frac{g[i]}{\sum_{t=0}^{k-1} g[t]} \quad (5)$$

$$K[i, j] = \bar{g}[i]\bar{g}[j], \quad i, j = 0, \dots, k-1 \quad (6)$$

$$Y_c(x, y) = \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} K[u, v] Y_c(x+u-m, y+v-m), \quad c = 1, \dots, C \quad (7)$$

#### Temporal Refinery module:

To minimize instability in segmentation across frames, we introduce a Temporal Refinery module that operates independently on the segmentation model’s output. Given the input image  $X_t \in \mathbb{R}^{B \times 3 \times H \times W}$  and the base model prediction  $Y_t \in \mathbb{R}^{B \times C \times H \times W}$ , we compute the optical flow  $F_{t-1 \rightarrow t} \in \mathbb{R}^{B \times 2 \times H \times W}$  from the previous frame to warp the previous refined prediction  $Z_{t-1}$  onto the current frame. This warping is defined as  $\tilde{Z}_{t-1} = W(Z_{t-1}, F_{t-1 \rightarrow t})$ , where  $W$  is the differentiable warping operator. The Temporal Refinery combines  $\tilde{Z}_{t-1}$  with  $Y_t$  through Dual Convolutions with an embedding size of 64 to produce refined logits. A tuning factor  $\alpha \in [0, 1]$  adjusts the contribution of the refined output to prevent it from overshadowing the original prediction, combining the logits via residual summation to yield the output  $Z_t$ . A smaller  $\alpha$  aligns closer to the base model, while a larger value emphasizes temporal smoothing. This lightweight module reduces frame-to-frame fluctuations and enhances segmentation consistency without modifying the main segmentation model.

### 3. Experiments

#### Datasets:

We validated our approach on two holistic scene segmentation datasets, namely EndoVis18 and CholecSeg8k. Details of each dataset are provided below.

**EndoVis18 Holistic** The EndoVis18 dataset was released as part of the Endoscopic Vision Challenge 2018 (Allan et al., 2020). The dataset consists of an official training set

comprising 15 procedural sequences, each with 149 frames, and a testing set comprising 4 sequences with 249 frames each. All frames are  $1280 \times 1024$  pixels in size. The dataset contains annotations for various instruments, instrument parts, and anatomical structures, forming a holistic scene segmentation usecase.

**CholecSeg8k** The CholecSeg8k dataset (Twinanda et al., 2016) contains frames from 17 cholecystectomy surgical videos. The annotations include various anatomy and two instruments. The official test split consists of videos 12 and 27 (1,040 images), the validation split of videos 17 and 52 (1,120 images), and the remaining videos constitute the training split (5,920 images). The images are of size  $854 \times 480$  pixels.

### Implementation Details:

For the Temporal Refinery module, we set the value of the tuning factor  $\alpha$  for the EndoVis18 dataset to 0.2, and for CholecSeg8k to 0.05, based on an empirical analysis presented in Section A. Optical flow information was extracted using SPyNet (Ranjan and Black, 2017), a lightweight algorithm that introduces minimal overhead, making it suitable for real-time applications. Experiments were performed on an A6000 GPU with 48GB RAM using the Adam optimizer with a learning rate of  $1 \exp -5$  and a Cosine Annealing scheduler. A combination of Tversky loss (Salehi et al., 2017) and Cross Entropy loss (Mao et al., 2023) was applied (Eq. 8) (Ahmed et al., 2025a). Models were trained on  $512 \times 512$  images for 100 epochs, with data augmentations including random rotations and flips. A fixed random seed ensured reproducibility. To train the Temporal Refinery, the architecture without the module was frozen, and the module was trained on consecutive frames for 50 epochs with early stopping, adding a temporal consistency loss (Eq. 9) to the combined segmentation loss to minimize sudden deviations between the current prediction  $Z_t$  and warped previous prediction  $\tilde{Z}_{t-1}$  while encouraging segmentation accuracy. Model performance was evaluated using well-established metrics, namely mean Intersection over Union (mIoU) (González et al., 2020), Dice Similarity Coefficient (Dice), and mean class Intersection over Union (mcIoU). A Laplacian value was added to all metrics to prevent penalizing absent classes. The testing results maintained the integrity of expert annotations by resizing predictions to match the dataset’s label sizes. Inference speed was measured on an NVidia RTX5000 GPU in FPS, excluding data transfer time.

$$\mathcal{L}_{seg} = \alpha \mathcal{L}_{tversky} + (1 - \alpha) \mathcal{L}_{CE} \quad (8)$$

$$\mathcal{L}_{temp} = \frac{1}{BCHW} \left\| Z_t - \tilde{Z}_{t-1} \right\|_1, \quad Z_t, \tilde{Z}_{t-1} \in \mathbb{R}^{B \times C \times H \times W} \quad (9)$$

## 4. Results

We compare the performance of FlowRefineSeg to various segmentation frameworks, including transformer-based SegFormer (Xie et al., 2021), MedT (Valanarasu et al., 2021), SSwin and STSwinCL (Jin et al., 2022), Swin-SPTCN (Grammatikopoulou et al., 2024), and SOTA model FASL-Seg (Abdel-Ghani et al., 2025). Additionally, CNN-based LSKANet (Liu et al., 2023b), and lightweight models LWANet (Ni et al., 2020), TinyUNet (Chen et al., 2024), and CFFANet (Mahmood et al., 2025) are included. The results for EndoVis18 and CholecSeg8k datasets are presented in Table 1 and 2, respectively. On both benchmarks, lightweight models show limited performance compared to complex transformer-based architectures. In contrast, FlowRefineSeg demonstrates a notable improvement over the baseline

performance on EndoVis18 (74% mIoU and 78% Dice) and exceeds SOTA model FASL-Seg’s performance on most classes. It also achieves SOTA performance on CholecSeg8k. FlowRefineSeg exceeds lightweight model performances by +11  $\rightarrow$  21% mIoU on both benchmarks. The consistently high performance reflects FlowRefineSeg’s strong segmentation capabilities for holistic scene segmentation.

Table 1: Testing Results on EndoVis18 in mIoU and Dice, Label Key: BT: Background Tissue ISh:Instrument Shaft, IC: Instrument Clasper, IW: Instrument Wrist, KP:Kidney Parenchyma, CK: Covered Kidney, SmInst: Small Intestine, SI: Suction Instrument, UP: Ultrasound Probe. Top overall result is shown in bold font.

Model	Mean Intersection Over Union (mIoU)												
	mIoU	BT	ISh	IC	IW	KP	CK	Thread	Clamps	Needle	SI	SmInt	UP
LWANet (Ni et al., 2020)	0.64	0.73	0.69	0.35	0.35	0.51	0.15	0.90	0.93	0.91	1.0	0.33	0.85
MedT* (Valanarasu et al., 2021)	0.65	0.40	-	0.54	-	0.16	0.44	0.82	0.96	0.90	0.61	0.78	0.84
SegFormer (Xie et al., 2021)	0.68	0.69	0.59	0.27	0.32	0.39	0.50	0.90	0.93	0.91	1.0	0.80	0.85
SSwin* (Jin et al., 2022)	0.62	-	0.87	0.58	0.64	0.66	0.40	0.13	0.77	-	-	0.54	0.24
STSwinCL* (Jin et al., 2022)	0.64	-	0.88	0.59	0.65	0.68	0.47	0.18	0.77	-	-	0.54	0.25
LSKANet* (Liu et al., 2023b)	0.66	-	-	-	-	-	-	-	-	-	-	-	-
CFFANet (Mahmood et al., 2025)	0.56	0.40	0.10	0.12	0.18	0.31	0.24	0.90	0.93	0.9	1.0	0.77	0.85
TinyUNet (Chen et al., 2024)	0.56	0.52	0.43	0.12	0.18	0.18	0.00	0.90	0.93	0.90	1.0	0.77	0.85
FASL-Seg (Abdel-Ghani et al., 2025)	0.73	0.80	0.86	0.52	0.67	0.64	0.23	0.76	0.89	0.91	0.85	0.81	0.87
FlowRefineSeg (Ours)	<b>0.74</b>	0.70	0.85	0.57	0.69	0.54	0.34	0.87	0.94	0.91	0.93	0.65	0.88
Model	Dice Similarity Coefficient (Dice)												
	Dice	BT	ISh	IC	IW	KP	CK	Thread	Clamps	Needle	SI	SmInt	UP
LWANet (Ni et al., 2020)	0.68	0.84	0.75	0.48	0.42	0.60	0.20	0.90	0.93	0.91	1.0	0.35	0.85
MedT* (Valanarasu et al., 2021)	0.68	0.56	-	0.66	-	0.26	0.44	0.82	0.96	0.90	0.62	0.78	0.84
SegFormer (Xie et al., 2021)	0.72	0.79	0.68	0.36	0.38	0.46	0.53	0.90	0.93	0.91	1.0	0.81	0.85
SSwin* (Jin et al., 2022)	0.70	-	-	-	-	-	-	-	-	-	-	-	-
STSwinCL* (Jin et al., 2022)	0.72	-	-	-	-	-	-	-	-	-	-	-	-
LSKANet* (Liu et al., 2023b)	0.75	-	-	-	-	-	-	-	-	-	-	-	-
TinyUNet (Chen et al., 2024)	0.59	0.67	0.54	0.12	0.18	0.27	0.01	0.90	0.93	0.90	1.0	0.77	0.85
CFFANet (Mahmood et al., 2025)	0.58	0.56	0.10	0.12	0.18	0.42	0.24	0.90	0.93	0.9	1.0	0.77	0.85
FASL-Seg (Abdel-Ghani et al., 2025)	0.77	0.88	0.90	0.64	0.76	0.70	0.28	0.77	0.90	0.91	0.85	0.84	0.89
FlowRefineSeg (Ours)	<b>0.78</b>	0.84	0.90	0.70	0.78	0.66	0.38	0.83	0.93	0.91	0.91	0.65	0.89

\* Based on reported results

In Figure 3, we visually compare inference by FlowRefineSeg against lightweight architectures LWANet and TinyUNet, transformer-based SegFormer, and the SOTA model FASL-Seg. Both LWANet and TinyUNet demonstrate suboptimal segmentation, producing incomplete outputs with missing elements such as tool segments. SegFormer achieves high precision in anatomical structures but struggles with tool segmentation. In contrast, FlowRefineSeg delivers segmentation that matches the outputs of FASL-Seg while operating at a significantly lower computational cost. This reinforces FlowRefineSeg’s standing as a robust and efficient lightweight segmentation model. We also analyze the impact of the Temporal Refinery on reducing noise in predictions derived from preceding inferences in Fig. 4 on consecutive frames from the EndoVis18 dataset. The Temporal Refinery effectively mitigates extraneous noise by leveraging temporal coherence from prior predictions, thereby improving segmentation accuracy. This underscores the module’s capability as a lightweight, yet powerful, mechanism for enhancing predictive performance in time-sequenced data, ultimately facilitating smoother video inference.

**Model Complexity and Inference Speed** We evaluate the complexity of FlowRefineSeg in relation to several SOTA architectures and lightweight models, focusing on key metrics such as the number of parameters, floating point operations (GFlops), and model inference speed in FPS. These metrics are compared to the model’s performance (mIoU)

Table 2: Testing Results on Cholec8K in mIoU and Dice, Label Key: Ft: Fat, Lv: Liver, GB: Gallbladder, AW: Abdominal Wall, GT: Gastrointestinal Tract, Gr: Grasper, LE: L-hook Electrocautery, Bld: Blood, HV: Hepatic Vein, CT: Connective Tissue, LL: Liver Ligament, CD: Cystic Duct. Top overall result is shown in bold font.

Model	Mean Intersection Over Union (mIoU)												
	mIoU	Ft	Lv	GB	AW	GT	Gr	LE	Bld	HV	CT	LL	CD
LWNet (Ni et al., 2020)	0.54	0.48	0.56	0.18	0.68	0.25	0.19	0.64	1.0	0.16	0.23	0.70	1.0
MedT *(Valanarasu et al., 2021)	0.50	0.69	0.25	0.41	0.21	0.17	0.35	0.67	1.0	0.70	0.62	0.0	0.89
SegFormer (Xie et al., 2021)	0.74	0.86	0.74	0.45	0.82	0.48	0.47	0.78	1.0	0.70	0.39	1.0	1.0
Swin-SPTCN *(Grammatikopoulou et al., 2024)	0.69	0.84	0.78	0.61	0.74	0.57	0.74	0.68	-	-	0.31	-	-
TinyUNet (Chen et al., 2024)	0.58	0.59	0.56	0.03	0.74	0.15	0.39	1.0	1.0	0.24	0.18	0.70	1.0
CFFANet (Mahmood et al., 2025)	0.64	0.74	0.65	0.08	0.82	0.15	0.39	1.0	1.0	0.46	0.41	0.70	1.0
FASL-Seg (Abdel-Ghani et al., 2025)	0.75	0.88	0.76	0.45	0.82	0.40	0.52	0.73	1.0	0.70	0.50	1.0	1.0
FlowRefineSeg (Ours)	<b>0.75</b>	0.83	0.79	0.31	0.83	0.52	0.56	0.95	1.0	0.76	0.56	0.70	1.0
Model	Dice Similarity Coefficient (Dice)												
	Dice	Ft	Lv	GB	AW	GT	Gr	LE	Bld	HV	CT	LL	CD
LWNet (Ni et al., 2020)	0.60	0.62	0.71	0.26	0.81	0.36	0.19	0.64	1.0	0.21	0.35	0.70	1.0
MedT *(Valanarasu et al., 2021)	0.57	0.81	0.39	0.56	0.34	0.25	0.48	0.71	1.0	0.70	0.69	0.0	0.89
SegFormer (Xie et al., 2021)	0.79	0.92	0.85	0.57	0.89	0.59	0.58	0.82	1.0	0.70	0.40	1.0	1.0
TinyUNet (Chen et al., 2024)	0.63	0.73	0.71	0.06	0.85	0.15	0.39	1.0	1.0	0.31	0.28	0.70	1.0
CFFANet (Mahmood et al., 2025)	0.68	0.84	0.78	0.08	0.90	0.15	0.39	1.0	1.0	0.46	0.54	0.70	1.0
FASL-Seg (Abdel-Ghani et al., 2025)	0.80	0.94	0.86	0.57	0.89	0.52	0.63	0.78	1.0	0.70	0.56	1.0	1.0
FlowRefineSeg (Ours)	<b>0.80</b>	0.90	0.88	0.39	0.91	0.61	0.64	0.95	1.0	0.80	0.67	0.70	1.0

\* Based on reported results

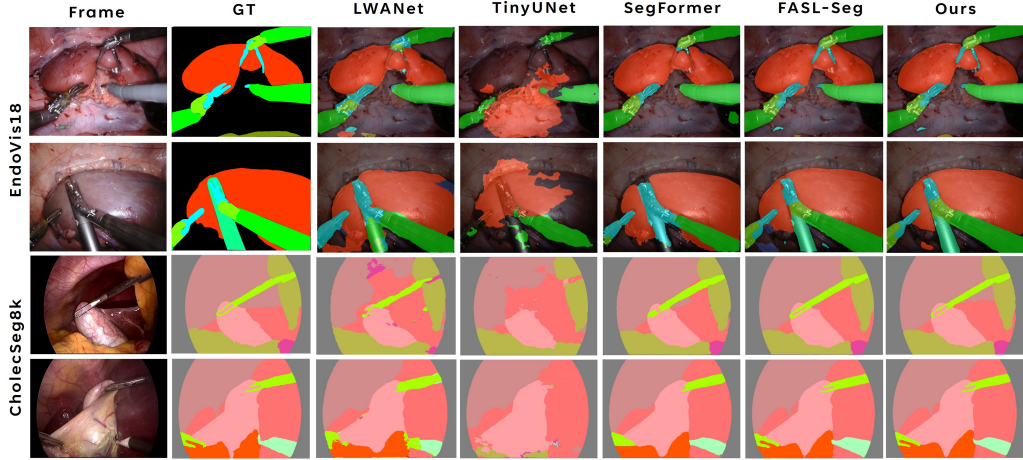


Figure 3: Comparison of Inference of FlowRefineSeg against SOTA models on frames from EndoVis18 and CholecSeg8k. GT stands for Ground Truth prediction.

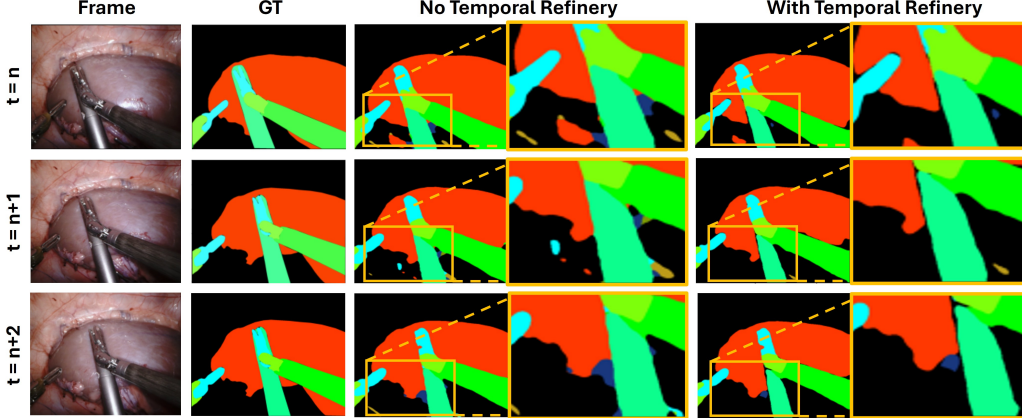


Figure 4: Comparison of inference on three consecutive frames from EndoVis18 dataset at time  $t=n$ ,  $n+1$  and  $n+2$ , with and without the Temporal Refinery module. GT stands for Ground Truth.

on the EndoVis18 dataset frames, as presented in Table 3. Our comparison reveals that FlowRefineSeg delivers superior performance while requiring less parameters and Flops. Furthermore, our model’s inference speed is comparable to lightweight CFFANet, positioning FlowRefineSeg as a robust segmentation model with real-time performance.

Table 3: Model Complexity in terms of Parameters, Floating Point Operations (GFlops), and Frames Per Second (FPS)

Model	Architecture	#Params	GFLOPs	FPS	mIoU
LWANet (Ni et al., 2020)	CNN	2.07M	1.7	45.8	0.64
STSwInCL (Jin et al., 2022)	Transformer	125.46M	922.72	-	0.64
LSKANet (Liu et al., 2023b)	CNN	72.23M	189.09	-	0.66
TinyUNet (Chen et al., 2024)	CNN	481.2K	7.72	53.9	0.56
CFFANet (Mahmood et al., 2025)	CNN	2.2M	2.88	32.6	0.56
FASL-Seg (Abdel-Ghani et al., 2025)	Transformer,CNN	83.63M	224.09	16.4	0.73
FlowRefineSeg (Ours)	CNN	24.87M	41.7	32.1	<b>0.74</b>

**Ablation Study** In our first ablation study, we compared lightweight and multi-head attention mechanisms with our proposed Linear multi-head Self-Attention (LSA). The attention types evaluated include Multi-Head Self-Attention (MHSA) (Vaswani et al., 2017), and Lightweight Multi-Head Channel Attention (MCA) (Zhang et al., 2026), as detailed in Table 4. Performance metrics clearly demonstrate that LSA achieves results that are within a mere 1  $\rightarrow$  2% mIoU and Dice of MCA and MHSA, while boasting remarkable efficiency—being over 80% more efficient in Flops compared to MCA and 90% more efficient than MHSA. This performance is attributed to its vectorized attention mechanism, which

delivers multi-head self-attention with minimal computational overhead. We further illustrate these findings through a visual comparison of each attention on final model activation heatmaps, presented in Section B. In the subsequent ablation, we introduced the Gaussian Refinement block and the Temporal Refinery. Each additional component leads to an incremental improvement in performance, underscoring the significance of these modules in attaining SOTA results within our model.

Table 4: Ablation Study on Proposed Architecture

Model	Attention Type			GaussRefine	TemporalRefinery	mIoU	Dice	#Params	GFlops
	MHSA	MCA	LSA						
Model 1	✓					0.6647	0.7082	25.84M	353.09
Model 2		✓				0.6608	0.7055	25.84M	197.84
Base Model			✓			0.6465	0.6921	24.85M	36.18
Base Model+GaussRefine			✓	✓		0.6820	0.7280	24.85M	36.18
FlowRefineSeg			✓	✓	✓	0.7390	0.7854	24.87M	41.65

## 5. Limitations and Ethical Considerations

FlowRefineSeg achieves new benchmark performance for lightweight models in holistic surgical scene segmentation. However, the presented results, while promising, remain preliminary, and extensive clinical validation, including user testing with surgeons, will be crucial before deployment. Furthermore, the work focuses on endoscopic surgical videos, but can be explored for other surgeries or medical imaging modalities. Integrating FlowRefineSeg into clinical workflows presents ethical concerns, especially the risk of over-reliance on its segmentation, which could lead to missed fluctuations or misclassifications that may result in surgical complications. Conducting clinical trials is essential to create an effective schema and user interface that allows surgeons to use the model while maintaining critical judgment. Comprehensive training and a strong mitigation process are necessary for successful integration and use of such advanced solutions.

## 6. Conclusion

To conclude, this paper presents FlowRefineSeg, a lightweight architecture for improved holistic surgical scene segmentation. The model features a Linear Self-Attention module for efficient low-level feature processing in the streaming decoder, improving instrument tips and edge-level segmentation. It further integrates a Gaussian Spatial Refinement block to smooth pixel-level inconsistencies, and a Temporal Refinery to ensure stable segmentation in surgical videos. Achieving new benchmark performance on the EndoVis18 dataset and SOTA performance on CholecSeg8k, it remains computationally efficient at 24.87 M parameters, making it a strong baseline for lightweight holistic scene segmentation and suitable for integration into downstream applications like proximity analysis and skill assessment. Future work will explore applying this model to other medical imaging modalities, such as Computer Tomography (CT) scans, and will investigate clinical validation to improve integration into clinical workflows and assess its applicability in different contexts.

## Acknowledgments

This work was supported by the Qatar Research Development and Innovation Council (QRDI), Grant no. ARG01-0522-230266, and partially supported by the Hamad Medical Corporation (HMC) Internal Research Grant no. MRC-01-22-308. The content reported through this research is solely the responsibility of the authors and does not necessarily represent the official views of Qatar Research Development and Innovation Council.

## References

- Muraam Abdel-Ghani, Mohamed Ali, Mahmoud Ali, Fatmaelzahraa Ahmed, Muhammad Arsalan, Abdulaziz Al-Ali, and Shidin Balakrishnan. Fasl-seg: Anatomy and tool segmentation of surgical scenes. In *Proceedings of the European Conference on Artificial Intelligence (ECAI) 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 1001–1008. IOS Press, 2025. doi: 10.3233/FAIA250908. URL <https://doi.org/10.3233/FAIA250908>.
- Fatimaelzahraa Ahmed, Muraam Abdel-Ghani, Muhammad Arsalan, Mahmoud Ali, Abdulaziz Al-Ali, and Shidin Balakrishnan. Surg-segformer: A dual transformer-based model for holistic surgical scene segmentation. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 1304–1309, 2025a. doi: 10.1109/CASE58245.2025.11163962.
- Fatimaelzahraa Ahmed, Muhammad Arsalan, Abdulaziz Al-Ali, Khalid Al-Jalham, and Shidin Balakrishnan. Clip-rl: Surgical scene segmentation using contrastive language-vision pretraining & reinforcement learning. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, 2025b.
- Fatimaelzahraa Ali Ahmed, Mahmoud Yousef, Mariam Ali Ahmed, Hasan Omar Ali, Anns Mahboob, Hazrat Ali, Zubair Shah, Omar Aboumarzouk, Abdulla Al Ansari, and Shidin Balakrishnan. Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review. *Artificial Intelligence Review*, 58(1):1, 2024.
- Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.
- Nicolás Ayobi, Alejandra Pérez-Rondón, Santiago Rodríguez, and Pablo Arbeláez. Matis: Masked-attention transformers for surgical instrument segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. doi: 10.1109/ISBI53787.2023.10230819.
- Junren Chen, Rui Chen, Wei Wang, Junlong Cheng, Lei Zhang, and Liangyin Chen. Tinyunet: Lighter yet better u-net with cascaded multi-receptive fields. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15009, pages 626–635. Springer Nature Switzerland, October 2024.

- Rohan Raju Dhanakshirur, Mrinal Tyagi, Britty Baby, Ashish Suri, Prem Kalra, and Chetan Arora. VideoCutMix: Temporal Segmentation of Surgical Videos in Scarce Data Scenarios . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15006. Springer Nature Switzerland, October 2024.
- Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Pyramidal convolution: Rethinking convolutional neural networks for visual recognition, 2020. URL <https://arxiv.org/abs/2006.11538>.
- C. González, L. Bravo-Sánchez, and P. Arbelaiz. Isinet: An instance-based approach for surgical instrument segmentation. In Anne L. Martel et al., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12263 of *Lecture Notes in Computer Science*. Springer, Cham, 2020. doi: 10.1007/978-3-030-59716-0\_57. URL [https://doi.org/10.1007/978-3-030-59716-0\\_57](https://doi.org/10.1007/978-3-030-59716-0_57).
- Maria Grammatikopoulou, Ricardo Sanchez-Matilla, Felix Bragman, David Owen, Lucy Culshaw, Karen Kerr, Danail Stoyanov, and Imanol Luengo. A spatio-temporal network for video semantic segmentation in surgical videos. *International Journal of Computer Assisted Radiology and Surgery*, 19(3):375–382, 2024. doi: 10.1007/s11548-023-02971-6. URL <https://link.springer.com/article/10.1007/s11548-023-02971-6>.
- Yueming Jin, Yang Yu, Cheng Chen, Zixu Zhao, Pheng-Ann Heng, and Danail Stoyanov. Exploring intra- and inter-video relation for surgical semantic scene segmentation. *IEEE Transactions on Medical Imaging*, 41(11):2991–3002, 2022. doi: 10.1109/TMI.2022.3177077.
- Tariq M. Khan, Muhammad Arsalan, Antonio Robles-Kelly, and Erik Meijering. Mkis-net: A light-weight multi-kernel network for medical image segmentation. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2022. doi: 10.1109/DICTA56598.2022.10034573.
- Rui Li, Jianlin Su, Chenxi Duan, and Shunyi Zheng. Linear attention mechanism: An efficient attention for semantic segmentation. *arXiv preprint arXiv:2007.14902*, 2020.
- Langming Liu, Liu Cai, Chi Zhang, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Yifu Lv, Wenqi Fan, Yiqi Wang, Ming He, et al. Linrec: Linear attention mechanism for long-term sequential recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 289–299, 2023a.
- Min Liu, Yubin Han, Jiazheng Wang, Can Wang, Yaonan Wang, and Erik Meijering. Lskanet: Long strip kernel attention network for robotic surgical scene segmentation. *IEEE Transactions on Medical Imaging*, TMI, 2023b. doi: 10.1109/TMI.2023.3335406.
- Tahir Mahmood, Ganbayar Batchuluun, Seung Gu Kim, Jung Soo Kim, and Kang Ryoung Park. A lightweight hierarchical feature fusion network for surgical instrument segmentation in internet of medical things. *Information Fusion*, 123:103303, 2025. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2025.103303>. URL <https://www.sciencedirect.com/science/article/pii/S1566253525003768>.

- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: theoretical analysis and applications. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.
- Wilson Matasyoh, Jane Muthoni, and Wei Zhang. Samsurg: Surgical instrument segmentation in robotic surgeries using vision foundation model. *IEEE Access*, 12:12345–12356, 2024. doi: 10.1109/ACCESS.2024.3520386.
- Zhen-Liang Ni, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Hu Zhou, Xiao-Liang Xie, and Zhen Li. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9939–9945, 2020. doi: 10.1109/ICRA40945.2020.9197425.
- Zhen-Liang Ni, Gui-Bin Bian, Zhen Li, Xiao-Hu Zhou, Rui-Qi Li, and Zeng-Guang Hou. Space squeeze reasoning and low-rank bilinear feature fusion for surgical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3209–3217, 2022. doi: 10.1109/JBHI.2022.3154925.
- Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.
- Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In Qian Wang, Yinghuan Shi, Heung-Il Suk, and Kenji Suzuki, editors, *Machine Learning in Medical Imaging*, pages 379–387, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67389-9.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.
- Liping Sun and Xiong Chen. Pixel-wise contrastive learning for multi-class instrument segmentation in endoscopic robotic surgery videos using dataset-wide sample queues. *IEEE Access*, 12:156867–156877, 2024. doi: 10.1109/ACCESS.2024.3476622.
- Adnan P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 35(7):1746–1758, 2016. doi: 10.1109/TMI.2016.2593967. URL <https://ieeexplore.ieee.org/document/7593967>.
- Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M. Patel. Medt: A transformer-based medical image segmentation approach. *Medical Image Analysis*, 75: 102321, 2021. doi: 10.1016/j.media.2021.102321.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

- Meng Wei, Miaoqing Shi, and Tom Vercauteren. Enhancing surgical instrument segmentation: Integrating vision transformer insights with adapter. *International Journal of Computer Assisted Radiology and Surgery*, 19:1313–1320, 2024. doi: 10.1007/s11548-024-03140-z.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL <https://proceedings.neurips.cc/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf>.
- Jiayi Zhang, Xinyu Zhang, Jiahe Tan, Rui Zhang, Wei Shi, Yechen Zhu, Suet To, Wenkui Wang, and Xin Gao. Lms-net: Light-weight multi-object segmentation network for laparoscopic surgical scene segmentation. *Biomedical Signal Processing and Control*, 111: 108264, 2026. ISSN 1746-8094. doi: <https://doi.org/10.1016/j.bspc.2025.108264>. URL <https://www.sciencedirect.com/science/article/pii/S174680942500775X>.
- Mengyu Zhou, Xiaoxiang Han, Zhoujin Liu, Yitong Chen, and Liping Sun. A lightweight segmentation network for endoscopic surgical instruments based on edge refinement and efficient self-attention. *PeerJ Computer Science*, 9:e1746, 2023. doi: 10.7717/peerj-cs.1746. URL <https://peerj.com/articles/cs-1746>.
- Zili Deng Dan Wang Ying Zhu Xiaoli Jin Lijun Zhang Tianxiang Chen Hanwei Zhang Mingliang Wang Zi Ye, Ru Zhou. A comprehensive video dataset for surgical laparoscopic action analysis. *Data Descriptor*, 2025.

## Appendix A. Ablation on Tuning Factor $\alpha$ in Temporal Refinery module

We experimented with a range of values for the tuning factor  $\alpha$  in the Temporal Refinery module to control the amount of tuning applied to the base model predictions with the temporal information. For each of EndoVis18 and CholecSeg8k datasets, values from 0 to 1 at 0.05 intervals were explored until the performance started to drop. The results are presented in Figure 5, showing that setting  $\alpha$  to 0.2 provides the greatest benefits to the full framework for EndoVis18, while for CholecSeg8k, 0.05 provides enough temporal information to improve the overall segmentation. As a result, these respective values were used in the remaining experiments.

## Appendix B. Activation Heat Map Analysis on Attention Mechanisms

To evaluate the impact of the attention mechanisms detailed in the ablation study on final predictions, we generated heat maps showing the gradients of the final convolution layer for each attention type: Multi-Class Attention (MCA), Multi-Head Self-Attention (MHSA), and our proposed Linear Self-Attention (LSA). The heat maps were created using the GradCAM (Selvaraju et al., 2020) PyTorch library. We focused on two classes—the instrument shaft, represented by semantic features in later encoder blocks, and the clasper, captured by low-level features in earlier blocks. The heat maps presented in Figure 6 indicate

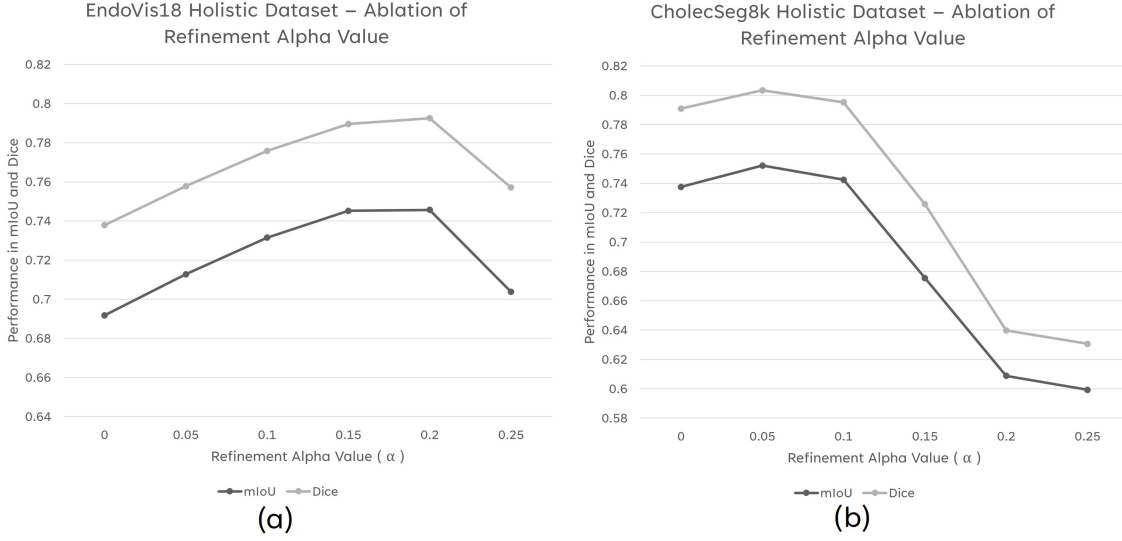


Figure 5: Ablation on tuning factor  $\alpha$  in Temporal Refinery on (a) EndoVis18 and (b) CholecSeg8k

that the model with LSA achieves high confidence in both feature types. Importantly, LSA enhances clasper features similar to MCA and MHSA, while operating at less than 20% of their computational cost in Flops. This supports LSA as an effective and efficient self-attention mechanism.

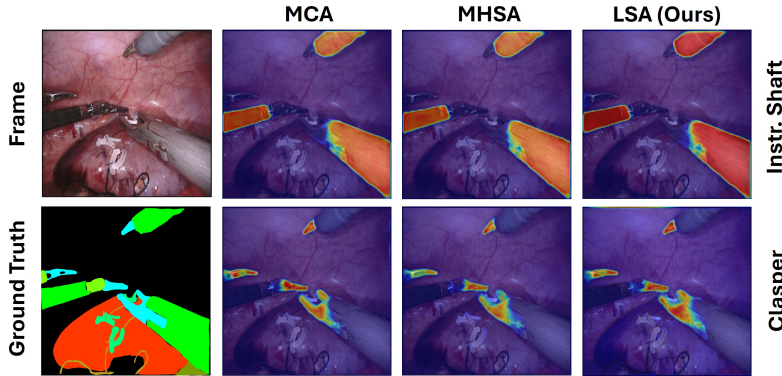


Figure 6: Activation Heat maps for final prediction of Instrument Shaft and Clasper classes when using MCA, MHSA and LSA (ours) attention mechanisms