Enhancing Surgical Instrument Segmentation: Integrating Vision Transformer Insights with Adapter

Meng Wei[1*], Miaojing Shi[2] and Tom Vercauteren[1]

[1*]School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom.

[2]College of Electronic and Information Engineering, Tongji University , Shanghai, China.

*Corresponding author(s). E-mail(s): meng.wei@kcl.ac.uk;
Contributing authors: mshi@tongji.edu.cn; tom.vercauteren@kcl.ac.uk;

Abstract

Purpose: In surgical image segmentation, a major challenge is the extensive time and resources required to gather large-scale annotated datasets. Given the scarcity of annotated data in this field, our work aims to develop a model that achieves competitive performance with training on limited datasets, while also enhancing model robustness in various surgical scenarios.
Methods: We propose a method that harnesses the strengths of pre-trained Vision Transformers (ViT) and data efficiency of Convolutional Neural Networks (CNN). Specifically, we demonstrate how a CNN segmentation model can be used as a lightweight adapter for a frozen ViT feature encoder. Our novel feature adapter uses cross-attention modules that merge the multi-scale features derived from the CNN encoder with feature embeddings from ViT, ensuring integration of the global insights from ViT along with local information from CNN.
Results: Our method outperforms current models in surgical instrument segmentation on the Robust-MIS 2019 dataset and showcases remarkable robustness through cross-dataset validation across five public datasets.
Conclusion: In this study, we presented a novel approach integrating ViT and CNN. Our unique feature adapter successfully combines the global insights of ViT with the local, multi-scale spatial capabilities of CNN. This integration effectively overcomes data limitations in surgical instrument segmentation. The source code is available at: https://github.com/weimengmeng1999/AdapterSIS.git.

Keywords: Vision Transformer, Adapter Network, Surgical Instrument Segmentation

1 Introduction

Detecting and tracking surgical instruments in laparoscopic videos is crucial for autonomous surgery and enhanced clinical support [1]. The trend in the field is towards the utilization of deep learning methodologies [2, 3]. Current models heavily depend on fully supervised learning, requiring extensive annotated data. However, acquiring such data, especially in surgical tool segmentation, is expensive and time-intensive, resulting in the lack of large-scale annotated datasets, a significant hurdle for precise model development. Additionally, biases in training datasets arise from outdated datasets, geographical diversity, and unverified clinical relevance, affecting the robustness needed for applications like autonomous surgery.

In light of the rapid advancements in large-scale ViTs [4] and their excellent ability to learn from extensive data, pre-trained ViT models [4, 5] offer promising potential for downstream tasks [6–8]. CNNs have revolutionized the medical image segmentation field. However, their localized convolution operations limit capturing global and long-range semantic interactions. Transformers provide global self-attention but might lack detailed localization abilities [9]. Merging CNN and ViT is a recent trend to leverage their strengths [9–12]. Yet, these methods, often starting from scratch, might not fully exploit pre-trained knowledge from large image datasets, a significant ViT strength. Moreover, while most of them focus on bridging the global and local information gaps between the two methods, they neglect the inherent advantages of each: CNNs exhibit better performance with limited datasets, whereas ViTs are superb with extensive data training.

Therefore, given the constraints imposed above, we harness the full potential of both ViT and CNN. We are particularly focused on capitalizing on the pre-trained general knowledge derived from ViT to enhance surgical image segmentation models, with an overarching goal of optimal both of the model performance and robustness within the complex and diverse domain of surgical images. Our main contributions are: 1) Adapting a pre-trained and frozen ViT based on DINOV2 [5] to a CNN backbone segmentation model optimized for scenarios with limited annotated data; 2) Introducing innovative adapter modules with cross-attention (CA) to integrate the global information from ViT and local features from CNN; 3) Enhancing the generalizability of the segmentation model across multiple datasets.

2 Related Work

2.1 Surgical Instrument Segmentation

The majority of surgical instrument segmentation works are CNN-based methods. For example, ISINet [3] proposes an instance-based surgical instrument segmentation CNN network that includes a temporal consistency module. OR-UNet [2] is introduced as an optimized 2D UNet [13] for instrument segmentation. There is a growing trend of exploring ViT-based methods. MATIS [14] is a fully transformer-based method that utilizes pixel-wise attention and masked attention modules. TraSeTR [15] introduces a Track-to-Segment transformer that leverages tracking cues to enhance surgical instrument segmentation.

2.2 Pre-trained Vision Transformers

Driven by extensive pretraining on large datasets, ViT [4] employs self-supervised learning for vision tasks. DINOV2 [5] improves the training of large-scale ViT models with 1B parameters and distils it into smaller models. The pre-trained ViTs are successfully applied to the downstream tasks such as image classification [6, 7], object detection [5], semantic segmentation [5, 6], and video action classification [7]. Research on fine-tuning cross-attention modules with pre-trained embeddings [16] aligns with our method of harnessing pre-trained knowledge from large-scale ViT models. Yet, there is no existing work that adapts pre-trained ViT features by a CNN adapter, crucial due to limited data availability [4].

2.3 Hybrid CNN and ViT Models

ViTs and CNNs inherently complement each other. Numerous studies fuse two architectures to address their limitations. For instance, TransUNet [9] hybrids in which ViT processes CNN-derived patches for global context. TransFuse [10] parallels ViT and CNNs for efficient global and multi-level spatial feature fusion. There are also works that simulate the characteristics of CNN in their ViT models [7, 16] or directly adopt the cross-attention mechanism to augment the CNN structure [17], but none of the existing work integrates cross-attention into a CNN model to serve as a lightweight adapter for a pre-trained ViT model.

3 Method

We present the three primary elements of our model with the detailed architecture illustrated in Fig. 1. The ViT feature encoder remains frozen, with only the adapter and the CNN backbone segmented undergoing training. The CNN decoder receives three distinct feature inputs: 1) Patch tokens from the ViT branch, encapsulating local information; 2) Output from the adapter, which combines local and global insights from both the ViT and CNN branches; 3)Feature maps from the CNN encoder, preserving the spatial information of the original image.

3.1 Vision Transformer Encoder

Our vision transformer encoder follows the established method inspired by ViT [4]. Given an input image, denoted as $I \in \mathbb{R}^{H \times W \times C}$, where H is the height and W is the width. The ViT encoder initially divides the image into patches, forming a sequence represented as $I = [I_1, \ldots, I_N] \in \mathbb{R}^{N \times P^2 \times C}$, where P corresponds to the patch size. The count of patches, N, is calculated as $N = \frac{HW}{P^2}$. Each of these individual patches is then converted into a 1D vector and linearly projected, resulting in a sequence of patch embeddings, denoted as $I_0 = [E_{I_1}, \ldots, E_{I_N}] \in \mathbb{R}^{N \times D}$, with the transformation matrix $E \in \mathbb{R}^{D \times (P^2 C)}$. To account for positional information, the ViT encoder introduces learnable position embeddings to combine with the patch sequence. The transformer encoder then maps the input sequence of embedded patches with position encoding to the output $x_{ViT} = \left[x_{ViT}^{patch} || x_{ViT}^{CLS}\right]$, a contextualized encoding sequence containing

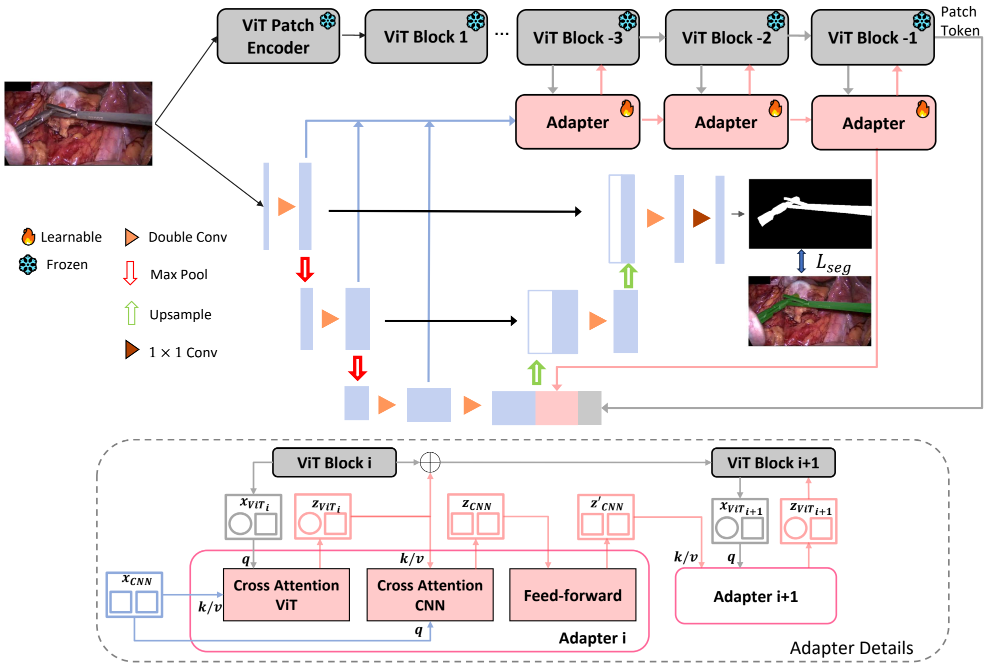

Fig. 1 An overview of our method. Our model includes two main parts: the top consists of a frozen pre-trained ViT feature encoder; the middle introduces adapter modules that enable CA integration between multi-scale features from CNN and pre-trained ViT features; the bottom is backbone segmenter tailored for instrument segmentation; q is query and k/v is key/value.

rich semantic information. To utilize pre-learned knowledge, we employed and froze the entire ViT model. However, we selectively integrated the ViT feature embeddings from the deeper layers into our backbone segmentation model using adapters. We opted not to utilize the shallower layers to optimize computational efficiency.

3.2 Feature Adapter

Building on the strengths of ViT and CNN highlighted in Section 1, our adapter integrates multi-scale features from the CNN backbone segmentation encoder with those from the pre-trained ViT feature encoder.

Cross Attention for ViT In our CA module for the ViT, we first utilize the patch token at the ViT branch, denoted as x_{ViT}^{patch}, which includes local information from the ViT pre-trained knowledge, as the query to exchange information among the multi-scale feature embeddings from the backbone segmentation encoder and then back project it to the ViT branch.

For visual clarity, Fig. 2 illustrates the CA module for ViT. Specifically, the multi-scale feature embeddings from the backbone segmenter encoder were initially aggregated. Several fully connected layers are applied at the end to project the feature maps to D dimensions, which equals the patch embedding size of the ViT branch. The multi-scale feature map from the backbone encoder, denoted as $x_{HW/S}$, $x_{HW/2S}$, and $x_{HW/4S}$, then comprises D-dimensional features at $1/S$, $1/2S$, and $1/4S$ resolutions

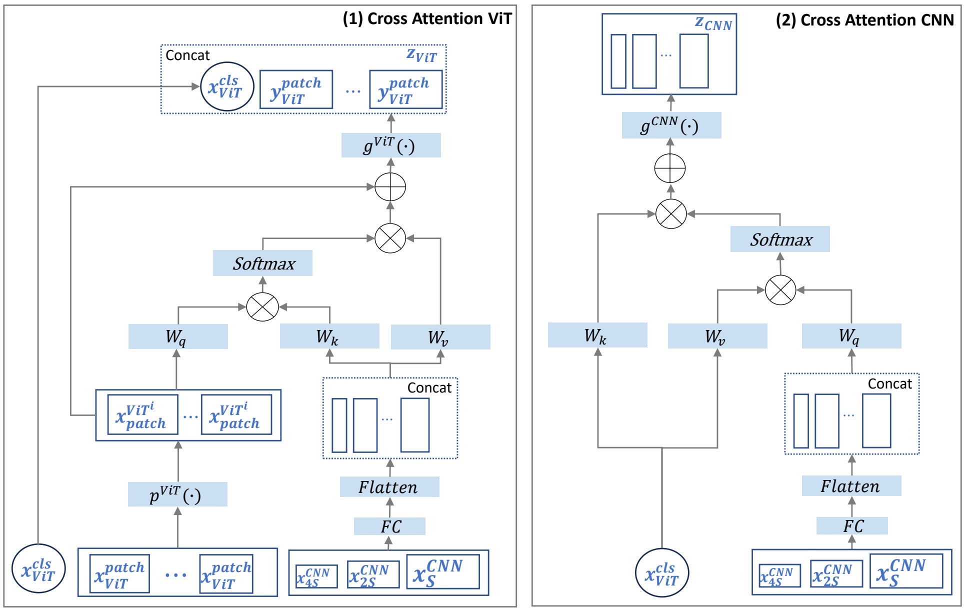

Fig. 2 Cross-attention module for the ViT branch and backbone segmentation model: **(1) Cross Attention ViT** The feature embedding from CNN serves as a query to interact with the patch tokens from the ViT branch; **(2) Cross Attention CNN** The CLS token of the ViT serves as a query token to interact with the feature map from CNN through attention.

of the original image, encompassing features with distinct receptive fields. Then we flatten and concatenate these feature maps, as illustrated in Eq. (1), serving as the key and value for the cross attention, where $||$ denotes the concatenation operation.

$$x_{CNN} = \text{Flatten}(\text{FC}(\left[x_{HW/4S}||x_{HW/2S}||x_{HW/S}\right])) \tag{1}$$

Here, S represents the reduction scaling factor of the feature map size from the first layer of the backbone segmenter to the original input size. By taking x_{ViT}^{patch}, the module then performs CA between x_{ViT}^{patch} and x_{CNN}. Mathematically, the CA can be expressed as:

$$q = x_{ViT}^{patch}W_q, \quad k = x_{CNN}W_k, \quad v = x_{CNN}W_v,$$
$$A = \text{softmax}\left(\frac{qk^T}{\sqrt{D/h}}\right), \quad CA(x_{CNN}) = Av$$

where $W_q, W_k, W_v \in \mathbb{R}^{D\times(D/h)}$ are learnable parameters, D and h are the embedding dimension and number of heads. Specifically, the output of the CA for ViT module, denoted as z_{ViT}, is defined by the input from ViT and CNN branches with projection operations and residual shortcut as follows:

$$y_{ViT}^{patch} = g^{ViT}(p^{ViT}(x_{ViT}^{patch}) + \text{CA}(x_{CNN})), \quad z_{ViT} = x_{cls}^{ViT}||y_{ViT}^{patch} \tag{2}$$

Where $p^{ViT}(\cdot)$ and $g^{ViT}(\cdot)$ are projections to align dimensions.

Cross Attention for CNN Our CA for CNN module is designed to facilitate information exchange between the global insights harnessed by the ViT branch and the localized details captured within the backbone segmentation encoder. The core mechanics of this process are akin to CA for ViT, albeit with a distinctive adjustment—here, the query and key/value roles are swapped.

More specifically, the multi-scale feature x_{CNN} in Eq. (1) now takes on the role of the query. For the key and value, we exclusively utilize the CLS token of the ViT feature embedding. The CLS token has already assimilated abstract information across all patch tokens within the ViT branch, constituting a global representation. This CA procedure can be concisely expressed as follows:

$$q = x_{CNN} W_q, \quad k = x_{ViT}^{CLS} W_k, \quad v = x_{ViT}^{CLS} W_v,$$

$$A = \text{softmax}(\frac{qk^T}{\sqrt{D/h}}), \quad CA(x_{ViT}^{CLS}) = Av$$

Note that the character definitions remain consistent with those in the CA for ViT module. Therefore, similar to the above, the output of the CA for CNN module with the residual shortcut can be defined as below:

$$z_{CNN} = g^{CNN}(x_{CNN} + \text{CA}(x_{ViT}^{CLS})) \tag{3}$$

where $g^{CNN}(\cdot)$ is the projection that aligns the dimension of the output feature map size to the input for the feed-forward network. This approach ensures that the size of the feature embeddings remains unchanged, while simultaneously amalgamating global insights from the ViT branch and local details from the CNN branch.

Feed Forward Network This module is a composite of key layers: convolution, activation, dropout for regularization, and a fully connected layer, working together to process and enhance the feature map to obtain z'_{CNN}. Their concerted efforts aim to extract vital features essential for the backbone segmentation model's decoder.

Data Flow The latter adapter takes the output from the previous adapter, z'_{CNN}, which has interacted with block i of the ViT branch, as its input for the subsequent CA for ViT module, engaging with the feature embedding $x_{ViT_{i+1}}$ from block $i+1$ of ViT. Note that the input of block $i+1$ is the sum of the output of CA for ViT in the previous adapter and the feature embedding x_{ViT_i} of block i, denoted as z_{ViT}. The final output of the last adapter, interfaced with the final ViT block, serves as the input for the backbone segmentation decoder.

3.3 Backbone Segmentation Model

For our backbone segmentation model, we use the UNet-like [13] structure. The encoder is constructed as a series of stride-2 3×3 convolutions and MaxPooling layers. The feature maps from each layer of the encoder are contacted to create multi-scale feature maps, subsequently fed into the adapter as shown in Section 3.2.

For the input to the backbone decoder, the ultimate feature map from the CNN encoder is combined with the output of the adapter which encompasses global insights from the ViT branch and local information from the CNN branch. Additionally, the patch tokens of the final feature embedding from the ViT branch were also contacted to preserve the contextual information of ViT.

Our backbone decoder is designed with a sequence of upsampling and convolutional layers. Significantly, we implement skip connections, a key feature that links feature maps at corresponding scales from the encoder to the decoder.

3.4 Implementation Details

Loss Function. In surgical image datasets, a substantial number of images predominantly comprise a background with no visible tools. Even in cases where tools are present, they often occupy a relatively small portion of the overall image. To address the class imbalance, we combine the Dice Loss with the Focal Tversky Loss for the assessment of our predictions against the ground truth segmentation map.

Model Configuration We construct our ViT feature encoder in three distinct sizes, denoted as ViT-T, ViT-S, ViT-B, and ViT-g , all pre-trained using the DINOV2 framework [5]. These models exhibit varying parameter counts for our adapters: 21M, 86M, 14.0M, and 300M, respectively. The number of attention heads is configured as 6, 6, and 12. In our setup, we chose a patch size of 14, resulting in a feature map scale of 1/14 for the ViT models. Additionally, for the CNN branch, the scaling factor S is set to 2, effectively leading to multi-scale feature maps with scales of 1/4, 1/8, and 1/16.

Hyper Parameters The input image is 588×588, considering the ViT branch's input requirement, and augmented with the image augmentation techniques presented in [4]. We adopt the SGD optimizer with a learning rate of 0.01 and momentum of 0.9. We applied the linear scaling rule to reduce the learning rate. The model is trained on 2 V100 GPUs, and the batch size is set to 16.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets Our binary segmentation experiments on the Robust-MIS 2019 [1] dataset utilized 5,983 annotated images for training, with three-stage testing, where stage 3 is from a procedure unseen during training. Multi-class segmentation was performed on EndoVis 2017 [18] and EndoVis 2018 [19]. Cross-dataset validation was conducted across the aforementioned datasets, along with CholecSeg8k [20] and AutoLaparo [21]. Each dataset was split into training and validation subsets at an 8:2 ratio with no patient overlap across folds.

Evaluation Metrics For the state-of-the-art comparison experiments on binary segmentation, we assessed our model using the metrics outlined in the Robust-MIS 2019 challenge [1], which includes Dice Similarity Coefficient and Normalized Surface Dice (NSD) [1]. Following the challenge's specifications [1], we adopted a 13-pixel tolerance for NSD. For the cross-dataset validation and ablation study, we also use the mean Intersection over Union (mIoU). For multi-class segmentation, we applied Ch_IoU, ISI_IoU, and mc_IoU following the evaluation metrics provided in [3, 22].

4.2 Results

Comparison to State-of-the-art In Table 1, we compare our model with several state-of-the-art models on Robust-MIS 2019 dataset for binary segmentation. Our model outperformed the CNN models designed for this task and the pre-trained ViT models for natural semantic segmentation downstream, indicating the success of merging the pre-trained knowledge with the CNN models. The existing hybrid approaches were trained for a shorter duration (smaller epochs) which signifies a potential for improvement. An essential takeaway here is that our proposed model exhibits superior efficiency: it requires minimal training to yield outstanding outcomes.

Table 1 Comparison on the Robust-MIS 2019 dataset between state-of-the-art models: above are the fully supervised CNN and ViT models for surgical segmentation task; the middle is the existing hybrid CNN-ViT models (all trained for 400 epochs); the bottom is the pre-trained ViT model for semantic segmentation downstream

Method	Whole Testing		Stage 1		Stage 2		Stage 3	
	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD
OR-Unet [2]	88.0	86.2	90.2	88.5	87.9	85.6	85.9	84.5
Robust-MIS 2019 winner [1]	90.1	88.9	92.0	92.7	90.2	88.6	89.0	86.4
ISINet [3]	88.9	86.3	90.9	87.6	89.6	86.5	86.2	84.7
TransUNet[9]	79.6	76.5	82.2	77.9	80.4	76.2	75.2	75.4
TransFuse [10]	80.1	78.6	82.2	79.1	81.3	79.0	76.8	77.7
Swin TransV2 [7]	82.9	78.6	84.6	80.2	84.0	79.9	80.1	75.7
MaskFormer [6]	84.1	80.5	87.2	84.3	85.9	80.2	79.2	77
Ours	92.9	91.5	94.2	92.4	92.6	91.4	91.9	90.7

For the multi-class segmentation task, we also compare our model with existing models including S3Net[22], TraSeTR[15], and MSLRGR [23]. Table 2 shows our model outperforms the state-of-the-art on the EndoVis 2018 dataset with +15.78 percentage point (pp) gain in mc_IoU. The improvements across both datasets demonstrate the multi-class segmentation capability of our model. Moreover, our model outperforms MSLRGR [23], which directly introduces global context into CNN, suggesting our approach of integrating the global information from pre-trained ViT is more effective than the state-of-the-art models.

Table 2 Comparison of our method with state-of-the-art methods on the EndoVis 2017 and EndoVis 2018 datasets for multi-class segmentation.

					EndoVis 2017					
Method	Ch_IoU	ISI_IoU	Bipolar Forceps	Prograsp Forceps	Large Needle Driver	Vessel Instrument	Grasping Applier	Monopolar Curved Scissors	Ultrasound Probe	mc_IoU
TraSeTR [15]	60.40	65.20	45.20	56.70	55.80	38.90	11.40	31.3	18.20	36.79
S3Net [22]	72.54	**71.99**	**75.08**	54.32	61.84	35.5	**27.47**	**43.23**	28.38	46.55
Ours	**73.96**	69.15	66.45	**67.56**	**70.52**	**42.68**	12.9	40.15	**29.12**	**47.06**
					EndoVis 2018					
TraSeTR [15]	76.20	-	76.30	53.30	46.50	40.60	13.90	86.30	17.50	47.77
S3Net [22]	75.81	74.02	77.22	50.87	19.83	50.59	0.00	**92.12**	7.44	42.58
MSLRGR [23]	-	-	69.66	43.56	0.15	34.71	3.87	87.16	12.03	35.88
Ours	**85.25**	**82.99**	**85.72**	**67.86**	**72.56**	**89.16**	**6.39**	91.07	**22.12**	**63.55**

Cross Dataset Validation We conducted experiments using a cross-dataset validation approach, where we trained the model on one dataset and validated it on another, shown as Table 3. We present comparative experiments between our model, the top-performing CNN model OR-Unet [2], and ViT based model MaskFormer [6].

Table 3 Cross dataset validation on EndoVis 2017, EndoVis 2018, CholecSeg8k, Robust-MIS 2019, and AutoLaparo datasets for OR-Unet [2], MaskFormer [6], and our method.

Train dataset	Model	Test Dataset									
		EndoVis 2017		EndoVis 2018		CholecSeg8k		Robust-MIS 2019		AutoLaparo	
		Mean Dice	mIoU	Mean Dice	mIoU	Mean Dice	mIoU	Mean Dice	mIoU	Mean Dice	mIoU
EndoVis 2017	OR-UNet [2]	92.4	81.3	73.0	62.4	74.3	65.7	59.2	10.3	74.5	56.7
	MaskFormer [6]	93.2	84.2	79.8	70.2	73.8	65.2	54.2	19.7	83.2	52.8
	Ours	98.9	96.2	94.2	85.8	85.9	80.7	88.4	80.6	89.9	69.7
EndoVis 2018	OR-UNet [2]	85.1	64.2	89.5	77.9	68.2	64.3	57.4	12.9	76.9	52.8
	MaskFormer [6]	84.3	72.2	88.2	81.8	74.8	61.9	56.7	31.9	77.9	65.9
	Ours	98.1	89.5	94.9	86.2	86.2	81.5	84.5	63.2	90.4	83.9
CholecSeg8k	OR-UNet [2]	82.3	71.4	69.9	53.2	82.7	75.4	51.5	8.2	69.7	61.4
	MaskFormer [6]	80.1	70.2	78.7	69.9	86.9	80.7	52.9	20.3	72.9	62.2
	Ours	95.9	88.6	92.1	82.8	91.9	86.6	90.1	83.5	90.2	82.4
Robust-MIS 2019	OR-UNet [2]	73.6	45.5	70.8	59.2	67.6	55.2	88.0	86.2	65.1	62.5
	MaskFormer [6]	86.4	79.0	81.8	70.1	77.2	62.7	84.1	80.5	71.8	65.2
	Ours	97.9	91.4	93.2	84.5	86.5	70.2	92.9	86.6	95.1	89.5
AutoLaparo	OR-UNet [2]	71.9	65.2	69.1	52.7	62.7	43.1	62.1	31.4	82.1	75.3
	MaskFormer [6]	85.1	73.6	79.0	60.8	76.4	63.2	60.5	37.4	92.7	84.9
	Ours	97.2	89.9	91.8	81.2	89.2	84.6	91.6	83.7	96.9	92.3

OR-UNet [2] and MaskFormer [6] experience significant performance drops when the training and testing datasets are different, while these variations are substantially reduced when they are trained and tested on the same dataset, yet the performance variability underscores their limited generalizability. Conversely, our model maintains consistent scores across different datasets, indicating its excellent robustness and accuracy. Some combinations, like training on EndoVis 2017 and testing on Robust-MIS 2019, show a more significant drop in performance than others, which hints at challenges the model faces when trained on a comparatively simpler dataset and tested on more complex, real-world data.

4.3 Ablation Study

Transformer Feature Encoder We conducted an ablation on the transformer feature encoder, and observed utilizing only the last layer resulted in a notable drop in both Dice scores and mIoU across datasets. However, by incorporating the last 3 layers, we observed performance metrics are close to that using all layers. Importantly, this configuration with the last 3 layers strikes a balance, offering near-optimal performance while being significantly more computationally efficient.

Adapter We conduct the ablation study with or without CA for ViT and CA for CNN as shown in Table 5. When CA modules are removed entirely, there's a substantial decrease in Dice and mIoU scores, highlighting their importance to the model's performance and robustness. The drop is less severe when CA is removed only for CNN, suggesting the importance of integrating patch tokens from the pre-trained ViT embeddings.

Table 4 Ablation studies on the transformer encoder when trained on Robust-MIS 2019 and tested on Robust-MIS 2019 and cross-dataset validated on CholecSeg8k dataset.

Transformer Encoder	Robust-MIS 2019		CholecSeg8k	
	Dice	mIoU	Dice	mIoU
All blocks	93.2	87.1	87.4	71.5
Last block	88.9	83.2	83.7	67.5
Last 3 blocks (ours)	92.9	86.6	86.5	70.2

Cross Attention for ViT We offer the ablation study for the adapter module in Table 5. For the CA for ViT module, we observe that 1) when solely relying on the single scale, there was a decrease of 3.4 pp in Dice scores on Robust-MIS 2019, indicating the significance of multi-scale features in capturing diverse spatial information; 2) Adopting the strategy of replacing the patch tokens with CLS token has led to some performance decreases, suggesting incorporating the global information from the CLS token, loses the local details that patch tokens offer; 3) Excluding the shortcut residuals leads to a drop in the Dice score by 2.3 pp for Robust-MIS 2019 gave the importance of residual shortcut to maintain information flow; 4) Even with variations in the ablation studies causing some drops in performance, the model's consistent decent scores on CholecSeg8k, underscores its superb generalization capability across datasets.

Table 5 Ablation studies for adapter when trained on Robust-MIS 2019 and tested on Robust-MIS 2019 and cross-dataset validated on CholecSeg8k dataset.

Adapter		Robust-MIS 2019		CholecSeg8k	
		Dice	mIoU	Dice	mIoU
× CA ViT&CNN		85.3	76.9	81.2	66.5
× CA ViT		88.9	80.1	82.6	65.7
× CA CNN		89.8	83.4	83.9	68.2
CA ViT	Single scale	89.5	81.9	83.2	68.2
	Patch → CLS	89.9	82.2	83.9	69.3
	× Residual	90.6	85.4	84.1	70.5
CA CNN	Single scale	91.3	87.4	82.7	68.6
	CLS → Patch	89.2	81.5	82.6	70.1
	× Residual	91.6	86.9	85.5	69.3
	Ours	92.9	86.6	86.5	70.2

Cross Attention for CNN In the context of the CA for CNN module shown in Table 5, using only the single scale results in a lesser decline in Dice score compared to that in CA for ViT, which suggests the output of CA for ViT already embodies multi-scale information, reducing its impact for the latter CA for CNN module. Opting to substitute the CLS token with patch tokens, despite being computationally costly, has observed a decrease in performance. This highlights the significance of integrating global information within the CNN branch.

5 Conclusion

In conclusion, our research presents an innovative approach to surgical image segmentation by combining ViT with a CNN used as a lightweight adapter module. Our work tackles the challenge of gathering large-scale annotated data and enhances the generalizability of different surgical scenarios. Our unique feature adapter, integrating cross-attention modules, facilitates the fusion of global and local, multi-scale spatial information from ViT and CNN, respectively. Our model achieves excellent accuracy and robustness across diverse surgical scenarios, as evidenced by our model's superior performance on the Robust-MIS 2019 dataset and across five other datasets. Our model has potential for applications in autonomous surgery, offering a solution that is both robust and adaptable to varying surgical environments.

Declarations

Supplementary information. Additional qualitative results and information are given in Online Resource 1.

Acknowledgements. This work was supported by core funding from the Wellcome/EPSRC[WT203148/Z/16/Z; NS/A000049/1]. MW is supported by the UKRI EPSRC CDT in Smart Medical Imaging [EP/S022104/1]. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Conflict of Interest. TV is a co-founder and shareholder of Hypervision Surgical. The authors declare that they have no other conflict of interest.

Ethical Approval / Informed Consent. This article only uses publicly available datasets. Their re-use did not require any ethical approval.

References

[1] Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Filimon, D.M., Scholz, P., Tran, T.N.: Robust medical instrument segmentation challenge 2019. arXiv preprint arXiv:2003.10299 (2020)

[2] Isensee, F., Maier-Hein, K.: Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images. ArXiv **abs/2004.12668** (2020)

[3] González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: MICCAI, pp. 595–605 (2020). Springer

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[5] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[6] Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS **34**, 17864–17875 (2021)

[7] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: CVPR, pp. 12009–12019 (2022)

[8] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR, pp. 1290–1299 (2022)

[9] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[10] Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: MICCAI, pp. 14–24 (2021). Springer

[11] Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: MICCAI, pp. 61–71 (2021). Springer

[12] Yuan, F., Zhang, Z., Fang, Z.: An effective cnn and transformer complementary network for medical image segmentation. Pattern Recognition **136**, 109228 (2023)

[13] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015). Springer

[14] Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: Matis: Masked-attention transformers for surgical instrument segmentation. ISBI, 1–5 (2023)

[15] Zhao, Z., Jin, Y., Heng, P.-A.: Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: ICRA, pp. 11186–11193 (2022). IEEE

[16] Gheini, M., Ren, X., May, J.: On the strengths of cross-attention in pretrained transformers for machine translation. ArXiv **abs/2104.08771** (2021)

[17] Liu, M., Yin, H.: Cross attention network for semantic segmentation. In: ICIP, pp. 2434–2438 (2019). IEEE

[18] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

[19] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)

[20] Hong, W.-Y., Kao, C.-L., Kuo, Y.-H., Wang, J.-R., Chang, W.-L., Shih, C.-S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)

[21] Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.-H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: MICCAI, pp. 486–496 (2022). Springer

[22] Baby, B., Thapar, D., Chasmai, M., Banerjee, T., Dargan, K., Suri, A., Banerjee, S., Arora, C.: From forks to forceps: A new framework for instance segmentation of surgical instruments. In: WACV, pp. 6191–6201 (2023)

[23] Seenivasan, L., Mitheran, S., Islam, M., Ren, H.: Global-reasoned multi-task learning model for surgical scene understanding. RA-L **7**(2), 3858–3865 (2022)