Parallel Cross Window Attention Transformer and CNN Model for Segmentation of Instrument during Surgery

Abdul Qayyum
Imperial College, London, United Kingdom, Email: a.qayyum@imperial.ac.uk

Steven Niederer
The Alan Turing Institute, London, United Kingdom

M. K. A. Ahamed Khan
Department of Mechanical and Mechatronics Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, Malaysia.
Email: mohamedkhan@ucsiuniversity.edu.my

Moona Mazher
Centre for Medical Image Computing, Department of Computer Science, University College London, UK
Email: m.mazher@ucl.ac.uk

Imran Razzak
University of New South Wales, Sydney, Australia

Mastaneh Mokayef
UCSI University, Faculty of Engineering, Malaysia

C. S. Hassan
UCSI University, Faculty of Engineering, Malaysia

Ridzuan, A
UCSI University, Faculty of Engineering, Malaysia

Abstract

Precise segmentation of surgical instruments is a fundamental component in the development of computer-aided surgery systems by assisting the surgeons to navigate the patient's body aiming to enhance the surgical precision and patient safety. Though real-time tracking of surgical instruments is critically important in invasive computer-assisted surgeries, it is challenging to achieve a highly sensitive and accurate system in complex surgical environment. Recently, synthetic data for instrument segmentation in surgery (Syn-ISS) challenge using synthetic datasets is organized to develop high performance methods for instrument segmentation. In this work, we present encoder and decoder-based hybrid parallel cross window attention-based transformer during the feature extraction, which consists of the multi-scale channel attention, convolutional layers, and Transformer layers, forming a unified block. Syn-ISS challenge dataset comprised of two tasks. In first task1, they need to develop deep learning-based method for binary instrument segmentation and in second task multiclass instrument segmentation is required. Experiments conducted on Syn-ISS dataset achieved 0.993 F-score for task 1 and 0.993, 0.975, and 0.951 F-score for shaft, wrist, and jaw segmentation respectively for Task 2.

Keywords: Deep Learning, Parallel Cross Window Attention, Transformer, 2D Instrument Surgery segmentation, Dense Net.

1. Introduction

Minimally invasive segmentation using optical imaging systems have gained popularity in modern healthcare due to their advantages, including reduced patient recovery time and lower mortality rates. Optical imaging has enabled the use of robotic platforms such as the da Vinci surgical system by Intuitive Surgery for complex minimally invasive surgeries [1]. Nevertheless, during endoscopic surgical suturing procedures, the presence of surgical instruments can impede surgeons' dexterity due to the confined working space and limited visual field-of-view. These visual obstructions elevate the risk of tissue scars and tears. Therefore, the crucial task is to transparently remove or mask the surgical instruments from the background and subsequently fill the masked region with appropriate background content. Automated segmentation of surgical instruments in MIS is currently a focal point of research due to its significant practical applications [2].

The challenges associated with surgical instrument segmentation are diverse and contingent upon factors such as the source of dataset acquisition, the type of surgical procedure, the specific instruments or tools in use, image resolution, dataset scale, tool characteristics, and challenging conditions like occlusions, rapid appearance alterations, specular reflections, smoke, blur, and blood spatter. Segmentation of surgical instruments has been formulated using both instance segmentation [2] and semantic and segmentation [2].

Recently, Syn- ISS (synthetic data for instrument segmentation in surgery) challenge using synthetic datasets is organized to develop high performance methods for instrument segmentation [3]. Recently, there is a different method has been proposed using medical imaging and signals [4], [5], [6], [7], [8] for classification and segmentation. Based on our previous work on segmentation [6], [7], [8], [9]. We presented encoder-decoder based hybrid transformer and CNN model for instrument segmentation with parallel cross window attention-based transformer block in encoder and 2D Dense Net layer CNN blocks in decoder. The main contribution in this work is:

i. Developed Parallel Cross Window Attention Transformer encoder block for 2D segmentation task.
ii. Proposed 2D Dense Net block at decoder side of proposed model using transformer-based encoder features.
iii. Compare performance on synthetic data for instrument segmentation in surgery for binary and multiclass surgery instrument segmentation.

2. Methodology

2.1. Parallel Cross Window Attention Transformer and CNN model for instrument segmentation

Due to the intrinsic locality of convolution, which is incapable of modelling long-range dependencies. In addition, Transformer generates single scale features with only token wise attention, and it ignores the relationship among channels, thus subpar to tackle situations such as segmenting multi scale lesion regions in medical images. Considering these issues, we introduce the hybrid Transformer block during the feature extraction, which consists of the multi-scale channel attention, convolutional layers, and Transformer layers, forming a unified block. Our proposed model is based on encoder and decoder layers, and we have proposed a hybrid transformer and CNN model for instrument segmentation. Our proposed model consisted of Parallel Cross window attention-based transformer block on the encoder side and 2D Dense Net layer CNN blocks on the decoder side. We have used Dense Net 201 based layers on the decoder side. The proposed model is shown in Figure 1 (a) and Figure 1 (b).

2.2. *Parallel Cross Window Attention Transformer Block:*

An efficient hybrid segmentation framework consisting of integration of convolutional neural network and learnable global attention heads using Efficient Parallel-Cross Attention module. We use the depth-wise separable convolution as an efficient version of convolution implemented by depth wise conv and pointwise conv, where the depth wise convolution gathers the spatial information while the pointwise convolution gathers along the channel dimension. Furthermore, we have concatenated features from multi-window transformer block with depth wise convolutional layer. In encoder side, we have used transformer-based block aided with cross attention window-based mechanism, however, we have used normal 2DCNN based module at decoder-side. The features are concatenated from window area partition and window partition and further these features are concatenated with depth wise convolutional layer features and then pass these features to next layer of the encoder block.

Unlike the vision transformer (ViT) that computes relationship between tokens at each step of self-attention module, swin transformer is based on computation of attention within partition of non-overlapping local windows of lower resolution feature map and original image. In contrast to the original swin transformer that uses patch merging layer to empower it for pixel level tasks, we used rectangular-paralleled-piped windows to accommodate non-square images using Parallel-Cross Attention approach (window area partition and window partition). To extract different feature maps from each convolutional block with parallel window-based transformer in encoder, each block consists of Parallel Cross Window Attention Transformer block, patch merging and depth wise [7] convolutional layer. The proposed block is shown in Figure 1 (b).

2.3 Convolutional Neural Network (CNN) block:

To better encode spatial location information and inject strong inductive bias, we adopt convolutional block to extract local spatial features. Specifically, given an input feature F_(i-1), we adopt the convolutional block to model local spatial features, which are shown as follows:

$$F_c^i = ConvBlocks_i(F_{i-1}), i \in \{1,2,3,4\} \qquad (1)$$

$where\ F_c^i \in R^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$ is the local features, which contains 2D spatial location information, making it possible to encode position information. The convolutional blocks in each decoder stage consisted of DensNet-201 based architecture is shown in Figure 1 (a).

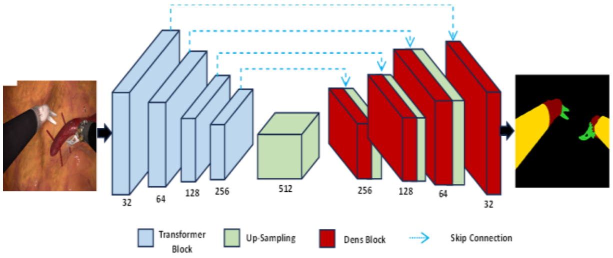


Fig. 1 (a) The proposed model based on hybrid transformer and Dense Net CNN for instrument and instruments parts segmentation

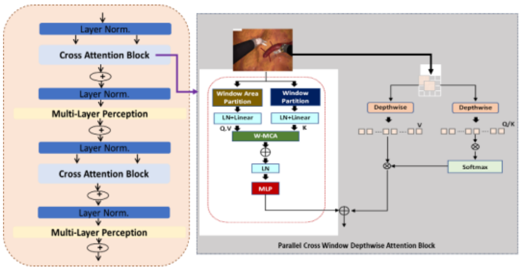


Fig. 1 (b) The proposed Transformer block based on parallel cross window depth wise attention module

In the encoder block, the spatial input size has been reduced with an increasing number of feature maps and on the decoder side, the input image spatial size has been increased using a 2D Conv-Transpose layer. The input features' maps that are obtained from every encoder block are concatenated with every decoder block feature map to reconstruct the semantic information. The spatial size doubled at every encoder block and feature maps are halved at each decoder stage of the proposed model. The feature concatenation has been done at every encoder and decoder block except the last 1x1 convolutional layer. The three-level deep-supervision techniques are applied to get the aggregated loss between ground truth and prediction.

3.0 Experiment

3.1 Dataset

Recently, Syn-ISS challenge [3] is introduce in MICCAI-2023 to develop high performance methods for instrument segmentation. Syn-ISS dataset is synthetic instrument segmentation datasets consist of two main tasks. The task-1 is a binary segmentation to annotate all pixels that contain an instrument and consist of 1200 instances of simulated scene along with computer generated corresponding masks. Task-2 further focuses on segmentation of pixels belonging to different parts of the instrument and consists of 1800 instances.

3.2 Network Setting

We adopted data augmentation of horizontal flips, vertical flips, and random rescales. The network is trained for 200 epochs using the Adam optimizer and the weight decay is 0.0001. We have used binary cross-entropy and dice loss used for training and optimization [10], [11]. [12]. The dataset has different spatial size, hence, we resized each image and label sample to 512x512, whereas we resized each sample to original input 2D image size by bilinear interpolation during inference time. The 25-batch size is used during training. The model is training on A6000 GPU machine with 4 GPUs and all model codes are developed from scratch using Pytorch Library.

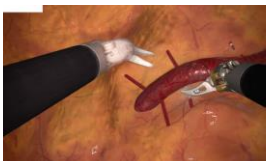
4.0 Result

The training dataset has been divided into 80 percent training and 20 percent validation. We have trained and validated our proposed model using 5-fold cross validation and based on the best validation score, the proposed model has been submitted for task 1 binary segmentation and task 2 multiclass segmentation. We have evaluated the performance using IOU, F-Score, Recall, Precision, and HD. The best score produced by our proposed model based on validation dataset is shown in Table.1 for Task 1 and Task 2.

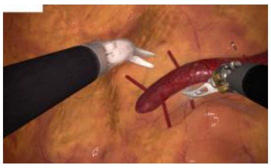
Table 1. The performance analysis of proposed solution for Task1 and Task2.

Algorithms	**Tasks**	**classes**	**IOU**	**F-Score**	**Recall**	**Precision**	**HD**
Proposed model	Task 2	shaft	0.9868	0.9930	0.9923	0.9920	14.23
		wrist	0.9497	0.9753	0.9753	0.9695	10.67
		jaw	0.9086	0.9506	0.9523	0.9357	12.88
Proposed model	Task 1	instrument	0.9868	0.9933	0.9934	0.9909	0

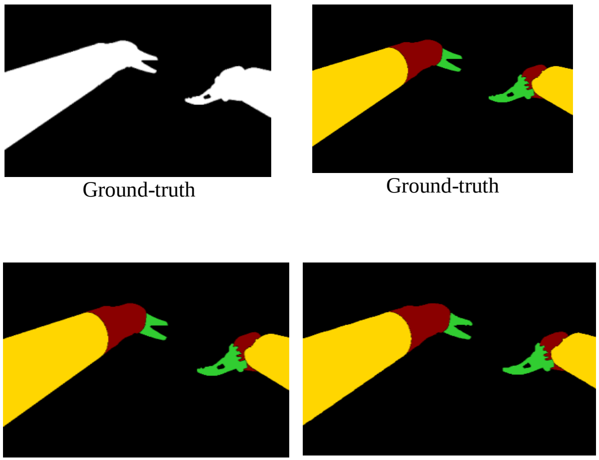
The visualization of validation input image for binary and multiclass segmentation is shown in Figure 2. Our proposed produced similar prediction masks as compared to ground-truth segmentation masks.



input



input



Ground-truth

Ground-truth

Prediction

Prediction

Fig. 2 The input, ground-truth, and predicted segmentation masks. shaft, wrist, and jaws are shown using a yellow, red, and green segmentation mask.

5.0 Conclusion

Ideally, surgical instrument segmentation is used in real time, to identify tools being used as a surgery is being performed. The integration of surgical instrument segmentation into computer-aided surgery systems offers numerous benefits, including real-time guidance, instrument tracking, and improved surgical outcomes. In this work, we presented an encoder and decoder-based hybrid parallel cross window attention-based transformer which consists of the multi-scale channel attention, convolutional layers, and Transformer layers. Experiments conducted on Syn-ISS challenge dataset achieved 0.993 F-score for task-1 and 0.993, 0.975, and 0.951 F-score for shaft, wrist, and jaw segmentation respectively for Task 2.

6. References

1. Cristina Gonz´alez, Laura Bravo-S´anchez, and Pablo Arbelaez. Isinet: an instancebased approach for surgical instrument segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 595–605. Springer, 2020.

2. Mobarakol Islam, VS Vibashan, and Hongliang Ren. Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery. In 2020 IEEE international conference on robotics and automation (ICRA), pages 8433–8439. IEEE, 2020.

3.https://www.synapse.org/#!Synapse:syn50908388/wiki/620516

4. Payette, Kelly, Hongwei Bran Li, Priscille de Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu et al. "Fetal brain tissue annotation and segmentation challenge results." Medical Image Analysis 88 (2023): 102833.

5. De Vente, Coen, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn et al. "AIROGS: Artificial Intelligence for robust glaucoma screening challenge." IEEE Transactions on Medical Imaging (2023).

6. Chen, Zhihao, Alain Lalande, Michel Salomon, Thomas Decourselle, Thibaut Pommier, Abdul Qayyum, Jixi Shi, Gilles Perrot, and Raphaël Couturier. "Automatic deep learning-based myocardial infarction segmentation from delayed enhancement MRI." Computerized Medical Imaging and Graphics 95 (2022): 102014.

7. Qayyum, Abdul, Mona Mazhar, Imran Razzak, and Mohamed Reda Bouadjenek. "Multilevel depth-wise context attention network with atrous mechanism for segmentation of COVID19 affected regions." Neural Computing and Applications (2021): 1-13.

8. Ahmad, Iftikhar, Abdul Qayyum, Brij B. Gupta, Madini O. Alassafi, and Rayed A. AlGhamdi. "Ensemble of 2D residual neural networks integrated with atrous spatial pyramid pooling module for myocardium segmentation of left ventricle cardiac MRI." Mathematics 10, no. 4 (2022): 627.

9. Ahmad, R. F., Malik, A. S., Kamel, N., Amin, H., Zafar, R., Qayyum, A., & Reza, F. (2014, November). Discriminating the different human brain states with EEG signals using Fractal dimension: A nonlinear approach. In 2014 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) (pp. 1-5). IEEE.

10. Qayyum, Abdul, Moona Mazher, Tariq Khan, and Imran Razzak. "Semi-supervised 3D-InceptionNet for segmentation and survival prediction of head and neck primary cancers." Engineering Applications of Artificial Intelligence 117 (2023): 105590.

11. Eisenmann, Matthias, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J. Adler, Patrick Godau et al. "Biomedical image analysis competitions: The state of current participation practice." arXiv preprint arXiv:2212.08568 (2022).

12. Qayyum, Abdul, Aamir Malik, Naufal M Saad, and Moona Mazher. "Designing deep CNN models based on sparse coding for aerial imagery: a deep-features reduction approach." European Journal of Remote Sensing 52, no. 1 (2019): 221-239.

Authors Introduction

Dr. Abdul Qayyum

He is currently working at National Heart and Lung Institute Imperial College London, UK. Previously, he was joined as lecturer at University of Bourgogne Franche-Comté France. He received his Ph.D in electrical & electronics engineering with specialization in deep learning and image processing in 2017 from Universiti Teknologi Petronas Malaysia. His area of interest is machine learning, deep learning and quantum machine learning for signal processing and bioemdical imaging.

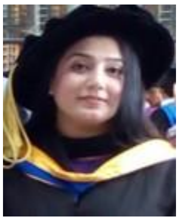
Prof. Steven Niederer

He completed his undergraduate degree in Engineering Science at the University of Auckland in 2003 and his DPhil at the University of Oxford in 2008. In 2023, he moved to Imperial College London as the Chair in Biomedical Engineering at the National Heart and Lung Institute. His current work is focused on reducing barriers to adopting digital twin technology, developing virtual patient cohorts for in-silico trials, mapping organ scale function through to cellular and molecular physiology, and using modelling and simulation to personalize and guide therapies.

Dr M. K. A. Ahamed Khan

He is currently working at UCSI University, Malaysia. He received his Ph.D in Robotics and controls from USA. His area of research is robotics, AI and controls. He has published more than100 papers. He is also an IEEE Senior member. He is also the past chair for IEEE RAS Malaysia chapter

Dr. Moona Mazher



She is a senior postdoc research fellow at Department of Computer Science, University College London. She received her Ph.D. from the University of Rovira i Virgili, Spain, 212 a specialization in Neuroscience from Universiti Teknologi PETRONAS, Malaysia in 2017.. Her areas of interest are machine learning, deep learning, medical imaging, signal processing, computer vision, and explainable AI.

Dr. Imran Razzak

He is a Senior Lecturer in Human-Centered Machine Learning in the School of Computer Science and Engineering at University of New South Wales, Sydney, Australia. Previously, he was as a Senior Lecturer in Computer Science at School of IT, Deakin University, Victoria. His area of research focuses on connecting language and vision for better interpretation of multidimensional data and spans over three broad areas: Machine Learning, Computer Vision, and Natural Language Processing.

Dr. Mastaneh Mokayef

He is currently working at National Heart and Lung Institute Imperial College London, UK. Previously, he was joined as lecturer at University of Bourgogne Franche-Comté France. He received his Ph.D in electrical & electronics engineering with specialization in deep learning and image processing in 2017 from Universiti Teknologi Petronas Malaysia. His area of interest is machine learning, deep learning and quantum machine learning for signal processing and bioemdical imaging.

Dr Cik Suhana Hassan

She currently works at UCSI University. She received her PhD from UTP Malaysia. She has been doing research in Mechanical Engineering and Materials Engineering. She is passionate about turning environmental waste into value-added products as part of her quest to live a more environmentally friendly life.

Ts Amar Ridzuan Bin Abd Hamid

He is a lecturer of Mechanical and Mechatronic programmes from Department of Mechanical Engineering, UCSI University, Malaysia. He has completed his Master Degree from Universiti Putra Malaysia, Postgraduate Diploma from UCSI, and a Bachelor Degree with Hounours in Mechanical Engineering (Automotive) from Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.