*Research Article*

# CST: A Multitask Learning Framework for Colorectal Cancer Region Mining Based on Transformer

**Dong Sui** [iD],[1] **Kang Zhang,**[1] **Weifeng Liu,**[1] **Jing Chen,**[2] **Xiaoxuan Ma,**[1] **and Zhaofeng Tian** [iD][2]

[1]*School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China*
[2]*Department of Laboratory and Diagnosis, Changhai Hospital, Navy Medical University, Shanghai 200433, China*

Correspondence should be addressed to Dong Sui; suidong@bucea.edu.cn and Zhaofeng Tian; tian_zhao_feng@163.com

Colorectal cancer is a high death rate cancer until now; from the clinical view, the diagnosis of the tumour region is critical for the doctors. But with data accumulation, this task takes lots of time and labor with large variances between different doctors. With the development of computer vision, detection and segmentation of the colorectal cancer region from CT or MRI image series are a great challenge in the past decades, and there still have great demands on automatic diagnosis. In this paper, we proposed a novel transfer learning protocol, called CST, that is, a union framework for colorectal cancer region detection and segmentation task based on the transformer model, which effectively constructs the cancer region detection and its segmentation jointly. To make a higher detection accuracy, we incorporate an autoencoder-based image-level decision approach that leverages the image-level decision of a cancer slice. We also compared our framework with one-stage and two-stage object detection methods; the results show that our proposed method achieves better results on detection and segmentation tasks. And this proposed framework will give another pathway for colorectal cancer screen by way of artificial intelligence.

## 1. Introduction

Colorectal cancer is a common malignancy tumour worldwide, which has ranked the third position as the most common cancer and the second cause of cancer-related deaths worldwide. Also, the 2021 analysis observes that the diagnosed patients are rising in the crowd younger than 50 years old [1]. In China, there are more than 480,000 new cases with a higher than 30% death percentage in 2020, which increases the incidence and mortality rates rank following lung cancer [2]. In the early stages, occult blood examination and medical images were employed for clinical detection and diagnosis. These methods exhibited a productive approach for the early colorectal cancer diagnosis and can improve the survival of these patients [3]. However, mere blood examination and colonoscopy inspection could not reveal the biological morphology and tumour statutes [4, 5]. In the past decades, imaging approaches such as computed tomography (CT) and magnetic resonance imaging (MRI)

have become an effective way for colorectal cancer diagnosis; doctors can get an overall scheme of the tumour region in a comprehensive way without invasion [2, 6]. In addition, medical imaging can get detailed information of the tumour region without any physical cathartic cleansing, which has become a prevalent screening guideline [7, 8].

Medical image processing and analysis have achieved remarkable progress in the past several years, especially the use of deep learning, but this field is still challenging for the difficulty of acquisition and annotation in medical imaging datasets [9, 10]. In this context, transfer learning is another pathway for handling the lack of annotated medical data with small-scale data training and becomes a common protocol instead of traditional supervised learning [9, 10]. As an effective way, the transfer learning protocol in medical image processing usually employed the ImageNet pretrained deep architecture, *e.g.*, ResNet, DenseNet, FCN, and U-Net family, and then, these models are fine-tuned on small-scale medical images to fit some certain tasks [11–15]. These

fine-tuned models yield much better results than bottom-up training strategy, especially when it confronts a small set of image samples.

However, despite 3D or 2D medicine having the same image structures, it is worth emphasizing that especially in tasks of computer vision, these medical images have distinct interest with the natural image benchmark datasets. In the case of medical data, most of the ROIs took part in a small region in the whole image, resulting in sophisticated yet hard example problems [16, 17]. Learning a medical image analysis network transferred from the natural scene usually leads to strong bias without considering the characteristic of the medical image. Thus, all of the above formations motivate developing a transfer learning protocol directly by a medical imaging dataset that can handle subtle variances in these two datasets. In this situation, ConvNets become a popular backbone in many medical image and nature image processing protocols. But for the 3D image tasks, the traditional ConvNets cannot perform long-range dependency modelling. In addition, applications of pyramid ConvNets and some attention mechanisms can facilitate the processing of these sequential image datasets [18]. However, most of these methods have not focused on long-range medical image pipeline on multitask tumour region analysis.

Transformer models such as BERT and DERT have successfully achieved the state of the art in nature language processing and computer vision fields [19–22]. Due to its ability to learn long-range dependencies from input tokens, the self-attention mechanism can model the dependency among the input tokens. The famous vision transformer (ViT) have achieved comparable performance with the traditional deep learning model such as the CNN model on image recognition tasks [23]. But all of these models have to be trained on a large-scale dataset. DeiT (Data-efficient Image Transformers) is the first transformer-based model adapted by midsized datasets [24].

In this work, we propose a novel transfer learning protocol, called *CST*, which is a union framework for colorectal cancer region detection and segmentation task based on the transformer model, which effectively constructs the cancer region detection and its segmentation jointly. To make a higher detection accuracy, we incorporate an autoencoder-based image-level decision approach that leverages the image-level decision of a cancer slice. First of all, we pretrain an encoder-decoder architecture for cancer/normal image slice representation that generates the encoding vectors of the original input image slices as the image-level label. Then, another transformer-based global to local architecture is pretrained by our colorectal medical image datasets for tumour region detection and segmentation. To validate the effectiveness of our proposed framework, we test the model output on the collected colorectal cancer MRI image series and achieved remarkable performances compared with other traditional methods.

In summary, our main contributions are as follows:

(i) We propose a novel framework for colorectal cancer region detection and segmentation. Our framework provides a more flexible pathway for tumour region mining

(ii) We combine the traditional autoencoder and transformer architecture together for the multitask framework for the final decision

(iii) We evaluate the proposed method on the colorectal cancer MRI image dataset, and our method has achieved a better result on tumour region detection and segmentation

## 2. Related Works

Convolution neural networks with their excellent feature representation ability have raised a revolution in the nature language processing field, as well as the computer vision and signal processing fields [25]. Position-sensitive tasks such as semantic segmentation that contains several parts of ROIs have been well represented by using convolutional encoder-decoder architectures [26, 27]. The main aim of convolutional operation is to catch local texture feature information by the convolution kernel, and more layers and stride kernel in the receptive field can extend the capture range during downsampling; in this way, the model can capture global to local information explicitly. However, the size and shape of these kernels are usually of fixed size and cannot adapt to all the input range [28–30].

Recent advance in transformer-based architecture with a self-attention mechanism and the ability of long-range modelling has achieved the state of the art in natural language processing and computer vision [31]. Vision transformer (ViT) can treat the whole image into several patches and feed into the transformer pipeline as tokens. The simple application of the transformer has shown excellent results compared with the traditional CNN model [23]. However, the computation cost and large-scale dataset are the fatal drawbacks to competing with the convolutional neuronal network.

In the medical image processing paradigm, few annotated clinical data cannot generate efficient models and have to use the ImageNet pretrained model for the downstream tasks [32]. In practice, most of the prevalent methods use their pretrained weights for the medical ROI detection, such as ResNet and DenseNet, and fine-tune the higher layers on some special tasks. But most of these strategies is limited to applying for new datasets [11–15]. In a word, the pretrained model on ImageNet and other datasets by fully supervised learning paradigm have to be severed with massive annotated datasets to fit the downstream transfer learning. In another way, a self-supervised learning framework can get a suitable result by using few or no need of labelled datasets; this has gained great attention in medical image analysis recently [33, 34]. Furthermore, the self-supervised learning paradigm has attracted great attention in the medical image analysis field [35]. The critical challenge for self-supervised learning is how to define a suitable proxy task from the unlabelled data. But most of these proxy tasks have exhibited less use on medical image-related tasks.

# 3. Material and Methods

*3.1. Dataset.* We construct a novel dataset for this research, which contains 375 cases of colorectal cancer tumour MRI image datasets from 2013 to 2020, which contains 289 CRM negatives and 86 CRM positives. For segmentation and detection tasks, we also collect 375 cases of colorectal cancer negative samples as the negative samples for deep architecture training. Our collaborator labels the image slices with mask and bounding box separately as the final ground truth. Then, the dataset was divided into training, testing, and validation sets for the network training and evaluation. The main aims of this dataset are to collect for colorectal cancer region detection and segmentation, and to follow this aim, we construct the framework in this manuscript to perform them and prepared for the clinical applications. Figure 1 shows the details of the labelled tumour region about CRM negative and positive, respectively.

*3.2. Multitask Framework.* The motivation of our work is to construct a multitask framework that combines tumour region detection and segmentation tasks. In this section, we illustrate the overall framework of our proposed CST framework as shown in Figure 2. Our framework is divided into two pipelines, the tumour region detector and the tumour segmentation pipeline. In the detection pipeline, we first generate the region proposal of the input images, and an encoder-decoder model is used for the position encoder as the DETR input. In the segmentation pipeline, we use image patches as the input and project to a sequence of embedding for the transformer, and the class embedding is used for the final mask prediction.

*3.2.1. Detector Pipeline.* In this part, we start from the region proposals generated from an input medical image with $H$ and $W$ in height and width as the initial image $x_{\mathrm{rpn}} \in \mathbb{R}^{3 \times H_0 \times W_0}$ with 3 channels, $H_0$ in height and $W_0$ width of the RPN. We choose a conventional CNN backbone to generate the lower resolution activation map $f \in \mathbb{R}^{C \times H \times W}$; typically, the values of $C = 2048$; $H'$ and $W'$ are resized as the initial input of the setting $H' = H/32$, $W' = W/32$. Unfortunately, the position encoding method in the original image only reflects the location of the pixels in the column and row, but the input position in our pipeline is from the random selected RPNs, so in this part, we pretrained an autoencoder for the position representation and we added this coding to the traditional position encoding with the anchor position together.

For the transformer encoder, we use a $1 \times 1$ convolution to reduce the dimension of the activation map into a $d$-dimension vector; for the input of the transformer, the feature map is collapsed into a 1-dimension vector with a $d \times HW$ feature map. Each encoder layer is adopting the standard setting as is stated in the DETR [19–22]. For the transformer decoder, it follows the standard architecture of the transformer, and the model can decode the tumour region at each decoder layer. Each object/RPN is transformed into an output embedding by the decoder. They are decoded into bounding box coordinates and tumour/nontumour class

labels by the following feed forward network (FFN). The FFN is a 3-layer perceptron with ReLU, hidden dimension $d$, and a linear projection layer. It can predict the normalized centre, height, and width of the tumour bounding box. In addition, the tumour and nontumour class of the detected bounding box is predicted by a SoftMax function.

The loss function of the tumour detection part is to optimize the lowest cost:

$$\widehat{\sigma} = \arg\min_{\sigma \in \wp_N} \sum_i^N L_{\mathrm{match}}\left(y_i, \widehat{y}_{\sigma(i)}\right), \quad (1)$$

$$L_{\mathrm{match}}\left(y_i, \widehat{y}_{\sigma(i)}\right) = L_{\mathrm{box}}\left(b_i, \widehat{b}_{\sigma(i)}\right) - \widehat{p}_{\sigma(i)}(c_i). \quad (2)$$

For efficient computing, we choose optimal assignment with the Hungarian algorithm to accelerate the training process. Here, we use Hungarian loss for all pairs matched, and the object detector loss is defined like similarly loss; the total Hungarian loss is defined as follows:

$$L_{\mathrm{Hungarian}}(y, \widehat{y}) = \sum_{i=1}^N \left[ -\log \widehat{p}_{\widehat{\sigma}(i)}(c_i) + L_{\mathrm{box}}\left(b_i, \widehat{b}_{\widehat{\sigma}}(i)\right) \right]. \quad (3)$$

$$L_{\mathrm{box}}\left(b_i, \widehat{b}_{\widehat{\sigma}}(i)\right) = \lambda_{\mathrm{iou}} L_{\mathrm{iou}}\left(b_i, \widehat{b}_{\sigma(i)}\right) + \lambda_{L1} \left\| b_i - \widehat{b}_{\sigma(i)} \right\|_1, \quad (4)$$

where for the prediction with index $\sigma(i)$, $y_i = (c_i, b_i)$ is the labelled ground truth, and we define the probability of the tumour region; we define $c_i$ as $\widehat{p}_{\widehat{\sigma}(i)}(c_i)$ and the predicted bounding box as $\widehat{b}_{\sigma(i)}$, $\lambda_{\mathrm{iou}} \in \mathbb{R}$ and $\mathbb{R}_{L1} \in \mathbb{R}$.

*3.2.2. Segmenter Pipeline.* The segmenter part is based on full transformer-based architecture for pixel-level class annotation. As shown in the upper part of Figure 2, we model the sequence of patches by using a transformer encoder and a point-wise linear mapping or a mask transformer. The whole pipeline is trained end-to-end with cross-entropy loss per pixel.

In the encoder part, we first split the input image into a sequence of identical size patches, and each patch is flattened into a 1-dimension vector to produce a sequence of patch embeddings. For the position information encoding, we treat each patch as a separated part from the whole image and finally add position information to the original patch position. After that, the traditional transformer encoder is employed for the sequential information encoding with a multihead self-attention block. For the decoder part, it first learns to map patch-level encodings from the encoder to patch-level class scores; following that, these scores are unsampled to pixel-level scores by bilinear interpolation. The whole mask transformer is illustrated in the lower part of Figure 2.

For the mask transformer, we use a set of $K$ learnable class embeddings in the decoder; in our pipeline, $K$ is 2. Each class embedding is randomly initialized and assigned to a single class so as to generate the class mask. At last, the class
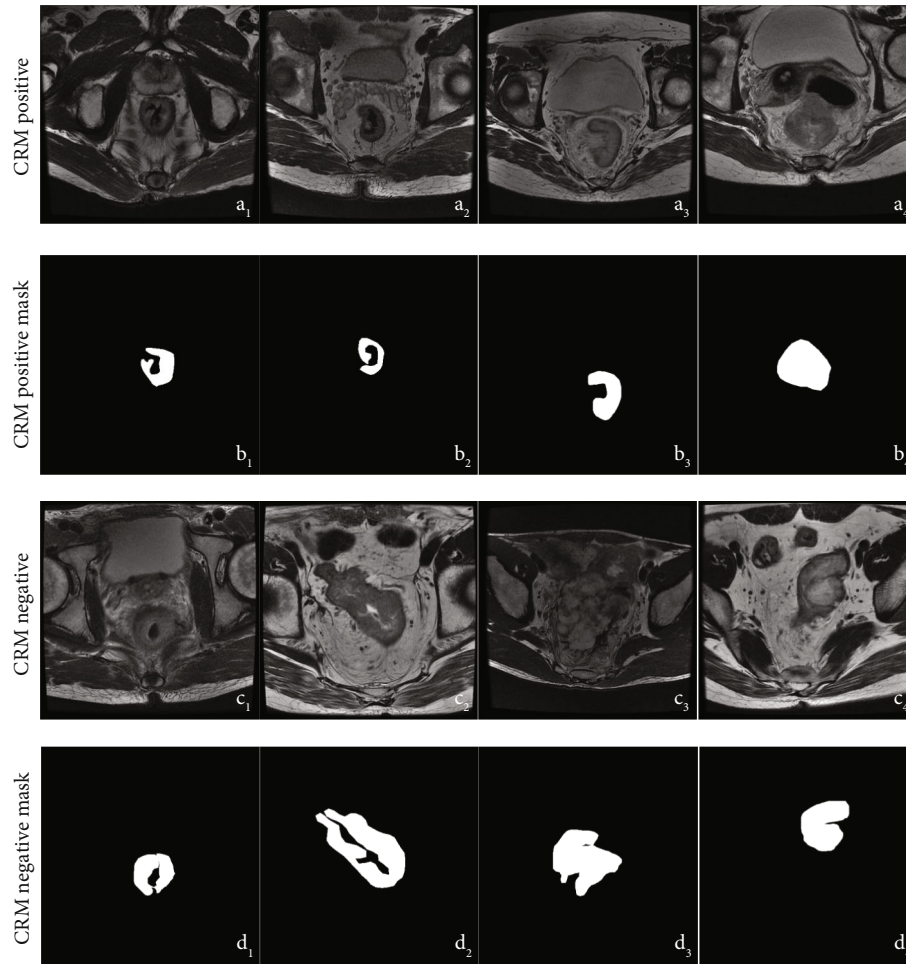
FIGURE 1: Examples of colorectal cancer in MRI images. (a1–a4, c1–c4) Are the original image slices from the MRI DICOM series; (b1–b4, d1–d4) are the tumour region mask labelled by doctors.
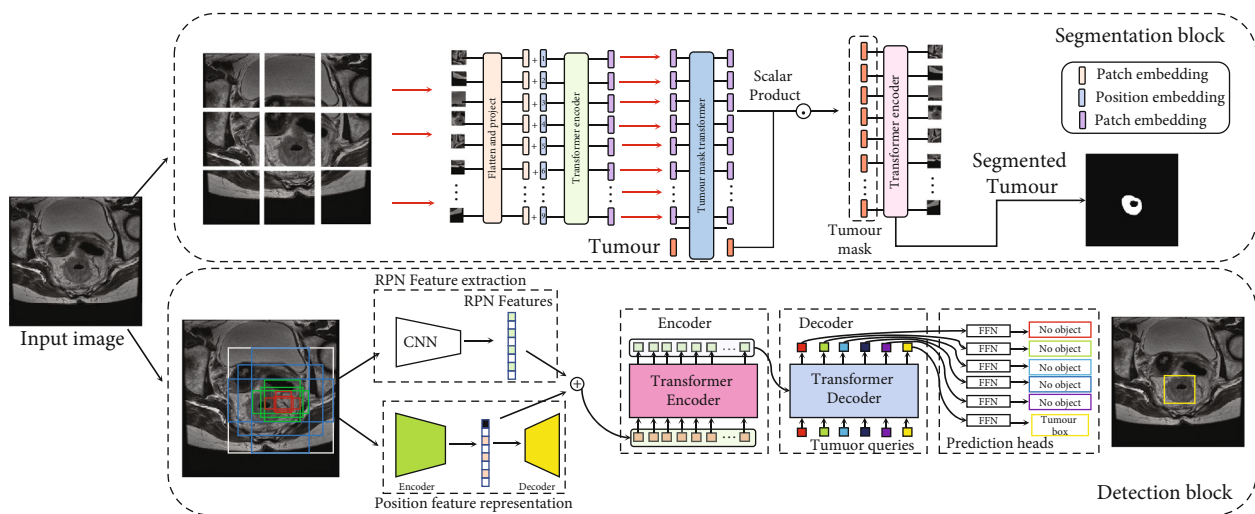


FIGURE 2: Schematic diagram of the proposed multitask learning framework for colorectal cancer region mining frame.
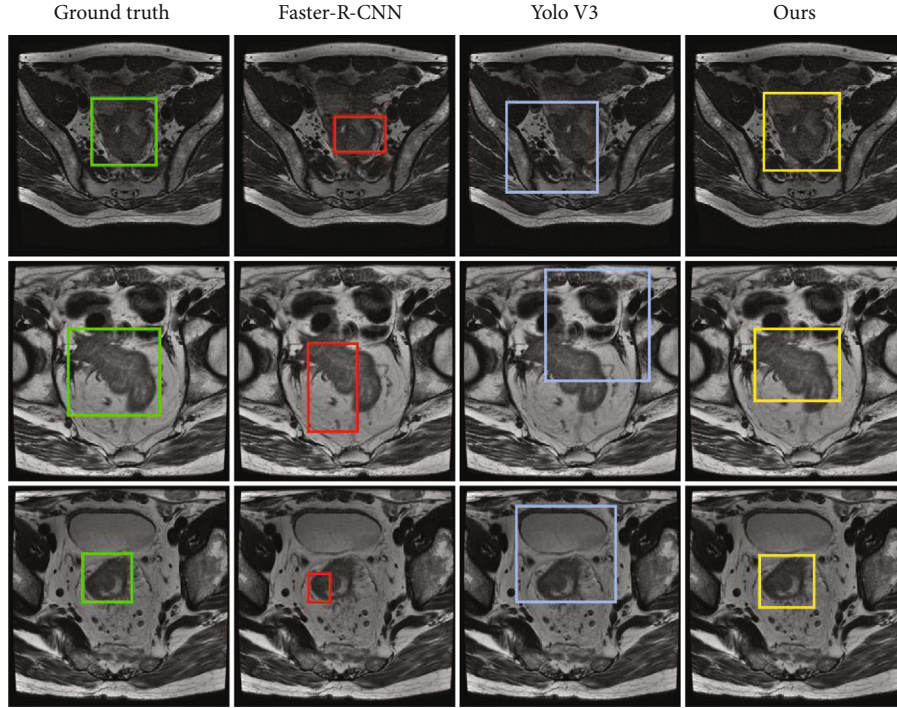
FIGURE 3: Tumour region detection results and its comparison results.

embeddings are processed jointly with patch encodings by the decoder as depicted in Figure 2. The total loss function is defined as follows:

$$L_{\text{tumour}} = -\sum_{i=1}^{n} t_i \log(p_i), \quad \text{for } k \text{ classes}, \tag{5}$$

where $t_i$ is the ground truth label and $p_i$ is the softmax probability for the $i^{\text{th}}$ class.

In this way, we combine these loss functions to form an end-to-end train protocol, for the total loss is defined as follows:

$$L_{\text{total}} = L_{\text{match}} + L_{\text{tumour}}. \tag{6}$$

Our proposed method employs a simple process to treat the patch and tumour region segmentation jointly during the decoding phase in the segmentator; in the whole framework, we address tumour region detection and segmentation jointly and combine them into a whole framework.

## 4. Results and Discussion

*4.1. Implementation Details.* In the tumour region detection pipeline, we train DETR with AdamW optimizer with the initial transformer's learning rate to $10^{-4}$, the backbone's to $10^{-5}$, and the weight decay to $10^{-4}$. We choose the transformer weights with Xavier in it and all backbone is ImageNet pretrained ResNet50 model for the basic architecture [36, 37].

In the tumour segmentation pipeline, the architecture is based on the vision transformer (ViT), and the head size of the multihead self-attention block is fixed to 64, other parameters are set as the default of the ViT model, and the input patches are with the same size [23, 24]. The segmentation model is pretrained on ImageNet; ViT is pretrained on ImageNet with random cropping. Following that, we fine-tune the pretrained models for the tumour region segmentation task and the pixel-wise cross-entropy loss without weight rebalancing. In the training phase, the SGD optimizer with a base learning rate 0.0001 and weight decay 0 is set in the initial training paradigm.

Here, we choose the standard evaluation method of tumour detection and segmentation. The Jaccard index is used for evaluating the ground truth bounding box and the predicted bounding box variances, and formally, the IoU measures the overlap between the ground truth box and the predicted box over their union. The total IoU is defined as follows:

$$\text{IoU}_{\text{pred}}^{\text{truth}} = \frac{\text{truth} \cap \text{pred}}{\text{truth} \cup \text{pred}}. \tag{7}$$

For comparisons with other methods, the results of our framework and other methods are reported in terms of recall, precision, and f1-measure values as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

Original image          Ground truth          U-net          FCN          Our method
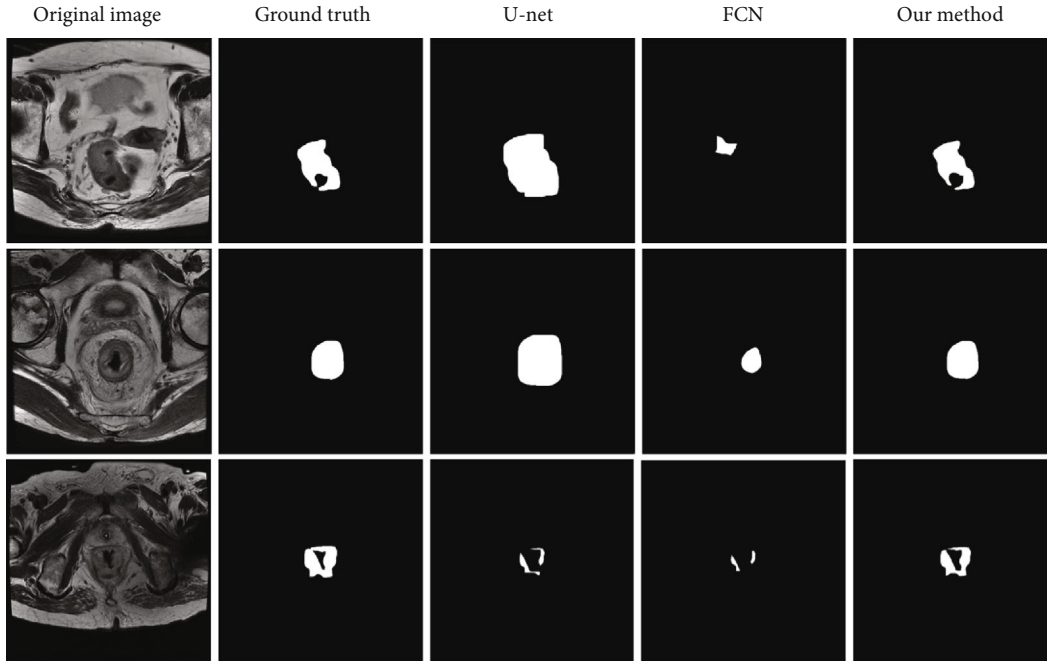
FIGURE 4: Basic rocket ship design. The rocket ship is propelled with three thrusters and features a single viewing window. The nose cone is detachable upon impact.

$$F1 = \frac{2PR}{P + R}. \tag{10}$$

Our framework is performed on the environment of Ubuntu 14.04 with Inter Core i9 platform, 32 GB RAM, and 2080Ti GPU×2 and is based on the PyTorch platform, CUDA v10.0, cuDNN v8.0.

*4.2. Experiment Results.* For the model training, we first divide the two pipelines separated, and training them individually, it takes about 72 hours and 103 hours for the detector and segmentator. For the restriction of our GPUs, we have not extended other comparisons with other transformer-based methods.

For tumour detection, we divide the dataset into training and validation sets. For the same baseline, these methods cover the one-stage and two-stage object detection pipelines. The most prevalent method Faster-RCNN and Yolo-V3 are chosen as the test bed for the final comparisons. Figure 3 shows the detection results of the three methods, our proposed pipeline can cover most of the tumour regions, and the bounding box can converge to the tumour boundary accurately. The Faster-RCNN model is a popular two-stage object detection method, which can catch the tumour region, but it is affected by the changes of the background; this is largely because the tumour regions have the same texture as other organs. The Yolo-v3 is a popular two-stage object detection method and has been employed for many object detection and location tasks. We have evaluated this method on this dataset, the results are shown in column 3 of Figure 3, it shows that Yolo-v3 has detected the tumour region, but the result exhibits that this method usually covers the tumour region and the neighbour organs together and

TABLE 1: Tumour region detection results covered by the study.

| Methods | CRM+ (%) | CRM- (%) | Average (%) | Total (%) |
|---|---|---|---|---|
| Faster-RCNN | 67.1 | 62.3 | 64.7 | 65.6 |
| Yolo-v3 | 43.4 | 37.6 | 40.5 | 41.2 |
| Ours | 87.5 | 89.1 | 88.3 | 88.6 |

TABLE 2: Tumour region segmentation accuracy covered by the study.

| Methods | CRM+ (%) | CRM- (%) | Average (%) | Total (%) |
|---|---|---|---|---|
| U-Net | 82.1 | 81.7 | 81.9 | 81.8 |
| FCN | 67.2 | 66.4 | 66.8 | 66.5 |
| Ours | 91.2 | 90.6 | 90.9 | 91.1 |

this is largely affected by the distinct boundary of these regions.

For the tumour segmentation, we have the same separation of dataset like the detection pipeline. We set the U-Net and FCN as the baseline for the comparison, the U-Net is a specific method for medical image analysis, and FCN have been greatly used in the nature image segmentation; for this reason, we have listed the comparison result with these two methods and to exhibit our framework's advantages. In this part, the U-Net model has better results than the FCN model, and it can catch the tumour region in high contract images but less in low contract slices. The FCN model usually needs an intensive training protocol on large image data, but in this program, the dataset is less than those, so it cannot get better results. Compared with these two methods, our proposed method has achieved an excellent result on the image segmentation tasks. And the results are shown in Figure 4.

For a better comparison, we also compared the accuracy of these two methods, and our proposed framework also achieves a better result on the dataset (see Tables 1 and 2).

## 5. Conclusions

In this paper, we propose a novel transfer learning framework, CST. We combine the colorectal cancer region detection and segmentation task jointly and fine-tuned a transformer-based model to perform these tasks. For higher accuracy, we incorporate image-level information into the final cancer region detection, the results demonstrate that the proposed framework can handle these tasks well, and the comparison results have shown that our method has achieved better accuracy than the traditional methods such as CNN. In this way, the proposed framework explores a new protocol for colorectal cancer information mining. In future works, we mainly focus on how to use few samples to achieve a better result.

## Data Availability

The DICOM data used to support the findings of this study were supplied by Zhaofeng Tian and Jing Chen under license and so cannot be made freely available. Requests for access to these data should be made to Dong Sui.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

Dong Sui and Jing Chen have equal contributions to this manuscript.

## Acknowledgments

## References

[1] H. Sung, J. Ferlay, R. L. Siegel et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] J. Jin, "Screening for colorectal cancer," *JAMA*, vol. 325, no. 19, p. 2026, 2021.

[3] J. S. Lin, L. A. Perdue, N. B. Henrikson, S. I. Bean, and P. R. Blasi, "Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force," *JAMA*, vol. 325, no. 19, pp. 1978–1998, 2021.

[4] S. J. Mehta, A. M. Morris, and S. S. Kupfer, "Colorectal cancer screening starting at age 45 years-ensuring benefits are realized by all," *JAMA Netw Open*, vol. 4, no. 5, p. e2112593, 2021.

[5] K. Ng, F. P. May, and D. Schrag, "US Preventive Services Task Force recommendations for colorectal cancer screening: forty-five is the new fifty," *JAMA*, vol. 325, no. 19, pp. 1943–1945, 2021.

[6] A. B. Knudsen, C. M. Rutter, E. F. P. Peterse et al., "Colorectal cancer screening: an updated modeling study for the US Preventive Services Task Force," *JAMA*, vol. 325, no. 19, pp. 1998–2011, 2021.

[7] S. Jensch, S. Bipat, J. Peringa et al., "CT colonography with limited bowel preparation: prospective assessment of patient experience and preference in comparison to optical colonoscopy with cathartic bowel preparation," *European Radiology*, vol. 20, no. 1, pp. 146–156, 2010.

[8] S. Wilson and J. D. Thompson, "Comparison of two Meglumine-Diatrizoate based bowel preparations for computed tomography colonography: comparison of patient symptoms and bowel preparation quality," *Radiography (Lond)*, vol. 26, no. 4, pp. e290–e296, 2020.

[9] S. B. Desai, A. Pareek, and M. P. Lungren, "Deep learning and its role in COVID-19 medical imaging," *Intelligence-Based Medicine*, vol. 3-4, p. 100013, 2020.

[10] H. M. Ahmad, M. J. Khan, A. Yousaf, S. Ghuffar, and K. Khurshid, "Deep learning: a breakthrough in medical imaging," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 16, no. 8, pp. 946–956, 2020.

[11] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against COVID-19: a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection," *Sustainable Cities and Society*, vol. 65, p. 102600, 2021.

[12] Y. D. Zhang, S. C. Satapathy, X. Zhang, and S. H. Wang, "COVID-19 diagnosis via DenseNet and optimization of transfer learning setting," *Cognitive Computation*, pp. 1–17, 2021.

[13] X. Li, Q. Dou, H. Chen et al., "3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Medical Image Analysis*, vol. 45, pp. 41–54, 2018.

[14] Y. Ma, P. Feng, P. He et al., "Segmenting lung lesions of COVID-19 from CT images via pyramid pooling improved Unet," *Biomedical Physics & Engineering Express*, vol. 7, no. 4, 2021.

[15] M. Lei, J. Li, M. Li, L. Zou, and H. Yu, "An improved UNet++ model for congestive heart failure diagnosis using short-term RR intervals," *Diagnostics (Basel)*, vol. 11, no. 3, 2021.

[16] T. Hassanzadeh, D. Essam, and R. Sarker, "2D to 3D evolutionary deep convolutional neural networks for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 712–721, 2021.

[17] V. Gupta, M. Demirer, M. Bigelow et al., "Performance of a deep neural network algorithm based on a small medical image dataset: incremental impact of 3D-to-2D reformation combined with novel data augmentation, photometric conversion, or transfer learning," *Journal of Digital Imaging*, vol. 33, no. 2, pp. 431–438, 2020.

[18] O. Petit, N. Thome, and L. Soler, "Iterative confidence relabeling with deep ConvNets for organ segmentation with partial labels," *Computerized Medical Imaging and Graphics*, vol. 91, p. 101938, 2021.

[19] D. Song, B. Fu, F. Li et al., "Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2392–2402, 2021.

[20] Z. Li, Z. Zhang, H. Zhao et al., "Text compression-aided transformer encoding," *IEEE Trans Pattern Anal Mach Intell*, p. 1, 2021.

[21] J. Wang, R. Chakraborty, and S. X. Yu, "Spatial transformer for 3D point clouds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2021.

[22] Z. Zhang, Y. Wu, J. Zhou, S. Duan, H. Zhao, and R. Wang, "SG-Net: syntax guided transformer for language representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020.

[23] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16519–16529, 2021.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *In International Conference on Machine Learning*, pp. 10347–10357, 2021.

[25] Y. Gao, C. Liu, and L. Zhao, "Multi-resolution path cnn with deep supervision for intervertebral disc localization and segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2019.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, Cham, October 2015.

[27] L. Yang, Y. Zhang, Z. Zhao et al., "Boxnet: deep learning based biomedical image segmentation using boxes only annotation," *ArXiv abs/1806.00593*, 2018.

[28] Y. Gao, R. Huang, Y. Yang et al., "FocusNetv2: imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images," *Medical Image Analysis*, vol. 67, p. 101831, 2021.

[29] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[31] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *In Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[32] H. C. Shin, H. R. Roth, M. Gao et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[33] P. Liu, M. R. Lyu, I. King, and J. Xu, "Learning by distillation: a self-supervised learning framework for optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[34] H. Xie, Y. Lei, T. Wang et al., "High through-plane resolution CT imaging with self-supervised deep learning," *Physics in Medicine & Biology*, vol. 66, no. 14, p. 145013, 2021.

[35] Q. Lu, Y. Li, and C. Ye, "Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks," *Medical Image Analysis*, vol. 72, p. 102094, 2021.

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings*, 2010.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Proceedings of the International Conference on Learning Representations (ICLR) 2019*, 2019.