MSF-TransUNet: transformer multi-scale fusion on U-Net for gastric cancer pathological image segmentation

Bing Bai*[a,b], Xiaoqi Zhang[a]
[a]Department of Information Engineering, Xuan Cheng Vocational & Technical College, Xuancheng 242099, Anhui, China; [b]School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China

ABSTRACT

To address the challenges in gastric cancer pathological images, such as varying sizes and shapes of lesion regions as well as blurry boundaries, we propose an enhanced U-Net architecture segmentation algorithm based on an affine cross-attention mechanism. Specifically, we introduce affine transformation modules into the up-sampling and down-sampling stages of the U-Net, replacing adjacent convolutional blocks to better capture variations in shape and size. Additionally, a cross-attention module is incorporated in the bridging phase to enhance feature utilization and mitigate mis-segmentation of healthy tissues. In contrast to the conventional U-Net, our algorithm demonstrates notable enhancements in terms of 8.21%, 6.87%, and 5.57% in Dice coefficient, Intersection over Union (IoU), and Accuracy (ACC), respectively. The effectiveness of our introduced modules is reinforced through ablation investigations. The segmentation performance of lesion regions in gastric cancer pathological images is augmented by the proposed algorithm, as shown by the experimental results, effectively reducing the false-positive rate in image diagnosis.

Keywords: Affine transformation, cross-attention, transformer, local features, global features, gastric cancer pathological images

1. INTRODUCTION

One of the most widespread malignancies in the digestive system, gastric cancer occupies a critical spot on the global cancer landscape. Consequently, the effective diagnosis and timely treatment of gastric cancer are paramount. The primary diagnosis of gastric cancer frequently hinges upon the examination of pathological images. Nevertheless, the process of diagnosing through pathological section screening is not only a time-intensive endeavor but also poses challenges for early detection, primarily due to the inconsistent quality of the pathological images, which can hinder accurate assessment. Therefore, the utilization of computer-aided diagnostic methods for classification and localization can aid in gastric cancer screening, saving time and improving efficiency.

Medical image segmentation algorithms serve as a crucial tool in computer-aided diagnosis. Expediting diagnosis through image segmentation algorithms enhances the likelihood of tumor detection, enabling clinicians to efficiently utilize their discovery time and improve patient prognosis[1]. At present, the majority of medical image segmentation algorithms are designed with a mirrored encoder-decoder framework that operates from a top-down approach. Among them, the UNet architecture and its variants are renowned, with their key lies in their fully convolutional nature[2].

Although the U-Net model, leveraging Convolutional Neural Networks (CNNs), has yielded promising outcomes in medical image segmentation, there persists an opportunity to enhance its performance to better aid clinicians in early disease detection. Transformer models, renowned for their proficiency in managing lengthy sequence dependencies in language tasks[3], have made significant strides. As a result, Transformer-based architectures, including Vision Transformer (ViT[4]), have outperformed CNNs in benchmark image processing evaluations. Recent advancements in ViT, exemplified by CvT[5], CCT[6], and Swin Transformer[7], have demonstrated that Transformers can excel even with modest parameter counts and limited data inputs, challenging the notion that vast amounts of data are necessary. Presently, ViT models incorporate spatial positioning into image patches through positional encoding, which is then processed by standard Transformer layers to capture long-range semantic relationships within the data.

*1230313030@stu.xaut.edu.cn; phone 1 836 518 6578; fax 0 563 3023305

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133951S · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048552

Utilizing the strengths of Convolutional Neural Networks (CNNs) and Transformer models within the domain of image segmentation, deep learning frameworks featuring fully convolutional encoder-decoders are proficiently employed to exploit long-range semantic connections within medical images for segmentation purposes. In pursuit of this objective, the inaugural fully convolutional Transformer designed exclusively for medical image segmentation is introduced. At the core of this innovative model lies the fully convolutional Transformer layer, functioning as the principal component, which comprises two essential elements: a convolutional attention module and a fully convolutional broad-focus module. Our key contributions are outlined as below:

(1) Across diverse datasets, the efficacy of skip connections is scrutinized, uncovering the inadequacy of straightforward independent replication.

(2) A fresh perspective is presented to elevate semantic segmentation performance, addressing the semantic and resolution disparities between low-level and high-level features through refined feature fusion methodologies and multi-scale channel cross-attention frameworks. This methodology adeptly captures intricate channel interdependencies, ultimately enhancing segmentation precision.

2. RELATED WORK

The application of the Visual Transformer (ViT)[8] has generated encouraging outcomes in ImageNet classification through the direct utilization of Transformers with global self-attention on complete images. Following the triumphant expansion of Transformers across diverse computer vision realms, TransUNet[9] surfaced as the pioneering Transformer-based medical image segmentation platform. In response to the constraint of limited data samples in medical imaging, Valanarasu et al. introduced a gated axial-attention model[10]. Drawing inspiration from the groundbreaking performance of the Swin Transformer[11], Swin-U-Net[12] established the precedent for an entirely Transformer-based U-shaped architecture, leveraging the Swin Transformer to substitute convolutional blocks in U-Net. Nevertheless, these methodologies primarily focus on rectifying the deficiencies of convolutional operations, neglecting the intrinsic restrictions of the U-Net architecture itself, which could potentially lead to structural inefficiencies and heightened computational expenses.

The mechanism of skip connections, initially conceived in UNet[2], endeavors to span the semantic divide between the encoder and decoder, demonstrating its proficiency in restoring intricate details of targeted objects[13]. As U-Net's influence grows, a plethora of innovative models have surfaced, including UNet++[14], Attention U-Net[15], DenseUNet[16], r2e-net[17], and UNet3+[18], each tailored to excel in medical image segmentation and exhibiting exceptional performance. UNet++ emphasizes the semantic disparities among feature maps of identical scale from encoder and decoder, presenting a nested configuration known as UNet++, which captures multi-scale attributes to further narrow the gap. Attention U-Net introduces a cross-attention module, harnessing coarse-grained features as gating cues to mitigate uncertainties stemming from distracting and noisy responses within skip connections. MultiResUNet[19] acknowledges the potential semantic discrepancy between corresponding encoder and decoder features bypassed, integrating residual architectures to fortify skip connections.

Semantic feature extraction in histopathological images, attempted through traditional image processing techniques, proves inadequate due to their inability to manually discern such features. Conversely, deep convolutional networks, unencumbered by handcrafted features, adeptly extract the most pertinent and descriptive feature information during model training, enabling more judicious decisions. Block-based classification networks, though compromising on boundary smoothness, achieve WSI image segmentation with minimized computational overhead. Among these, U-Net is a preeminent choice for block-based pathological image segmentation, leveraging its prowess in capturing global features during contraction and precise localization during expansion. Nevertheless, U-Net's oversight of local pixel dependencies necessitates enhancements. In this regard, a fusion framework has been devised to bolster tumor edge segmentation accuracy14. Conditional random fields offer a means to model long-range dependencies, serving as a post-processing tool for semantic segmentation predictions. However, this method is computationally demanding and necessitates a substantial amount of expert annotations for effective training.

3. METHOD

3.1 Overall architecture

In the realm of medical image segmentation, acquiring multi-scale features holds paramount importance for tackling intricate scale variations. Given these challenges, there arises a necessity to bridge the semantic comprehension gap

between the encoder and decoder, achieved by fusing multi-scale channel information that encapsulates non-local semantic dependencies. This research introduces a Transformer-based multi-scale cross-fusion U-Net architecture, designated as MSF-TansUnet, tailored for gastric cancer pathological image segmentation. Its objective lies in ameliorating the feature connectivity issue between the encoder and decoder. Channels often prioritize distinct semantic features, and the adaptive fusion of an ample number of channel features facilitates more intricate medical image segmentation. We commence by presenting the Multi-Scale Cross-Fusion Transformer (MST), which integrates multi-scale contexts and cross-attention from a channel-centric perspective. This architecture endeavors to discern local cross-channel interactions and accomplish the effective fusion of multi-scale channel features, possibly harboring semantic disparities, via multi-scale learning as opposed to isolated connections.

Furthermore, we introduce a Multi-head Cross-Attention (MCA) module, which incorporates the converged multi-scale features with those from the decoder stage, aimed at resolving semantic inconsistencies. This MCA module, by examining multi-scale global contexts, fosters relationships between the encoder and decoder, thereby augmenting the original skip connections to bridge the semantic gap and elevate segmentation proficiency. Both of these presented modules can seamlessly integrate into U-shaped networks tailored for medical image segmentation endeavors. The comprehensive framework of this investigation is depicted in Figure 1.

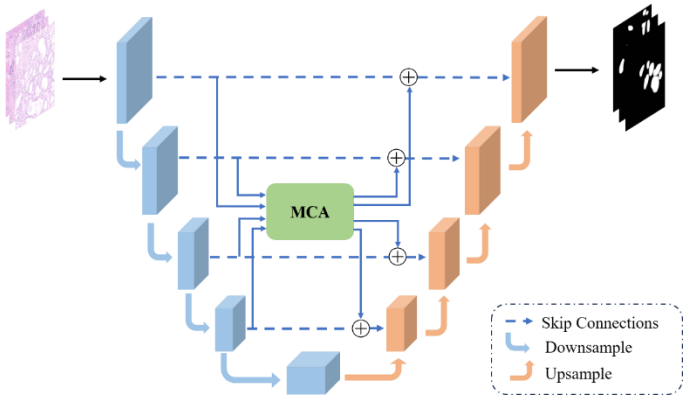


Figure 1. Overall structure of MSF-TransUNet.

3.2 Transformer multi-scale fusion on U-Net (MSF-TransUNet)

Figure 1 illustrates a U-Net-based Multi-head Cross-Attention (MCA) segmentation model that excels in the task of segmenting lesion regions in pathological images. Within this framework, the skip connections effectively preserve the original lesion areas, while the MCA module facilitates comprehensive segmentation of lesion regions by fusing multi-scale information while retaining contextual semantic information.

3.3 Scale affine layer (SAL)

The encoder and decoder are responsible for capturing fundamental image features, comprising four downsampling or upsampling blocks. The previous sampling blocks consisted of a single convolutional and pooling layer, aimed at altering the size of data features. However, this approach was insufficient for extracting sufficient semantic and contextual information. Inspired by spatial feature transformation (SFT)[20], this study introduces an innovation in convolutional neural networks. We have decided to incorporate a Scale Affine Layer (SAL) between two adjacent convolutional layers, as depicted in Figure 2. This SAL primarily consists of two 1×1 convolutional layers, where we specifically introduce a GELU activation function to enhance the model's nonlinear processing capabilities. The design of this layer involves inserting a scale affine layer with a GELU activation function between convolutional layers to optimize network performance.

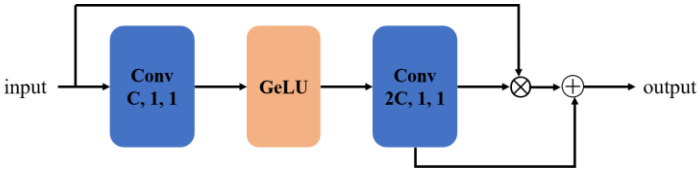


Figure 2. Structure diagram of the scale affine layer (SAL) module.

3.4 Multi-head cross-attention block (MCA)

To mitigate the problem of contextual semantic loss encountered in skip connections, this research endeavors to devise a Transformer-based multi-head cross-attention module, with the objective of overcoming the deficiency of contextual semantic information in pathological image segmentation endeavors. Figure 3 showcases the architecture of this innovative module.

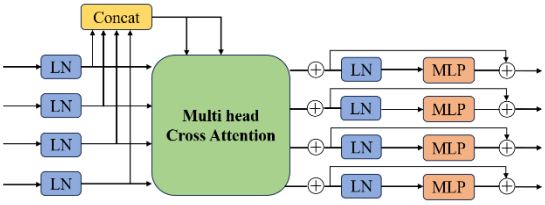


Figure 3. MCA block.

The input sequence for the MCA block comprises downsampled features F_1, F_2, F_3, and F_4. Initially, these tokens are subjected to a multi-head cross-attention module, which captures the interplay of information across various channels. Following this, the tokens undergo refinement through a multi-layer perceptron (MLP) equipped with a residual architecture, further embedding the associations and interdependencies among channels. After encoding by the MLP, we utilize the outputs of these MLPs to extract multi-scale features from each U-Net encoder level, refining the representation of the features. Ultimately, we leverage these multi-scale features to optimize and enhance the expression of features, thereby improving the model's performance. The entire process is designed to fully utilize the channel information of tokens and enhance the model's feature representation capabilities through the fusion of multi-scale features. The model takes five inputs, including F_i, and Formula (1) expresses the relationship between F_i and K, V:

$$Q_i = F_i W_{Q_i}, K = F_\Sigma W_K, V = F_\Sigma W_V \tag{1}$$

where W represents weights, $Q_i \in \mathbb{R}^{C_i \times d}, K \in \mathbb{R}^{C_i \times d}, V \in \mathbb{R}^{C_\Sigma \times d}$. Through the cross-attention (CA) mechanism, a similarity matrix M_i is generated and used to weight the values V, as shown in Equation (2):

$$CA_i = M_i V^F = \sigma\left[\psi\left(\frac{Q_i^F K}{\sqrt{C_\Sigma}}\right)\right] V^F = \sigma\left[\psi\left(\frac{W_{Q_i}^F F_i^F F_\Sigma W_K}{\sqrt{C_\Sigma}}\right)\right] W_V^F F_\Sigma^F \tag{2}$$

where $\psi(\cdot)$ and $\sigma(\cdot)$ represent instance normalization and softmax functions, respectively.

The model employs instance normalization on the similarity map to allow gradients to propagate smoothly. In the context of N-head attention, the computation of the output subsequent to the multi-head cross-attention process is performed as demonstrated in Equation (3), with the results being determined in a passive manner.

$$MCA_i = (CA_i^1 + CA_i^2 +, \cdots, + CA_i^N)/N \tag{3}$$

The number of input heads is represented by N, and the ultimate output of the MCA module is formulated as shown in Equation (4), with the expression being presented in a passive voice structure.

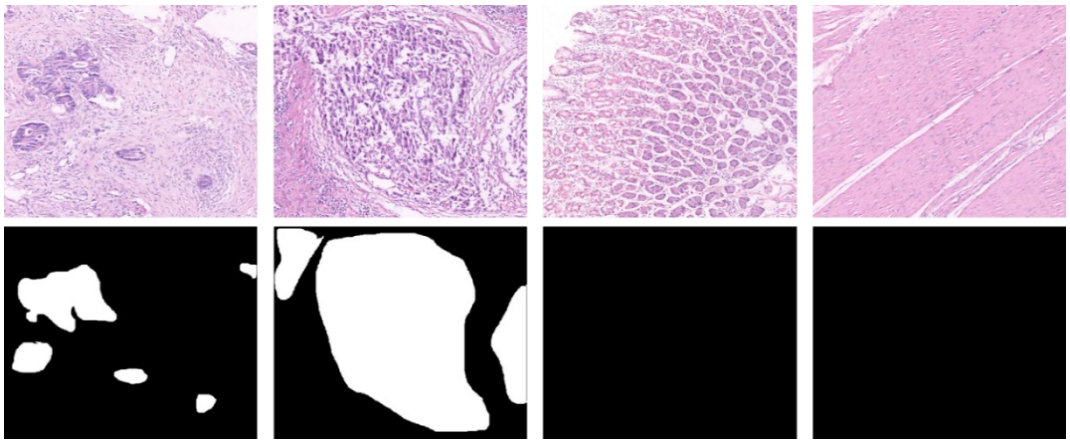
$$O_i = MCA_i + MLP(Q_i + MCA_i) \tag{4}$$

Outputs O_1, O_2, O_3, O_4 from the downsampling layers are reconstructed through upsampling operations and convolutional layers, and concatenated with the decoder features accordingly.

4. EXPERIMENTAL SETUP

4.1 Dataset

Gastric cancer cells typically feature enlarged nuclei, irregular shapes, and reduced cytoplasm, raising nucleus-to-cytoplasm ratios. These traits are key to identifying diseased regions in pathological images. Our study evaluates GasHis-Transformer1's segmentation performance using the H&E-stained HRCF dataset. H&E stains nuclei purple-blue and cytoplasm red, distinguishing features. The dataset contains 560 cancerous and 140 normal images in 2048×2048 resolution PNG format. Figure 4 contrasts normal (ordered nuclei, low nucleus-to-cytoplasm ratio) and abnormal (large, irregular nuclei) gastric histopathological images.



(A) Diseased images (B) Non-diseased images

Figure 4. Pathological images of gastric cancer.

4.2 Training, validation, and test data setup

The NVIDIA RTX 4090 GPU, equipped with 24 GB of memory, is utilized in this experiment, and the algorithmic models are implemented utilizing PyTorch. Specifically, the training process involves 280 images, while 140 images are designated for testing purposes. To mitigate the risk of overfitting, the training data is enriched through the application of various data augmentation strategies, including horizontal and vertical flipping, as well as random rotation. Notably, the proposed model is trained from scratch, without relying on any pre-trained weights. Furthermore, to expedite the convergence process, the Adam optimizer is adopted, with an initial learning rate initialized at 0.001. Our network's training employs a hybrid loss function, which combines cross-entropy loss and dice loss.

Given the substantial size of the original images, which poses the risk of a significant increase in model parameters, we adopt a sliding window technique with dimensions of 512×512 to segment both the original images and their corresponding mask images into locally overlapping segments, which are subsequently normalized. To bolster the model's generalization capabilities and mitigate overfitting, the training images undergo augmentation processes such as 90° and 270° rotations, horizontal flipping, and vertical flipping, effectively multiplying the training set by a factor of five, resulting in a total of 56,000 512×512 training images. As outlined in Table 1, 44,800 of these images, inclusive of 8,960 images allocated for validation, are utilized for training purposes, while the remaining 11,200 images are designated for testing.

Table 1. The establishment of training, validation, and test sets.

Dataset	Training	Validation	Test	Sum
Insulator images	35840	8960	11200	56000

4.3 Evaluation metrics

The segmentation performance of the proposed model is assessed through a comparative analysis of the experimental outcomes, incorporating both subjective and objective evaluations. Subjectively, the emphasis is placed on visually assessing the overall segmentation quality of the images, along with the delineation of intricate edges. Objectively, the evaluation is carried out utilizing Accuracy, Intersection over Union (IoU), and Dice Score as the quantitative metrics. The formulas utilized for calculating these metrics are outlined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$IoU = \frac{TP}{TP+FP+FN} = \frac{|X \cap Y|}{|X \cup Y|} \tag{6}$$

$$Dice = \frac{2TP}{2TP+FP+FN} = \frac{2|X \cap Y|}{|X|+|Y|} \tag{7}$$

Wherein, *TP*, *TN*, *FP*, and *FN* signify the counts of accurately categorized cell pixels, accurately categorized background pixels, erroneously categorized background pixels, and erroneously categorized cell pixels, respectively. *X* and *Y* denote

the ground truth and predicted values, respectively. *Acc*, being the fundamental metric, represents the fraction of accurately segmented pixels within the total pixel count. *IoU*, a benchmark metric for semantic segmentation, quantifies the proportion of intersection to the union of the ground truth and predicted values, with a value of 1 indicating perfect concordance. *Dice*, a crucial metric in medical image segmentation, assesses the degree of similarity between the ground truth and predicted values, with a higher value indicating a stronger resemblance.

5. RESULTS AND DISCUSSION

5.1 Comparison of segmentation effects

Figure 5 showcases examples of pathological images segmented by various models, where the last image in each row represents the Ground Truth (GT) of the segmented lesion area in the pathological image, and the five intermediate images are the segmentation results of different networks. As seen in the figure, the lesion areas of different gastric cancer pathological slices exhibit characteristics such as uneven contrast, blurred edges, and varying sizes and shapes. In the second and third images, FCN and U-Net utilize different-sized convolution kernels to extract features, resulting in a single receptive field size at each network layer. This inability to adequately capture the features contained in small lesion areas leads to missed segmentations. However, both R2U-Net and TransU-Net exhibit issues of over-segmentation, segmenting non-lesion areas of the model, and the accuracy of the segmentation results at the edges does not fully meet the needs of physicians for analysis. The reason lies in the multi-level changes in contrast near the lesion in the third image, which can easily lead the attention mechanism to partially learn the stronger contrast areas while ignoring the weaker contrast parts as background. Additionally, the similar morphological appearance of some lesion areas with adjacent healthy tissues can easily cause the network to misclassify the lesion areas as false positives or false negatives. This paper utilizes a cross-attention guidance mechanism to mine spatial information from shallow features to alleviate the issue of inaccurate pixel localization in regions of interest during upsampling. The specificity outcomes reveal that the proposed method exhibits a superior capability in addressing the challenge of false-positive misclassification in comparison to alternative segmentation algorithms.

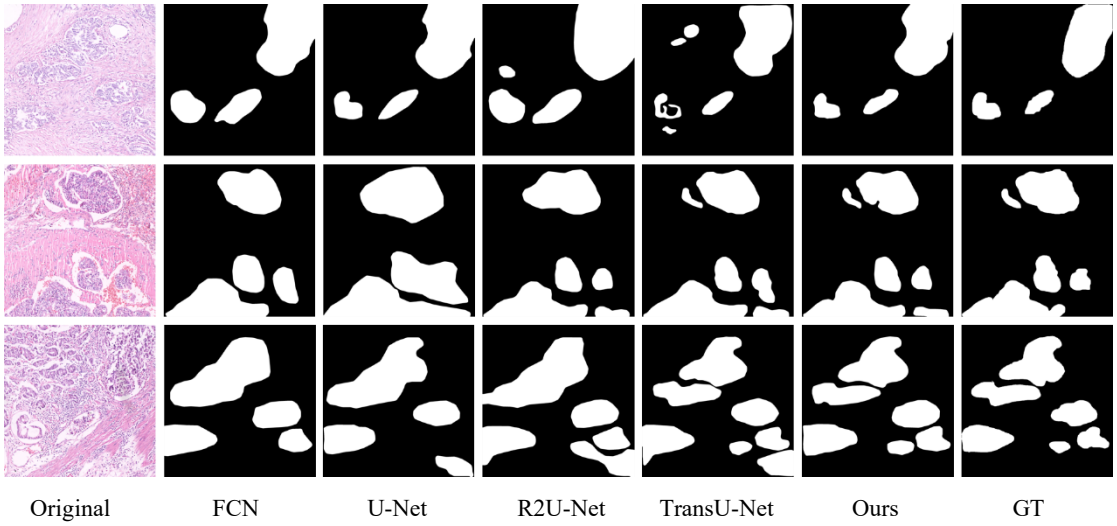


Figure 5. Segmentation results of various models.

5.2 Comparison of segmentation results

Visual inspection, though direct, is prone to subjective biases, highlighting the need for quantitative evaluation of segmentation models. Table 2 shows the proposed model outperforms others in Accuracy, IoU, and Dice Score, indicating its effectiveness and robustness in segmenting details. The FCN model struggles with detailed information recovery during upsampling, impacting its performance. In contrast, the U-Net and R2U-Net models enhance segmentation by compensating for lost details and leveraging recursive neural networks, respectively. The TransU-Net, with its transformer encoder, facilitates advanced feature extraction but may lose global features with deeper encoding-decoding. Compared to TransU-Net, the proposed model improves across metrics, excelling at segmenting low-contrast cells and faint edges while adapting better to brightness, noise, and other factors, resulting in superior segmentation.

Table 2. An analysis is conducted quantitatively to compare various segmentation models.

Model	Acc	IoU	Dice
FCN	0.8921	0.6423	0.8064
U-Net	0.9014	0.7985	0.8462
R2U-Net	0.9119	0.8034	0.8739
TransU-Net	0.9462	0.8441	0.9136
Ours	0.9571	0.8672	0.9283

5.3 Ablation study

To assess the efficacy of MSF-TransUNet's modules, ablation experiments were individually conducted on each, with outcomes detailed in Table 3. This study's baseline is the U-Net model. Firstly, substitution of the encoder with the SAL module resulted in a 0.50% improvement in IoU over U-Net. Following this, integrating the MCA module between the encoder and decoder further bolstered IoU by 0.87% above U-Net. Lastly, implementation of TTA post-processing led to a 1.62% enhancement in IoU compared to U-Net. These findings affirm the proficiency of the proposed modules in augmenting segmentation accuracy.

Table 3. Ablation experimental results of MSF-TransUNet.

Method	ACC	IoU	Dice
U-Net	0.9014	0.7585	0.8462
U-Net+SAL	0.9202	0.8296	0.9013
U-Net+MCA	0.9286	0.8463	0.9085
Ours	0.9571	0.8672	0.9283

6. CONCLUSION

This paper tackles the intricacies of segmenting lesion regions in gastric cancer pathological images, characterized by varied shapes, irregular sizes, and inconsistent contrasts, by introducing a cross-attention-based multi-scale U-Net segmentation model. Central to this model is the SAL (Scale-Attentive Layer) unit, which facilitates the extraction of lesion-specific features and endows each layer with the capability to access varied receptive field sizes, thereby enhancing the capture of lesion shape and size details. Furthermore, to mitigate the problem of contextual semantic loss in skip connections, an MCA (Multi-scale Context Attention) module is incorporated, effectively reducing false positives. The bridging stage employs a non-direct approach, further optimizing feature utilization. Comparative assessments against prevalent segmentation networks underscore the superiority of our approach, achieving noteworthy Dice and IoU coefficients of 92.83% and 86.72%, respectively, in gastric cancer pathological image segmentation. This advancement paves the way for more precise region delineation, facilitating enhanced analysis of gastric cancer pathological images.

ACKNOWLEDGMENTS

This research is supported by the Anhui Provincial Department of Education, with the project type of "Characteristic High-level Major", project name of "Digital Media Technology Characteristic High-level Major", project number of 2021tszy071; and the 2023 Xuancheng Vocational & Technical College's College-level Scientific Research Revitalization Plan Project: Research on Image Recognition Algorithm Based on Image Edge Detection.

REFERENCES

[1] Yan, J., Chaitanya, K., Tom, L., Roderick, M. S. and Ibrahim, H., "Prediction of weaning from mechanical ventilation using convolutional neural networks," Artificial Intelligence in Medicine 117, 102087 (2021).

[2] Olaf, R., Philipp, F. and Thomas, B., "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical Image Computing and Computer-Assisted Intervention, 234-241 (2015).

[3] Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., Lukasz, K. and Illia, P., "Attention is all you need," Advances in Neural Information Processing Systems, 5998-6008 (2017).

[4] Alexey, D., Lucas, B., Alexander, K., Dirk, W., Zhai, X., Thomas, U., Mostafa, D., Matthias, M., Georg, H., Sylvain, G., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv arXiv:2010.11929, (2020).

[5] Wu, H., Xiao, B., Noel, C., Liu, M., Dai, X., Yuan, L. and Zhang, L., "CvT: Introducing convolutions to vision transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision, 22-31 (2021).

[6] Ali, H., Steven, W., Nikhil, S., Abulikemu, A., Li, J. and Humphrey, S., "Escaping the big data paradigm with compact transformers," arXiv arXiv:2104.05704, (2021).

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022 (2021).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., "An image is worth 16x16 Words: Transformers for image recognition at scale," arXiv arXiv:2010.11929, (2020).

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

[10] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. and Patel, V. M., "Medical transformer: Gated axial-attention for medical image segmentation," arXiv arXiv:2102.10662, (2021).

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv arXiv:2103.14030, (2021).

[12] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M., "Swin-Unet: Unet-like pure transformer for medical image segmentation," arXiv arXiv:2105.05537, (2021).

[13] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q., "Densely connected convolutional networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261-2269 (2017).

[14] Zhou, Z., Siddiquee, M. R., Tajbakhsh, N. and Liang, J., "UNet++: A nested U-Net architecture for medical image segmentation," [Lecture Notes in Computer Science], Springer, Germany, (2018).

[15] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B. and Rueckert, D., "Attention U-Net: Learning where to look for the pancreas," arXiv arXiv:1804.03999, (2018).

[16] Li, X., Chen, H., Qi, X., Dou, Q., Fu, C. W. and Heng, P. A., "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," IEEE Transactions on Medical Imaging 37(12), 2663-2674, (2018).

[17] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M. and Asari, V. K., "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," arXiv arXiv:1802.06955, (2018).

[18] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W. and Wu, J., "UNet 3+: A full-scale connected UNet for medical image segmentation," arXiv arXiv:2004.08790, (2020).

[19] Rahman, M. S., "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," Neural Networks 121, 74-87 (2020).

[20] Wang, X., et al., "Recovering Realistic texture in image super-resolution by deep spatial feature transform," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2018).