SOSEGFORMER: A CROSS-SCALE FEATURE CORRELATED NETWORK FOR SMALL MEDICAL OBJECT SEGMENTATION

Wei Dai[1,2], Zixuan Wu[1,2], Rui Liu[1,2], Junxian Zhou[1,2], Min Wang[1,2], Tianyi Wu[1,2], Jun Liu[1,2]

[1] Centre for Robotics and Automation, City University of Hong Kong, Hong Kong, China
[2] Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China

ABSTRACT

A mild syndrome with a small infected region is an ominous warning and is foremost in the early diagnosis of diseases. Recently, deep learning algorithms, such as convolutional neural networks (CNN), have been successfully applied to segment natural or medical objects, yielding promising results. However, the analysis of medical objects with small area occupation in images remains largely underexplored. This task poses a significant challenge due to information loss caused by convolution and pooling operations in CNN, particularly for small medical objects. To tackle these challenges, we propose a novel small-object segmentation with transformer (SoSegFormer) network for accurate small-object segmentation in medical images. Quantitative experimental results demonstrate the top-level performance of SoSegFormer, achieving the best mIoU, mDice, MAE, and F2 Score. Notably, it achieved 87.02%, 80.91%, and 65.17% in mDice for segmenting liver tumour, polyp, and sperm objects, which occupy less than 1% of the image areas in ATLAS, PolypGen, and SemSperm datasets.

***Index Terms*—Small medical object, cross-scale feature instruction, vision transformer, medical image segmentation**

I. INTRODUCTION

Semantic segmentation plays a crucial role in analysing natural and medical images, enabling distinguishment among different types of objects. This process allows for a detailed analysis of every pixel in a given image. Recently, deep learning algorithms have demonstrated significant potential for medical diagnosis through recognising medical images [1]–[6]. However, these methods often neglect the detection of small medical objects.

In the early stages of diseases, the affected regions, such as tumour [7] or skin lesions [8], are relatively small. Early detection can aid in the discovery of potential diseases, thereby increasing patient survival rates. Cell-level imaging analysis is a cutting-edge topic with various clinical applications, such as human reproduction [10]. Moreover, a considerable number of medical images contain numerous lesions that occupy less than 10% of the total image area. For example, 74% of tumours in ATLAS [7] and 63% of polyps in PolypGen [9]. Therefore, developing a practical method

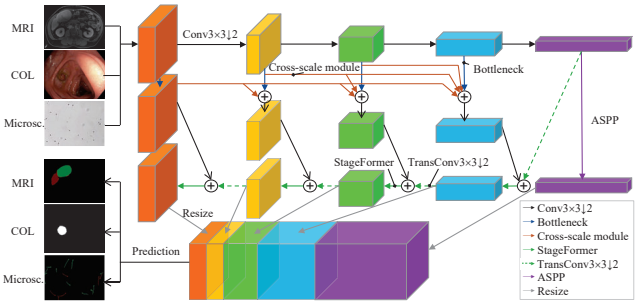


Fig. 1: Architecture of small-object segmentation with transformer (SoSegFormer) network. MRI is magnetic resonance imaging; COL is colonoscopy; Miscosc. is microscopy imaging.

for detecting small medical objects is paramount. A small medical object refers to an object that occupies a relatively small area in the image, presenting a significant challenge to deep learning methods. Convolution and pooling operations used in deep learning algorithms generate lower resolution of image features, leading to the loss of the morphology characteristics of medical objects [4]. For instance, only 3×3 pixels remain for a 45×45 medical object in a 512×512 image that has undergone two convolution operations with 3×3 kernel size, which is particularly problematic for small medical objects.

To address the issue of diminished image resolution and information loss, cross-scale feature aggregation has been proposed to harness the hierarchical semantic feature of objects [5], [11]. Moreover, attention mechanisms have been employed to learn the global information of medical objects [3], [4], [6]. In medical applications, researchers have leveraged attention mechanisms through convolution or pooling to extract the global feature of polyp images [3], [4]. However, these methods employ relatively small feature maps (ranging from 11×11 to 44×44) to bridge area and boundary cues, which may not adequately capture the structural details of minuscule objects. Most recently, the vision transformer (ViT) has been introduced to process sequences of image patches to learn the inter-patch representations, which has shown immense potential in aggregating and preserving the features of small objects [6].

In this study, we present a novel convolution-transformer hybrid architecture, termed SoSegFormer, to address the

challenges in medical object segmentation. SoSegFormer integrates two core techniques, cross-scale feature instruction and vision transformer (StageFormer), to extract features and learn global positional and morphological information of small medical objects. Qualitative evaluations on three benchmark datasets, using five metrics, demonstrate that our SoSegFormer significantly enhances the accuracy of small medical object segmentation across multiple modalities (*i.e.*, MRI, COL, and Microsc.). For example, SoSegFormer surpassed other tested methods, achieving the highest mDice of 87.02% and the lowest MAE of 1.9×10^{-3} in distinguishing liver and tumour covering less than 1% regions in abdominal slices.

II. METHODOLOGY

This section elucidates the primary methodologies and specifics of the small-object segmentation with transformer (SoSegFormer) model, as detailed in Sec. II-A–II-C.

II-A. Overall Framework

This subsection introduces the SoSegFormer model, specifically designed to segment small objects. The SoSegFormer network, schematically depicted in Fig. 1, comprises two main components: cross-scale feature map instruction in Sec. II-B and the convolution with vision transformer in Sec. II-C. The cross-scale feature map instruction module is designed to enhance the performance of extracting the features of small medical objects (*i.e.*, orange arrow in Fig. 1). Moreover, to improve the identification of medical object features and their correlations, the correlated features are further processed by convolutions integrated with transformers (*i.e.*, green arrow in Fig. 1). In addition, every bottleneck in Fig. 1 (indicated by blue arrows) comprises a 1×1 convolution, followed by a 3×3 convolution and subsequent 1×1 convolution operations. Besides, each TransConv 3×3↓2 in Fig. 1 (denoted by green dashed arrows) consists of three convolution units, each formed by a 1×1 convolution, followed by a 3×3 transposed convolution and a final 1×1 convolution. To expand the receptive field for extracting features in multi-scale, atrous spatial pyramid pooling (ASPP) [11] was introduced after the final stage of our model.

II-B. Cross-scale Feature Instruction

The features of tiny medical objects, such as sperms, are challenging to preserve following several convolution operations with a stride of 2 (strided convolution). The smaller the medical object, the less information remains after a strided convolution. Therefore, leveraging the higher resolution features at the earlier model stages is essential. This study applies cross-scale feature maps to guide the latter stages of learning small object features. Given n input tensors, $R_s, s = 1, 2, ..., n$, the output is presented by R' and can be computed as follows:

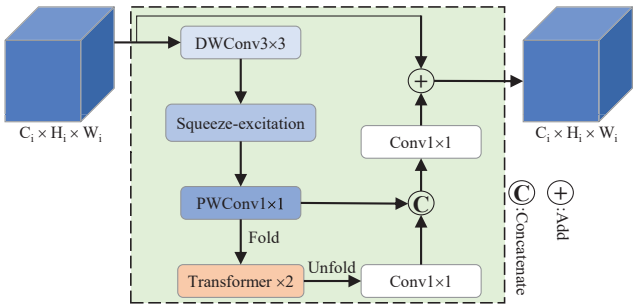


Fig. 2: An illustration of stage transformer (StageFormer). StageFormer correlates local and global features using convolution with transformer operations.

$$R' = f_{1r}\left(R_1\right) + f_{2r}\left(R_2\right) + ... + f_{nr}(R_n) \tag{1}$$

where the transformation $f_{xr}\left(R_s\right)$ involves $(r - x)$ 3×3 strided convolutions. The function is schematically presented by the orange arrows in Fig. 1.

II-C. Convolution with Vision Transformer

The proposed stage transformer (StageFormer), detailed in Fig. 2, incorporates a depthwise separable convolution with a squeeze-excitation layer, followed by two transformer blocks and several convolution operations. The StageFormer module is based on the convolution-transformer hybrid structure, capable of learning the local and global representations of an input medical image simultaneously. The functionalities of convolution and transformer operations differ. Convolution operations are tasked with learning local and general features, including corners, edges, angles, and colours of medical objects. Conversely, the transformer module extracts global information such as morphology, depth, and colour distribution of medical objects using multi-head self-attention (MHSA). Furthermore, the transformer module also learns positional relationships of medical objects, such as those between a tumour and kidney, kidney and abdomen, and tumour and abdomen in an MRI slice image. Additionally, a skip connection and a concatenation are incorporated to mitigate the information loss of small medical objects and enhance the SoSegFormer's ability to recognise these objects.

III. EXPERIMENTAL RESULTS

This section introduces the implementation of the proposed methods and discusses the experimental data.

III-A. Dataset and Evaluation Metric

To validate the efficacy of SoSegFormer, we tested it alongside six other state-of-the-art (SOTA) models for small medical object segmentation on three benchmark datasets, ATLAS [7], PolypGen [9], and SemSperm. The ATLAS dataset includes 90 T1 CE-MRI scans of livers featuring two types of medical objects: the liver and tumour. For experimental comparison, the ATLAS dataset was sliced into sequences of 2-D images. The PolypGen dataset, used for

Table I: Details of ATLAS, PolypGen, and SemSperm datasets.

Dataset	Number of Image (train + test)	Object area ratio	Number of object		
			ultra small	small	all
ATLAS [7]	997 + 249	0.001% $\sim$ 25.826%	274	1084	1464
PolypGen [9]	1230 + 307	0.003% $\sim$ 85.850%	81	895	1411
SemSperm	118 + 30	0.042% $\sim$ 0.651%	1456	1456	1456

polyp segmentation, contains data from six different hospitals, focusing on one type of medical object, the polyp. The SemSperm is a private dataset comprising low-dimension sperm images (640×480, 96 dpi) separated from videos from the Prince of Wales Hospital in Hong Kong. These sperm images were carefully annotated into normal and abnormal categories by experienced fertility doctors. Further details of the tested datasets are presented in Tab. I. As there is no strict definition for the size of a small object in an image, this study considers medical objects with area ratio below 1% (ultra-small) and 10% (small) as small medical objects.

The metrics used to assess the performance of semantic segmentation include the mean Intersection over Union (mIoU), mean Dice Coefficient (mDice), Mean Absolute Error (MAE), Precision, and F2 Score.

III-B. Implementation Details

In this study, the mini-batch size was set to 4. Data augmentation strategies applied to pre-process the input images included random horizontal flip, random crop with a resolution of 512×512, Gaussian blur, distortion, and rotation. The AdamW [12] optimiser and cross-entropy loss were utilised with the learning rate decaying from 5×10^{-5} to 1×10^{-6}. The total training process spanned 100 epochs. The results were calculated by averaging the outcomes of three times of training and testing processes. The experiments were conducted using an RTX3090 GPU with an Intel Xeon Platinum 8375C CPU.

III-C. Results and Analysis

The experimental results in ATLAS and PolypGen datasets are presented in Tab. II. It is apparent that the SoSegFormer outperforms other SOTA methods in all used metrics covering ultra-small, small, and all medical object segmentation in both datasets. For instance, SoSegFormer achieved the highest segmention mDice with 87.02% & 80.91%, 87.73% & 90.98%, and 89.15% & 93.06% in ultra-small, small, and all medical object segmentation in ATLAS and PolypGen datasets. Moreover, SoSegFormer delivered up to 2.42% & 13.24%, 3.63% & 7.83%, and 4.35% & 9.13% better F2 Score covering ultra-small, small, and all medical object segmentation in ATLAS and PolypGen datasets. Because the F2 Score is the harmonic mean of Precision and Recall, such results demonstrate that SoSegFormer is comparatively robust in medical object segmentation. Closer inspection of the first column of Tab. I shows SoSegFormer gained the best mIoU of 83.51%, MAE of 0.19×10^{-4}, and Precision of 93.88% to segment ultra-small size compared to small and all sizes of livers and tumours in ATLAS dataset, revealing SoSegFormer's efficacy of discriminating ultra-small medical objects.

Furthermore, Tab. III shows that the SoSegFormer network secured the first place in sperm segmentation on the SemSperm dataset, attaining 54.45% in mIoU, 65.17% in mDice, 4.25×10^{-4} in MAE, 68.25% in Precision, and 63.70% F2 Score. Interestingly, the performance of SoSegFormer became notably superior in ultra-small medical object segmentation, surpassing other SOTA models by at least 4.29% in mIoU, 5.45% in mDice, 4.47% in Precision, and 5.12% in F2 Score. Conversely, those values are less than 2% in ATLAS and PolypGen datasets. Such a variance is likely due to SemSperm's relatively abundant number of ultra-small objects, with 1456 sperms used for training and testing. Besides, the values of mIoU, mDice, Precision, and F2 Score gained in SemSperm are significantly lower than those in ATLAS and PolypGen for all tested models, with a gap of $> 10\%$, because all sperms in SemSperm have an area lower than 1% with limited learnable features and more challenging to be differentiated.

The examples of prediction results of small medical objects are presented in Fig. 3. As observed in the first two rows of Fig. 3, the proposed SoSegFormer network can not only distinguish all tumours, livers, and polyps' positions but preferably recover the morphologies of these medical objects in the image. In contrast, other SOTA methods either mistakenly categorise tumour regions as liver regions [see pink dashed regions in Fig. 3 (c)(e)] or struggle to differentiate ultra-small polyps and tumours from the background [see Fig. 3 (b-g)]. Furthermore, the last bottom row of the figure highlights that SoSegFormer accurately located all sperms' positions and effectively recognised the region of the broken-line tail of an abnormal sperm. In contrast, other SOTA methods either missed the broken-line shape or misclassified the abnormal sperm head as the normal sperm head. Because the SoSegFormer won the first in segmenting medical images with three different modalities (*i.e.*, MRI, COL, Microsc.), the SoSegFormer has huge potential to be applied to general small medical object recognition regardless of modality.

IV. CONCLUSION

In this paper, we introduce a novel network, SoSegFormer, to enhance the performance of segmenting small clinical objects in medical applications. The experimental results indicate that the SoSegFormer network is capable of effectively differentiating ultra-small, small, and all sizes of medical objects. The SoSegFormer outperformed SOTA methods and achieved 87.02%, 80.91%, and 65.17% mDice for segmenting objects with less than 1% images area in ATLAS, PolypGen and SemSperm datasets. Visualisation results demonstrate that the SoSegFormer could accurately detect all medical object locations and morphologies. These results demonstrate the superior ability of SoSegFormer to distinguish small medical objects.

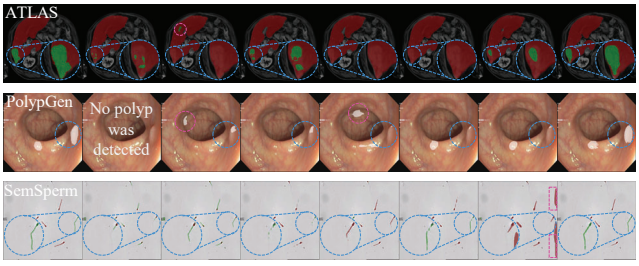
Table II: Quantitative results in the ATLAS and PolypGen datasets. Results vary across area ratios of medical objects below 1% (ultra-small), 10% (small), and 100% (all). The best results are highlighted.

	Methods	mIoU			mDice			MAE ($\times 10^{-4}$)			Precision			F2 Score		
		ultra small	small	all	ultra small	small	all	ultra small	small	all	ultra small	small	all	ultra small	small	all
ATLAS	U-Net (MICCAI'15) [1]	79.36	76.84	77.72	81.44	84.18	85.39	0.60	0.83	0.95	80.33	87.21	87.90	82.68	83.20	84.34
	U-Net++ (MICCAI'18) [2]	80.31	77.97	77.23	82.80	84.65	84.46	0.39	0.65	0.77	82.79	87.52	88.14	82.80	83.79	83.34
	PraNet (MICCAI'20) [3]	82.12	80.57	80.79	85.32	87.12	87.62	0.30	0.58	0.73	86.64	89.35	89.77	84.69	86.08	86.62
	HRNet (TPAMI'20) [11]	80.14	77.43	80.02	82.56	84.07	87.13	0.41	0.62	0.72	81.96	89.03	91.35	83.01	82.62	85.57
	CaraNet (JMI'23) [4]	82.23	78.09	79.93	85.45	85.31	87.29	0.27	0.74	0.84	87.82	87.43	88.88	84.48	84.31	86.60
	CFANet (PR'23) [5]	81.94	79.46	80.13	85.05	86.50	87.42	0.24	0.78	0.93	88.19	87.55	87.49	83.92	85.92	87.38
	SoSegFormer (Ours)	**83.51**	**81.20**	**82.52**	**87.02**	**87.73**	**89.15**	**0.19**	**0.57**	**0.70**	**93.88**	**91.42**	**92.54**	**85.10**	**86.25**	**87.69**
PolypGen	U-Net (MICCAI'15) [1]	70.92	74.94	76.60	73.22	83.50	85.45	2.51	1.20	1.59	71.96	84.99	90.96	74.97	82.67	82.98
	U-Net++ (MICCAI'18) [2]	72.42	77.85	80.21	75.71	85.93	88.15	2.53	1.05	1.35	73.51	86.78	91.86	78.96	85.44	86.31
	PraNet (MICCAI'20) [3]	75.50	83.23	86.36	79.96	90.00	92.30	1.45	0.72	0.90	77.21	92.44	94.94	82.83	88.68	90.91
	HRNet (TPAMI'20) [11]	68.17	76.32	81.59	69.82	84.73	89.15	7.37	1.25	1.31	69.94	82.67	90.32	70.80	86.12	88.49
	CaraNet (JMI'23) [4]	75.33	83.55	86.83	79.75	90.22	92.60	1.57	0.72	0.89	76.81	91.95	94.20	83.06	89.26	91.72
	CFANet (PR'23) [5]	74.15	83.87	86.82	78.21	90.45	92.60	2.06	0.69	0.88	75.28	92.48	94.77	82.14	89.33	91.43
	SoSegFormer (Ours)	**76.22**	**84.63**	**87.56**	**80.91**	**90.98**	**93.06**	**1.37**	**0.67**	**0.83**	**77.92**	**91.79**	**94.80**	**84.04**	**90.50**	**92.11**

Table III: Quantitative results in the SemSperm dataset. All sperms in SemSperm datasets occupy below 1% area in images. The best results are highlighted.

	Methods	mIoU	mDice	MAE*	Precision	F2 Score
SemSperm	U-Net (MICCAI'15) [1]	45.85	53.26	4.89	63.52	51.87
	U-Net++ (MICCAI'18) [2]	47.89	56.05	5.11	64.74	55.50
	PraNet (MICCAI'20) [3]	45.56	53.87	4.78	62.27	51.02
	HRNet (TPAMI'20) [11]	50.16	59.72	4.84	62.90	58.58
	CaraNet (JMI'23) [4]	46.27	54.94	4.50	59.69	53.07
	CFANet (PR'23) [5]	42.65	59.62	5.69	52.23	48.47
	SoSegFormer (Ours)	**54.45**	**65.17**	**4.25**	**68.26**	**63.70**

*MAE unit: $\times 10^{-4}$



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3: Visualisation of segmentation (a) ground truth and results using (b) U-Net, (c) U-Net++, (d) PraNet, (e) HRNet, (f) CaraNet, (g) CFANet, and (h) SoSegFormer (ours).

V. REFERENCES

[1] 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*. Springer, 2015, pp. 234–241.

[2] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *8th ML-CDS Workshop on International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2018, pp. 3–11.

[3] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020, pp. 263–273.

[4] A. Lou, S. Guan, and M. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," *Journal of Medical Imaging*, vol. 10, no. 1, p. 014005, 2023.

[5] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, and D. Shen, "Cross-level feature aggregation network for polyp segmentation," *Pattern Recognition*, vol. 140, p. 109555, 2023.

[6] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, "Deeply supervised skin lesions diagnosis with stage and branch attention," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 719-729, 2024.

[7] F. Quinton, R. Popoff, B. Presles, S. Leclerc, F. Meriaudeau *et al.*, "A tumour and liver automatic segmentation (ATLAS) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma," *Data*, vol. 8, no. 5, p. 79, 2023.

[8] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 168–172.

[9] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro *et al.*, "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *Scientific Data*, vol. 10, no. 1, p. 75, 2023.

[10] C. Dai, Z. Zhang, J. Huang, X. Wang, C. Ru *et al.*, "Automated non-invasive measurement of single sperm's motility and morphology," *IEEE Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2257–2265, 2018.

[11] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint:1711.05101*, 2017.