Multi-scale cross fusion attention for few-shot intestinal polyp image semantic segmentation

Yuxiu Kang*[a], Yaling Zhu[b], Jundi Wang[b]
[a]Clinic of Logistics Management, Lanzhou Institute of Technology, Lanzhou 730050, Gansu, China;
[b]School of Computer Science & Artificial Intelligence, Lanzhou Institute of Technology, Lanzhou 730050, Gansu, China

ABSTRACT

The few-shot intestinal polyp image semantic segmentation aims to segment unseen targets in the query image with only a few pixel-by-pixel annotated support images. However, existing few-shot intestinal polyp image semantic segmentation methods mainly mine valuable guidance information from the support branch, ignoring the role played by the task target information in the query branch in improving model performance. In this paper, a few-shot intestinal polyp image semantic segmentation method based on multi-scale cross fusion attention is proposed. Firstly, the pre-trained convolutional neural network is used to map the images on both branches into the same feature space, and the multi-scale information on different branches is extracted respectively. Then, cross attention is used to establish the scale fusion between the multi-scale information of the support branch and the query branch, promoting the semantic alignment of features between branches. Finally, the similarity values between the encoding features at each position on the prototype set and the query image are calculated using a parameter-free metric method, and the unseen target area in the query image is segmented according to the similarity value. Extensive experiments on open-source intestinal polyp image dataset demonstrate the superiority of the designed method.

Keywords: Intestinal polyp image semantic segmentation, multi-scale information, cross attention

1. INTRODUCTION

The mortality rate of colon cancer has been increasing annually, with the primary cause being intestinal polyps that gradually deteriorate over time[1]. Therefore, early detection of colon cancer can help provide patients with effective and timely treatment. Currently, the detection of colon polyps mainly relies on colonoscopy, and judgments are based solely on the experience of experts, which can easily lead to missed detections and misjudgments[2]. As a result, using computer-aided diagnostic technology to assist professional physicians in detecting and segmenting intestinal polyps has become a hot research topic in the field of medical image semantic segmentation in recent years.

The performance of traditional deep learning models heavily depends on data samples with pixel-by-pixel annotations[3]. However, annotating intestinal polyp images pixel by pixel is time-consuming and labor-intensive, and there are few annotated samples available for new types of intestinal polyp lesions[4,5]. Few-shot learning can leverage a few annotated images to guide the segmentation of unseen target areas in query images of the same category. In few-shot learning networks[6], there are typically two branches: the support branch, which takes annotated intestinal polyp images as input, and the query branch, which takes the intestinal polyp images to be segmented as input. For example, Fan et al. have proposed a parallel backward attention network for the intestinal polyp image semantic segmentation, capturing the global encoding features of intestinal polyps in high-level feature maps and employing a reverse attention module to mine boundary clues of the polyp regions, effectively enhancing the contextual semantic association between regional and boundary clues[7]. Kim et al. introduced a contextual attention enhancement network for the semantic segmentation of intestinal polyp images, using an improved U-Net as the backbone network and introducing an attention mechanism between layers to enhance the encoding ability of local and global features, demonstrating the superiority of their designed method on open datasets[8]. Zhang et al. utilized a parallel dual-branch network of Transformer and CNN (Convolutional Neural Network) to capture the global and local encoding features of intestinal polyp images, respectively, and fused the encoding features of Transformer and CNN through a cross-attention fusion network, significantly improving the segmentation performance of intestinal polyps by fully leveraging the advantages of each

*1323226853@qq.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133952X · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048130

feature layer and reconciling inconsistencies between them[9]. Ling et al. proposed a novel semantic segmentation method for intestinal polyp images based on PETNet, guiding the semantic fusion of intestinal polyp encoding features with Gaussian probability and using a binary mask-supervised decoder to achieve the segmentation of the intestinal polyp regions to be segmented[10].

Although the methods mentioned above have been successful, the performance of these models still relies on a large dataset with pixel-by-pixel annotated images of intestinal polyps. Additionally, existing methods primarily extract global encoding features from intestinal polyp images to guide the mask prediction of unseen target areas in query images. However, this approach can easily overlook low-level texture and edge clues, as well as high-level feature classification information, which hinders the improvement of intestinal polyp image segmentation performance. To tackle this issue, we propose a few-shot intestinal polyp image semantic segmentation method based on multi-scale cross-fusion attention. This method constructs a multi-scale feature set for intestinal polyp images on both the support and query branches by extracting different layer features encoded by the backbone network. It also utilizes a cross-attention mechanism to establish feature interaction between the support and query branches, thereby enhancing the robustness and reliability of feature representation.

2. PROBLEM SETUP

The goal of few-shot intestinal polyp image semantic segmentation is to utilize a few-shot pixel-by-pixel annotated support images to guide the segmentation of unseen regions in the query image. Unlike traditional image semantic segmentation tasks[11], the few-shot intestinal polyp image semantic segmentation includes two disjoint datasets, namely the Base set and the Novel set. All classes as C with $C_B \bigcap C_N = \varnothing$ and $C=C_B \bigcup C_N$. The images in the Base set are from the class C_B and are used for model training, while the images in the Novel set are from the class C_N and are used for model testing. Here, the mainstream Episode training mechanism in the current few-shot image semantic segmentation field is adopted[12], which divides the original task into multiple subtasks. The generalization performance of the model is trained and evaluated using the support set $S=\{(x_k, y_k)\}_{k=1}^{K}$ and query set $Q=\{(x, y)\}$ within the subtask. Here, x represents the image, y represents the corresponding true mask, and K denotes the number of samples of support images. Notably, y in the query set is only used during the model training phase.

3. METHODOLOGY

Figure 1 illustrates the network architecture of the few-shot intestinal polyp image semantic segmentation model proposed in this paper, which comprises three modules: multi-scale feature extraction, feature interaction, prototype generation, and mask prediction. In the multi-scale feature extraction stage, a backbone network pre-trained on the ImageNet dataset is employed to encode support and query images into a common feature space, extracting low, middle, and high-level features from the encoded feature maps to form a multi-scale feature set. A feature interaction module based on a cross-attention mechanism is then used to facilitate semantic alignment between different features across branches by interacting with the multi-scale feature sets. Global average pooling is applied to the resulting integrated feature map to create a prototype set. This set is then used to calculate similarity scores between the encoded features of the intestinal polyp images to be segmented and the prototype set.

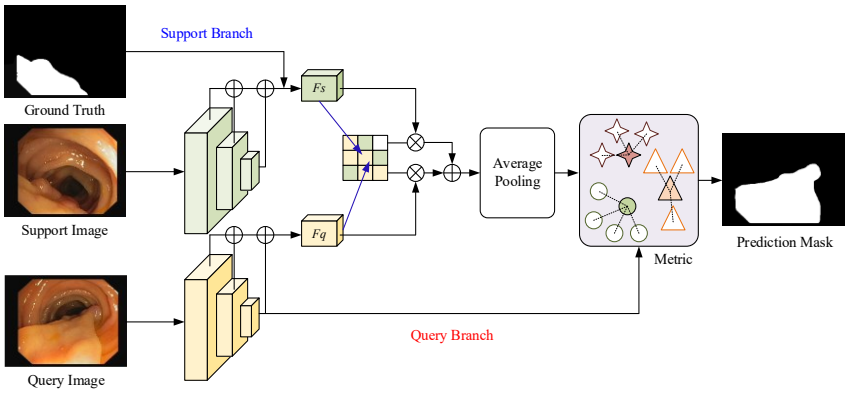

Figure 1. Model structure for few-shot intestinal polyp image semantic segmentation.

3.1 Multi-scale feature extraction

Using pre-trained feature extractors to map images into deep feature space has become the standard procedure in computer vision tasks[13]. Existing methods for intestinal polyp image semantic segmentation primarily utilize global features to construct prototype sets. However, different layers into which images are mapped as deep features contain different types of semantic information[14]. Where the low feature maps focus on textures and edges, the high-level feature maps focus on class information. Here, to fully exploit the semantic information of deep encoded features in both low and high-level feature spaces, features from low, middle, and high layers are extracted respectively, forming a multi-scale feature set. The multi-scale features on the support branch can be defined as equation (1).

$$Fs = \lambda_{1-l}(x_s^i) + \lambda_{l-m}(x_s^i) + \lambda_{m-h}(x_s^i) \tag{1}$$

Similarly, the multi-scale feature set on the query branch can be defined as equation (2).

$$Fq = \lambda_{1-l}(x_q) + \lambda_{l-m}(x_q) + \lambda_{m-h}(x_q) \tag{2}$$

where $\lambda(\cdot)$ represents the backbone network pre-trained on ImageNet, while l, m, and h denote the number of layers corresponding to low, middle, and high-level features, respectively.

3.2 Feature interaction

Different support images play different roles in guiding the segmentation of unseen targets in query images. In order to make good use of the semantic correlations between the multi-scale feature sets on the support branch and the query branch, a cross-attention mechanism as shown in Figure 2 is used to establish semantic relationships between branches, promoting the exchange of information across branches. The cross-attention computation is defined by equation (3).

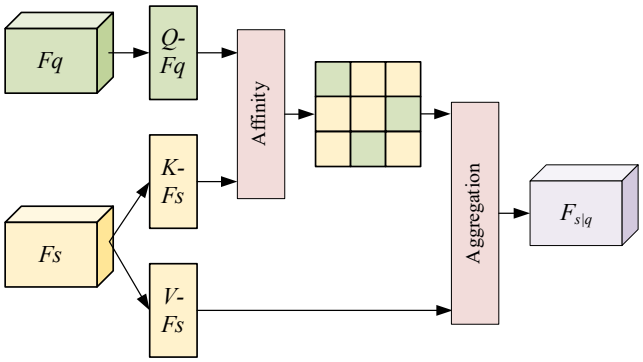

Figure 2. Cross-attention interaction process.

$$F_{s|q} = \text{softmax}\left(\frac{Fs \cdot Fq^{\mathrm{T}}}{\sqrt{d}}\right)(Fs + Fq) \tag{3}$$

where $F_{s|q}$ represents the cross-fused features, T denotes the transpose operation, and d indicates the feature dimension. Establishing the cross-attention between branches can help to facilitate the exchange of information between different feature maps across branches, enhancing the robustness and reliability of feature representations.

3.3 Prototype generation and mask prediction

The prototype vector is an abstract feature representation used to express the semantic information of classes in visual space. Here, inspired by the work[15], we obtain a prototype representation that guides the unseen target in the query image. The computation is defined as equation (4).

$$P = \frac{\sum_{x,y} F_{s|q}^{(x,y)} \otimes y_i^{(x,y)}}{\sum_{x,y} y_i^{(x,y)}} \tag{4}$$

where (x, y) denotes location information, and P represents the prototype set.

To match the prototypes with the target region features in the query image on a per-pixel basis, a non-parameter metric approach is adopted, namely, the cosine similarity function, which calculates the similarity between the feature at each location and the prototype set, as specifically calculated by equation (5).

$$s^{(x,y)}=\sum_{x,y}^{h,w}\frac{Fq^{(x,y)}\cdot P_i}{|Fq^{(x,y)}||P_i|}, P_i \in P \tag{5}$$

where $s^{(x,y)}$ represents the similarity score between the prototype set and the query feature map at location (x,y). h and w denote the length and width of the feature map respectively, and the semantic class information at the current position is determined based on the maximum score. The maximum similarity score calculation is defined as equation (6).

$$\text{max-}s^{(x,y)} = \sum_{x,y} \text{argmax}(s^{(x,y)}) \tag{6}$$

where $\text{max-}s^{(x,y)}$ represents the maximum similarity score at location (x,y). Finally, the cross-entropy loss between the predicted mask and the true mask at each location is computed, the prototype set is optimized in an end-to-end manner.

4. EXPERIMENTS

4.1 Implementation details and evaluation metric

All experiments in this paper are conducted on the Ubuntu 18.04 operating system with an NVIDIA GPU V100 32GB, using the PyTorch deep learning framework and Python as the programming language. The initial learning rate is set to 0.0025, which decreases to 0.1 every 50 cycles, and the batch size is set to 8. The Adam optimizer is adopted, with a momentum parameter of 0.9. In addition, the input image size is set to 224×224.

To evaluate the superiority of the designed method, the mainstream mIoU (Mean Intersection over Union) and FB-IoU (Foreground & Background Intersection over Union) metrics in the field of few-shot intestinal polyp image semantic segmentation are selected.

4.2 Datasets

The current classical intestinal polyp image datasets are used to evaluate the proposed model, namely Kvasir-SEG, EndoTect, CVC-ColonDB, and CVC-ClinicDB, with 1000, 612, 200, and 380 images respectively selected. These datasets are divided into Base and Novel sets in a 7:3 ratio. The intestinal polyp images on different datasets are shown in Figure 3.

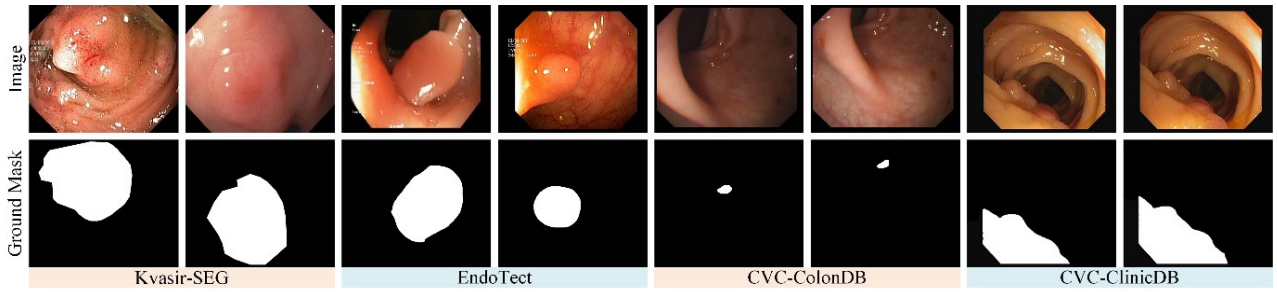

Figure 3. Example images on Kvasir-SEG, EndoTect, CVC-ColonDB, and CVC-ClinicDB.

4.3 Experimental results and analysis

To validate the effectiveness of the proposed method, the classic segmentation model U-Net in the medical image field is chosen as the baseline model, and mIoU and FB-IoU are calculated on the four datasets: Kvasir-SEG, EndoTect, CVC-ColonDB, and CVC-ClinicDB. The detailed results are shown in Table 1.

From Table 1, it can be seen that the proposed method outperforms the baseline model U-Net in both mIoU and FB-IoU on the four datasets. Specifically, on the Kvasir-SEG dataset, compared to the baseline model U-Net, the mIoU of this method improved by 8.5%, on the EndoTect dataset, this method improved by 1.8%, on the CVC-ColonDB and CVC-ClinicDB datasets, this method improved by 9.5% and 4.7% respectively. The above experimental results verify that the proposed method has better segmentation performance, which is due to the fact that this method fully exploits the limited

information of dual-branch input images during the feature extraction stage, and secondly, the use of a cross-attention mechanism can effectively promote feature alignment between branches. Figure 4 provides a visualization of the segmentation results of this method and the baseline model U-Net, showing that this method is superior to the baseline model U-Net in terms of the overall positioning of edges and target areas.

Table 1. Performance comparison of different methods.

Datasets	Methods	mIoU	FB-IoU
Kvasir-SEG	UNet	0.746	0.875
	Ours	0.831	0.892
EndoTect	UNet	0.825	0.906
	Ours	0.843	0.913
CVC-ColonDB	UNet	0.444	0.618
	Ours	0.539	0.636
CVC-ClinicDB	UNet	0.755	0.884
	Ours	0.802	0.893

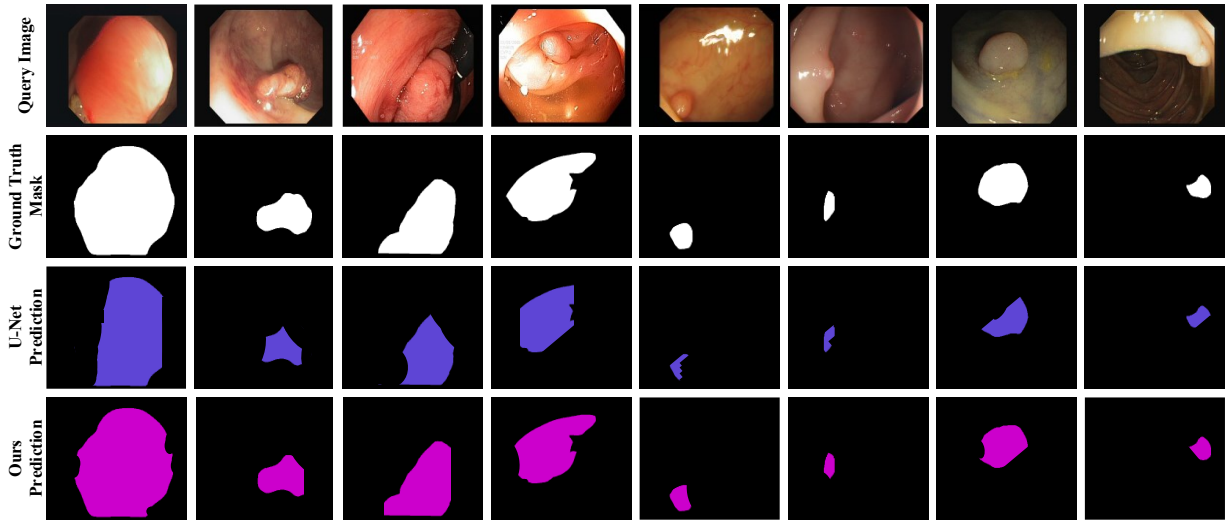

Figure 4. Segmentation visualization results.

4.4 Ablation experiment

To analyze the role of the multi-scale and cross-attention fusion modules in the performance improvement of the intestinal polyp image segmentation model proposed in this paper, several ablation experiments are designed as follows, with detailed results shown in Table 2.

Table 2. Ablation experiment results

Datasets	Methods	mIoU	FB-IoU
Kvasir-SEG	w/o Multi-scale	0.749	0.855
	w/ Multi-scale	0.802	0.881
	+ Cross attention	0.831	0.892
EndoTect	w/o Multi-scale	0.784	0.862

Datasets	Methods	mIoU	FB-IoU
	w/ Multi-scale	0.811	0.900
	+ Cross attention	0.843	0.913
CVC-ColonDB	w/o Multi-scale	0.495	0.586
	w/ Multi-scale	0.513	0.602
	+ Cross attention	0.539	0.636
CVC-ClinicDB	w/o Multi-scale	0.733	0.842
	w/ Multi-scale	0.785	0.871
	+ Cross attention	0.802	0.893

5. CONCLUSION

A few-shot intestinal polyp image semantic segmentation method based on multi-scale cross-integration attention has been proposed, which effectively improves the segmentation performance for intestinal polyp regions. On one hand, by extracting the multi-scale features of support and query images in the deep feature space, the full utilization of features is effectively enhanced. On the other hand, establishing information exchange between the support branch and the query branch helps to promote the alignment of features between branches, enhancing the reliability of feature expression. Extensive experiments on open-source intestinal polyp image datasets have also demonstrated the superiority of the proposed method.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Fund of Gansu Province, China (Grant No. 23JRRA1650, 22JR5RA384), the Support Project for Higher Education Industry of Gansu Province, China (Grant No. 2021CYZC-35)

REFERENCES

[1] Zheng, X., Gong, W., Yang, R., et al., “Image segmentation of intestinal polyps using attention mechanism based on convolutional neural network,” International Journal of Advanced Computer Science and Applications, 14(1), (2023).
[2] Amin, J., Sharif, M., Gul, E., et al., “3D-semantic segmentation and classification of stomach infections using uncertainty aware deep neural networks,” Complex & Intelligent Systems, 8(4), 3041-3057 (2022).
[3] Yang, K., Chang, S., Tian, Z., et al., “Automatic polyp detection and segmentation using shuffle efficient channel attention network,” Alexandria Engineering Journal, 61(1), 917-926 (2022).
[4] Hwang, M., Qian, Y., Wu, C., et al., “A local region proposals approach to instance segmentation for intestinal polyp detection,” International Journal of Machine Learning and Cybernetics, 14(5), 1591-1603 (2023).
[5] Dumitru, R., Peteleaza, D., Craciun, C., “Using DUCK-Net for polyp image segmentation,” Scientific Reports, 13(1), 9803 (2023).
[6] Yue, G., Li, S., Cong, R., et al., “Attention-guided pyramid context network for polyp segmentation in colonoscopy images,” IEEE Transactions on Instrumentation and Measurement, 72, 1-13 (2023).
[7] Fan, D., Ji, G., Zhou, T., et al., “Pranet: Parallel reverse attention network for polyp segmentation,” International Conference on Medical Image Computing and Computer-Assisted Intervention, 263-273 (2020).
[8] Kim, T., Lee, H. and Kim, D., “Uacanet: Uncertainty augmented context attention for polyp segmentation,” The 29th ACM International Conference on Multimedia, 2167-2175 (2021).
[9] Zhang, Y., Liu, H. and Hu, Q., “Transfuse: Fusing transformers and CNNS for medical image segmentation,” Medical Image Computing and Computer Assisted Intervention, 14-24 (2021).
[10] Ling, T., Wu, C., Yu, H., et al., “Probabilistic modeling ensemble vision transformer improves complex polyp

segmentation," International Conference on Medical Image Computing and Computer, 572-581 (2023).
[11] Yang, Y., Chen, Q., Feng, Y., et al., "Mianet: aggregating unbiased instance and general information for few-shot semantic segmentation," The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7131-7140 (2023).
[12] Song, Y., Wang, T., Cai, P., et al., "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," ACM Computing Surveys, 55 (13s), 1-40 (2023).
[13] Cao, Q., Chen, Y., Ma, C., et al., "Few-shot rotation-invariant aerial image semantic segmentation," IEEE Transactions on Geoscience and Remote Sensing, 62, 1-13 (2023).
[14] Kayabaşı, A., Tüfekci, G. and Ulusoy, İ., "Elimination of non-novel segments at multi-scale for few-shot segmentation," The IEEE/CVF Winter Conference on Applications of Computer Vision, 2559-2567 (2023).
[15] Chang, Z., Lu, Y., Wang, X., et al., "Mgnet: Mutual-guidance network for few-shot semantic segmentation," Engineering Applications of Artificial Intelligence, 116, 105431 (2022).