scientific reports

OPEN

A differential network with multiple gated reverse attention for medical image segmentation

Shun Yan, Benquan Yang & Aihua Chen

UNet architecture has achieved great success in medical image segmentation applications. However, these models still encounter several challenges. One is the loss of pixel-level information caused by multiple down-sampling steps. Additionally, the addition or concatenation method used in the decoder can generate redundant information. These limitations affect the localization ability, weaken the complementarity of features at different levels and can lead to blurred boundaries. However, differential features can effectively compensate for these shortcomings and significantly enhance the performance of image segmentation. Therefore, we propose MGRAD-UNet (multi-gated reverse attention multi-scale differential UNet) based on UNet. We utilize the multi-scale differential decoder to generate abundant differential features at both the pixel level and structure level. These features which serve as gate signals, are transmitted to the gate controller and forwarded to the other differential decoder. In order to enhance the focus on important regions, another differential decoder is equipped with reverse attention. The features obtained by two differential decoders are differentiated for the second time. The resulting differential feature obtained is sent back to the controller as a control signal, then transmitted to the encoder for learning the differential feature by two differential decoders. The core design of MGRAD-UNet lies in extracting comprehensive and accurate features through caching overall differential features and multi-scale differential processing, enabling iterative learning from diverse information. We evaluate MGRAD-UNet against state-of-theart (SOTA) methods on two public datasets. Our method surpasses competitors and provides a new approach for the design of UNet.

Keywords Medical image segmentation, Multi-scale feature extraction, Differential feature

Medical image segmentation plays a pivotal role in clinical applications. With the advancement of convolutional neural networks (CNNs)[1], various segmentation models have been developed, such as UNet[2], FCNs[3], etc. Among them, UNet has been widely used in medical image segmentation. It adopts symmetric encoder-decoder components with skip connections to accomplish medical image segmentation tasks. The encoder extracts deep features through convolution and downsampling, while the decoder upsamples and fuses encoder features from different scales to mitigate low-level features information loss.

With such excellent structural design, UNet has achieved great success in various medical imaging applications. Later, various improved algorithms based on UNet have been proposed. The improvements are reflected in following aspects: Some methods[4–7] have utilized multi-scale feature fusion mechanism. Unet++[7] introduces a novel architecture that improves feature fusion through modified skip connections and cascaded UNet modules. It decreases the semantic disparity between encoder and decoder feature maps. Res2Net[5] introduces a novel building block for convolutional neural networks, enhancing multi-scale feature representation within a single residual block. Several methods[8–10] have employed attention mechanism. PraNet[9] utilizes a reverse attention mechanism to capture contextual information effectively and refines polyp boundaries and internal structures. Attention UNet[10] introduces attention gates within the UNet architecture, enabling adaptive modulation of feature map relevance during processing. Some methods[11–14] have utilized transformer technique. TransUnet[11] innovates by integrating transformer-based global context extraction with CNN feature maps, enhancing semantic segmentation accuracy through precise localization. Swin-UNet[12] combines the Swin Transformer with the UNet architecture, addressing the computational resource consumption and performance degradation issues faced by traditional UNet models in handling large medical images. Some methods[15–18] have utilized double decoders or double encoders. DC-UNet[16] adopts a structural design with dual decoders, which facilitate the

School of Electronic and Information Engineering, Taizhou University, Taizhou 318000, Zhejiang, China. email: yangbq266@sina.com; chen_1216@163.com

extraction of multi-scale features and effectively solves the problem of single decoder being unable to balance precise localization and fine boundary refinement. DDU-Net[15] introduces of a dual-encoder structure, which enhances feature extraction and representation by incorporating additional encoders. A number of approaches as ASPP[19], DenseASPP[20] concentrate on extracting intra-layer multi-scale information using the spatial pyramid pooling module in their networks, gradually incorporate the semantic context and intricate texture information from various scale representations.

However, previous methods often fuse encoder features using simple operations like addition or concatenation. These methods generate redundant information, weaken specific-level features, affecting both localization and boundary refinement. They overlook the importance of differences between levels, resulting in a decrease in segmentation performance.

In this paper, we propose a novel multi-gated reverse attention multi-stage differential network (MGRAD-UNet) from the perspective of learning differential features at multiple scales and between the double decoder and encoder for general medical image segmentation. Firstly, we place emphasis on differential features. We design two differential decoders with the differential process (DP) applied to each pair of neighboring levels. This highlights useful differences between features and eliminates interference from redundant parts. Then, we utilize a pyramid-style differential process to capture cross-level information. Next, we aggregate specific level features and multi-level differential features. We use them as gate signals to feed into another decoder, providing richer information for another differential decoder. Meanwhile, another differential decoder is equipped with gate and reverse attention mechanisms, as well as the pyramid differential process from the first decoder. These features from both decoders undergo another round of differentiation, enhancing the differential features, and then using them as gate signals to pass into gate-control port. Finally, excellent segmentation results are recursively obtained from this network.

Our main contributions to this research are as follows:

- We propose an efficient and versatile multi-gated reverse attention multi-scale differential network (MGRAD-UNet) for multifarious medical image segmentation. With multi-gated reverse attention mechanism and multi-scale differential module, the network can efficiently acquire differential features, thereby comprehensively enhancing the perception of organs or lesions.
- We propose a new perspective, which involves learning from differential features generated by two differential decoders.
- To better integrate the differential features into the network, we propose Multi-Scale Differential Decoder. It replaces traditional addition or concatenation feature fusion with an efficient differential aggregation. And we propose gate-control reverse attention differential decoder. It integrates the differential features of MSD, resulting in higher segmentation accuracy thereafter.
- MGRAD-UNet has conducted extensive experiments on two publicly available medical image segmentation datasets, and the results show that this method outperforms the current state-of-the-art methods.

Related works

Medical image segmentation

Medical image segmentation can be described as a dense prediction task that involves classifying pixels of lesions or organs in endoscopy, CT, MRI, etc[21]. The UNet architecture, introduced by Ronneberger et al.[2], has established itself as a cornerstone in medical image segmentation. Its encoder-decoder structure, featuring skip connections for enhanced feature aggregation, has become a bedrock for segmentation tasks. In light of the exceptional performance consistently demonstrated by numerous variants inspired by UNet, it becomes evident that this architectural paradigm has a lasting influence and proven effectiveness. UNet++[22] uses nested encoder-decoder sub-networks that integrate long and short connections to reduce the semantic gap between encoder and decoder feature mappings. MC-Net+[22] utilizes three decoders with different structures and enforces output consistency. For attention UNet[10], each transition layer and decoder block is embedded with an attention gate to automatically learn to focus on different shapes and sizes of target structures. Recently, the Transformer[23] architecture has achieved success in many natural language processing tasks. Some researches[11,24], have explored its effectiveness in medical visual tasks. UTNet[24] is a simple but powerful hybrid transformer architecture that minimizes the cost of capturing long-range dependencies between encoder and decoder. TransUNet[11] is a typical transformer-based model that is worth mentioning, which encodes image features as a sequence of global contexts and utilizes a U-shaped hybrid design to combine low-level CNN features. Swin-Unet[12] is a transformer-centric architecture built upon the Swin transformer[25]. Unlike conventional approaches, Swin-Unet integrates transformers into both the encoder and decoder. However, contrary to expectations, this dual implementation fails to yield performance enhancements.

We can see that the majority of medical image segmentation methods employ rich feature representation, multi-scale information extraction, and cross-level feature aggregation. These models rely on a large number of aggregation or concatenation operations for feature fusion, emphasizing consistency among features and therefore weakening the differential feature components. In contrast, our MGRAD-UNet focuses on extracting multi-scale differential features and learning differences between decoders, resulting in more efficient segmentation.

Multi-scale feature extraction

Multi-scale feature extraction refers to the use of multiple scales of features to extract information from an image, thereby improving the model's perception and segmentation accuracy. In medical image segmentation, multi-scale feature extraction can help the model better capture contextual information and features of lesions at different scales, thereby improving the accuracy of the segmentation results. Zhao et al.[26] introduced M^2

SNet, which utilizes scale cues to play an important role in capturing contextual information of objects. As the scale-space theory, a widely validated and theoretically sound framework, continues to inspire, an increasing number of multi-scale methods are being introduced[27]. This can be achieved by operating on multiple levels of the feature encoder and decoder, or by using multiple feature extraction modules to obtain features at different scales. Multi-scale methods include inter-layer multi-scale structures and intra-layer multi-scale structures. Inter-layer multi-scale structures achieve multi-scale by progressively aggregating features with different scales in the decoder, such as U-shaped[2,9,23,28–30] architectures. Intra-layer multi-scale structures obtain multi-scale features by using different dilation rates in feature extraction modules as some ASPP[30] modules, such as DenseASPP[20] and FoldASPP[31]. Differing from the previous methods, we simultaneously introduced inter-layer and intra-layer multi-scale. This allows us to utilize multi-scale differential information in our proposed method. The intra-layer differential process focuses on exploring the self-difference characteristics of feature pairs from pixel-pixel to region-region.

Experiments

Datasets

To verify the effectiveness of proposed framework on medical segmentation tasks, we have conducted tests on multiple organ CT segmentation challenges Synapse[11] and automatic heart diagnosis challenges ACDC[11] datasets.

Synapse for multi-organ CT segmentation

The Synapse dataset contains 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal CT images. Following the experimental protocol of TransUNet[11], we split the dataset into 18 scans for training, and 12 for testing. We extracted 2D slices from CT scans and segmented 8 abdominal organs, including the aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM), and evaluated the results using the Dice similarity coefficient.

ACDC for automated cardiac segmentation

The ACDC dataset contains Cine MR images from 100 patients, acquired with different 1.5T and 3.0T MR scanners and different temporal resolutions. For each patient, manual annotations of the right ventricle (RV), left ventricle (LV), and myocardium (MYO) are provided at both end-diastole (ED) and end-systole (ES) phases. In the following, we subjectively divide the ACDC dataset into three groups of 70%, 15%, and 15%, respectively, for training, validation, and testing.

Implementation details

The MGRAD-UNet is achieved based on Python 3.8 and Pytorch 1.11.0. Besides, we use a single NVIDIA RTX 4090 GPU with 24GB of memory to train all models. For all training cases, data augmentations such as flips and rotations are used to increase data diversity. The input image size is set as 224 × 224. The primary learning speed is put to 0. 01, the default optimize procedure is SGD, the momentum to 0.9 and the weight fall off is made to 1e−4. We train each model for a maximum of 300 epochs with a batch size of 8 for multi-organ segmentation in Synapse. For heart organ segmentation in ACDC, we use a batch size of 12 and train each model for a maximum of 300 epochs.

Experiment results

Experiment results on synapse

Based on results of multi-organ segmentation shown in Table 1, it can be seen that our proposed MGRAD-UNet outperforms all previous state-of-the-art CNN and transformer-based 2D medical image segmentation methods on the Synapse multi-organ CT dataset. Our MGRAD-UNet achieved an average Dice score of 83.33%, which is notably higher than the reported Dice scores of TransUNet and SwinUNet, which are 5.72% and 5.75% lower, respectively. When compared to the recent top-performing method, TransCASCADE (Dice of 82.68%), MGRAD-UNet shows a 0.65% improvement on this dataset. Furthermore, when comparing the 95% Hausdorff Distance(HD95) of all methods, it is observed that MGRAD-UNet has the lowest HD95 distance (16.67), which is 10.23 lower than TransUNet (HD95 of 26.90) and 0.67 lower than TransCASCADE (HD95 of 17.34). Moreover, the Fig. 2 visually represent the segmentation outcomes of models.

From our study of Dice scores of individual organs, we can see that our proposed MGRAD-UNet performs significantly better than other methods on five out of eight organs. We can also conclude that MGRAD-UNet performs better in both large and small organs, although the improvement is greater for small organs. We believe that the reason why MGRAD-UNet achieves better segmentation results is because it uses a multi-scale differential decoder and combines differential features from two decoders, and by reusing these features, the model's learning ability is strengthened.

Experiment results on ACDC

Table 2 reports cardiac organ segmentation results on the MRI data modality of the ACDC dataset. Our proposed MGRAD-UNet outperforms all other SOTA methods with better Dice scores. Compared to TransUNet and SwinUNet, MGRAD-UNet shows improvements of 2.42% and 4.06%, respectively. MGRAD-UNet also achieves the highest Dice scores in RV (90.42%), Myo (90.00%), and LV (95.98%) segmentation. Therefore, we can infer that by considering differential features as feedback signals and feeding them back to the encoder, along with multi-scale differential learning, the robustness of the encoder is effectively enhanced, and redundant information is eliminated.

Model parameters comparison

We further conducted experiments to compare parameters and computational complexities of various models. As shown in Fig. 3, although our computational complexity is the highest compared to other models, parameters is the lowest and our results are also improved. We observe that Dice is the highest in performance on the Synapse dataset. The reason behind this increase in computational complexity is the use of 3-time loop to fully utilize the differential information. These observations validate the efficiency and effectiveness of MGRAD-UNet.

Ablation study

Effectiveness of different modules

We have conducted an ablation study on the Synapse dataset and the ACDC dataset to evaluate the effectiveness of different components in our proposed MGRAD-UNet. We remove modules such as RFB and SideConv from the MGRAD-UNet and compare the results.

The results on the Synapse dataset, as shown in Table 3, clearly demonstrate that the RFB and SideConv modules contribute to improved performance. When the SideConv module is removed, average Dice and mIoU values decrease by 0.33% and 0.11% respectively. Average HD95 and ASD values increase by 2.13 and 0.23 respectively. Similarly, when the RFB module is removed, average Dice and mIoU values decrease by 0.72% and 1.13% respectively. Additionally, average HD95 and ASD values increase by 2.87 and 0.18 respectively.

When both modules are removed simultaneously, average Dice and mIoU values decrease by 0.93% and 1.23% respectively. Average HD95 and ASD values increase by 1.91 and 0.25 respectively. The best performance is achieved when both the RFB and SideConv modules are used, as it consistently outperforms other variations across all test datasets.

The results on the ACDC dataset, as shown in Table 4. When the SideConv module is removed, average Dice values decrease by 0.43%. Additionally, Dice values of RV, Myo and LV values decrease by 0.33%, 0.38% and 0.59% respectively. Similarly, when the RFB module is removed, average Dice values decrease by 0.71%. Additionally, Dice values of RV, Myo and LV values decrease by 0.37%, 0.26% and 1.11% respectively. When both modules are removed simultaneously, average Dice values decreased by 1.01%. Additionally, Dice values of RV, Myo and LV values decrease by 0.59%, 0.55% and 2.10% respectively. The best performance is achieved when both the RFB and SideConv modules are used, as it consistently outperforms other variations across all test datasets.

Effectiveness of multi-gated reverse attention mechanism

To evaluate the effectiveness of our proposed multi-gated reverse attention mechanism, we conduct a comparative analysis of MGRAD-UNet with and without reverse attention on the Synapse multi-organ dataset. It can be observed from Fig. 4 that the heatmaps generated with multi-gated reverse attention mechanism are more detailed, clearer, and richer in information compared to those generated without it. This demonstrates the remarkable effectiveness of the multi-gated reverse attention mechanism in improving feature extraction and model performance.

Effectiveness of aggregated loss

In our experiments, we used aggregated loss, which takes all predictions mapping from different stages of the network as input and sums up the losses of the prediction mappings generated from the non-empty subsets of n prediction mappings in $2^n - 1$. From Table 5, we can see When we replaced it with weighted loss, we found that the average values of Dice and mIoU decreased by 0.19% and 0.05% respectively, while the value of HD95 increased by 6.49.

In addition, to compare the convergence properties of the two loss functions, we provide visualizations of the corresponding losses in Fig. 5. It is precisely because aggregated loss can generate and combine prediction maps from subsets, resulting in more combined prediction maps, that its effectiveness has been proven.

Discussion

We adopt a differential process that is distinct from previous addition or aggregation operations in the multi-scale module. This differential process reduces redundancy among different levels in the resulting feature while significantly enhancing their scale-specific properties. Compared to single-scale designs, the multi-scale differential approach enables the network to gather more complementary information at both the pixel and neighborhood levels. Additionally, we further explore the potential of learning differentials and propose a dual-differential decoder structure, which leverages the differential information obtained through training with different decoders. In contrast to a single decoder, multiple decoders not only generate diverse prediction results, enhancing the accuracy and robustness of segmentation results, but also can capture features at different scales and levels. Our dual-differential decoder structure addresses scale information extraction and feature aggregation challenges. We believe that this new paradigm can drive further research on differential operations in the future.

Conclusion

In this paper, we reflect on the traditional methods of addition or connection and consider the differences among multiple decoders. On this basis we propose a new and effective differential network method MGRAD-UNet for more efficient medical image segmentation. Based on the proposed differential process, adjacent layers are differentially aggregated to extract complementary features from both low-level and high-level representations to enhance the multi-scale feature representation. Meanwhile, the network uses differences between two decoders as feedback signals and sends them to the gating unit. Unlike previous methods that emphasize filtering inconsistent regions, our approach is more effective. Specifically, it involves training differential decoders, followed by a learning process guided by resulting differences. In addition, we use an aggregated loss method to supervise the

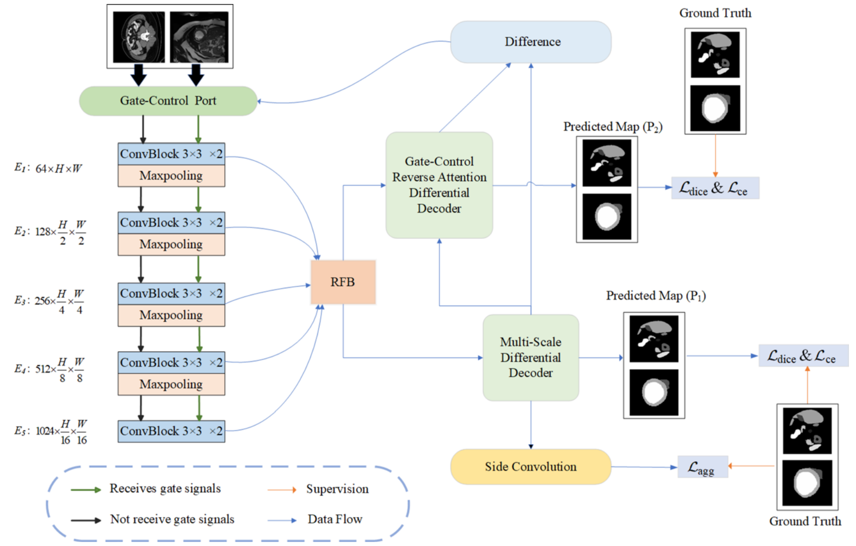

Fig. 1. Illustration of the proposed multi-gated reverse attention multi-scale differential network. On the left of the figure is the encoder network, and on the right are the dual differential decoder network and the prediction results generated by them. We perform differential processing on the features from decoders and provide differential results as supplementary information to the encoder for learning.

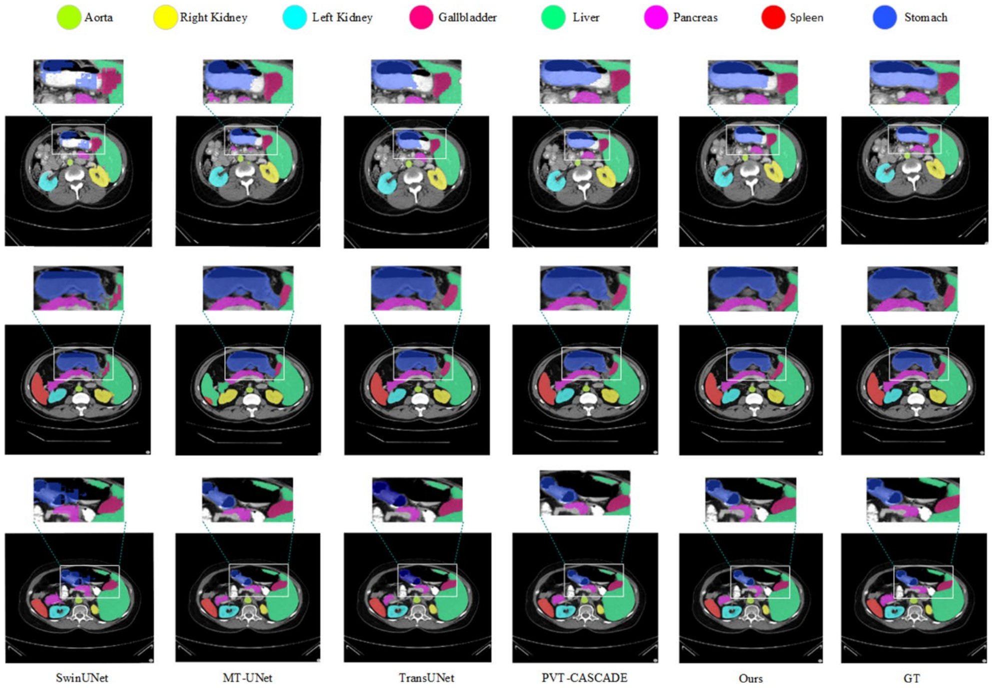

Fig. 2. The qualitative results of different methods on the Synapse Multi-Organ dataset include the Ground Truth (GT), Ours, SwinUNet, MT-UNet, TransUNet, and PVT-CASCADE. We overlay the segmentation maps on top of the original image/slice. We use white bounding boxes to highlight our excellent segmentation results.

prediction. Experimental results on medical segmentation tasks demonstrate that the proposed model outperforms various state-of-the-art methods.

Method

Overview

The MGRAD-UNet overall architecture is illustrated in Fig. 1. The MGRAD-UNet consists of a gate-control port, an encoder, a multi-scale differential decoder (MSD) and a gate-control reverse attention mechanism differential

Architectures	Average			Aorta	GB	KL	KR	Liver	PC	SP	SM
	Dice (%)↑	HD95 (voxel)↓	mIoU (%)↑								
UNet[2]	70.11	44.69	59.39	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet[10]	71.70	34.47	61.38	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet[11]	74.68	36.87	–	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet[11]	75.47	36.97	–	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
TransUNet[11]	77.61	26.90	67.32	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
MT-UNet[33]	78.59	26.59	–	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer[34]	81.96	18.20	–	86.99	68.65	85.21	82.00	94.41	**65.67**	91.92	80.81
SwinUNet[12]	77.58	27.32	66.88	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
PolypPVT[35]	78.08	25.61	67.43	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.40
SSFormerPVT[36]	78.01	25.72	67.23	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
HiFormer[37]	80.69	19.14	–	87.03	68.61	84.23	78.37	94.07	60.77	90.44	82.03
PVT-CASCADE[32]	81.06	20.23	70.88	83.01	70.59	82.23	80.37	94.08	64.43	90.10	**83.69**
TransCASCADE[32]	82.68	17.34	73.58	86.63	68.48	**87.66**	84.56	94.43	65.33	90.79	83.52
MGRAD-UNet (Ours)	**83.33**	**16.67**	**74.68**	**89.31**	**76.41**	85.22	**84.94**	**94.53**	61.17	**93.39**	81.66

Table 1. Evaluation metrics for the Synapse multi-organ segmentation. The evaluation metrics for UNet, AttnUNet, SSFormerPVT, and PolypPVT are taken from[32]. We reproduced the results for TransUNet, SwinUNet, HiFormer, PVT-CASCADE, TransCASCADE with a batch size of 8 and an input resolution of 224 × 224. Additionally, ↑ (↓) indicates that a higher (lower) value is better. The best-performing results are highlighted in bold.

Methods	Dice (%)			
	Average	RV	Myo	LV
R50+UNet[11]	87.55	87.10	80.63	94.42
R50+AttnUNet[11]	86.75	87.58	79.20	93.47
ViT+CUP[11]	81.45	81.46	70.71	92.18
R50+ViT+CUP[11]	87.57	86.07	81.88	94.75
TransUNet[11]	89.71	86.67	87.27	95.18
SwinUNet[12]	88.07	85.77	84.42	94.03
MT-UNet[33]	90.43	86.64	89.04	95.62
MISSFormer[34]	90.86	89.55	88.04	94.99
PVT-CASCADE[32]	91.46	89.97	88.90	95.50
TransCASCADE[32]	91.63	90.25	89.14	95.50
MGRAD-UNet (Ours)	**92.13**	**90.42**	**90.00**	**95.98**

Table 2. Qualitative results of the ACDC dataset. We report Dice scores for the left ventricle (LV), right ventricle (RV), myocardium (Myo), and the average Dice score. ↑ (↓) indicates that a higher (lower) value is better. The best performance is highlighted in bold.

decoder (GRAD). The input data $X \in \mathbb{R}^{3\times224\times224}$. Initially, we employ convolutional blocks comprised of 3 × 3 convolutions, BatchNorm function, and ReLU activation to extract features from each layer, denoted as E_i, $i \in \{1, 2, 3, 4, 5\}$. Next, we adjust the number of channels to 64 through the RFB module to decrease subsequent parameter count. Subsequently, these features from different layers are input into the MSD, resulting in five complementary enhanced differential features M_i, $i \in \{1, 2, 3, 4, 5\}$ being output. Then M_i are input into the GRAD to reinforce the edge features, and obtain the output feature G_i. To better utilize the differences between feature maps generated by two differential decoders, we will perform a differential operation on M_i and G_i, and absorb these differences recursively. In the current iteration, the differential information obtained from the previous iteration will be adopted as auxiliary information and integrated into the encoder.

Multi-scale differential decoder

The architecture of the Multi-Scale Differential Decoder is illustrated in Fig. 6. It involves five input features R_i, five output features M_i, addition operations, the difference process as DP. The DP composed of ConvBlocks and subtractions between adjacent layers. The process can be expressed as follows:

$$DP = Conv\left(\left|Conv(Up(R_{i+1})) \ominus Conv(R_i)\right|\right) \tag{1}$$

where R_i and R_{i+1} represent feature maps from adjacent levels, $Conv(\cdot)$ denotes the ConvBlock with a convolutional layer with a kernel size of 3 and padding of 1, a batch normalization layer and ReLU. $Up(\cdot)$ denotes the

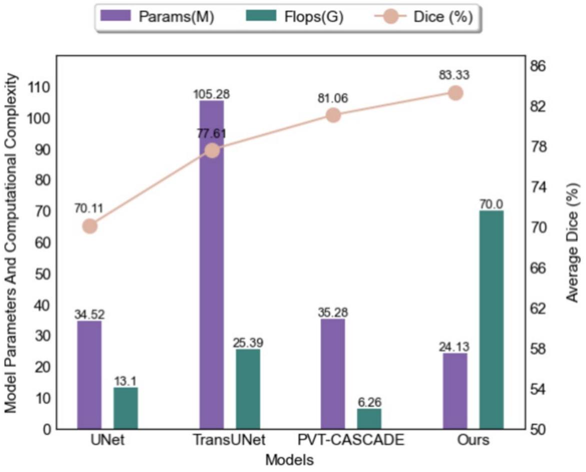

Fig. 3. This figure shows Average Dice, the number of parameters, and computational complexities of different models.

Compenents		Average			
RFB	**SideConv**	**Dice↑**	**mIoU↑**	**HD95↓**	**ASD↓**
✓	✓	**83.33**	**74.68**	**16.67**	**3.26**
✓	×	83.00	74.57	18.80	3.49
×	✓	82.61	73.55	19.54	3.44
×	×	82.40	73.45	18.58	3.51

Table 3. Ablation studies were conducted on the Synapse multi-organ segmentation dataset. The modules“RFB” and “SideConv” represent different modules used in the study. Significant values are in (bold).

Compenents		Dice (%)			
RFB	**SideConv**	**Average↑**	**RV↑**	**Myo↑**	**LV↑**
✓	✓	**92.13**	**90.42**	**90.00**	**95.98**
✓	×	91.70	90.09	89.62	95.39
×	✓	91.62	90.05	89.74	94.87
×	×	91.12	89.83	89.45	93.88

Table 4. Ablation studies were conducted on the ACDC for automated cardiac segmentation dataset. The modules “RFB” and “SideConv” represent different modules used in the study. Significant values are in (bold).

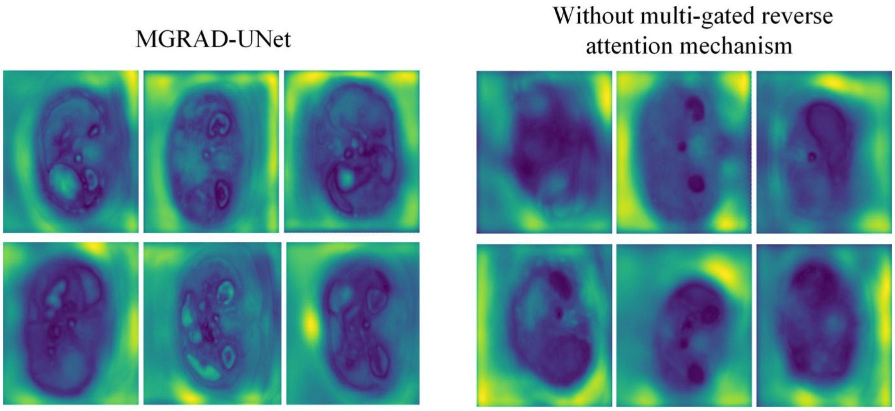

Fig. 4. This figure shows feature heatmaps of MGRAD-UNet with and without multi-gated reverse attention.

Loss	Dice	HD95	mIoU
Weighted loss	83.14	23.16	74.68
Aggregated loss (Ours)	**83.33**	**16.67**	**74.73**

Table 5. Performance comparisions of different loss functions on Synapse multi-organ segmentation dataset. Significant values are in (bold).

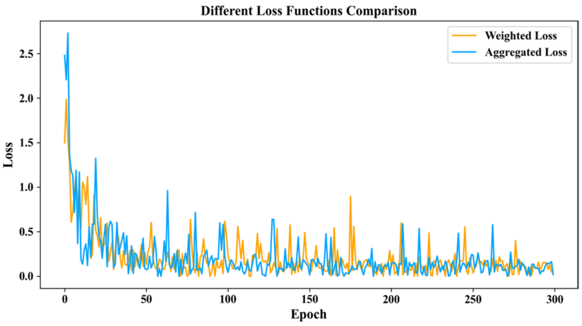

Fig. 5. Trends in aggregated loss and weighted loss during training on the Synapse dataset.

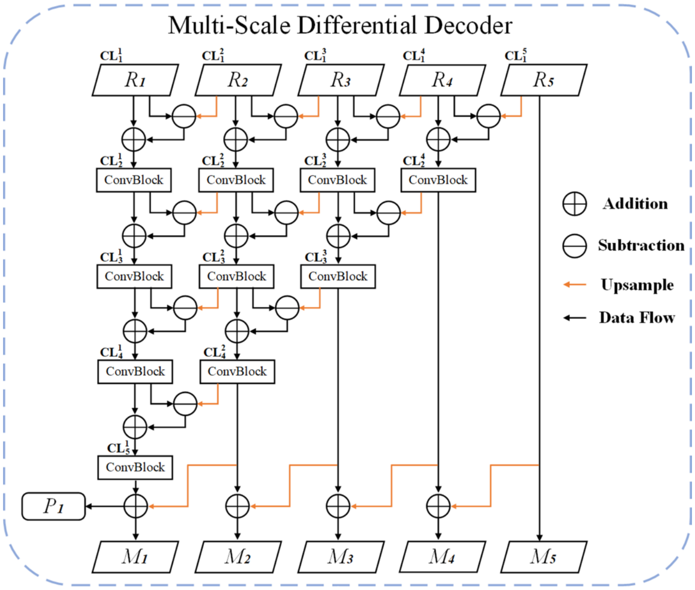

Fig. 6. Details of the introduced MSD. It designed with a pyramid structure, extracts features from each level and performs differential processes between layers, resulting in 5 complementary enhanced features and one prediction result.

upsample function with a scale factor of 2. $|\cdot|$ denotes the absolute value computation. The *DP* can obtain and highlight differential features between adjacent levels, thus supplying more affluent information for subsequent decoding.

To enhance the complementarity of features at each level, establishing differential relationships at both pixel and structural levels. We employ *DP* to obtain differential features between adjacent levels and across layers, meanwhile we connect differential features in both horizontal and vertical directions, calculating differential features of different levels. Then aggregated the features CL_n^i of the corresponding layer and differential features $CL_{n\neq 1}^i$ of any other layers, to generate complementary enhanced features as output features M_i. The entire process can be formulated as:

$$M_i = Conv(\sum_{n=1}^{6-i} CL_n^i) \quad i \in (1, 2, 3, 4, 5) \tag{2}$$

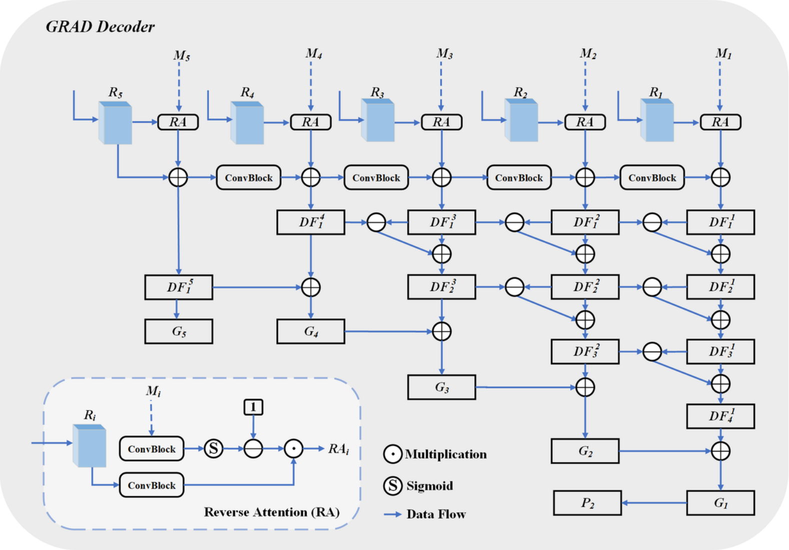

Fig. 7. Details of proposed GRAD. It receives complementary enhanced features from MSD and performs RA operations with features from each level, followed by differential processes between layers, resulting in 5 output features and one prediction result.

To ensure that the features passed into GRAD are robust and accurate, we design a supervision block called SideConv (SConv):

$$SConv = Conv_{1\times1}(Up^{n=i-1}(M_i)) \tag{3}$$

where n represents the number of upsampling times.

Gate-control reverse attention differential decoder

As illustrated in Fig. 7. In order to better utilize multi-scale contextual information to enhance differential features, we design a gate-control reverse attention method. The gate-control mechanism adjusts the propagation path based on the gate signal, thereby reinforcing the edge features. The reverse attention mechanism as RA_i, strengthens the focus on edge features by adjusting the weights of R_i, thereby improving the accuracy of segmentation. RA_i can be formulated as follows:

$$RA_i = (1 - \sigma(M_i)) \cdot R_i \quad i \in (1, 2, 3, 4, 5) \tag{4}$$

where $\sigma(\cdot)$ represents the Sigmoid function. By Equation 4, we can focus more on key areas in the image, enhanced the utilization of contextual information.

Subsequently, we further enhance the complementarity between different layers. The weights of the features from RFB are adjusted and weighted with the features from each layer from the RA_i :

$$DF_1^i = \sum^{6-i} RA_i \quad i \in (1, 2, 3, 4, 5) \tag{5}$$

This process obtains strengthened features focusing on key regions and edge. Following that, the features were passed through DP operation, resulting in the scale-specific feature (DF_1^i) and the cross-scale differential features ($DF_{n\neq1}^i$) between corresponding level and every other level. Aggregate them to generate complementary enhanced features as out features (G_i) of the GRAD. This process can be formulated as follows:

$$G_i = Conv(\sum_{n=1}^{6-i} DF_n^i) \quad i \in (1, 2, 3, 4, 5) \tag{6}$$

Gate-control feedback differential encoder

To enhance the robustness of feature extraction for low-level features, such as texture, color, and edges in traditional UNet. We propose a gate-control feedback differential encoder. Specifically, we obtain differential features by differencing the decoding features of two decoders at the same resolution. Then we feed these differential features as control signals into the next round encoder.

As shown in Fig. 1. When no signal is received from the control port, we perform normal encoding following the black arrows. When the control signal is received, we perform differential encoding following the green arrows. The differential encoding can be formulated as follows:

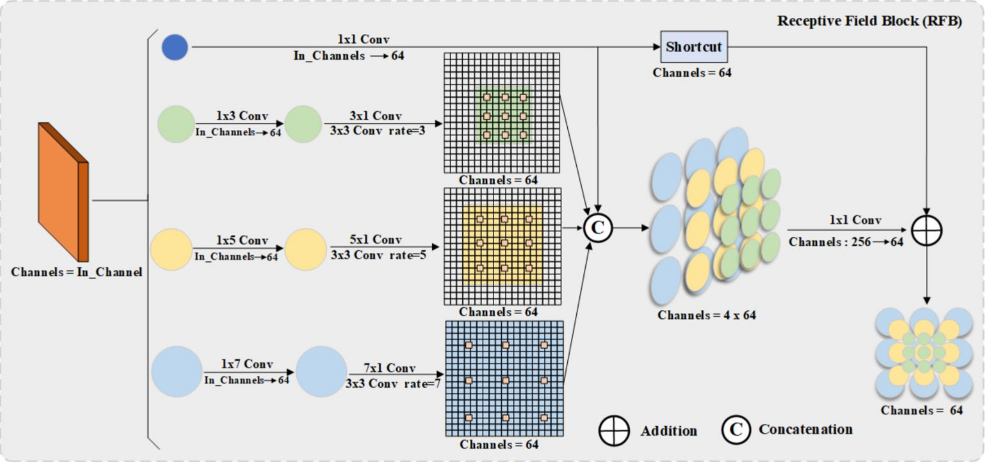

Fig. 8. Details of RFB module. The RFB module processes the encoder features through four branches, concatenates the results, and reduces the dimensionality with a 1×1 convolution. Additionally, a shortcut connection adds feature maps to the output, preserving spatial information.

$$E_i^{'} = E_i + \lambda \cdot (G_i - M_i) \tag{7}$$

where E_i represents the normal encoding features, $E_i^{'}$ represents the differential encoding features, λ represents the initial value for the coefficient weight, set to $1e-3$.

In order to enhance the model's capability of capturing multi-scale information. We introduce the receptive field block (RFB)[38]. As illustrated in Fig. 8. RFB offers an efficient means to expand the receptive field and adjust the number of channels to 64 to improve computational efficiency. This process can be formulated as follows:

$$R_i = RFB(E_i \; or \; E_i^{'}) \tag{8}$$

where R_i represents the feature generated from each E_i or $E_i^{'}$ after passing through the RFB module.

Loss function

In our proposed model, we utilize two loss strategies, specifically aggregate loss and weighted loss, alongside a defined loss function. The expression of the loss function is as follows:

$$\mathcal{L}oss = \lambda_1 \cdot \mathcal{L}_{dice} + \lambda_2 \cdot \mathcal{L}_{ce} \tag{9}$$

where $\mathcal{L}_{dice}$ represents dice coefficient loss[39], $\mathcal{L}_{ce}$ represents cross-entropy loss[3]. λ_1 and λ_2 represent the weight coefficients, with values assigned as 0.3 and 0.7 respectively.

Aggregated loss

In the proposed model, the total training loss can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{agg}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{agg}$ represent supervised loss and aggregated loss. The supervised loss is calculated between result (P_i) from the differential decoders and ground truth (GT), which can be written as:

$$\mathcal{L}_{sup} = \sum_{i=1}^{2} \mathcal{L}oss(P_i, GT) \tag{10}$$

We introduce an effective multi-stage differential feature mixing loss aggregation strategy for image segmentation. Our goal is to create new prediction maps by combining the existing prediction maps. To achieve this, we take all feature maps from different stages of the network as input and aggregate the losses of the new prediction maps generated from non-empty subsets of these prediction maps. We then aggregate these losses with the losses of main maps. Aggregated loss can be written as:

$$\mathcal{L}_{agg} = \sum_{1}^{2^n - 1} \mathcal{L}oss(\tilde{G}_i, G) \tag{11}$$

where $\tilde{G}$ represents $2^n - 1$ non-empty subsets of n prediction maps generated by all predicted mappings from n stages of the network. This aggregation strategy does not require additional parameter calculations, and can be used in conjunction with multi-stage image segmentation. Algorithm 1 presents the steps for generating new prediction maps and aggregating the losses.

```
Input: $G$; the ground truth;
A list $M_i, i \in \{0, 1, 2, 3, 4\}$, where each element is a prediction map
Output: $L_{agg}$; the aggregated loss
$\mathcal{L}_{agg} \leftarrow 0.0$;
$LS \leftarrow$ locate all subsets of the predicted map indices that are not empty;
for $ls \in LS$ do
    $\tilde{G} \leftarrow 0.0$; // $\tilde{G}$ is a newly predicted map.
    for $i \in ls$ do
        $\tilde{G} \leftarrow \tilde{G} + M_i$;
    end for
    $\mathcal{L}_{agg} \leftarrow \mathcal{L}_{sf}(\tilde{G}, G)$; //$\mathcal{L}_{sf}(\cdot)$ is any loss function such as Dice, CrossEntropy;
end for
```

Algorithm 1. Aggregated Loss Algorithm

Weighted loss

Additionally, we also use another efficient weighted loss ($\mathcal{L}_w$), which directly weights the losses of multi-stage differential features and each main graph, as follows:

$$\mathcal{L}_w = \mathcal{L}_{sup} + \mathcal{L}_{aux} \tag{12}$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{aux}$ are the supervised loss and auxiliary loss, respectively. The auxiliary loss $\mathcal{L}_{aux}$ is calculated between the result S_i from the *SConv*, and ground truth *GT*, which can be formulated as:

$$\mathcal{L}_{aux} = \sum_{i=1}^{5} w_i \cdot \mathcal{L}oss(S_i, GT) \tag{13}$$

where w_i represents the weight coefficient.

Data availility

The ACDC dataset originates from https://www.creatis.insalyon.fr/Challenge/acdc/. The Synapse dataset originates from https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

Received: 16 May 2024; Accepted: 26 August 2024
Published online: 31 August 2024

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25** (2012).
2. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* 234–241 (Springer, 2015).
3. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3431–3440 (2015).
4. Amer, A., Lambrou, T. & Ye, X. Mda-unet: A multi-scale dilated attention u-net for medical image segmentation. *Appl. Sci.* **12**(7), 3676 (2022).
5. Gao, S.-H. *et al.* Res2net: A new multi-scale backbone architecture. *IEEE Trans. Patt. Anal. Mach. Intell.* **43**(2), 652–662 (2019).
6. Zhao, X., Zhang, L. & Lu, H. Automatic polyp segmentation via multi-scale subtraction network. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24* 120–130 (Springer, 2021).
7. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imag.* **39**(6), 1856–1867 (2019).
8. Cai, Y. & Wang, Y. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. in *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, vol. 12167, 205–211 (SPIE, 2022).
9. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J. & Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. in *International Conference on Medical Image Computing and Computer-assisted Intervention* 263–273 (Springer, 2020).
10. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y. & Kainz, B. *et al.* Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
11. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L. & Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
12. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. in *European Conference on Computer Vision* 205–218 (Springer, 2022).
13. Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T. & Soler, L. U-net transformer: Self and cross attention for medical image segmentation. in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12* 267–276 (Springer, 2021).

14. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M.: Medical transformer: Gated axial-attention for medical image segmentation. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24* 36–46 (Springer, 2021).
15. Cheng, J. *et al.* Ddu-net: A dual dense u-structure network for medical image segmentation. *Appl. Soft Comput.* **126**, 109297 (2022).
16. Lou, A., Guan, S. & Loew, M. Dc-unet: Rethinking the u-net architecture with dual channel efficient CNN for medical image segmentation. in *Medical Imaging 2021: Image Processing*, vol. 11596, 758–768 (SPIE, 2021).
17. Xie, M., Li, Y., Xue, Y., Huntress, L., Beckerman, W., Rahimi, S. A., Ady, J. W. & Roshan, U. W. Two-stage and dual-decoder convolutional u-net ensembles for reliable vessel and plaque segmentation in carotid ultrasound images. in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* 1376–1381 (IEEE, 2020).
18. Zeng, Q., Xie, Y., Lu, Z., Lu, M. & Xia, Y. Discrepancy matters: Learning from inconsistent decoder features for consistent semi-supervised medical image segmentation. arXiv preprint arXiv:2309.14819 (2023)
19. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Patt. Anal. Mach. Intell.* **40**(4), 834–848 (2017).
20. Yang, M., Yu, K., Zhang, C., Li, Z. & Yang, K. Denseaspp for semantic segmentation in street scenes. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3684–3692 (2018).
21. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. Feature pyramid networks for object detection. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2117–2125 (2017).
22. Wu, Y. *et al.* Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **81**, 102530 (2022).
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
24. Gao, Y., Zhou, M. & Metaxas, D. N. Utnet: A hybrid transformer architecture for medical image segmentation. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24* 61–71 (Springer, 2021).
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
26. Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., Sun, W. & Lu, H. M^{2}snet: Multi-scale in multi-scale subtraction network for medical image segmentation. arXiv preprint arXiv:2303.10894 (2023).
27. Sinha, A. & Dolz, J. Multi-scale self-guided attention for medical image segmentation. *IEEE J. Biomed. Health Inform.* **25**(1), 121–130 (2020).
28. Ahmad, P. *et al.* Mh unet: A multi-scale hierarchical based architecture for medical image segmentation. *IEEE Access* **9**, 148384–148408 (2021).
29. Qin, X. *et al.* U2-net: Going deeper with nested u-structure for salient object detection. *Patt. Recognit.* **106**, 107404 (2020).
30. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., & Heng, P.-A.: R3net: Recurrent residual refinement network for saliency detection. in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* 684–690 (AAAI Press Menlo Park, 2018).
31. Zhao, X., Pang, Y., Zhang, L., Lu, H. & Zhang, L.: Suppress and balance: A simple gated network for salient object detection. in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 35–51 (Springer, 2020).
32. Rahman, M. M. & Marculescu, R. Medical image segmentation via cascaded attention decoding. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 6222–6231 (2023).
33. Jha, A., Kumar, A., Pande, S., Banerjee, B. & Chaudhuri, S. Mt-unet: A novel u-net based multi-task architecture for visual scene understanding. in *2020 IEEE International Conference on Image Processing (ICIP)* 2191–2195 (IEEE, 2020).
34. Huang, X., Deng, Z., Li, D. & Yuan, X. Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
35. Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H. & Shao, L. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
36. Vázquez, D. *et al.* A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* **2017**, 4037190 (2017).
37. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J. & Merhof, D. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 6202–6212 (2023).
38. Liu, S. & Huang, D., et al: Receptive field block net for accurate and fast object detection. in *Proceedings of the European Conference on Computer Vision (ECCV)* 385–400 (2018)
39. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. in *2016 Fourth International Conference on 3D Vision (3DV)* 565–571 (IEEE, 2016).

Author contributions

This research was conducted in collaboration with all authors. Conceptualization, S.Y.; methodology, S.Y.; sofware, S.Y.; Provision of study material: A.C., B.Y., S.Y.; Collection and/or assembly of data: A.C., B.Y., S.Y.; Data analysis and interpretation: SY.; Manuscript writing: A.C., B.Y., S.Y.; Manuscript review: A.C., B.Y., S.Y.; Manuscript revision: A.C., B.Y., S.Y.; All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 6197022977.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Y. or A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.