Heliyon 10 (2024) e35698

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A frequency selection network for medical image segmentation

Shu Tang [a,*], Haiheng Ran [a], Shuli Yang [a], Zhaoxia Wang [b,**], Wei Li [c,***], Haorong Li [a], Zihao Meng [a]

[a] *Chongqing University of Posts and Telecommunications, No.2 Road of Chongwen, Nanan District, 400000, Chongqing,China*
[b] *Chongqing Emergency Medical Center, Chongqing University Central Hospital, School of Medicine, Chongqing University, Chongqing, China*
[c] *Children's Hospital of Chongqing Medical University, China*

ARTICLE INFO

Keywords:
Medical image segmentation
Frequency selection
Global-local aggregation
Feature filter
Deformable convolution

ABSTRACT

Existing medical image segmentation methods may only consider feature extraction and information processing in spatial domain, or lack the design of interaction between frequency information and spatial information, or ignore the semantic gaps between shallow and deep features, and lead to inaccurate segmentation results. Therefore, in this paper, we propose a novel frequency selection segmentation network (FSSN), which achieves more accurate lesion segmentation by fusing local spatial features and global frequency information, better design of feature interactions, and suppressing low correlation frequency components for mitigating semantic gaps. Firstly, we propose a global-local feature aggregation module (GLAM) to simultaneously capture multi-scale local features in the spatial domain and exploits global frequency information in the frequency domain, and achieves complementary fusion of local details features and global frequency information. Secondly, we propose a feature filter module (FFM) to mitigate semantic gaps when we conduct cross-level features fusion, and makes FSSN discriminatively determine which frequency information should be preserved for accurate lesion segmentation. Finally, in order to make better use of local information, especially the boundary of lesion region, we employ deformable convolution (DC) to extract pertinent features in the local range, and makes our FSSN can focus on relevant image contents better. Extensive experiments on two public benchmark datasets show that compared with representative medical image segmentation methods, our FSSN can obtain more accurate lesion segmentation results in terms of both objective evaluation indicators and subjective visual effects with fewer parameters and lower computational complexity.

1. Introduction

Deep neural network (DNN) can learn complex representation relationship and realize pixel-level classification of the input image with its powerful learning ability. Therefore, medical image segmentation methods based on DNN have become a research hotspot in the field of medical image processing. Among these methods, convolutional neural network (CNN) and Transformer have shown excellent performance in various medical image segmentation tasks, especially for the segmentation of lesion areas. CNN-based methods [1–8] sually employ an encoder-decoder structure to perform layer-wise encoding and features restoration, which

[*] Corresponding author.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: tangshu@cqupt.edu.cn (S. Tang), zhaoxia_wang103@163.com (Z. Wang), liwei0111@cqu.edu.cn (W. Li).

https://doi.org/10.1016/j.heliyon.2024.e35698
Received 7 May 2024; Received in revised form 18 July 2024; Accepted 1 August 2024
Available online 5 August 2024

facilitates high quality segmentation results. Recently, Transformer-based medical image segmentation methods have become a research hotspot, and have outperformed most CNN-based methods with excellent performance. Transformer-based methods usually take vision Transformers [9–14] as a backbone (encoder) because of their abilities of capturing long-range dependencies, and mainly design the decoder [15–21]. However, most of existing CNN-based and Transformer-based methods either only consider feature extraction and information processing in the spatial domain or ignore the influence of semantic gaps between shallow features and deep features and directly conduct concatenation and fusion between features of different levels, resulting in poor discriminations between lesion features and background features, and thus make it hard to segment accurate lesion contours and hinder model performance. Therefore, there are still much rooms for improvement in the performance of existing medical image segmentation methods.

According to above analysis, we propose a novel frequency selection segmentation network (FSSN) which performs the fusion of spatial domain local features and frequency domain global information, as well as the adaptive sifting of different frequency components in the frequency domain. Specifically, we first propose a global-local aggregation module (GLAM) to realizes complementary fusion of local detailed features and global semantic information. Secondly, we propose a feature filter module (FFM) to bridge semantic gaps between shallow features and deep features so that our FSSN can discriminatively preserve the important features for accurate lesion segmentation. Finally, deformable convolution is introduced to accurately extract associated features in the local range so that our FSSN can focus on relevant image content better, and further enhances the capabilities of feature capture and expression. Our FSSN can be put forward by end-to-end training, and achieves more accurate lesion area segmentation with fewer parameters and lower computational complexity. In summary, the main components our method are listed as follows:

(1) Inspired by the global characteristic of Fourier theory [22], we propose a GLAM. GLAM consists of a spatial domain branch and a frequency domain branch. Spatial domain branch captures local detailed features through a set of convolution layers with different receptive fields, while frequency domain branch realizes the exploitation of all frequency components by using Fourier transform. And by fusing spatial domain branch and frequency domain branch, the complementary learning of local detail features in spatial domain and global semantic information in frequency domain is realized.
(2) In medical image segmentation tasks, semantic gaps between features of different levels often bring interferences and affect segmentation results. Therefore, we propose an FFM to bridge semantic gaps between shallow features and deep features by discriminatively treating the frequency components of fusion features. And FFM makes our FSSN can discriminatively determine which frequency information should be protected for accurate lesion segmentation, and greatly reduces the adverse impact caused by semantic gaps when we conduct cross-level features fusion.
(3) In order to mitigate the influence caused by irrelevant information, we employ deformable convolution (DC) to accurately extract pertinent features so that our FSSN can focus on pertinent image contents better for more powerful features capture and network representation.

2. Related work

Among the existing methods, Transformer-based methods usually perform better, but the complexity of the model is usually higher than that of CNNs. Transformer-based methods can capture long-range dependencies, while CNNs are good at extracting local features. Table 1 shows their differences.

2.1. CNN-based methods

In 2014, Long et al. proposed a fully convolutional network (FCN) [23] and pioneered CNN-based image segmentation. Subsequently, Ronneberger et al. proposed a U-shaped fully convolutional neural network U-Net [1] based on FCN for medical image segmentation. Zhou et al. [2], Huang et al. [3] and Feng et al. [4], introduced the idea of multi-scale feature fusion into their works. Gu et al. [5] and Lou et al. [24] won a larger receptive field by using dilated convolution. Oktay et al. [6], Cai et al. [7] and Fan et al. [8] introduced attention mechanism into medical image segmentation, so that the network can filter out redundant information and focus more attention on the foreground region of interest. In Jin et al.'s work, deformable convolution [25,26] was used to learn the desired region. Furthermore, Fan et al. have shown that using a backbone network [27–30] often results in higher performance than training a network from scratch.

2.2. Transformer-based methods

In recent years, Transformer has achieved significant success in medical image segmentation tasks due to its ability of capturing

Table 1
Comparison of different types of methods.

	Local Feature	Long-range	Low-complexity	Performance
CNN-based	✓		✓	
Transformer-based		✓		✓
Ours	✓	✓	✓	✓

long-range dependencies and global information extraction (i.e., multi-scale self-attention mechanism [31]: MSA). Huang et al. [15] 's work explored global dependencies and local context by re-integrating local and global information. Zhang et al. [16] adopted a hybrid architecture and cooperatively encoded feature information by using ViT [9] branch and CNN branch. Wang et al. [32] achieved significant performance gains through multi-scale channel-wise information fusion. Wang et al. [17] 's work deeply mined local information by upsampling the outputs of encoder from each layer to same resolution and conducting cascading fusion. Yue et al. [21] directly fused deep features and shallow features to guide shallow features, and aggregated feature information of all layers to achieve better segmentation. Rahman et al. [18] used attention mechanism [6,33–35] to refine deep features and shallow features. Duc et al. [19] designed a residual axial reverse attention (RA-RA) to analyze localization information and multi-scale features. Zhou et al. [36] simultaneously used convolution kernels of different sizes to mine similar information in multi-scale features and conducted the fusion of features of different scales.

2.3. *Frequency-based methods*

The work by Liu et al. [37] performed a sifting of frequency information in the frequency domain and used filtered frequency information to guide the decoding process. Nam et al. [38] simultaneously encoded the high frequency, low frequency and original features. Chi et al. [39] proposed a fast frequency domain convolution to obtain global receptive fields by transforming features into frequency domain and processing them in global and local ranges respectively. Li et al. [40] proposed a global-frequency-domain network (GFUNet) to correctly identify small areas by differently treating frequency components and restraining irrelevant features. Wang et al. [41] proposed g fast Fourier convolutional ResNet (FFC-ResNet) by combining spatial and frequency domain learning. Zou et al. [42] proposed a Fourier channel attention, which improves segmentation performance by encoding frequency-domain information of different channels. Wang et al. [43] proposed an effective foreground masking pre-training strategy to generate better segmentation representations by modeling high and low frequency components of medical images separately. Han et al. [44] introduced Fourier transform to enrich the information of RGB images and achieved good segmentation performance. Li et al. [45] analyzed the frequency domain features in terms of amplitude and phase, and fused shallow and deep information based on the graph structure. Yang et al. [46] introduced deformable convolution to extract localized features, but they ignored the semantic gaps between cross-level features.

In summary, the existing DNN-based medical image segmentation methods, either only consider feature extraction and information processing in the spatial domain [1,2,4–8,15–20,24,32,36,47–50], or ignore semantic gaps between shallow features and deep features, and directly fuse features of different levels [1,3,7,15–17,37–40,42–47,49], resulting in poor discriminations between lesion features and background features. Therefore, there are still much rooms for improvement in the performance of existing medical image segmentation methods.

3. Methods

3.1. *Overall architecture*

As Fig. 1 illustrated, our FSSN consists of four parts: features encoding (encoder: PVT-v2 [14]), fusion of local features and global features (GLAM), discriminative integration of shallow features and deep features (FFM and convolutional block attention module: CBAM [51]), and pertinent features extraction in the local range (DC). We use a four-stage network named PVT-v2-B2 [14] as our encoder. And the outputted features of four stages are denoted as F_{s1}, F_{s2}, F_{s3} and F_{s4} respectively. Where F_{s2}, F_{s3} and F_{s4} are the inputs of $GLAM_1$, $GLAM_2$ and $GLAM_3$ respectively. As we discussed above, the main components of our FSSN are: GLAM, FFM and DC. Therefore, in this section, we will discuss these modules in detail.

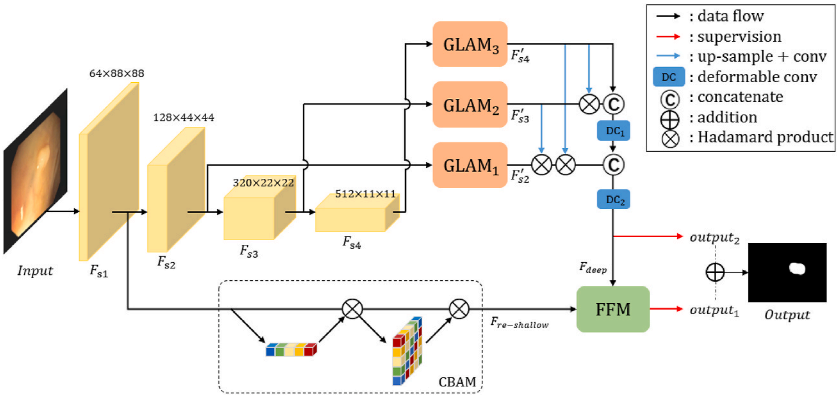

Fig. 1. Our proposed frequency selection segmentation network (FSSN).

3.2. Global-local aggregation module (GLAM)

Han et al. [52] have proved that a well-trained deep neural network usually contains rich or even redundant feature maps to ensure a comprehensive understanding of the input data. Therefore, the features with partial channels may have similar expressive ability as full features. Inspired by the work of Han et al. as shown in Fig. 2, our proposed GLAM consists of a local spatial branch (LSB) and a global frequency branch (GFB) for local details features capture and global frequency information exploitation respectively. Denoting F_{si} as the input of $GLAM_{i-1}$, where $i \in [2,4]$. We averagely split F_{si} into two parts along channel dimension: $F_{si\text{-}LSB}$ (first n/2 channels) and $F_{si\text{-}GFB}$ (last n/2 channels) as the inputs of LSB and GFB.

For the LSB, Fan et al. [53] have proved that a set of various sized receptive fields help to highlight the area of small/local space, which helps to incorporate more discriminative feature representations during the searching stage. Ding et al. [54] have proved that by decomposing a $k \times k$ standard convolution into a $k \times 1$ convolution and a $1 \times k$ convolution in parallel rather than serial can enrich the feature space during training. And the depthwise separable convolution (DS-conv) proposed by Howard et al. [55] can reduce the number of parameters and computational complexity. Based on these designs, our LSB includes three parallel branches with three different receptive fields respectively. From left to right: a 7×1 DS-conv, a 3×3 DS-conv, and a 1×7 DS-conv. Then, the outputs of three parallel branches are pixel-wise added and further refined by a 1×1 convolution, so that FSSN can aggregate features at different scales. Finally, two identity shortcut branches are multiplied and added in, respectively. LSB are similar to the spatial attention (SA) [34], it can be formulated as:

$$F^{ds}_{m\times n} = Conv(ds_{m\times n}(F_{si-LSB})), \tag{1}$$

$$attn = \sigma\left(Conv_{1\times 1}\left(F^{ds}_{7\times 1} + F^{ds}_{3\times 3} + F^{ds}_{1\times 7}\right)\right), \tag{2}$$

$$F^{si}_{LSB} = F_{si-LSB} \odot attn + F_{si-LSB}, \tag{3}$$

where $ds_{m\times n}(\bullet)$ denotes the DS-conv with a receptive field of $m \times n$. $Conv_{m\times n}(\bullet)$ denotes a $m \times n$ convolutional layer. $\odot$ is Hadamard product and $\sigma(\bullet)$ is Sigmoid function.

For the GFB, according to Fourier theory [22], processing information in Fourier space is capable of capturing global frequency representation in the frequency domain. And by directly processing information in the frequency domain, different image components can be exploited better, such as low-frequency contours and high-frequency texture details. Therefore, in our GFB, we first use a 1×1

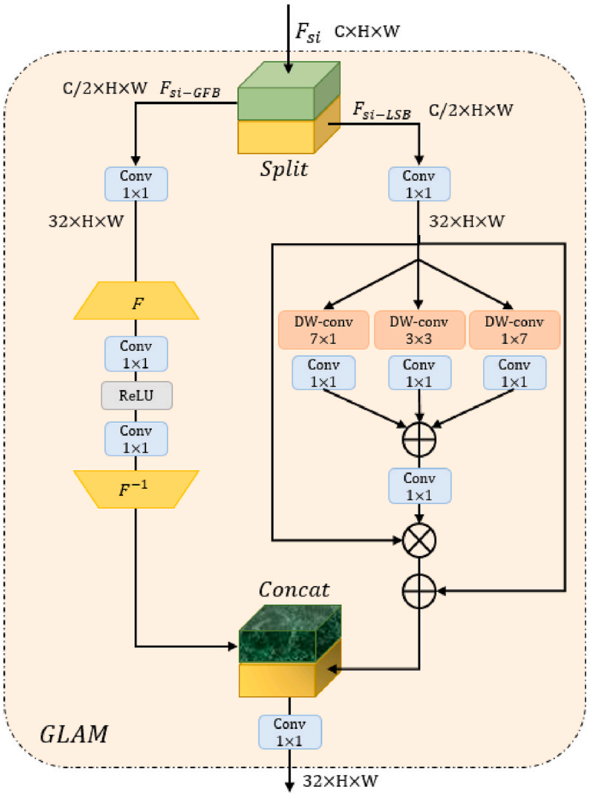

Fig. 2. The proposed global-local aggregation module (GLAM).

convolution to process $\text{F}_{\text{si-GFB}}$, and then adopt Fourier transform to convert it to the Fourier space as $F_{si\text{-}Fourier}$ by Equation (4). To process frequency-domain representation $\text{F}_{\text{si-Fourier}}$, we adopt two 1 × 1 convolutions and a ReLU [56] activation function to exploit global frequency information and finally invert it back to the spatial domain. GFB can be formulated as Equation (6):

$$F_{si-Fourier} = \mathscr{F}(Conv_{1\times1}(F_{si-GFB})), \tag{4}$$

$$MLP(x_{input}) = Conv_{1\times1}(ReLU(Conv_{1\times1}(x_{input}))), \tag{5}$$

$$F_{GFB}^{si} = \mathscr{F}^{-1}(MLP(F_{si-Fourier})), \tag{6}$$

where, $\mathscr{F}$ (•) and $\mathscr{F}^{-1}$ (•) denote 2D discrete Fourier transform and 2D inverse discrete Fourier transform, respectively. $ReLU$(•) denotes ReLU activation function, and x_{input} denotes the input of MLP(•).

Finally, we fuse the outputs of LSB and GFB by concatenation operation and a 1 × 1 convolution, which can be formulated as:

$$F_{si}' = GLAM(F_{si}) = Conv_{1\times1}(©(F_{GFB}^{si}, F_{LSB}^{si})), \tag{7}$$

where, ©(•) denotes concatenate operation. From Equations (1)–(7) we can see that, by capturing multi-scale local features in the spatial domain and exploiting global frequency information in the frequency domain simultaneously, our proposed GLAM achieves the complementary fusion of local details features and global frequency information well. Therefore, in our FSSN, the representation capability of F_{s2}, F_{s3} and F_{s4} are further improved by using $GLAM_1$, $GLAM_2$ and $GLAM_3$ respectively. The effectiveness of our GLAM will be verified in Sec. 4.

3.3. Feature filter module (FFM)

Aggregating multi-level features is essential for precise segmentation. However, the improvement by directly fusing shallow features and deep features becomes limited as the semantic gaps between them increases [57]. Therefore, we propose a FFM to mitigate semantic gaps when we conduct cross-level features fusion. As shown in Fig. 3, our FFM contains two parts: frequency selection and multi-scale features refinement. In FFM, the frequency selection part is mainly responsible for the mitigation of semantic gaps. And it's different from GLAM, FFM aggregates global and local information serially because we wish them to interact with each other, whereas GLAM needs to independently mine useful global and local information before fusing them together.

Before FFM, we first use CBAM [51] to rescale the shallow feature F_{s1}. CBAM consists of a channel attention (CA) branch and a spatial attention (SA) branch and it can be formulated as Equation (8):

$$F_{re-shallow} = CBAM(F_{s1}) = SA(CA(F_{s1})), \tag{8}$$

$$\begin{aligned} CA(F_{s1}) = F_{s1} \odot \sigma(\quad & MLP(P_m(F_{s1})) + MLP(P_a(F_{s1}))), \\ & F_c = CA(F_{s1}), \end{aligned} \tag{9}$$

$$SA(F_c) = F_c \odot \sigma(Conv_{1\times1}(C(P_m(F_c), P_a(F_c)))), \tag{10}$$

where CA(•) and SA(•) denote the CA operation and SA operation, respectively. And P_m(•) and P_a(•) represent maximum pooling and average pooling operations, respectively. And F_c is denoted as the output features of CA branch.

As shown in Fig. 1, we denote the outputs of DC_2 and CBAM as F_{deep} and $F_{re\text{-}shallow}$ respectively. In FFM, we first concatenate F_{deep} and $F_{re\text{-}shallow}$ and coarsely fuse them as F_{fuse} by Equation (11). Similar to GLAM, we convert F_{fuse} to the Fourier space for global receptive field, then FSSN can adaptively sift significant information in the global range, and invert sifted results back to the spatial domain. Finally, an identity shortcut branch is added to the result of inverse Fourier transform. The frequency selection part can be formulated as:

$$F_{fuse} = Conv_{1\times1}(©(F_{deep}, F_{re-shallow})), \tag{11}$$

$$F_f = \mathscr{F}^{-1}(W(\mathscr{F}(F_{fuse}))) + F_{fuse}, \tag{12}$$

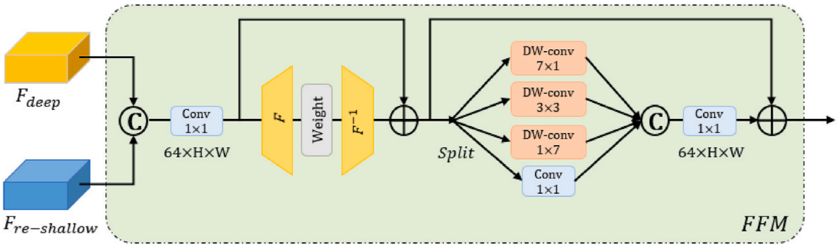

Fig. 3. The proposed feature filter module (FFM).

where, $W(\bullet)$ is a learnable weight map with same resolution as F_{fuse}.

As we discussed above (Sec. 3.2), the combination of features with various sized receptive fields helps to incorporate more discriminative feature representations, thereby highlighting regions of small/local space. Therefore, in our FFM, we again adopt four parallel convolution layers with different sized receptive fields to refine F_f. Then, the outputs of four parallel branches are concatenated along channel dimension and further refined by a 1 × 1 convolution. Finally, an identity shortcut branch is added in, which can be formulated as:

$$F_{cat} = ©\left(ds_{7\times1}\left(F_f\right), ds_{1\times7}\left(F_f\right), Conv_{1\times1}\left(F_f\right)\right), \tag{13}$$

$$FFM\left(F_{deep}, F_{re-shallow}\right) = Conv_{1\times1}\left(F_{cat}\right) + F_f, \tag{14}$$

from Equations (11)–(14) we can see that, by sifting frequency components of F_{fuse} in the frequency domain and mining multi-scale local features in the spatial domain, our proposed FFM can effectively mitigate semantic gaps between cross-level features and facilitate more accurate lesion segmentation. The effectiveness of our FFM will be verified in Sec. 4.

3.4. Deformable convolution (DC)

It is well-known that, in medical images, the distinction between lesion and background is small. Therefore, accurate lesion features extraction is crucial, especially at the boundary of lesion area. In existing medical image segmentation methods, almost all works just used standard convolution with different sizes kernels (e.g. 3 × 3 or 5 × 5) to extract local information. However, because of fixed receptive field and sampling locations, the learned features by standard convolution will be influenced by irrelevant information when we conduct standard convolution operation at the boundary of lesion area. Therefore, as shown in Fig. 1, in order to mitigate the influence caused by irrelevant features, we employ two deformable convolutions (DC_1 and DC_2) [25,26] to extract pertinent features in the local range. We use $x(p)$ and $y(p)$ to denote the feature at location p in the input feature maps x and output feature maps y, respectively [26], then the deformable convolution can be formulated as:

$$y(p) = \sum_{k=1}^{K} \omega_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{15}$$

where, K and k denote the number of sampling locations and the k-th sampling location (e.g. K = 9 and $p_k \in$ {(-1, −1), (−1, 0), …, (1, 1)}), respectively. ω_k, p_k and Δp_k denote the weight, original location and learnable offset for k-th sampling location, respectively [26]. Δm_k is a learnable modulation scalar for each sampling location. And whole process of two deformable convolutions can be formulated as:

$$F''_{s3} = F'_{s3} \odot Conv_{3\times3}\left(up_2\left(F'_{s4}\right)\right), \tag{16}$$

$$F''_{s2} = F'_{s2} \odot Conv_{3\times3}\left(up_2\left(F'_{s3}\right)\right) \odot Conv_{3\times3}\left(up_4\left(F'_{s4}\right)\right), \tag{17}$$

where, $up_2(\bullet)$ and $up_4(\bullet)$ denote bilinear interpolation upsampling operations by the factors of 2 and 4, respectively. And $\odot$ is Hadamard product.

$$F_{deep} = DC_2\left(©\left(F''_{s2}, up_2\left(DC_1\left(©\left(F''_{s3}, up_2\left(F'_{s4}\right)\right)\right)\right)\right)\right), \tag{18}$$

where DC_1 and DC_2 denote two 3 × 3 deformable convolution operations. The effectiveness of our two deformable convolutions will be investigated in Sec.4. The pseudocode of our FSSN is shown in **Algorithm 1**.

Algorithm 1. FSSN Architecture

```
function FSSN($Input$)
$F_{s1}, F_{s2}, F_{s3}, F_{s4} \leftarrow PVTv2(Input)$
$output_1, output_2 \leftarrow Decoder(F_{s1}, F_{s2}, F_{s3}, F_{s4})$
$Output \leftarrow output_1, output_2$
return $Output$
end function
function DECODER($F_{s1}, F_{s2}, F_{s3}, F_{s4}$ )
$F'_{s2} \leftarrow GLAM_1(F_{s2})$
$F'_{s3} \leftarrow GLAM_2(F_{s3})$
$F'_{s4} \leftarrow GLAM_3(F_{s4})$
$F_{re-shallow} \leftarrow CBAM(F_{s1})$

$F''_{s3} \leftarrow F'_{s3} \odot Convolution(up_2(F'_{s4}))$
$F''_{s2} \leftarrow F'_{s2} \odot Convolution(up_2(F'_{s3})) \odot Convolution(up_4(F'_{s4}))$
$F_{deep} \leftarrow DeformConv(©(F''_{s2}, up_2(DeformConv(©(F''_{s3}, up_2(F'_{s4}))))))$
```

(continued on next page)

(*continued*)

```
$output_1 \leftarrow Convolution(F_{deep})$
$output_2 \leftarrow FFM(F_{deep}, F_{re-shallow})$
return $output_1, output_2$
end function
function GLAM $(F_{si})$
$F_{si-GFB}, F_{si-LSB} \leftarrow Split(F_{si})$
$F'_{si-GFB} \leftarrow GlobalBranch(F_{si-GFB})$
$F'_{si-LSB} \leftarrow LocalBranch(F_{si-LSB})$
$F'_{si} \leftarrow Convolution(©(F'_{si-GFB}, F'_{si-LSB}))$
return $F'_{si}$
end function
function FFM $(F_{deep}, F_{re-shallow})$
$F_{fuse} \leftarrow Convolution(©(F_{deep}, F_{re-shallow}))$
$F_f \leftarrow GlobalFilter(F_{fuse}) + F_{fuse}$
$F_{cat} \leftarrow MultiScale(F_f)$
$output_2 \leftarrow Convolution(F_{f+}Convolution(F_{cat}))$
return $output_2$
end function
```

4. Experiments

To demonstrate the effectiveness of our method, extensive experiments are performed on a PC with a NVIDIA Geforce RTX 3090 GPU, a Intel Core I9–10900K CPU, and PyTorch platform 1.11.0 Library. We train our models on two datasets: Polyp datasets and ISIC-2018 dataset.

Polyp datasets contain two seeable datasets named CVC-ClinicDB [58], Kvasir and three unseen datasets, namely CVC-300, CVC-ColonDB [59] and ETIS-LaribDB [60]. Specifically, CVC-ClinicDB contained 612 polyp images and corresponding labels extracted from 29 different endoscopic video clips. Kvasir contains 1000 polyp images and corresponding labels from Kvasir-SEG [61]. CVC-300 is a subset of EndoScene [62], which refers to the remaining part after subtracting 612 samples of EndoScene from CVC-ClinicDB. CVC-ColonDB and ETIS-LaribDB contain 380 and 192 polyp images and corresponding labels, respectively. ISIC-2018 dataset contains 2594 training images and 100 test images and their corresponding labels.

4.1. *Implement details*

Following [18], for Polyp datasets, we use 1450 images for training and 798 images for testing. The epoch is set to 50 with batch size of 16. For ISIC-2018 dataset, we use 2594 images for training and 100 images for testing, and the epoch is set to 50 with batch size of 16. No validation set for all experiments. The random flip, random rotation and a scaling strategy $\{0.75, 1.0, 1.25\}$ [8] are used.

We train models end-to-end by using AdamW [63] optimizer with an initial learning rate 1e-4 gradually reduced to 1e-6, and we kept the model that performs best. For fair comparison, we standardized the size of images in each dataset. For Polyp datasets, the size of training image is 352 × 352 and that of ISIC-2018 dataset is 224 × 224. In our experiments, we use dice coefficient (Dice), mean intersection over union (mIoU), Mean Absolute Error (MAE) to evaluate the performance of our method.

We use combined weighted IoU and weighted binary cross entropy (BCE) as our loss function for all experiments, which can be formulated as Equations (19) and (20):

$$\mathscr{L}_{total} = \mathscr{L}^{w}_{IoU} + \mathscr{L}^{w}_{BCE}, \tag{19}$$

$$\mathscr{L}oss = \mathscr{L}_{total}(output_1) + \mathscr{L}_{total}(output_2), \tag{20}$$

where $\mathscr{L}^{w}_{IoU}$ represents the weighted IoU loss [8], and $\mathscr{L}^{w}_{BCE}$ represents the weighted BCE loss [8]. $output_1$ and $output_2$ represent two outputs of our FSSN respectively, and the final segmentation mask of our FSSN is the sum of them.

4.2. *Ablation study*

To demonstrate the effectiveness of each component of our FSSN, we first perform ablation experiments on Polyp datasets. In all ablation experiments, the scores we obtained were averaged over multiple experiments.

4.2.1. *Effectiveness of GLAM*

As we discussed in Section 3.2, in our GLAM, LSB employs three parallel depthwise separable convolutions with different receptive fields to capture local details features, and GFB adopts Fourier transform to achieve global frequency information exploitation. Therefore, to fairly verify the effectiveness of our proposed GLAM, we compare our proposed FSSN with models FSSN-NoFT, FSSN-

Sdsconv, FSSN-NoGLAM and FSSN-Full. Here, FSSN-NoFT refers to the model that only removes Fourier transform and inverse Fourier transform from GFB and keeps the others untouched. FSSN-Sdsconv refers to the model that, in the LSB, replaces three parallel DS-convs with three serial DS-convs of the same receptive field sizes, and the rest parts of FSSN-Sdsconv are untouched. FSSN-NoGLAM refers to the model that simultaneously removes Fourier transform and inverse Fourier transform from GFB and replaces the parallel DS-convs with serial DS-convs in the LSB, and keeps the others untouched. FSSN-Full refers to the model that does not split features along channel dimension in GLAM, and inputs all channels into LSB and GFB respectively.

From Table 2 we can see that our proposed GLAM plays an important role in improving the performance: compared with FSSN, the average Dice values of FSSN-NoFT, FSSN-Sdsconv and FSSN-NoGLAM are reduced by 1.53%, 1.12% and 1.56%, respectively. Moreover, FSSN performs 0.03% better than FSSN-Full, and FSSN slightly reduces the parameters (from 25.234M to 25.203M) and computational complexity (from 4.133G to 4.127G at the resolution of 224 × 224). Therefore, Table 2 demonstrates the effectiveness of LSB, GFB and GLAM.

4.2.2. *Effectiveness of FFM*

To verify the effectiveness of FFM, we conduct ablation experiments specifically for our learnable weight map $W(\cdot)$ in Equation (12) and FFM's serial structure. For fair comparison, we compare our proposed FSSN with FSSN-NoWeight and FSSN-Parallel. Here, FSSN-NoWeight refers to the model removing the $W(\cdot)$ from Equation (12), which means that here is no any information sifting process. And FSSN-Parallel refers to the model that feeds forward the two parts of FFM in parallel. As shown in Table 3, compared with FSSN, the average Dice value of FSSN-NoWeight and FSSN-Parallel are reduced by 1.56 % and 0.32 %, which demonstrates the effectiveness of learnable weight map and serial structure (different from GLAM). To further demonstrate the effectiveness of our learnable weight map $W(\cdot)$, in Figs. 4 and 5, we visualize the heat maps of FFM with and without the $W(\cdot)$ respectively. From Fig. 4 we can see that, without $W(\cdot)$, the small lesion area cannot be learned by using FSSN-NoWeight. And from Fig. 5 we can see that, without $W(\cdot)$, FSSN-NoWeight cannot locate lesion area correctly. By contrast, our FSSN can not only learn all lesion areas (both large lesion area and small lesion area), but also can achieve accurate lesion area location. Table 3, Figs. 4 and 5 demonstrate the effectiveness of our FFM well.

4.2.3. *Effectiveness of DC*

The third component of our FSSN is DC. So, in order to verify the effectiveness of deformable convolution, we compare our proposed FSSN with FSSN-RegularConv. Here, FSSN-RegularConv refers to the model that replaces two 3 × 3 deformable convolutions with two 3 × 3 standard convolutions. As shown in Table 4, compared with FSSN, the average Dice value of FSSN-RegularConv is reduced by 0.36 %. Fig. 6 visualizes the receptive field and sampling locations of DC_2 for two pixels in representative locations. From Fig. 6 we can see that, in the local range, compared with standard convolution, our DC can not only reduce the influence caused by different image components greatly, but also can enlarge receptive field and pay more attention to the fusion of pertinent image contents. And our DC almost only extracts the same image components related to the reference pixel, even at the boundary of lesion area. Tables 2–4 and Figs. 4–6 demonstrate the effectiveness of our components very well.

4.3. *Comparisons with representative medical image segmentation methods*

To demonstrate the superiority of our method, we compare our proposed FSSN with representative CNN-based and Transformer-based medical image segmentation methods [1,8,15–21,24,32,36,38,40,50] on Polyp and ISIC-2018 datasets respectively.

4.3.1. *Objective metrics results on polyp datasets*

The performance of FSSN and other methods on Polyp datasets are shown in Tables 5 and 6. From Tables 5 and 6 we can see that, our proposed FSSN has competitive performance. On the Kvasir, CVC-300 and ETIS-LaribDB datasets, in term of Dice, our FSSN outperforms methods [18] (the second best method on Kvasir) [19], (the second best method on CVC-300) and [21] (the second best method on ETIS-LaribDB) by 0.27 %, 0.03 % and 2.18 %, respectively. And, on the unseen ETIS-LaribDB dataset, we outperform other methods by a significant advantage of more than 2.18 %.

In addition, from Tables 6 and 7 we can see that, first, our FSSN not only can achieve the highest average Dice and average IoU: for all five Polyp datasets, our FSSN surpasses CASCADE [18] (the second best method) 0.23 % and 0.39 % in terms of both average Dice and average IoU respectively. Second, compared with representative medical image segmentation methods, our FSSN has competitive or better performance, and has fewer parameters and lower computational complexity. Although the Dice values of our FSSN are 0.0017 and 0.0144 lower than those of [18] on CVC-ClinicDB and CVC-ColonDB datasets respectively, our FSSN reduces the number of

Table 2
Ablation studies on GLAM.

	Fourier transform	Parallel DS-convs	Full-channels	avg-Dice↑
FSSN-NoGLAM				0.8667
FSSN-NoFT		✓		0.8670
FSSN-Sdsconv	✓			0.8711
FSSN-Full	✓	✓	✓	0.8820
FSSN (Ours)	✓	✓		**0.8823**

Table 3
Ablation studies on FFM.

	Learnable Weight Map	Serial Structure	avg-Dice↑
FSSN-NoWeight		✓	0.8667
FSSN-Parallel	✓		0.8791
FSSN (Ours)	✓	✓	**0.8823**

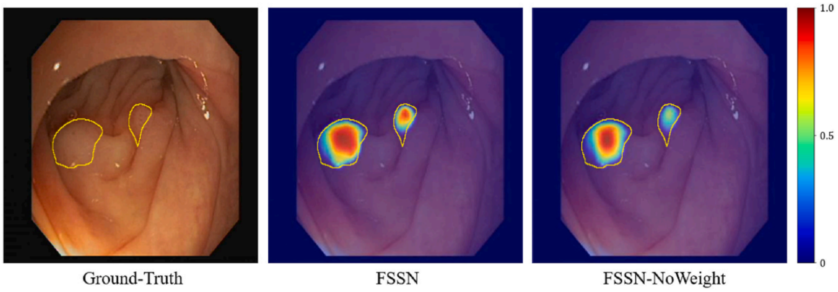

Fig. 4. The heat maps with and without the learnable weight map. The yellow curve shows the ground truth boundary outline of the lesion area. Obviously, FSSN-NoWeight misses out the small lesion area. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

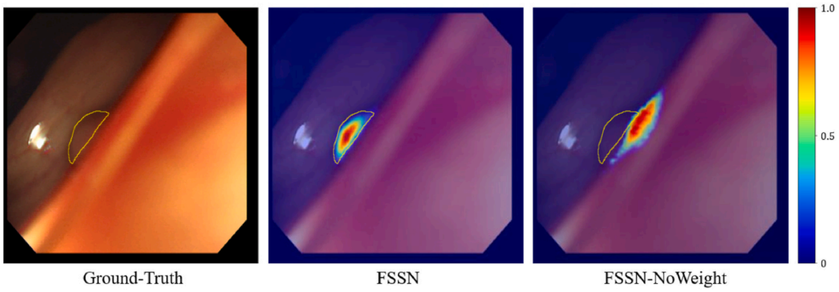

Fig. 5. The heat maps with and without the learnable weight map. The yellow curve shows the ground truth boundary outline of the lesion area. Obviously, FSSN-NoWeight incorrectly locates the lesion area. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Ablation study on DC.

	Deformable convolution	avg-Dice↑
FSSN-RegularConv	✓	0.8787
FSSN (Ours)		**0.8823**

parameters and computational complexity by 28.55 % and 33.83 % respectively.

4.3.2. Objective metrics results on ISIC-2018 dataset

The results of objective evaluation indicators of our FSSN and other methods on ISIC-2018 are shown in Table 8. As shown in Table 8, although the MAE of our FSSN is 0.0009 lower than that of TransFuse-L [16], the Dice and mIoU of our FSSN are 0.38 % and 0.65 % higher than those of TransFuse-L [16] (the second best method). And as shown in Table 9, compared with TransFuse-L, our FSSN reduces the number of parameters and computational complexity by 82.50 % and 93.35 % respectively, which proves that our FSSN can achieve good performance while maintaining a low number of parameters.

Finally, considering that Fourier transform is a complex-valued transform, so it may slow down the inference significantly in some scenarios. Therefore, we compare the inference time of FSSN with that of [1,8,15–18,32,36,38,40]at the input resolution of 224 × 224 (single image). As shown in Table 10, the inference speed of CNN-based methods is quite fast, but their performance tends to be low. Even though our FSSN is about 0.01 s slower than most Transformer-based methods, it outperforms other methods in performance-overhead balance.

Taken together, inference delay within 0.01 s is clearly an acceptable range. It is also worth noting that M3FPolypSegNet [38] and

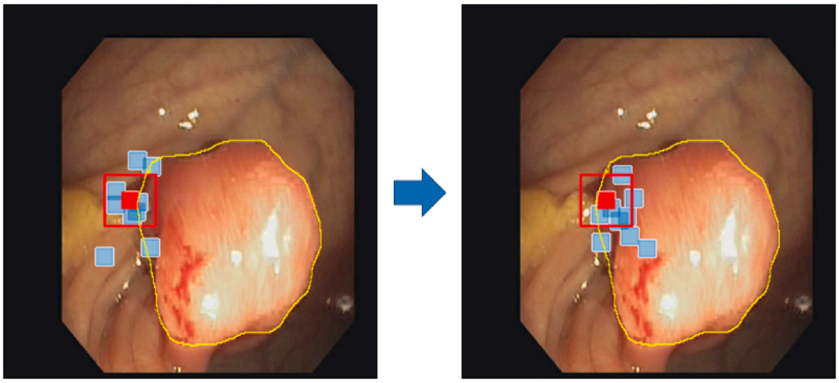

Fig. 6. The visualization of the sampling locations (3×3 = 9 blue points in each image) for two pixels in different locations (red points) respectively. From left to right: the pixel on the boundary (yellow curve) of lesion area (located in the background), and the pixel on the boundary of lesion area (located in lesion area). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Results of all methods on CVC-ClinicDB, Kvasir and CVC-300 datasets. The best, second best and third best performance are marked , and colors re spectively.

red

green

blue

Method	Type	CVC-ClinicDB		Kvasir		CVC-300	
		Dice↑	mIoU↑	Dice↑	mIoU↑	Dice↑	mIoU↑
U-Net [1]	CNN	0.8230	0.7550	0.8180	0.7460	0.7100	0.6270
GFUNet [40]	CNN	0.8280	0.7642	0.7993	0.7282	0.7258	0.6227
M3FPolypSegNet [38]	CNN	0.8820	0.8551	0.8647	0.8230	0.7747	0.7313
PraNet [8]	CNN	0.8990	0.8490	0.8980	0.8400	0.8510	0.7970
CFA-Net [36]	CNN	0.9330	0.8830	0.9150	0.8610	0.8930	0.8270
TransFuse-L [16]	Hybrid	0.9340	0.8860	0.9180	0.8680	0.9040	0.8380
CaraNet [24]	CNN	0.9360	0.8870	0.9180	0.8650	0.9030	0.8380
SSFormer-L [17]	ViT	0.9288	0.8827	0.9111	0.8601	0.8946	0.8268
PPNet [50]	ViT	0.9210	0.8780	0.9200	0.8740	0.8990	0.8390
BiDFNet [20]	ViT	0.9370	0.8900	0.9180	0.8660	0.8880	0.8150
APCNet-v2 [21]	ViT	0.9300	0.8840	0.9160	0.8650	0.8910	0.8260
ColonFormer-L [19]	ViT	0.9320	0.8840	0.9240	0.8760	0.9060	0.8420
CASCADE [18]	ViT	0.9434	0.8998	0.9258	0.8776	0.9047	0.8379
FSSN (Ours)	ViT	0.9417	0.8971	0.9285	0.8833	0.9063	0.8439

GFUNet [40] both introduce Fourier transform to their model, but they lack the design of interaction between frequency information and spatial information. This lack may be responsible for the poor performance of them. GFUNet [40] performed well only on ISIC-2018 dataset and poorly for datasets with insignificant lesion areas (Polyp datasets). Even though they may have lower parameters and computational complexity (especially GFUNet [40]), our FSSN clearly strike a better balance.

In order to further verify the effectiveness of our FSSN, we use the Mean Absolute Error (MAE) to evaluate the performance of our method. The comparison results are shown in Table 11. From Table 11 we can see that, our FSSN can achieve the best MAE values on all five datasets, which again demonstrates the superiority of our FSSN.

Table 6
Results of all methods on CVC-ColonDB, ETIS-LaribDB. And the average Dice and average IoU of all methods for all five Polyp datasets. The best, second best and third best performance are marked , and colors respectively.

red

green

blue

Method	Type	CVC-ColonDB		ERIS-LaribDB		avg-Dice ↑	avg-IoU ↑
		Dice ↑	mIoU ↑	Dice ↑	mIoU ↑		
U-Net [1]	CNN	0.5120	0.4440	0.3980	0.3350	0.6522	0.5814
GFUNet [40]	CNN	0.6017	0.5075	0.3689	0.3180	0.6616	0.5881
M3FPolypSegNet[38]	CNN	0.6712	0.6128	0.4621	0.4203	0.7302	0.6885
PraNet [8]	CNN	0.7120	0.6400	0.6280	0.5670	0.7976	0.7386
CFA-Net [36]	CNN	0.7430	0.6650	0.7320	0.6550	0.8432	0.7782
TransFuse-L[16]	Hybrid	0.7440	0.6760	0.7370	0.6610	0.8474	0.7858
CaraNet [24]	CNN	0.7730	0.6890	0.7470	0.6720	0.8554	0.7902
SSFormer-L[17]	ViT	0.7934	0.7063	0.7803	0.7010	0.8616	0.7954
PPNet [50]	ViT	0.7910	0.7260	0.7840	0.7160	0.8630	0.8066
BiDFNet [20]	ViT	0.8130	0.7290	0.7990	0.7220	0.8710	0.8044
APCNet-v2[21]	ViT	0.8220	0.7470	0.8020	0.7230	0.8722	0.8090
ColonFormer-L[19]	ViT	0.8110	0.7330	0.8010	0.7220	0.8748	0.8114
CASCADE [18]	ViT	0.8254	0.7453	0.8007	0.7258	0.8800	0.8173
FSSN (Ours)	ViT	0.8110	0.7360	0.8238	0.7459	0.8823	0.8212

Table 7
The parameters and computational complexity of methods [1,8,16–19,21,36,38,40] and our FSSN. The size of input image is 352 × 352.

Method	Params (M)	FLOPs (G)	Journals/Conferences
U-Net [1]	7.76	25.99	MICCAI 2015
M3FPolypSegNet [38]	22.39	68.71	ICIP 2023
GFUNet [40]	3.94	5.57	CIBM 2023
PraNet [8]	30.50	13.11	MICCAI 2020
CFA-Net [36]	25.24	55.36	PR 2023
TransFuse-L [16]	43.39	62.10 (192 × 256)	MICCAI 2021
SSFormer-L [17]	66.20	34.60	MICCAI 2022
APCNet-v2 [21]	33.11	16.33	IEEE TIM 2023
ColonFormer-L [19]	52.94	22.94	IEEE Access 2022
CASCADE [18]	35.27	15.40	WACV 2023
FSSN (Ours)	**25.20**	**10.19**	–

4.3.3. *Subjective visual comparisons*

We also show visual comparisons of methods [8,15–18,32,36] and our FSSN on Polyp and ISIC-2018 datasets in Figs. 7–8, respectively. From Figs. 7–8 we can see that, the above methods suffer from various different degrees flaws. By contrast, our FSSN can obtain the most accurate lesion segmentation results: closest to the ground-truth masks. Tables 5–11 and Figs. 7–8 demonstrate the superiority of our method in terms of both qualitative evaluation and quantitative metrics.

5. Conclusion

In this paper, we propose a novel frequency selection segmentation network (FSSN) for high quality lesion segmentation. Inspired by Fourier theory, GLAM and FFM are proposed for the complementary fusion of local details features and global frequency information, better interaction between features, and the adaptive sifting of different frequency components respectively. In addition, we employ deformable convolution (DC) to make our FSSN focus on relevant image contents better for more powerful performance. All these designs effectively improve the abilities of features capture and network representation, and make our FSSN more suitable for lesion segmentation. Extensive experiments on both public benchmark datasets show that our proposed FSSN can achieves better segmentation results than existing representative medical image segmentation methods in terms of both objective evaluation

Table 8
Results of all methods on ISIC-2018. The best, second best and third best performance are marked , and colors respectively.

red

green

blue

Method	Type	Dice ↑	mIoU ↑	MAE ↓
U-Net [1]	CNN	0.8704	0.8127	0.0729
M3FPolypSegNet [38]	CNN	0.8854	0.8171	0.9323
MISSFormer [15]	ViT	0.8860	0.8262	0.0596
UCTransNet [32]	ViT	0.8869	0.8229	0.0617
PraNet [8]	CNN	0.8914	0.8150	0.0570
CASCADE [18]	ViT	0.8966	0.8223	0.0525
GFUNet [40]	CNN	0.8997	0.8340	0.0557
DermoSegDiff-A [64]	Diffusion	0.9005	-	-
TransFuse-L [16]	Hybrid	0.9057	0.8380	0.0477
FSSN (Ours)	ViT	0.9095	0.8445	0.0486

Table 9
The parameters and computational complexity of the methods [1,8,15,16,18,32,38,40,64] and our FSSN. The size of the input image is 224 × 224.

Method	Params (M)	FLOPs (G)	Journals/Conferences
M3FPolypSegNet [38]	22.39	27.83	ICIP 2023
GFUNet [40]	3.94	2.26	CIBM 2023
U-Net [1]	7.76	10.52	MICCAI 2015
MISSFormer [15]	35.45	7.28	TMI 2023
UCTransNet [32]	66.24	32.98	AAAI 2022
PraNet [8]	30.50	5.33	MICCAI 2020
CASCADE [18]	35.27	6.24	WACV 2023
DermoSegDiff-A [64]	–	–	PRIME 2023
TransFuse-L [16]	143.39	62.10 (192 × 256)	MICCAI 2021
FSSN (Ours)	**25.10**	**4.13**	–

Table 10
Comparison of inference time at 224 × 224 resolution.

Method	Type	Inference Time (s)
U-Net [1]	CNN	0.041
PraNet [8]	CNN	0.067
CFA-Net [36]	CNN	0.071
M3FPolypSegNet [38]	CNN	0.056
GFUNet [40]	CNN	0.041
TransFuse-L [16]	Hybrid	0.071 (192 × 256)
SSFormer-L [17]	ViT	0.083
CASCADE [18]	ViT	0.063
MISSFormer [15]	ViT	0.064
UCTransNet [32]	ViT	0.067
FSSN (Ours)	ViT	**0.073**

indicators and subjective visual effects, while has fewer parameters and lower computational complexity. However, In GLAM, our FSSN does not focus on phase and amplitude information separately (e.g. Ref. [45]), and compared to lightweight models (e.g. Ref. [40]), FSSN is slightly large. In the future, we will investigate how to handle the frequency domain information in a more comprehensive way, and further reduce the model size and computational complexity while achieving better performance. And extending our FSSN to 3D medical image segmentation is also our future work.

Table 11
The MAE results of all methods on CVC-ClinicDB, Kvasir and CVC-300 datasets. The best is marked color respectively.

red

Method	MAE ↓				
	CVC-ClinicDB	Kvasir	ETIS-LaribDB	ISIC-2018	CVC-300
U-Net [1]	0.019	0.055	0.036	0.073	0.022
GFUNet [40]	-	-	-	0.056	-
M3FPolypSegNet [38]	-	-	-	0.932	-
PraNet [8]	0.009	0.030	0.031	0.057	0.010
CFA-Net [36]	0.007	0.023	0.014	-	0.008
TransFuse-L [16]	-	-	-	0.0477	-
CaraNet [24]	0.007	0.023	0.017	-	0.007
PPNet [50]	0.008	0.024	0.013	-	0.006
BiDFNet [20]	0.006	0.023	0.016	-	0.010
APCNet-v2 [21]	0.008	0.024	0.016	-	0.007
CASCADE [18]	-	-	-	0.0525	-
MISSFormer [15]	-	-	-	0.0596	-
UCTransNet [32]	-	-	-	0.0617	-
FSSN (Ours)	0.006	0.021	0.012	0.046	0.006

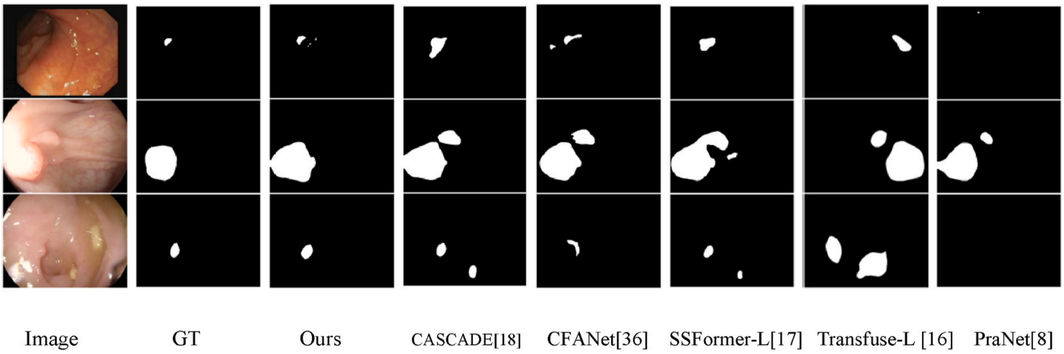

Fig. 7. The visual comparisons of representative methods and our FSSN on Polyp datasets.

CRediT authorship contribution statement

Shu Tang: Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Haiheng Ran:** Visualization, Software, Methodology. **Shuli Yang:** Writing – original draft. **Chaoxia Wang:** Data curation. **Wei Li:** Investigation. **Haorong Li:** Visualization, Validation. **Zihao Meng:** Validation, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Shu Tang reports financial support was provided by Natural Science Foundation of China (61601070) Chongqing Natural Science Foundation(CSTB2023NSCQ-MSX0680) The Major Project of Science and Technology Research of Chongqing Education Commission (KJZD-M202300101). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

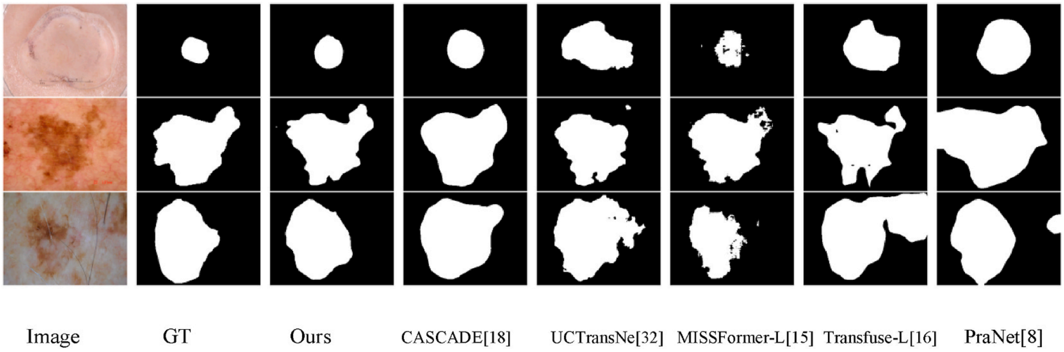

Fig. 8. The visual comparisons of representative methods and our FSSN on ISIC-2018 dataset.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networksfor biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), Part III 18, Springer, 2015, pp. 234–241.

[2] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++:A nested u-net architecture for medical image segmentation, in: DeepLearning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham, Switzerland, 2018, pp. 3–11.

[3] H. Huang, et al., Unet 3+: a full-scale connected unet for medical imagesegmentation, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Process.(ICASSP), IEEE, 2020, pp. 1055–1059.

[4] S. Feng, et al., CPFNet: context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imag. 39 (10) (2020) 3008–3018.

[5] Z. Gu, et al., Ce-net: context encoder network for 2d medical imagesegmentation, IEEE Trans. Med. Imag. 38 (10) (2019) 2281–2292.

[6] O. Oktay, et al., Attention U-Net: Learning where to Look for the Pancreas, 2018 arXiv preprint arXiv:1804.03999.

[7] Y. Cai, Y. Wang, Ma-unet: an improved version of unet based on multiscale and attention mechanism for medical image segmentation, Proceedings of the Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021) 12167 (2022) 205–211. SPIE.

[8] D.-P. Fan, et al., Pranet: parallel reverse attention network for polypsegmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assist.Intervent. (MICCAI), Springer, 2020, pp. 263–273.

[9] A. Dosovitskiy, et al., An Image Is Worth 16×16 Words: Transformersfor Image Recognition at Scale, 2020 arXiv:2010.11929.

[10] W. Wang, et al., Pyramid vision transformer: a versatile backbone fordense prediction without convolutions, in: Proc. IEEE/CVF Int. Conf.Comput. Vis. (ICCV), 2021, pp. 568–578.

[11] J. Guo, et al., Cmt: convolutional neural networks meet vision transformers, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR), 2022, pp. 12175–12185.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, Segformer: simple and efficient design for semantic segmentation withtransformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[13] Z. Liu, et al., Swin transformer: hierarchical vision transformer usingshifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.

[14] W. Wang, et al., Pvt v2: improved baselines with pyramid vision transformer, Computational Visual Media 8 (3) (2022) 415–424.

[15] X. Huang, Z. Deng, D. Li, X. Yuan, Y. Fu, MISSFormer: an effectivetransformer for 2d medical image segmentation, IEEE Transactions onMedical Imaging 42 (5) (2023) 1484–1494, https://doi.org/10.1109/TMI.2022.3230943.

[16] Y. Zhang, H. Liu, Q. Hu, Transfuse: fusing transformers and cnns formedical image segmentation, in: Proc. Int. Conf. Med. Image Comput.Comput.-Assist. Intervent. (MICCAI), Springer, 2021, pp. 14–24. Part I 24.

[17] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, S. Song, Stepwise featurefusion: local guides global, in: Proc. Int. Conf. Med. Image Comput.Comput.-Assist. Intervent. (MICCAI), 2022, pp. 110–120.

[18] M.M. Rahman, R. Marculescu, Medical image segmentation via cascaded attention decoding, in: Proceedings of the IEEE/CVF WinterConference on Applications of Computer Vision, 2023, pp. 6222–6231.

[19] N.T. Duc, N.T. Oanh, N.T. Thuy, T.M. Triet, V.S. Dinh, Colonformer: an efficient transformer based method for colon polyp segmentation, IEEE Access 10 (2022) 80575–80586.

[20] S. Tang, J. Qiu, X. Xie, H. Ran, G. Zhang, BiDFnet: Bi-decoder andfeedback network for automatic polyp segmentation with vision transformers, in: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2022, pp. 16–27.

[21] G. Yue, S. Li, R. Cong, T. Zhou, B. Lei, T. Wang, Attention-guidedpyramid context network for polyp segmentation in colonoscopy images, IEEE Trans. Instrum. Meas. 72 (2023) 1–13.

[22] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, Y. Wang, Intriguing findings offrequency selection for image deblurring, Proc. AAAI Conf. Artif. Intell. 37 (2023) 1905–1913.

[23] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks forsemantic segmentation, in: Proc. IEEE Conf. Comput. Vis. PatternRecognit. (CVPR), 2015, pp. 3431–3440.

[24] A. Lou, S. Guan, M. Loew, Caranet: context axial reverse attentionnetwork for segmentation of small medical objects, Journal of MedicalImaging 10 (1) (2023) 14005.

[25] J. Dai, et al., Deformable convolutional networks, in: Proc. IEEE Int.Conf. Comput. Vis. (ICCV), 2017, pp. 764–773.

[26] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: more deformable, better results, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, 9308–9316.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, CVPR, 2016, pp. 770–778.

[28] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P.T. Res2net, A new multi-scale backbone architecture 43 (2019) 652–662, https://doi.org/10.1109/TPAMI.

[29] H. Zhang, et al., Resnest: split-attention networks, in: Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2022, pp. 2736–2746.

[30] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proc. IEEE Conf. Comput.Vis. Pattern Recognit. (CVPR), 2017, pp. 1492–1500.

[31] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] H. Wang, P. Cao, J. Wang, O.R. Zaiane, Uctransnet: rethinking the skipconnections in u-net from a channel-wise perspective with transformer, Proc. AAAI Conf. Artif. Intell. 36 (2022) 2441–2449.

[33] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc.IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformernetworks, Adv. Neural Inf. Process. Syst. 28 (2015).

[35] J. Schlemper, et al., Attention gated networks: learning to leveragesalient regions in medical images, Med. Image Anal. 53 (2019) 197–207.
[36] T. Zhou, et al., Cross-level feature aggregation network for polyp segmentation, Pattern Recogn. 140 (2023) 109555.
[37] G. Liu, Z. Chen, D. Liu, B. Chang, Z. Dou, FTMF-net: a fouriertransform-multiscale feature fusion network for segmentation of smallpolyp objects, IEEE Trans. Instrum. Meas. (2023).
[38] J.-H. Nam, S.-H. Park, N.S. Syazwany, Y. Jung, Y.-H. Im, S.-C. Lee, M3FPolypsegnet: segmentation network with multi-frequency featurefusion for polyp localization in colonoscopy images, in: Proc. 2023 IEEEInt. Conf. Inf. Process. (ICIP), IEEE, 2023, pp. 1530–1534.
[39] L. Chi, B. Jiang, Y. Mu, Fast fourier convolution, in: Advances in NeuralInformation Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 4479–4488.
[40] P. Li, R. Zhou, J. He, S. Zhao, Y. Tian, A global-frequency-domainnetwork for medical image segmentation, Computers in Biology andMedicine 164 (2023) 107290.
[41] M.P. Paing, C. Pintavirooj, Adenoma dysplasia grading of colorectalpolyps using fast fourier convolutional resnet (ffc-resnet), IEEE Access11 (2023) 16644–16656.
[42] F. Zou, Y. Liu, Z. Chen, Z. Karl, D. Jin, Fourier channel attentionpowered lightweight network for image segmentation, IEEE Journal ofTranslational Engineering in Health and Medicine 11 (2023) 252–260.
[43] W. Wang, J. Wang, C. Chen, J. Jiao, L. Sun, Y. Cai, S. Song, J. Li, Fremim: fourier transform meets masked image modeling for medicalimage segmentation, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7845–7855.
[44] Q. Han, H. Wang, M. Hou, T. Weng, Y. Pei, Z. Li, G. Chen, Y. Tian, Z. Qiu, Hwa-segnet: multi-channel skin lesion image segmentation network with hierarchical analysis and weight adjustment, Computers inBiology and Medicine 152 (2023) 106343.
[45] Y. Li, Z. Zheng, W. Ren, Y. Nie, J. Zhang, X. Jia, Frequency aware andgraph fusion network for polyp segmentation, in: ICASSP 2024 - 2024IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1586–1590, https://doi.org/10.1109/ICASSP48485.2024.10446687.
[46] C. Yang, Z. Zhang, Pfd-net: pyramid fourier deformable network formedical image segmentation, Comput. Biol. Med. 172 (2024) 108302, https://doi.org/10.1016/j.compbiomed.2024.108302.
[47] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, Sa-unet:Spatial attention u-net for retinal vessel segmentation, in: Proc. Int.Conf. Pattern Recognit. (ICPR), IEEE, 2021, pp. 1236–1242.
[48] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt:Polyp segmentation with pyramid vision transformers (2021) arXiv preprintarXiv:2108.06932.
[49] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, DUNet: adeformable network for retinal vessel segmentation, Knowl. Base Syst. 178 (2019) 149–162.
[50] K. Hu, W. Chen, Y. Sun, X. Hu, Q. Zhou, Z. Zheng, PPNet: pyramidpooling based network for polyp segmentation, Computers in Biologyand Medicine 160 (2023) 107028.
[51] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional blockattention module, in: Proc. Eur. Conf. Comput. Vis, ECCV, 2018, pp. 3–19.
[52] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: more features from cheap operations, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577–1586.
[53] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 6024–6042.
[54] X. Ding, Y. Guo, G. Ding, J. Han, Acnet: strengthening the kernelskeletons for powerful cnn via asymmetric convolution blocks, in: Proc.IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1911–1920.
[55] A.G. Howard, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017) arXiv preprint arXiv:1704.04861.
[56] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificialintelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
[57] Y. Pang, Y. Li, J. Shen, L. Shao, Towards bridging semantic gap to improve semantic segmentation, in: Proc. IEEE/CVF Int. Conf. Comput.Vis. (ICCV), 2019, pp. 4230–4239.
[58] J. Bernal, F. Sánchez, G. Fernández-Esparrach, D. Gil, C.R. Miguel, F. Vilariño, Wm-dova maps for accurate polyp highlighting incolonoscopy: validation vs. saliency maps from physicians, Computerized medical imaging and graphics : the official, journal of the Computerized Medical Imaging Society 43 (2015) 99–111.
[59] N. Tajbakhsh, S. Gurudu, J. Liang, Automated polyp detection incolonoscopy videos using shape and context information, IEEE Trans. Med. Imag. 35 (2) (2016) 630–644.
[60] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embeddeddetection of polyps in wce images for early diagnosis of colorectal cancer, Int. J. Comput. Assist. Radiol. Surg. 9 (2) (2014) 283–293.
[61] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H.D. Johansen, Kvasir-seg: a segmented polyp ddataset, in: MultiMedia Modeling, Springer International Publishing, Cham, 2020, pp. 451–462.
[62] D. Vázquez, J. Bernal, F. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdzal, A.C. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, Journal of Healthcare Engineering (2016).
[63] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proceedings of the International Conference on Learning Representations, 2019.
[64] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, D. Merhof, Dermosegdiff: a boundary-aware segmentation diffusion model for skin lesion delineation, in: Proceedings of the International Workshop on PRedictive Intelligence in MEdicine, Springer, 2023, pp. 146–158.