Applied Intelligence (2025) 55:103
https://doi.org/10.1007/s10489-024-06029-1

A swin-transformer-based network with inductive bias ability for medical image segmentation

Yan Gao[1] · Huan Xu[1] · Quanle Liu[1] · Mei Bie[1] · Xiangjiu Che[1]

Accepted: 16 September 2024

Abstract

Accurately segmenting organs, or diseased regions of varying sizes, is a challenge in medical image segmentation tasks with limited datasets. Although Transformer-based methods have self-attention mechanisms, excellent global modelling abilities and can effectively focus on crucial areas in medical images, they still face the intractable issues of computational complexity and excessive reliance on data. The Swin-Transformer has partly addressed the problem of computational complexity, however the method still requires large amounts of training data due to its lack of inductive bias capabilities. When applied to medical image segmentation tasks with small datasets, this leads to suboptimal performances. In contrast, the CNN-based methods can compensate for this limitation. To address this issue further, this paper proposes Swin-IBNet, which combines the Swin-Transformer with a CNN in a novel manner to imbue it with inductive bias capabilities, reducing its reliance on data. During the encoding process of Swin-IBNet, two novel and crucial modules, the feature fusion block (FFB) and the multiscale feature aggregation block (MSFA), are designed. The FFB is responsible for propagating the inductive bias capability to the Swin-Transformer encoder. Different from the previous use of multiscale features, MSFA efficiently leverages multiscale information from different layers through self-learning. This paper not only attempts to analyse the interpretability of the proposed Swin-IBNet but also performs more verifications on the public Synapse, ISIC 2018 and ACDC datasets. The experimental results show that Swin-IBNet is superior to the baseline method, Swin-Unet, and several state-of-the-art methods. Especially on the Synapse dataset, the DSC of Swin-IBNet surpasses that of Swin-Unet by 3.45%.

Keywords Medical image segmentation · Convolutional neural network · Deep learning · Transformer

1 Introduction

Medical image segmentation tasks have extensive application and research value in medical studies, clinical diagnoses, pathological analyses, surgical planning and computer-assisted surgical procedures. However, compared to the semantic segmentation tasks of natural images, medical image segmentation tasks possess unique characteristics. First, medical image datasets are generally smaller, and positive pathology data representing instances of the targeted pathology or disease are often scarce. Second, annotating medical images is challenging because it requires prior knowledge from specialized medical professionals. For example, expert physicians need to accurately mark the affected areas, which may exhibit variations in interpretation due to differences in expertise, experience or subjective judgement. Third, medical image segmentation tasks impose a demand for high precision, presenting a significant challenge in this field. Finally, medical image segmentation tasks often encounter patient privacy concerns, which further act as a barrier to advancement in this area. In the face of the rapidly evolving field of artificial intelligence, various advanced methods and techniques have been applied to address these challenges.

✉ Xiangjiu Che
chexj@jlu.edu.cn

Yan Gao
gyan19@mails.jlu.edu.cn

Huan Xu
xuhuan20@mails.jlu.edu.cn

Quanle Liu
qlliu18@mails.jlu.edu.cn

Mei Bie
bie-mei@163.com

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, Jilin, China

Published online: 09 December 2024

CNNs have made significant progress in the field of computer vision, and there are notable networks for classification [1–3], detection [4, 5] and segmentation, such as Mask R-CNN [6], FCN [7], the Deeplab series [8], U-Net [9], improved Mask R-CNN [10] and ConvUNeXt [11]. CNN-based networks are well suited for dense prediction tasks due to their translation invariance, equivariance, scale invariance and rotation invariance properties. However, while pooling operations in CNNs prevent overfitting, maintain invariance, reduce computational complexity and extract deep semantic information, excessive use of these operations makes deeper networks inefficient in improving the performance of dense prediction tasks. This is because pooling leads to a substantial loss of information that cannot be recovered. Several approaches have been proposed to address this limitation in CNNs, with the most effective method being the use of skip connections. These connections involve incorporating features from corresponding stages of the encoder during the decoding process to restore feature resolution. However, due to the use of small-window convolutional kernels in CNNs for feature extraction, it is challenging to directly model global semantic and contextual information.

The Transformer [12] is an emerging structure that has gained popularity in recent years, and is considered a potential alternative to standard convolutional neural networks (CNNs). It initially achieved remarkable success in the field of natural language processing (NLP) and was subsequently introduced into the computer vision domain. Although Transformers have powerful global modelling abilities, their high computational complexity has become a major obstacle to their popularization for computer vision tasks. In contrast to text processing in NLP, computer vision samples are typically large in size. Transformers, which rely on global modelling and require computations involving all the data, face challenges when directly applied to computer vision tasks. To reduce the influence of the computational complexity of Transformers, many approaches [13–23] rely on reducing dimensionality, using CNNs first, resulting in the loss of valuable information from the shallow layers of the CNNs, and suboptimal performances of Transformer global modelling. One groundbreaking work is ViT [24], which first divides an image into nonoverlapping patches and then applies linear projection to each patch to obtain “tokens”. Inspired by token handling in NLP transformers, ViT utilizes a series of multihead attention and feedforward layers to process all the “tokens”. In contrast to Transformers in the NLP domain, ViT uses attention layers to model the global relationships between tokens, which is a key distinction from CNNs. However, ViT sacrifices a portion of the Transformer’s global modelling capability and incurs a significant computational cost by dividing the sample into numerous small patches. Liu et al. [25] proposed the Swin-Transformer, which incorporates hierarchical and shifted windows to capture both local and global information effectively, resulting in reduced computational complexity compared to ViT’s patch-based approach, especially for larger images. In addition, numerous Transformer-based approaches have been proposed that address the sample size challenge in various tasks. These include methods for classification [26–28], detection [29–31] and segmentation [32–38]. However, Transformers lack the inductive bias present in CNNs, so they require more samples than CNNs to learn certain local properties of visual data. Therefore, in the context of medical image segmentation tasks, which typically involve small datasets, CNNs remain the preferred model. Nevertheless, given the powerful capabilities of Transformers, there is still a desire to apply them to visual tasks in the medical field. Additionally, the utilization of multiscale features often involves a process of spatial reintegration through concatenation and convolution operations after unified resizing. How to effectively utilize the strengths of both methods and fully exploit multiscale features to improve the segmentation performances of medical images is a focal point of this study.

Medical image segmentation tasks require high precision, and while CNN-based methods have achieved significant success, they lack the ability to perform global modelling, which limits the improvement possibilities of the model. On the other hand, Transformers possess powerful global modelling capabilities but lack inductive biases, making Transformer-based methods highly data intensive during training. This poses a significant challenge, particularly considering the scarcity of the medical data resources. Given these considerations, it becomes apparent that combining the strengths of both approaches can be highly beneficial, allowing mutual reinforcement and enabling the network to achieve desirable results in medical image segmentation tasks. In this paper, we propose Swin-IBNet, a network that offers the following advantages:

1) Building upon the inherent self-attention mechanism and global modelling capabilities of the Swin-Transformer, we reduce its reliance on the massive amounts of data in the encoder by introducing the inductive bias capabilities of CNNs. This enhancement allows it to maintain its advantages in medical image segmentation even with smaller datasets, thus improving its segmentation capabilities in this area.
2) This paper proposes a novel feature fusion block (FFB) involving two branches, and its outputs are instructive for the use of multiscale features. More importantly, FFB is responsible for enhancing the inductive bias ability of the Swin-Transformer encoder.
3) Another crucial contribution of this paper is the novel method proposed for multiscale feature aggregation (MSFA). Different from the previous use of multiscale

features, MSFA efficiently leverages multiscale information from different layers through self-learning.

4) This paper not only conducts abundant experiments to verify the effectiveness of Swin-IBNet but also attempts to analyse the interpretability of the individual components within Swin-IBNet.

2 Related work

2.1 Vision transformers and CNNs

The Transformer was initially introduced by A. Vaswani [12] for machine translation tasks. However, due to its computational cost, its application in computer vision tasks has been limited. Initially, attention mechanisms in computer vision were mainly used to replace certain components of CNNs, or inserted into specific parts of CNNs to improve performance [39–41]. Nevertheless, owing to the powerful global modelling capabilities of Transformers, the exploration of Transformers' applications in computer vision has never ceased, with ongoing research investigating the possibility of replacing CNNs with Transformers. However, when processing images using a Transformer, it is necessary to use self-attention to calculate the relationship between each pixel and other pixels, which leads to a quadratic cost for the number of pixels. Many methods [42–47] reduce the dimension first by using CNNs, and then use Transformers to obtain the global relationships between the pixels. However, these methods fail to leverage the advantages of the global modelling ability of Transformers. Subsequently, a ground-breaking work called ViT [24] was proposed, which is a method that utilizes a pure Transformer as the backbone for computer vision tasks. To address the issue of computational complexity, this approach first divides the image into nonoverlapping patches. Each patch is then linearly projected to obtain a "token", and attention layers are used to model the global relationships between these tokens, thereby leveraging the global modelling capabilities of the Transformer to some extent. Following ViT, more Transformer-based approaches [25, 48, 49] have been proposed. Among them, the Swin-Transformer [25] is one of the most influential works, which introduces a hierarchical Transformer. This method reduces the computational complexity from the original quadratic cost for the number of pixels to a linear computational complexity by employing a shifted windowing scheme, while also allowing for cross-window connections. The Swin-Transformer is also employed as a fundamental framework in this paper. However, positional relationships within different windows are still not effectively captured. In contrast, the convolutional kernel of the CNN slides over feature maps with overlaps, which effectively avoids the issue of the positional relationships within different windows being missed in the Transformer.

Researchers have found that Transformer-based methods perform significantly worse on smaller datasets than CNNs. Furthermore, studies [50–52] have shown that the reason behind this is that most Transformer-based approaches lack the inductive bias found in CNNs. As a result, CNNs require a larger number of samples to learn certain local properties of visual data, whereas CNNs embed these properties into their structural design. Initially, CNNs and Transformers were combined, using CNNs for feature extraction and dimensionality reduction, followed by the application of Transformers to appropriately sized feature maps. This is due to the computational characteristics of Transformers, as their strong global modelling capability comes at the cost of quadratic computational complexity with respect to all the pixels in an image. ViT [24] solves the problem of computational complexity to some extent and facilitates a further integration of CNNs and Transformers. To enhance ViT's ability to capture local features, numerous methods [53, 54] introduce convolutional operations in the Transformer block in different forms. Additionally, approaches such as [13, 25, 55, 56] not only incorporate CNNs locally but also design them in a hierarchical structure. The advantage of combining CNNs and ViT lies in the ability of CNNs to emphasize the local features of the image content, while ViT can model global dependencies. In theory, this combination should yield improved performances. However, ViT sacrifices a portion of the Transformer's global modelling capability and incurs a significant computational cost by dividing the sample into numerous small patches. The Swin-Transformer not only reduces computational complexity based on ViT but also inherits a significant extent of the global modelling capability. Nevertheless, as we have consistently emphasized, Transformer-based methods lack inductive bias abilities, which results in a heavy reliance on data. In this paper, we explore the respective advantages of CNNs and Transformers from different perspectives and propose a novel network. The integration of CNNs and Transformers in our network is not limited to local connections. Instead, we aim for mutual reinforcement and inspiration, thereby enhancing the network's segmentation capabilities for medical images. Additionally, we introduce a novel pixelwise multiscale feature aggregation method, which is highly advantageous for medical image segmentation.

2.2 Methods of multiscale features

Compared to object detection and classification tasks, image segmentation tasks require more detailed information. However, CNN-based image segmentation methods involve a process of continuous downsampling through pooling operations during the extraction of semantic information. This process leads to the loss of a significant amount of spatial information, posing challenges for segmenting small objects or objects at different scales. To address this issue, the FPN [57] was proposed. In this method, CNNs are initially utilized to extract feature maps at various scales, and subsequently, a feature pyramid is generated by incorporating a top-down pathway and lateral connections into the CNN backbone. FPN has been widely used in many object detection frameworks, such as Faster R-CNN [58] and Mask R-CNN [6]. It has been proven to be effective in improving the accuracy of object detection or segmentation tasks with significant scale variations. With the rise of Transformers, the integration of multiscale features has been widely explored in various Transformer-based methods [13, 25, 59, 60]. Among them, pyramid ViT (PVT) [13] is particularly notable because it introduces a hierarchical design for ViT. It incorporates a progressively shrinking pyramid and spatial-reduction attention to effectively capture multiscale information. Subsequently, this approach was quickly adopted in the design of other methods. PVTv2 [59] improved PVT by using a linear complexity attention layer, overlapped patch embedding and convolutional feedforward networks. Methods such as [25, 61] adopted a hierarchical structure similar to CNNs. However, the connections between different stages are sequential, without considering the relationships between feature maps of different scales. In most of these methods, after obtaining features at different stages, upsampling and concatenation are commonly used, which leads to the same contribution of all the pixels within the same stage, and potentially underutilizing the multiscale feature information. Instead, we aim for the network to leverage feature maps of different scales to acquire more detailed information and learn more accurate target features.

3 Proposed method

The overall structure of Swin-IBNet proposed in this paper still follows an encoder-decoder form. The encoding stage consists of two main branches, a CNN branch based on convolutional blocks, and a Swin-Transformer branch. To leverage the strengths of each branch, and compensate for their weaknesses, unlike previous approaches that simply combined CNNs and Transformers independently, there is an information interaction between the features from the two branches at each stage of the entire feature extraction process. Section 3.1 provides an overview of the proposed Swin-IBNet, followed by the FFB in Section 3.2 and the MSFA block in Section 3.3.

Algorithm 1 Calculation process of an image passing through Swin-IBNet.

```
Input: Image $I \in \mathbb{R}^{(B\times Channel\times H\times W)}$, Up-sampling layer number $L_{\text{Up}}$, Down-sampling stage number $L_{\text{Down}}$.
Output: $I' \in \mathbb{R}^{(B\times Class\times H\times W)}$
Perform image preprocessing before passing through Swin-Branch.
Perform image preprocessing before passing through CNN-Branch.
for $i = 1$ to $L_{\text{Down}}$ do
    $x_{\text{swin}}$ is obtained through the $i$-th stage of Swin-Branch.
    if this is not the last stage then
        $x_{\text{CNN}}$ is obtained through the $i$-th stage of CNN-Branch.
    else
        $x_{\text{CNN}}$ is obtained through ASPP.
    end if
    if this is not the first stage then
        if this is not the last stage then
            The outputs of both the current stage and all previous stages of the CNN-Branch, as well
            as the output of the previous stage FFB, serve as inputs to MSFA_CNN. Calculate the
            output of MSFA_CNN according to equation (6).
        else
            The outputs of ASPP and all previous stages of the CNN-Branch, as well as the output
            of the previous stage FFB, serve as inputs to MSFA_CNN. Calculate the output of
            MSFA_CNN according to equation (6).
        end if
        The outputs of both the current stage and all previous stages of the Swin-Branch, as well as
        the output of the previous stage FFB, serve as inputs to MSFA_Swin. Calculate the output
        of MSFA_Swin according to equation (6).
        Calculate the output of FFB according to equations (1), (2), (3), (4), and (5) based on the
        outputs of MSFA_CNN and MSFA_Swin.
    else
        Calculate the output of FFB according to equations (1), (2), (3), (4), and (5) based on $x_{\text{swin}}$
        and $x_{\text{CNN}}$.
    end if
    if this is not the last stage then
        $x_{\text{res\_concat}} = \text{concat}(x_{\text{CNN}}, \text{the output of FFB})$
        $x_{\text{res\_conv}} = conv_{1*1}(x_{\text{res\_concat}})$
        $x_{\text{CNN}} = x_{\text{CNN}} + x_{\text{res\_conv}}$
        $x_{\text{swin\_concat}} = \text{concat}(x_{\text{swin}}, \text{the output of FFB})$
        $x_{\text{swin\_conv}} = conv_{1*1}(x_{\text{swin\_concat}})$
        $x_{\text{swin}} = x_{\text{swin}} + x_{\text{swin\_conv}}$
    else
        The output of FFB undergoes tensor reshaping and normalization.
    end if
end for
for $l = 1$ to $L_{\text{Up}}$ do
    if this is the first stage then
        Directly Up-sample the final output of decoder.
    else
        $F$ = concat(The output of the upper stage of the decoder, output of FFB at the same stage)
        $F$ = Up-sampling($F$)
    end if
end for
$I'$ is obtained after tensor reshaping and normalizing of $F$.
```

3.1 Overview of the proposed swin-IBNet

The proposed Swin-IBNet is illustrated in Fig. 1. This network is a hybrid of a CNN and a Swin-Transformer. Given a medical image, one branch directly passes through a CNN network composed of convolutional blocks, while the other branch first divides the original image into several patches and then sends them to the Swin-Transformer encoder. The CNN branch can utilize existing backbones such as ResNet50. In the final stage, an ASPP (atrous spatial pyramid pooling) module is added. Typically, these network frameworks consist of several stages, each of which reduces the spatial dimensions of the previous stage's output by half for feature extraction. In this case, we ensure that the number of stages remains consistent with that of the Swin-Transformer branch for subsequent fusions. Furthermore, unlike a conventional CNN-based feature extraction, the output features of each stage are not directly used as inputs for the next stage. Instead, they are combined with the features from the Swin-Transformer branch through residual connections before further processing in the next stage. The Swin-Transformer branch should have the same feature dimensions as the corresponding stages of the CNN branch. Similarly, the features from each stage of this branch are not directly passed as inputs to the next stage. Like the CNN branch, they are combined with the features from the corresponding stages of the CNN branch using residual connections before being passed to the next stage. The specific network structures for both branches are detailed in Table 1, and the pseudocode for Swin-IBNet is also provided in Algorithm 1.

From Fig. 1, we observe that, as the network deepens, we obtain a feature map F_i for each stage, where i $\in (1, 2, \ldots, N)$ represents the stage index. The size of the output feature map at each stage satisfies $F_i = \frac{F}{2^{i-1}}$, where F is the initial image feature. Naturally, obtaining features at different scales leads us to consider using multiscale features to further enhance the network's performance. Unlike previous approaches, this paper utilizes feature maps from different stages before each stage in the feature extraction process. Additionally, the proposed MSFA and FFB methods are designed to determine which information should not be discarded. These blocks serve as compensatory information, which is fused with the output feature of the current stage and passed as input to the next stage. For the CNN branch, since the compensatory information includes features from the Swin-Transformer branch, leveraging the powerful global modelling capability of the Transformer helps the network learn global information and compensates to some extent for the information loss caused by the pooling operations in the CNN. Furthermore, since the fusion module in this

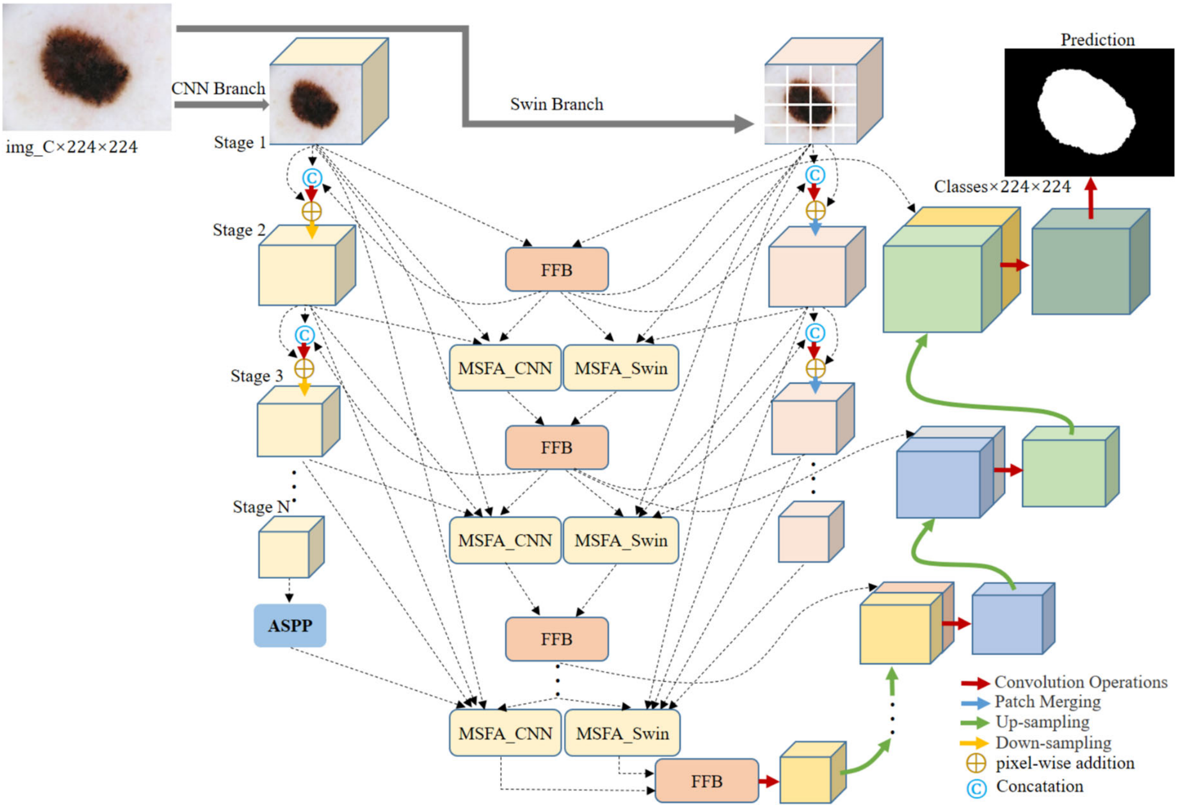

Fig. 1 Overview of Swin-IBNet with the ability to induce bias and global modeling

Table 1 Encoder structure details of the proposed Swin-IBNet

		Stage 1	Stage 2	Stage 3	Stage 4
CNN_Branch (Resnet_50)	Output size	256×56×56	512×28×28	1024×14×14	2048×7×7
	Component	7×7 (×1) Pooling (×1) 1×1, 64 (×3) 3×3, 64 (×3) 1×1, 256 (×3) 1×1 Conv (×1) Skip-con (×1)	1×1, 128 (×4) 3×3, 128 (×4) 1×1, 512 (×4) 1×1 Conv (×1) Skip-con (×1)	1×1, 256 (×6) 3×3, 256 (×6) 1×1, 1024 (×6) 1×1 Conv (×1) Skip-con (×1)	1×1, 256 (×3) 3×3, 256 (×3) 1×1, 1024 (×3) 1×1 Conv (×1) Skip-con (×1)
Swin_Branch	Output size	96×56×56	192×28×28	384×14×14	768×7×7
	Component	Embedding (×1) Swin-block (×1) 1×1 Conv (×1) Skip-con (×1)	Embedding (×1) Swin-block (×1) 1×1 Conv (×1) Skip-con (×1)	Embedding (×1) Swin-block (×1) 1×1 Conv (×1) Skip-con (×1)	Embedding (×1) Swin-block (×1)
FFB	Output size	96×56×56	192×28×28	384×14×14	768×7×7
	Component	1×1 Conv (×1) 3×3 Conv (×2) MH Block (×1) Skip-con (×1) BN (×1)	1×1 Conv (×1) 3×3 Conv (×2) MH Block (×1) Skip-con (×1) BN (×1)	1×1 Conv (×1) 3×3 Conv (×2) MH Block (×1) Skip-con (×1) BN (×1)	1×1 Conv (×1) 3×3 Conv (×2) MH Block (×1) Skip-con (×1) BN (×1)
MSFA_CNN	Output size	–	512×28×28	1024×14×14	2048×7×7
	Component	–	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)
MSFA_Swin	Output size	–	192×28×28	384×14×14	768×7×7
	Component	–	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)	Upsampling (×1) Linear (×1) Pooling (×1) RELU (×2) 3×3 Conv (×2) 1×1 Conv (×2) BN (×1)

Conv refers to *Convolution*, *Skip-con* refers to *Skip-connection*, *Embedding* refers to *Linear Embedding* [25], *Swin-block* denotes *Swin-Transformer Block* [25], *MH block* refers to Multi-head *self-attention block*, *BN* denotes *Batch Normalization*

paper also incorporates features from the CNN, the Swin-Transformer branch can benefit from the inductive bias of the CNN. Moreover, the influence of features at different scales on each stage depends on the specific stage. For example, as the first stage of the CNN branch at the lowest level, the output feature of this stage only contains information from the Swin-Transformer branch and does not include features from other scales. The fused information from both branches can directly serve as the residual input for the second stage. The output of the second stage includes not only information from the Swin-Transformer branch but also features from the first stage of the CNN branch. Similarly, the features of the final stage include information from the Swin-Transformer branch as well as features from the different scales in previous stages of the CNN. It is also applicable to the Swin-Transformer branch, where each stage contains the multiscale characteristics of the previous stage. The fusion of features from both branches, and how each stage utilizes multiscale features from previous stages, are the focus of this paper, which will be detailed in the following sections.

3.2 Structure of FFB

The primary motivation of this module is to achieve initial aggregation of features from the CNN and Swin-Transformer branches, serving as a crucial module for the interaction of information between the two branches. It enables the features learned by CNNs, which possess inductive biases, to propagate through this module and influence the Swin-Transformer. Conversely, it also facilitates the integration of the features learned by the Swin-Transformer branch, which exhibits global modelling capabilities, with the features learned by CNNs. As a result, the entire encoder composed of these two branches achieves a balance between local and global features through complementary processes. As shown in Fig. 2, the FFB is designed to integrate features from two different branches. There are two inputs to this module, namely, the feature information from the CNN and the Swin-Transformer branches. Since the network architecture ensures that the feature map sizes of both branches are consistent at the corresponding stages, the two inputs can be concatenated initially. After passing through a *conv_block*, the result is denoted as F^1:

$$F^1 = \text{Conv\_Block}\left(\text{Concat}\left(F_{CNN}, F_{Swin}\right)\right), \tag{1}$$

where *conv_block* includes two layers of 3 × 3 convolution and one layer of 1 × 1 convolution. F^1 undergoes downsampling as needed, with the downsampling factor being $2^{(i-1)}$, where i is the *i-th* stage. As a result, we can obtain the representation of M as shown in (2):

$$M = \text{Reshape}\left(\text{Down}\left(F^1\right) + P\left(\text{Down}\left(F^1\right)\right)\right), \tag{2}$$

where P represents positional encoding [12]. Adding positional information can mitigate the network's disruption of distance awareness and assist the subsequent attention mechanism in maximizing its effect. Then, we can obtain F^2:

$$F^2 = Up\left(\text{Linear}\left(\text{Concat}\left(M_1', M_2', \ldots, M_j'\right)\right)\right), \tag{3}$$

where Up refers to upsampling, with the purpose of obtaining features consistent with the size of the input feature. In this paper, $j \in (1, 2, 3)$, M_j' in (3) satisfies (4):

$$M_j' = \text{Reshape}\left(\delta\left(\frac{\left(W_{Qj}M\right)\left(W_{Kj}M\right)^T}{\sqrt{D}}\right)\left(W_{Vj}M\right)\right), \tag{4}$$

where W represents learnable parameters and D aligns with the dimensions of $(W_{Vi}M)$. The output of the FFB is eventually expressed as follows:

$$F = F^1 + F^2, \tag{5}$$

where "+" denotes elementwise addition. The output of this module serves two purposes. First, it provides residual information for both the CNN and Swin-Transformer branches. This enables the CNN branch to compensate for the global information from the Swin-Transformer branch, and allows the Swin-Transformer branch to benefit from the inductive

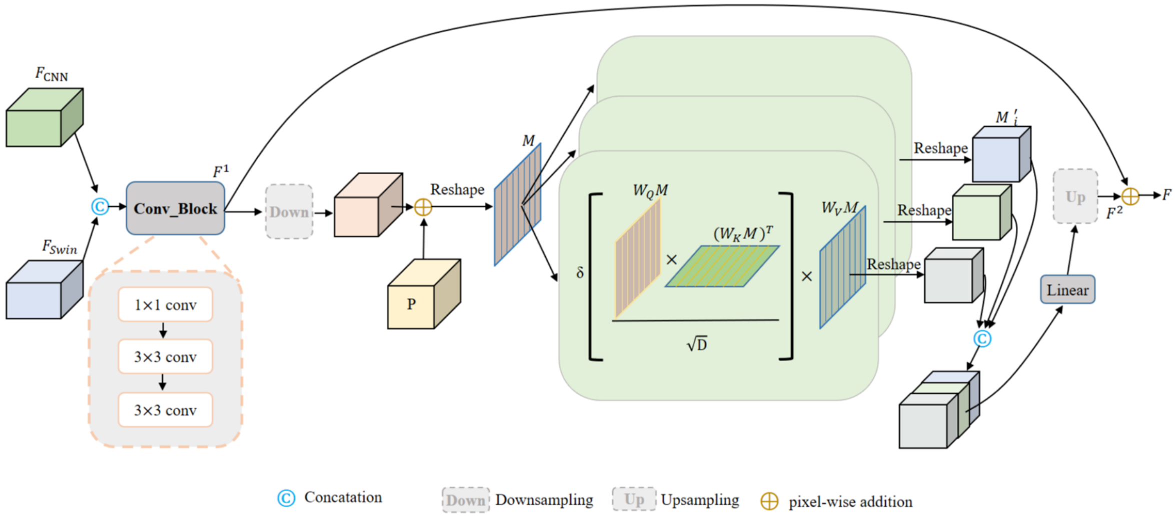

Fig. 2 Overview of the feature fusion block (FFB) for fusing features from different branches

bias of the CNN. Second, because the fused feature information, F, is more comprehensive, it provides guidance for each branch when fusing their respective multiscale features.

3.3 Structure of the MSFA block

The proposed MSFA module in this paper primarily addresses two issues. First, as the feature map scales vary across multiple stages, how can we effectively utilize and leverage the advantages of features at different scales? Second, how should the combined features be aggregated? Unlike traditional approaches for fusing multiscale features, the aggregation module of this paper operates at the pixel level, and the contributions of pixels at different positions within each layer to the next stage vary. Moreover, these contributions are learned through the network itself.

The most common approach to utilizing multiscale features is to upsample low-resolution features to the same size and concatenate them. However, this approach has limitations, as it introduces positional biases for features corresponding to the same object across different scales, which may be unfavourable for segmentation tasks. In this paper, when using multiscale information, the feature maps are aligned by considering features at the same relative positions across different scales. Specifically, to achieve this, the feature maps at different scales are first normalized in terms of their coordinates. For instance, in Fig. 3, P_q corresponds to the normalized coordinates of the reference point p. Despite the varying sizes of the feature maps from different stages, the same relative position after coordinate normalization represents the same object feature. This achieves the goal of aligning the feature maps. Coordinate normalization offers an advantage for feature maps with different scales, which minimizes the influence of the scale and position changes introduced by downsampling.

After determining the approach for obtaining multiscale features, the second issue is how to aggregate them. The MSFA module shown in Fig. 3 addresses this issue. The inputs to this module are variable, as the number of input feature stages increases with the deepening of the feature extraction network. However, these inputs can be divided into two categories, the outputs of the FFB module and the multiscale features. During the multiscale feature fusion in the $N-th$ stage, the output from the FFB module of the $(N-1)-th$ stage aligns with the feature map size of $stage_{N-1}$. The features of $stage_N$ first need to be upsampled to match the size of $stage_{N-1}$, and then concatenated with the output from the FFB module. Subsequently, $Conv\_Block$ is applied to obtain F_{fusion} . F_{fusion} learns the weights indicating the contribution of each pixel in the different scale features

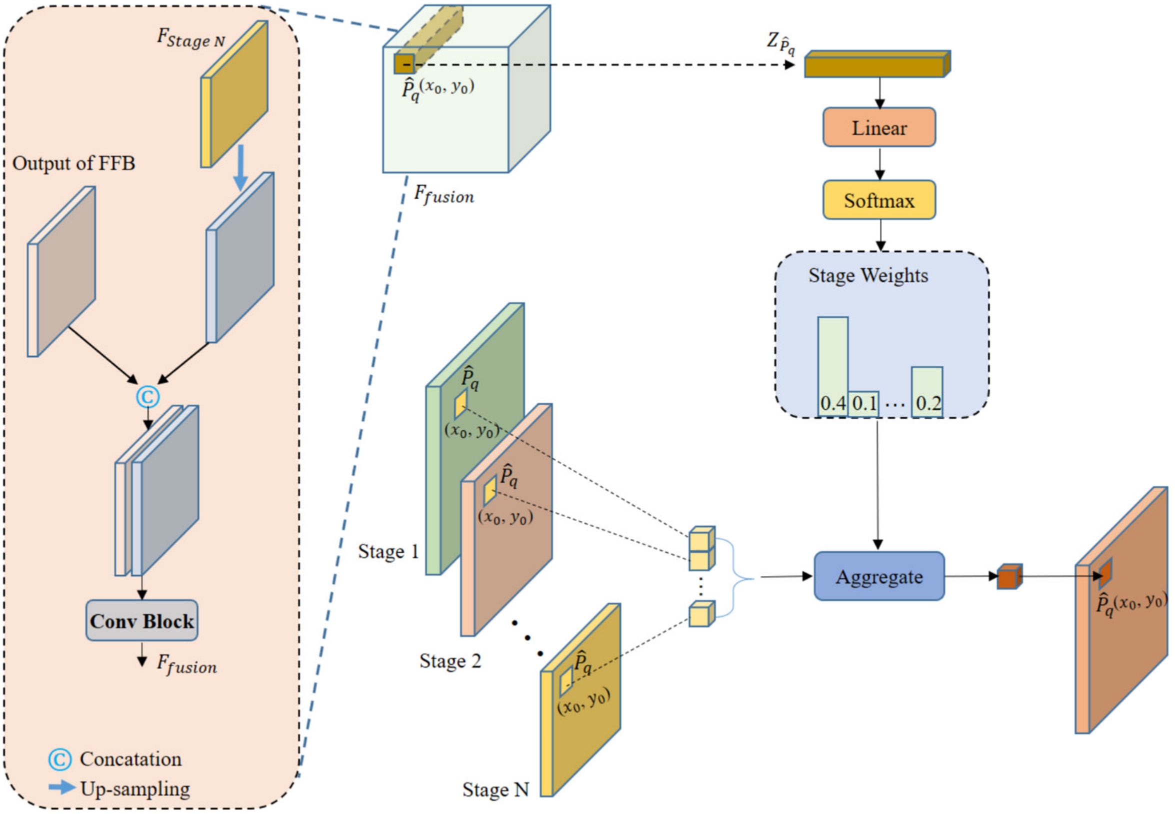

Fig. 3 Overview of the multiscale feature aggregation (MSFA) block

for the next stage. Specifically, after coordinate normalization, considering a reference point $\hat{\mathrm{P}}_{\mathrm{q}}$ at coordinates (x_0, y_0) in the feature maps, the set of this reference point in different channels is denoted as $Z_{\hat{P}_q}$. This vector passes through a linear layer and a softmax layer to generate attention weights. The sum of all the attention weights is 1, and the number of attention weights matches the number of multiscale feature maps. In Fig. 3, $stage_1$, $stage_2$, ... and $stage_L$ correspond to the multiscale feature maps at the different stages of the CNN or Swin-Transformer backbone network. Let $\left\{x^l\right\}_{l=1}^{L}$ be the input multiscale feature maps, where $x^l \in \mathbb{R}^{C\times H_l \times W_l}$. Given a reference point $\hat{P}_q$ at the same relative position in the different scale feature maps, the aggregated output at that point can be represented as:

$$\mathrm{MSFA\_Attn}\left(\mathrm{Z}_{\hat{\mathrm{P}}_q}, \hat{\mathrm{P}}, \left\{x^l\right\}_{l=1}^{L}\right) = \sum_{l=1}^{L} A_{lq} W x^l \left(\emptyset_l\left(\hat{\mathrm{P}}_q\right)\right). \tag{6}$$

where l indexes the stage of the input feature. A_{lq} denotes the attention weight of the q^{th} reference point at the l^{th} feature level. The scalar attention weight A_{lq} is normalized by $\sum_{l}^{L} A_{lq} = 1$. $\hat{\mathrm{P}}_q \in [0, 1]^2$ is the normalized coordinate of reference point p. $\emptyset_l(\bullet)$ is the function that rescales the normalized coordinates $\hat{\mathrm{P}}_q$ to the input feature map of the l-th stage.

Because deeper stages of features possess richer semantic information, while shallow layers of the feature maps contain more texture features, the accuracy of object boundary segmentation in segmentation tasks often depends on the texture features. However, these texture features are often massively discarded during the CNN feature extraction process. Therefore, through the designed aggregation module in this paper, the CNN network can compensate for texture information from the Swin-Transformer branch, and likewise, the Swin-Transformer can benefit from the inductive bias of the CNN branch, mitigating overreliance on the training dataset. This reciprocal compensation and enhancement between the two branches is achieved. Additionally, a novel multiscale feature fusion module is proposed in this paper. This module utilizes the self-learning capability of the network to provide attention coefficients for aggregating multiscale features, enabling pixel-level differentiated aggregation.

4 Experiments

This section begins with the introduction of three datasets, namely, Synapse, ISIC 2018 and ACDC. Then, the experimental implementation details are presented. To validate the effectiveness of each component in the proposed Swin-IBNet, ablation experiments are conducted. Finally, a comparison with the state-of-the-art methods is provided to further substantiate the efficacy of the proposed method.

4.1 Data preparation and implementation details

The Synapse dataset consists of 3779 axial contrast-enhanced abdominal CT image volumes from 30 subjects, including eight abdominal organs, which are the aorta (AT), gallbladder (GB), left kidney (KL), right kidney (KR), liver (LV), pancreas (PC), spleen (SP) and stomach (SM). Each CT scan consists of 85-198 slices with a size of 512 × 512 pixels. Following the settings of [33], CT images of 18 subjects were used as training samples, and CT images of 12 subjects were used as test samples. The ISIC 2018 dataset [62] contains 2594 original dermoscopy images and 2594 corresponding binary masks. A uniform size operation is necessary before using these samples. During the experiment, 80% of the samples are used for training, 10% for verification and the remaining 10% for testing. The ACDC dataset is a cardiac 3D magnetic resonance imaging dataset that consists of 100 cardiac MR images collected from different patients. Each scan includes the right ventricle (RV), left ventricle (LV) and myocardium (Myo). Following the settings of [33], 70 cases are used for training, 10 cases for verification and 20 cases for testing.

The proposed Swin-IBNet is implemented in a PyTorch library using an NVIDIA RTX 3090 (24 GB) GPU. The training and testing are based on the Ubuntu 18.04 system with four NVIDIA graphics cards. In our experiments, we adopt the SGD optimizer with a momentum of 0.9, and the initial learning rate is set to 0.05 with a weight decay of 0.1 to train the model. The batch size in our experiments is 24. In all the experiments, the resolutions of the training, validation and test samples are resized to 224 × 224. The data augmentation methods used in this paper are the same as those of Swin-Unet [33]. These methods include random flip, rotation, Gaussian noise, scaling and contrast transformation.

4.2 Experimental results on the synapse dataset, the ACDC Dataset and the ISIC 2018 dataset

4.2.1 Experimental results on the synapse dataset

This section primarily reports experiments on the Synapse dataset and compares the results with those of the current state-of-the-art methods. The experimental results are shown in Table 2. Table 2 includes comparisons with representative CNN-based methods such as U-Net and Att-UNet, and Transformer-based methods such as ViT, TransUnet, Swin-Unet, MISSFormer, MTMUnet, AFTer-UNet and TC-UNet. The bold data in Table 2 represent the best values for each respective class or metric. We observe that our method outperforms other methods in terms of the important overall

Table 2 Comparisons with other methods on the Synapse test set

Methods	DSC (%)	HD	AT (%)	GB (%)	KL (%)	KR (%)	LV (%)	PC (%)	SP (%)	SM (%)
U-Net [9]	76.26	36.62	86.95	64.76	80.98	75.06	93.09	50.24	86.05	72.93
Att-UNet [63]	78.07	32.10	88.41	**69.45**	83.14	74.55	93.46	54.80	86.61	74.16
R50 ViT [32]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [32]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [33]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MTMUnet [64]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
TransUNet+ [19]	81.57	–	88.70	67.57	82.48	81.42	94.20	65.73	90.55	81.95
AFTer-UNet [65]	81.02	–	**90.91**	64.81	**87.90**	**85.30**	92.20	63.54	90.99	72.48
TC-UNet [66]	78.09	–	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
CoTr [67]	78.46	–	87.06	63.65	82.64	78.69	94.06	57.86	87.95	75.74
MISSFormer [68]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	**91.92**	80.81
Swin-IBNet	**82.58**	**17.46**	87.96	67.99	84.12	84.64	**94.51**	**66.64**	91.12	**83.68**

The best values are marked in bold.

DSC metric, achieving 82.58%. Additionally, our method exhibits outstanding performance in measuring the HD metric, with a value of only 17.46. For relatively small targets, such as the aorta and gallbladder, the CNN-based methods still have a significant advantage. For instance, U-Net and Att-UNet achieve accuracies of 86.95% and 88.41% on Aorta, respectively, and Att-UNet even achieves the highest score of 69.45% on Gallbladder. This is mainly attributed to the CNN's ability to perform local feature extraction. However, it cannot capture the dependence of longer distances, which leads to an unsatisfactory segmentation effect for larger targets. The results in Table 2 also show that pure Transformer-based methods such as ViT, TransUnet and Swin-Unet perform relatively well in modelling larger targets globally. However, their lack of local modelling and application of multiscale features limits their overall performance. Recently, proposed methods such as AFTer-UNet, TC-UNet, CoTr and MISSFormer have all been introduced as improvements to address this issue. The Swin-IBNet model proposed in this paper is also designed to address this issue. Among these methods, Swin-IBNet, not only outperforms others in terms of the overall DSC metric but also achieves higher scores on relatively smaller targets such as Aorta and Gallbladder, surpassing TransUnet and Swin-Unet. Moreover, compared to advanced methods such as TC-UNet (Aorta 85.87%, Gallbladder 61.38%) and CoTr (Aorta 87.06%, Gallbladder 63.65%), Swin-IBNet also shows a superior performance. Most importantly, for the liver, pancreas and stomach, we achieved scores of 94.51%, 66.64% and 83.68%, respectively, surpassing all the other methods in Table 2. In summary, this experiment further demonstrates that Swin-IBNet, by combining the strengths of a CNN and a Transformer, can enhance the segmentation capability for medical images.

In addition, we visualize the segmentation results of various organs using different comparative methods in Fig. 4. Figure 4 is generated using the cancer imaging phenomics toolkit (CaPTK) tool [69] to present three-dimensional renderings with depth information loaded, showcasing segmentation outcomes from multiple angles for organs. We compare our method with the reimplemented Swin-Unet, U-Net and the recent MISSFormer method. Due to potential overlaps between certain organs, for clarity in displaying the segmentation results, we group the observations as outlined in Table 3. The segmentation result visualizations in Fig. 4 are presented based on these groupings, with each group including visualizations from both an axial and sagittal perspective. From Fig. 4, it can be observed that the organs segmented using our method are closer to the ground truth, particularly in the liver, pancreas and stomach.

4.2.2 Experimental results on the ACDC dataset

To further demonstrate the effectiveness of Swin-IBNet, this study conducted an additional validation on the ACDC dataset. Table 4 presents comparisons with relevant methods on the test set. The results in Table 4 reveal that our Swin-IBNet achieves a Dice similarity coefficient (DSC) of 91.54%. Notably, accuracies of 89.12% for Myocardium and 95.63% for Ventricle (L) are obtained. Compared to recent outstanding methods such as MISSFormer and MTMUnet, our approach exhibits improvements, further substantiating the efficiency of the proposed method.

Furthermore, to provide a clearer visualization of the segmentation performance of Swin-IBNet on the ACDC dataset, we visualize the segmentation results as shown in Fig. 5. We compare these results with those of the reimplemented Swin-Unet, U-Net and the recent work MISSFormer. It can

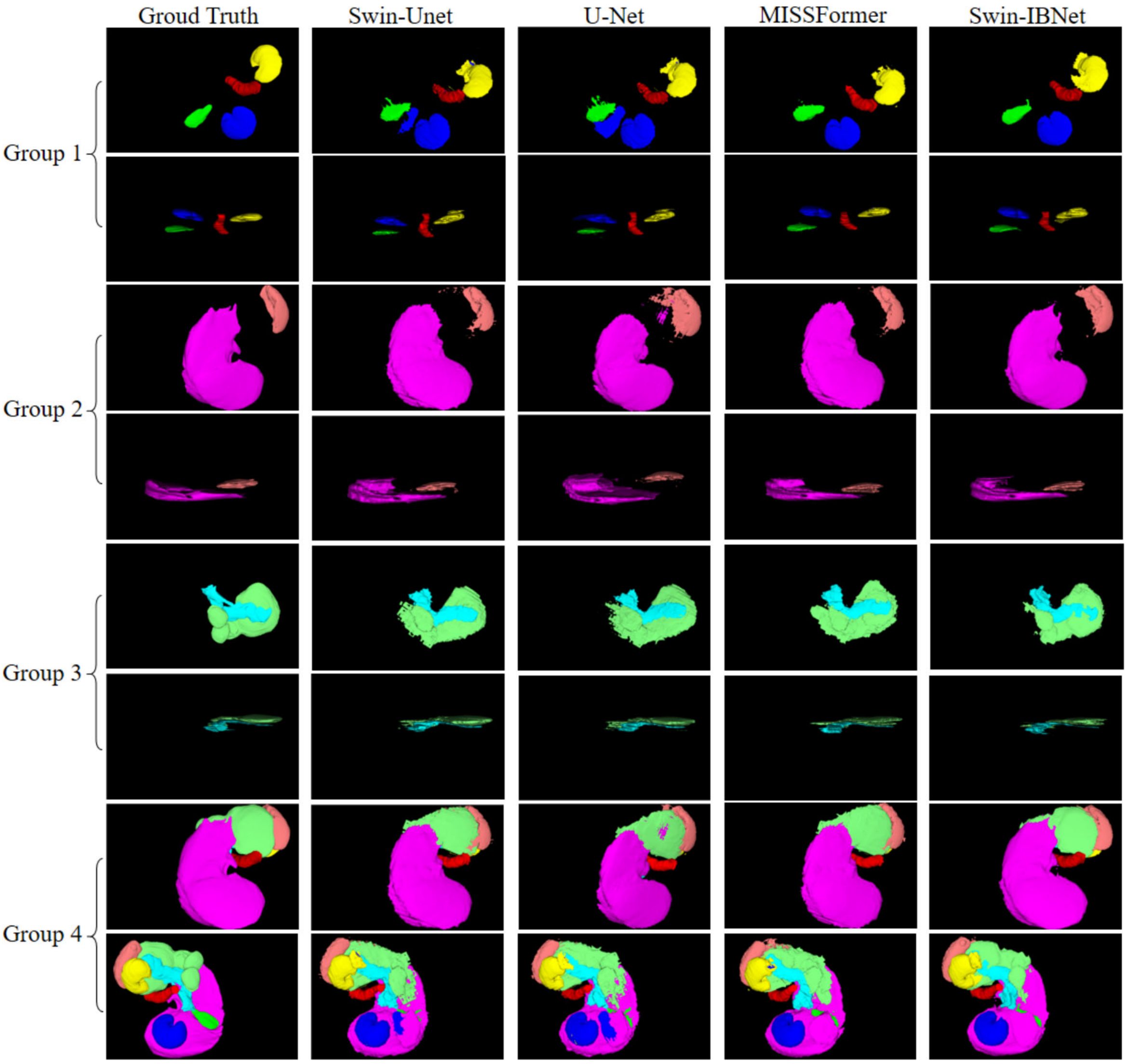

Fig. 4 Visualization of segmentation results on the Synapse test set

be observed that U-Net performs relatively insufficiently in segmenting larger targets, such as the red segmented region depicted in the second row of Fig. 5. This is because U-Net lacks long-range modelling capabilities. Conversely, our method demonstrates superior segmentation accuracy, as shown in the last row.

Table 3 Groups of segmentation targets for no-overlap visualization

Groups	AT	GB	KL	KR	LV	PC	SP	SM
Group 1	✓	✓	✓	✓	X	X	X	X
Group 2	X	X	X	X	✓	X	✓	X
Group 3	X	X	X	X	X	✓	X	✓
Group 4	✓	✓	✓	✓	✓	✓	✓	✓

4.2.3 Experimental results on the ISIC 2018 dataset

We conduct further experiments on the ISIC 2018 dataset. The results presented in Table 5 indicate that Swin-IBNet achieves a Dice similarity coefficient (DSC) of 90.7%, an average intersection over union (mIoU) of 84.3%, a recall of 88.93% and a precision of 93.48%. Compared with Swin-Unet, Swin-IBNet outperforms it by 2.4% in terms of DSC. Furthermore, compared to the recent methods, MSRF-Net and DCSAU-Net, Swin-IBNet achieves DSC improvements of 2.46% and 0.3%, respectively. In terms of recall, our method surpasses these two approaches by 3.67% and 0.4%, respectively, highlighting our model's ability to accurately capture the actual boundaries of skin lesions. Additionally, we visualize the segmentation results in Fig. 6.

Table 4 Comparisons with other methods on the ACDC test set (%)

Methods	DSC	RV	Myo	LV
R50 U-Net [9]	90.16	88.67	86.88	94.92
R50Att-UNet [63]	91.03	**89.90**	87.98	95.21
TransUnet [32]	90.44	88.82	87.54	94.95
Swin-Unet [33]	90.41	88.41	87.71	95.13
MISSFormer [68]	91.19	89.85	88.38	95.34
MTMUnet [64]	90.43	86.64	89.04	95.62
TransUNet+ [19]	90.42	89.15	87.98	94.12
UNet++ [70]	87.97	86.36	85.66	91.88
ViT-CUP [24]	87.57	86.07	81.88	94.75
Swin-IBNet	**91.54**	89.89	**89.12**	**95.63**

The best values are marked in bold.

The visualization results in Fig. 6 reveal that U-Net's segmentation results for certain complex samples are suboptimal due to its limited long-range modelling capability. In the third column, the segmentation results of U-Net include only a fraction of the central features. Despite Swin-Unet having a slightly lower overall DSC value compared to DoubleU-Net, it also performs well in terms of the visualization results for these samples. The last column in the middle row displays the segmentation results of our proposed method. From Fig. 6, it is evident that Swin-IBNet's result aligns closely with the ground truth. We attribute this to the combination of incorporating FFB and MSFA into the fusion of the CNN and the Swin-Transformer. This approach not only equips the network with long-range modelling capabilities but also enhances the extraction of local features. We also observe the presence of hair in the ground truth of the sample in row three. This variation arises from differences among physicians in annotating the samples, which presents a significant challenge in medical image research.

4.3 Analysis of the interpretability

In this section, we adopt a visual approach to study the interpretability of Swin-IBNet. We explore the features extracted by the two branches and the role of our modules from a visualization perspective. During training, we record the features from the CNN branch, the Swin branch and the output features, after passing through our modules every 20 iterations.

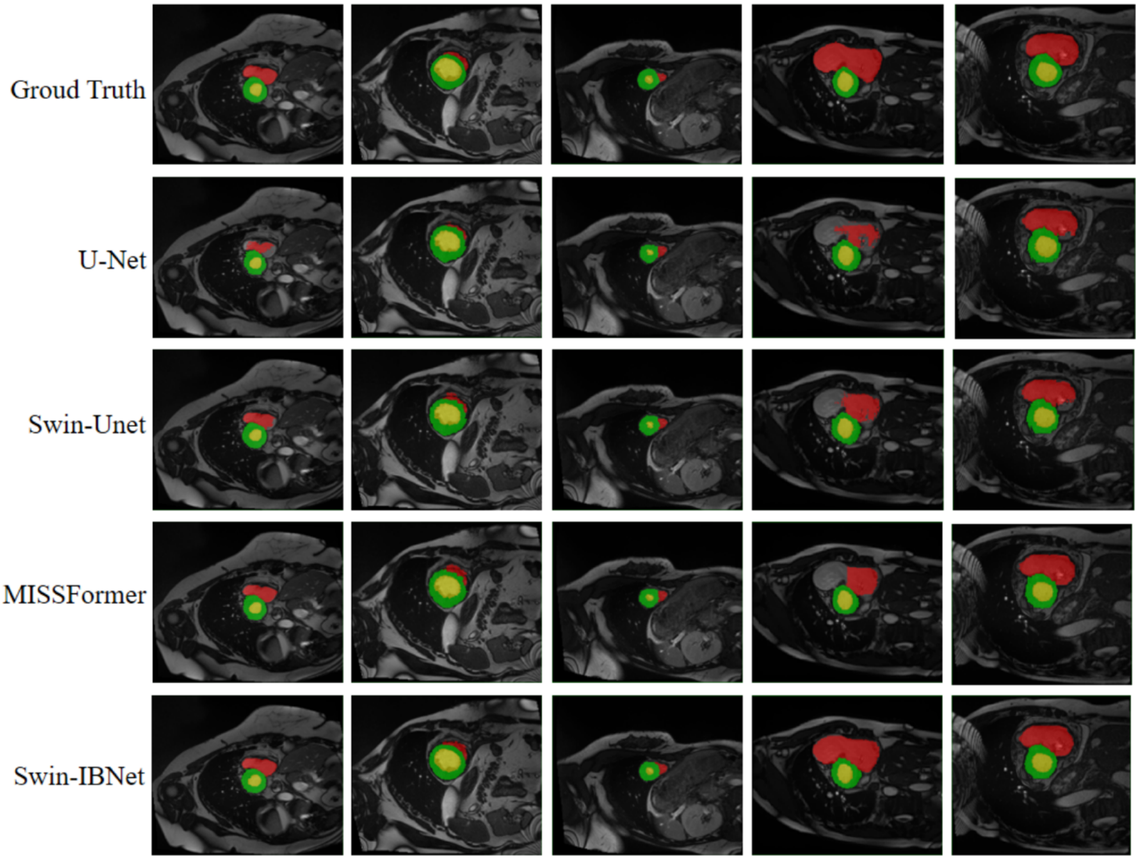

Fig. 5 Visualization of the segmentation results on the ACDC test set

Table 5 Comparisons with other methods on ISIC 2018 test set

Methods	DSC (%)	mIou (%)	Recall (%)	Precision (%)
U-Net [9]	85.54±18.48	78.47±20.94	82.04±21.86	**94.74±12.96**
UNet++ [70]	80.94±22.61	72.88±24.52	78.66±23.69	90.84±22.22
ResUNet++ [71]	85.57±20.14	81.35±22.10	88.01±23.20	86.76±15.62
Deeplabv3_Xcep [8]	87.72±14.65	81.28±18.06	86.81±17.92	92.72±13.602
Deeplabv3_Mob [8]	87.81±13.71	82.36±17.11	88.30±17.25	92.44±13.17
HRNetV2-W48 [72]	86.67±24.53	81.09±26.30	85.84±29.36	91.55±27.55
DoubleU-Net [73]	89.38±13.62	82.12±16.59	87.80±15.73	94.59±13.53
ResUNet++_CRF [74]	86.88±17.19	82.09±19.71	88.26±20.63	87.36±15.40
MSRF-Net [75]	88.24±16.02	83.73±18.18	88.93±18.89	93.48±14.88
Swin-Unet [33]	88.3±14.0	81.1±16.6	90.3±13.4	90.0±15.3
TransUnet [32]	84.9±17.8	77.0±20.3	89.8±18.5	84.7±18.6
LeViT-U [76]	88.3±16.1	81.7±18.5	90.8±17.6	89.6±15.2
DCSAU-Net [77]	90.4±12.8	84.1±15.8	92.2±13.9	91.7±12.7
Swin-IBNet	**90.7±13.6**	**84.3±16.1**	**92.6±14.0**	93.6±13.1

The best values are marked in bold.

We randomly select images from different stages of a specific iteration for visualization, as shown in Fig. 7. With the increase in the number of stages, the size of the images shown in Fig. 7 is gradually reduced by half. To ensure a clearer visualization, the images are appropriately enlarged in this study. From Fig. 7, it can be observed that the CNN branch and Swin branch can extract different features from the input images. Meanwhile, the feature maps obtained after passing through the proposed Swin-IBNet demonstrate that the modules proposed in this study achieve complementary feature extraction capabilities from both the CNN and Swin branches.

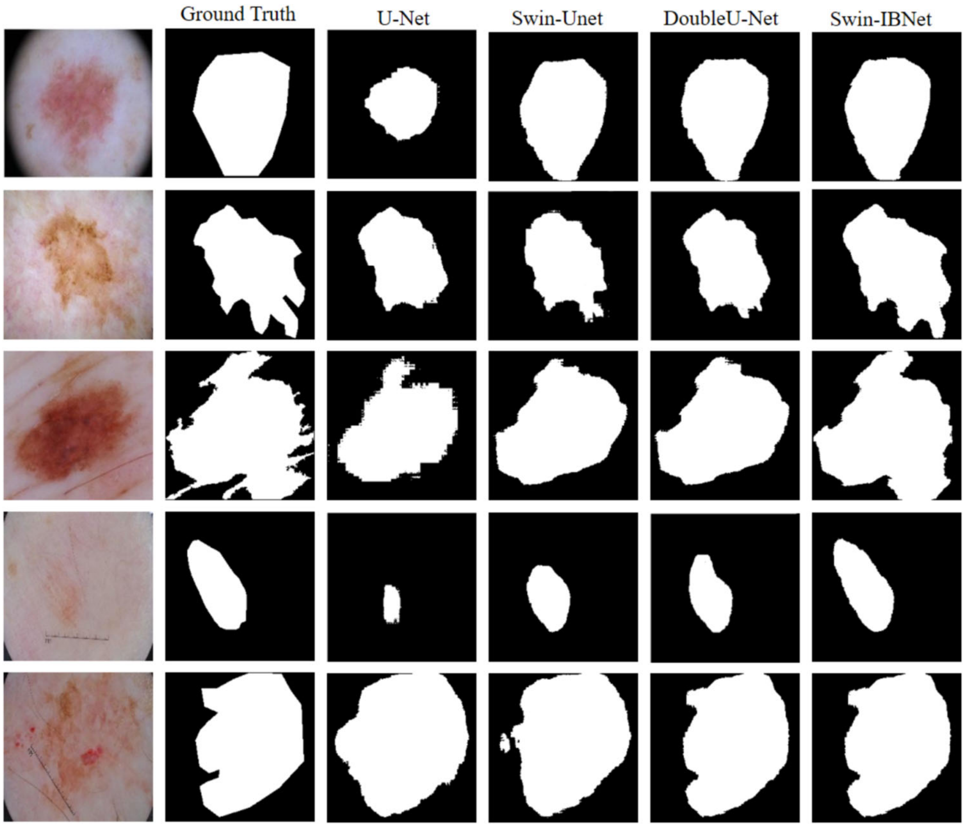

Fig. 6 Visualization of the segmentation results on the ISIC 2018 test set

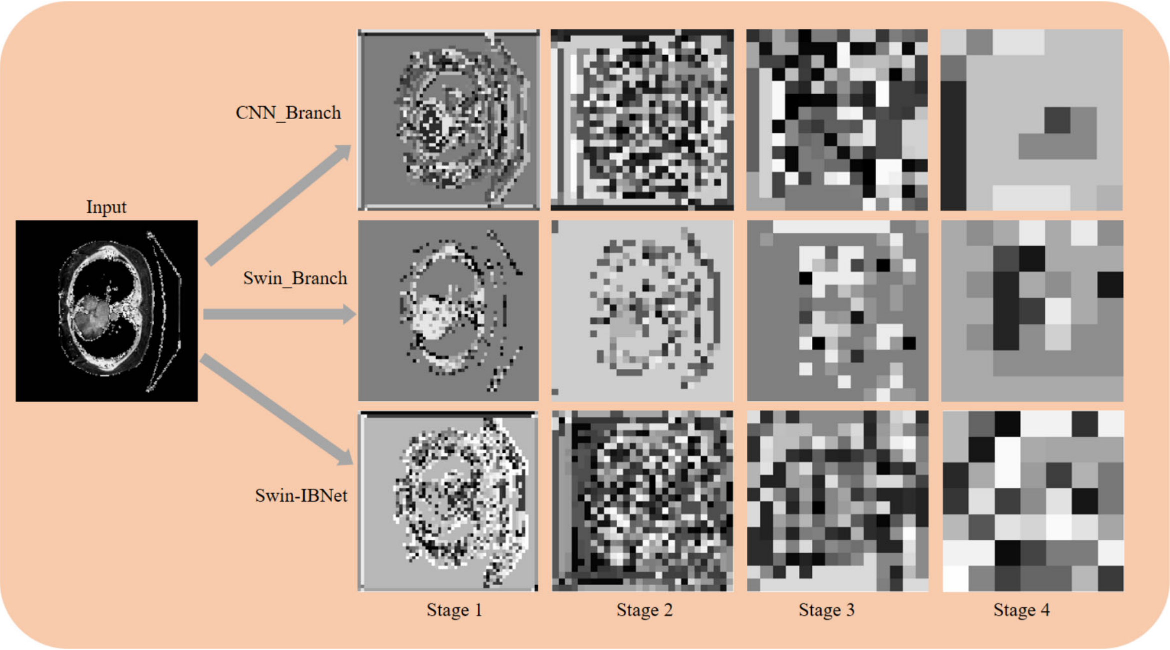

Fig. 7 Visualization of feature maps extracted from the CNN branch, Swin branch and Swin-IBNet

4.4 Ablation studies

This study conducts ablation experiments on the Synapse and ACDC datasets to thoroughly validate the effectiveness of the proposed Swin-IBNet and its key components. In this experiment, the number of stages of the two branches in Swin-IBNet is set to 4. The CNN branch employs ResNet50 as its backbone. Swin-UNet and a ResNet50-based U-Net are used as the baseline methods for comparison. To ensure a fair comparison, all the parameter settings are kept consistent. The evaluation metrics used for comparison include the Dice similarity coefficient (DSC) and Hausdorff distance (HD).

4.4.1 Ablation studies of the CNN branch, Swin-Unet branch and Swin-IBNet

The experiments in this section focus on individually using the CNN branch or the Swin-Unet branch to analyse their impact on the performance of medical image segmentation, thereby validating the necessity of their combination. All three networks have the same decoder, which consists of the decoding part of Swin-Unet. The first network utilizes a CNN encoder obtained by removing the Swin-branch part of Swin-IBNet, leaving only the remaining model with a ResNet50 backbone. The second network is the baseline model Swin-Unet [25], which is obtained by removing the CNN encoder with inductive bias from Swin-IBNet. The third network is the Swin-IBNet network proposed in this study. All three networks are validated on the Synapse and ACDC datasets, and the experimental results are presented in Table 6.

From the experimental results in Table 6, it can be observed that using only one branch results in a limited ability to enhance medical image segmentation. However, the proposed Swin-IBNet, in this study, demonstrates the ability to improve medical image segmentation by complementing the two branch networks.

4.4.2 Ablation Studies of FFB and MSFA

The experiments in this section are different from those in Section 4.2.1. In Section 4.2.1, the experiments involved using the CNN branch and Swin-Unet branch separately as individual encoders. Section 4.2.2 primarily validates the functions of FFB and MSFA and needs to fuse the CNN branch and the Swin branch together. If the features from the CNN branch and the Swin branch are simply concatenated, their respective advantages cannot be fully utilized. Therefore, in this study, an FFB is proposed that incorporates a

Table 6 Ablation studies of the CNN encoder alone and the Swin-Transformer encoder alone

	Synapse		ACDC
Branches	DSC (%)	HD	DSC (%)
CNN(U-Net_resnet50)	76.15	35.24	90.05
Swin-Unet	79.13	21.55	90.41
Swin-IBNet	82.58	17.46	91.54

multihead self-attention block to learn the weighting coefficients of the features from two branches to obtain local or global information. On the other hand, MSFA is designed in a novel manner to aggregate multiscale features. The experiments in this section are designed in the following way, without FFB and MSFA, with FFB but without MSFA, without FFB but with MSFA and with both FFB and MSFA. Table 7 presents the comparative results of these experiments on the Synapse and ACDC datasets.

The results shown in Table 7 indicate that even without using either the FFB or MSFA modules, the obtained results outperform those of Swin-Unet (79.13% in Table 6) and U-Net (76.15% in Table 6) on both the Synapse and ACDC datasets. It achieves DSC values of 80.01% and 90.48%, respectively. This suggests that the combination of a CNN and a Swin-Transformer can enhance the network's performance. We speculate that the network has learned the crucial factors from both the CNN branch and the Swin-Transformer branch that significantly impact the final segmentation results. By using only single-scale features at each stage (i.e., not using MSFA but adopting FFB), DSC values of 80.93% and 90.50% are achieved on the two datasets. Compared to the concatenation approach, the use of FFB leads to 0.92% and 0.02% improvements on the two datasets, respectively. MSFA is mainly used to fuse multiscale features. In contrast, W/O MSFA refers to using only the single-scale features of each stage, without considering the multiscale features of previous stages. From Table 7, it can be observed that using MSFA without FFB yields DSC values of 81.27% and 90.82% on the Synapse and ACDC datasets, respectively. Compared to using single-scale features, significant improvements are achieved in both accuracy and positional accuracy. This demonstrates that this study's proposed MSFA can greatly enhance the performance of medical image segmentation tasks. The most remarkable achievement is that, when FFB and MSFA are used simultaneously, the DSC values significantly improve to 82.58% and 91.54% on the Synapse and ACDC datasets, respectively. This is a significant improvement compared to the performance of Swin-Unet, which achieves DSC values of 79.13% and 90.41%. Particularly on the Synapse dataset, the proposed approach outperforms Swin-Unet by 3.45% in terms of DSC. In the above experiments, the gradual addition of the two designed modules consistently improved the network's performance in medical image segmentation tasks, which strongly supports the view that FFB and MSFA play a significant role in enhancing the network's performance in medical image segmentation.

Table 7 Ablation studies on Feature Fusion methods: Comparing Concatenation and Feature Fusion Module (FFB), and evaluating Single-Scale Feature versus Multi-Scale Feature Aggregation (MSFA)

	Synapse		ACDC
Combinations of FFB or MSFA	DSC (%)	HD	DSC (%)
W/O FFB and W/O MSFA	80.01	20.05	90.48
W FFB and W/O MSFA	80.93	19.61	90.50
W/O FFB and W MSFA	81.27	19.02	90.82
W FFB and W MSFA	82.58	17.46	91.54

5 Conclusion

In this paper, we propose Swin-IBNet, a network that combines the Swin-Transformer with a CNN in a novel manner to imbue it with inductive bias capabilities. The encoder of Swin-IBNet includes two novel and crucial modules, FFB and MSFA. The FFB is responsible for propagating the inductive bias capability to the Swin-Transformer encoder, and it initially fuses features from different branches, providing guidance signals for MSFA, and complementary information for each branch. Unlike conventional multiscale feature usage, our approach utilizes different scales of information at various stages in the feature extraction process, providing a novel way for the network to acquire contextual and global information. The proposed network possesses both the global modelling capability from the Transformer, and inductive bias from the CNN, enabling the network to be less dependent on the size of the dataset. We not only attempt to analyse the interpretability of the proposed Swin-IBNet but also perform more verifications on the public Synapse, ISIC 2018 and ACDC datasets. The experimental results validate the Swin-IBNet, and it outperforms other state-of-the-art methods on the three datasets.

Although there has been some progress in applying Transformer-based methods to computer vision tasks, the issue of computational complexity remains unresolved thoroughly. This challenge may require improvements in hardware performance and the development of more efficient methods. Given the current limitations in hardware performance, all Transformer-based methods applied to computer vision tasks inevitably need to address computational complexity. Currently, the Swin-Transformer method offers the most effective solution to this problem. This paper also conducts research based on this method. However, upon closer analysis, it becomes apparent that the Swin-Transformer method compromises to some extent between computational complexity and global modeling capability. In fact, compared to the original Transformer, its global modeling capability is somewhat diminished. Therefore, further exploration is needed to preserve global modeling capability to a greater extent while reducing computational complexity. Additionally, the combination of Transformer-based methods with CNNs, which lack inductive bias, still holds potential for exploration.

Acknowledgements This work was financially supported by the National Natural Science Foundation of China under Grant 62172184, the Science and Technology Development Plan of Jilin Province of China under Grants 20200401077GX and 20200201292JC and the Jilin Educational Scientific Research Leading Group (ZD21003, JS2338).

Author Contributions Yan Gao conceived and designed the research, conducted the experiments and authored the paper, which was followed by iterative revisions of the manuscript. Professor Xiangjiu Che acted as the corresponding author, conducting manuscript checks, providing valuable insights, and managed communications with the journal, including correspondence and responses. Huan Xu assisted in manuscript reviews and proofreading. Quanle Liu contributed to graphical content suggestions. Mei Bie participated in the review of the reference materials.

Data available statement All the data used in this work are publicly available to the researchers.

Declarations

Ethical and informed consent for data used Ethical and informed consent for data used.

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Liu C, Xie H, Zha Z, Yu L, Chen Z, Zhang Y (2019) Bidirectional attention-recognition model for fine-grained object classification. IEEE Trans Multimedia 22(7):1785–1795
2. Min S, Yao H, Xie H, Zha Z, Zhang Y (2020) Domain-oriented semantic embedding for zero-shot learning. IEEE Trans Multimedia 23:3919–3930
3. Min S, Yao H, Xie H, Zha Z, Zhang Y (2020) Multi-objective matrix normalization for fine-grained visual recognition. IEEE Trans Image Process 29:4996–5009
4. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inform Process Syst pp 91–99
5. Barroso-Laguna A, Mikolajczyk K (2022) Key. net: Keypoint detection by handcrafted and learned cnn filters revisited. IEEE Trans Pattern Anal Mach Intell 45(1):698–711
6. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proc IEEE ICCV, pp 2980–2988
7. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conf comput vis pattern recognit pp 3431–3440
8. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc 15th Eur conf pp 833–851
9. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention., pp. 234–241
10. Tian Y, Yang G, Wang Z et al (2020) Instance segmentation of apple flowers using the improved mask r-cnn model. Biosys Eng 193:264–278
11. Han Z, Jian M, Wang G-G (2022) Convunext: An efficient convolution neural network for medical image segmentation. Knowl-based Syst 253
12. Vaswani Aea (2017) Attention is all you need. Advances in neural information processing systems., 6000–6010
13. Wang W et al (2021) Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 568–578
14. Yang Y, Zhang L, Ren L, Wang X (2023) Mmvit-seg: A lightweight transformer and cnn fusion network for covid-19 segmentation. Comput Methods Programs Biomed 230:106365
15. Li X et al (2023) Attransunet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. Comput Biol Med 152:106365
16. Gao C, Ye H, Cao F, Wen C, Zhang Q, Zhang F (2021) Multiscale fused network with additive channel-spatial attention for image segmentation. Knowl-Based Syst 214:106754
17. Lin F, Liang Z, Wu S, He J, Chen K, Tian S (2023) Structtoken: Rethinking semantic segmentation with structural prior. IEEE Transactions on circuits and systems for video technology
18. Park K-B, Lee JY (2022) Swine-net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer. J Comput Des Eng 9(2):616–632
19. Liu Y, Wang H, Chen Z, Huangliang K, Zhang H (2022) Transu-net +: Redesigning the skip connection to enhance features in medical image segmentation. Knowl-Based Syst 256:109859
20. Tang P et al (2022) Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl-Based Syst
21. Qi M et al (2022) Ftc-net: Fusion of transformer and cnn features for infrared small target detection. IEEE Journal of selected topics in applied earth observations and remote sensing. 15:8613–8623
22. Gao G, Xu Z, Li J et al (2023) Ctcnet: A cnn-transformer cooperation network for face image super-resolution. IEEE Trans Image Process pp 1978–1991
23. Li W, Xue L, Wang X et al (2023) Convtransnet: A cnn-transformer network for change detection with multi-scale global-local representations. IEEE Trans Geosci Remote Sens 61
24. Dosovitskiy A et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations (ICLR)
25. Liu Z et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
26. Sun L, Zhao G, Zheng Y et al (2022) Spectral-spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–14
27. Hong D, Han Z, Yao J et al (2021) Spectralformer: Rethinking hyperspectral image classification with transformers. IEEE Trans Geosci Remote Sens 60:1–15
28. Touvron H, Bojanowski P, Caron M et al (2022) Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Trans Pattern Anal Mach Intell 45:5314–5321
29. Remote sensing image change detection with transformers (2021) Chen H, SZ. Qi Z. IEEE Trans Geosci Remote Sens 60:1–14
30. Li K, Wang Y, Zhang J et al (2023) Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Trans Pattern Anal Mach Intell 45:12581–12600
31. Li Y, Yao T, Pan Y et al (2022) Contextual transformer networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 45:1489–1500
32. Chen J al (2021) Transunet: Transformers make strong encoders for medical image segmentation. CoRR. **abs/2102.04306**, pp 1–13
33. Cao H al (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision (ECCV), pp 205–218
34. Wang L, Li R, Zhang C et al (2022) Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J Photogramm Remote Sens 190:196–214

35. Zhu Z, He X, Qi G et al (2023) Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. Inform Fusion 91:376–387
36. Yuan F, Zhang Z, Fang Z (2023) An effective cnn and transformer complementary network for medical image segmentation. Pattern Recogn 136:109228
37. Zhang C, Jiang W, Zhang Y et al (2022) Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–20
38. Ding W, Wang H, Huang J et al (2023) Ftranscnn: Fusing transformer and a cnn based on fuzzy logic for uncertain medical image segmentation. Inform Fusion 99: 101880
39. Zhao Z, Li Q, Zhang Z et al (2021) Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. Neural Netw 141:52–60
40. Mi Z, Jiang X, Sun T et al (2020) Gan-generated image detection with self-attention mechanism against gan generator defect. IEEE J Sel Top Signal Process 14:969–981
41. Zeng W, Li M (2020) Crop leaf disease recognition based on self-attention convolutional neural network. Comput Electron Agric 172:105341
42. Rao D, Xu T, Wu X (2023) Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. IEEE Trans Image Process
43. Yu H, Xu Z, Zheng K et al (2022) Mstnet: A multilevel spectral-spatial transformer network for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–13
44. Wu H, Zhang M, Huang P et al (2024) Cmlformer: Cnn and multi-scale local-context transformer network for remote sensing images semantic segmentation. IEEE J Sel Top Appl Earth Obs Remote Sens pp 1–10
45. Geng Z, Chen Z, Meng Q et al (2021) Novel transformer based on gated convolutional neural network for dynamic soft sensor modeling of industrial processes. IEEE Trans Industr Inf 18:1521–1529
46. Song R, Feng Y, Cheng W et al (2022) Bs2t: Bottleneck spatial-spectral transformer for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–17
47. Xie X, Wu D, Xie M et al (2024) Ghostformer: Efficiently amalgamated cnn-transformer architecture for object detection. Pattern Recogn 148:110172
48. Kang J, Guan H, Ma L et al (2023) Waterformer: A coupled transformer and cnn network for waterbody detection in optical remotely-sensed imagery. ISPRS J Photogramm Remote Sens 206:222–241
49. Wang C, Xu M, Jiang Y et al (2022) Translution-snet: A semisupervised hyperspectral image stripe noise removal based on transformer and cnn. IEEE Trans Geosci Remote Sens 60:1–14
50. Zhang Q, Xu Y, Zhang J et al (2023) Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. Int J Comput Vision 131:1141–1162
51. Sartran L, Barrett S, Kuncoro A et al (2022) Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. Trans Assoc Comput Linguist 10:1423–1439
52. Hao S, Li N, Ye Y (2023) Inductive biased swin-transformer with cyclic regressor for remote sensing scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens 16:6265–6278
53. Graham B et al (2021) Levit: A vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 12259–12269
54. Zhang Q, Yang Y-B (2021) Rest: An efficient transformer for visual recognition. Adv Neural Inform Process Syst 34:15475–15485
55. Heo B, Yun S, Han D, Chun S, Choe J, Oh SJ (2021) Rethinking spatial dimensions of vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 11916–11925
56. Zhang Z et al (2022) Nested hierarchical transformer: Towards accurate data-efficient and interpretable visual understanding. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 3417–3425
57. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognitio (CVPR), pp 2117–2125
58. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1440–1448
59. Wang W et al (2022) Pvtv 2: Improved baselines with pyramid vision transformer. Comput Vis Media 8(3):1–10
60. Xu W, Xu Y, Chang T, Tu Z (2021) Co-scale conv-attentional image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 9981–9990
61. Chen C-F, Fan Q, Panda R (2021) Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer visio (ICCV), pp 357–366
62. Codella NCF, Gutman D, Celebi ME et al (2018) Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018). IEEE, pp 168–172
63. Fu H, Xu Y, Lin S, Wong DWK, Liu J (2016) Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In: Medical image computing and computer-assisted intervention–MICCAI 2016: 19th international conference, pp 132–139
64. Wang H et al (2022) Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2390–2394
65. Yan X, Tang H, Sun S, Ma H, Kong D, Xie X (2022) After-unet: Axial fusion transformer unet for medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3971–3981
66. Chang Y, Menghan H, Guangtao Z, Xiao-Ping Z (2022) Transclaw u-net: claw u-net with transformers for medical image segmentation. In: 2022 5th IEEE International conference information communication signal processing (ICICSP), pp 280–284
67. Xie Y, Zhang J, Shen C, Xia Y (2021) Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, pp 171–180
68. Huang X, Deng Z, Li D, Yuan X, Fu Y (2022) Missformer: An effective transformer for 2d medical image segmentation. IEEE Trans Med Imaging
69. Center for Biomedical Image Computing & Analytics. https://www.med.upenn.edu/cbica/captk. Accessed 16 Sept 2023
70. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp 3–11
71. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2020) Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39(6):1856–1867
72. Wang J et al (2020) Deep high-resolution representation learning for visual recognition. IEEE Transactions on pattern analysis and machine intelligence, pp 5686–5696

73. Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD (2020) Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS), pp 558–564
74. Jha D et al (2021) A comprehensive study on colorectal polyp segmentation with resunet++ conditional random field and test-time augmentation. IEEE J Biomed Health Inform 25(6):2029–2040
75. Srivastava A et al (2022) Msrf-net: A multi-scale residual fusion network for biomedical image segmentation. IEEE J Biomed Health Inform 26(5):2252–2263
76. Xu G et al (2022) Levit-unet: Make faster encoders with transformer for biomedical image segmentation. In: Chinese conference on pattern recognition and computer vision (PRCV)
77. Xu Q, Ma Z, He N, Duan W (2023) Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. Comput Biol Med 154:106626

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

Yan Gao received the B.Sc. degree in Electronic and Information Engineering from Hebei Normal University of Science & Technology, in 2010, and the master's degree in Pattern Recognition from Nanjing University of Posts and Telecommunications, in 2014. In 2024, she received her Ph.D. in computer software and theory with the College of Computer Science and Technology, Jilin University. Her research interest includes artificial intelligence and deep learning.

Huan Xu received her master's degree from the College of Computer Science and Technology, Jilin University, where she is currently a Ph.D. candidate. Her research interests include deep learning, graph representation learning, and graph neural networks, with a particular focus on heterophily graph analysis and the development of advanced methods for analyzing heterophily graphs.

Quanle Liu received the Ph.D. degree in Computer Software and Theory from Jilin University, Changchun, China, in 2023. He is currently a lecturer with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. His research interests include deep learning, pattern recognition, data visualization and data analysis.

Mei Bie received the Ph.D. degree from Jilin University in the college of computer science and technology, Jilin University, in 2023. She is currently an Associate Professor at Changchun Normal University. Her research interest includes the information technology and its application.

Xiangjiu Che received the Ph.D. degree from Jilin University, Changchun, China, in 2003. He is currently a Professor with the College of Computer Science and Technology, Jilin University. At present, he is a council member of China Computer Federation; a member of CAD/CG professional committee, China Computer Federation; a member of Association for Computing Machinery. His research interests include network media and transmission, computer vision, pattern recognition, data visualization and data analysis.