J BONE ONCOL 49 (2024) 100647

Contents lists available at ScienceDirect

Journal of Bone Oncology

journal homepage: www.elsevier.com/locate/jbo

A novel adjunctive diagnostic method for bone cancer: Osteosarcoma cell segmentation based on Twin Swin Transformer with multi-scale feature fusion

Tingxi Wen, Binbin Tong, Yuqing Fu *, Yunfeng Li, Mengde Ling, Xinwen Chen

College of Engineering, Huaqiao University, Quanzhou 362021, China

HIGHLIGHTS

- Proposed a Twin Swin Transformer based on Swin Transformer, achieving efficient osteosarcoma cell segmentation with multi-scale feature fusion.
- Replaced the original Swin block with the Twin Swin Transformer block to enhance feature interactions across stages.
- Added channel attention mechanism, improving segmentation accuracy.
- Extracted detailed morphological and spatial information, aiding in personalized treatment strategies.

ARTICLE INFO

Keywords:
Diagnosis of bone cancer
Osteosarcoma
Cell segmentation
Twin Swin Transformer

ABSTRACT

Background: Osteosarcoma, the most common primary bone tumor originating from osteoblasts, poses a significant challenge in medical practice, particularly among adolescents. Conventional diagnostic methods heavily rely on manual analysis of magnetic resonance imaging (MRI) scans, which often fall short in providing accurate and timely diagnosis. This underscores the critical need for advancements in medical imaging technologies to improve the detection and characterization of osteosarcoma.

Methods: In this study, we sought to address the limitations of current diagnostic approaches by leveraging Hoechst-stained images of osteosarcoma cells obtained via fluorescence microscopy. Our primary objective was to enhance the segmentation of osteosarcoma cells, a crucial step in precise diagnosis and treatment planning. Recognizing the shortcomings of existing feature extraction networks in capturing detailed cellular structures, we propose a novel approach utilizing a twin swin transformer architecture for osteosarcoma cell segmentation, with a focus on multi-scale feature fusion.

Results: The experimental findings demonstrate the effectiveness of the proposed Twin Swin Transformer with multi-scale feature fusion in significantly improving osteosarcoma cell segmentation. Compared to conventional techniques, our method achieves superior segmentation performance, highlighting its potential utility in clinical settings.

Conclusion: The development of our Twin Swin Transformer with multi-scale feature fusion method represents a significant advancement in medical imaging technology, particularly in the field of osteosarcoma diagnosis. By harnessing advanced computational techniques and leveraging high-resolution imaging data, our approach offers enhanced accuracy and efficiency in osteosarcoma cell segmentation, ultimately facilitating better patient care and clinical decision-making.

1. Introduction

Osteosarcoma is one of the most common malignant primary bone tumours. It originates from the malignant proliferation of primitive mesenchymal cells and is predominantly found in adolescents, with a second peak incidence in older adults [1,2]. The late-stage survival rate of patients with osteosarcoma is only 20 % [3].

Currently, the leading medical images used to diagnose

* Corresponding author.
E-mail address: fuyq@hqu.edu.cn (Y. Fu).

https://doi.org/10.1016/j.jbo.2024.100647
Received 28 May 2024; Received in revised form 21 October 2024; Accepted 24 October 2024
Available online 1 November 2024

osteosarcoma are electronic computed tomography (CT) images, X-ray images, and MRI [4]. Among them, MRI images have become the primary tool for physicians to diagnose osteosarcoma due to their good representativeness of soft tissue components such as tumours, muscles, and blood vessels and safety. In addition, the current diagnosis of osteosarcoma patients mainly relies on manual identification by physicians. This manual diagnostic method has the following urgent problems: The diagnosis of osteosarcoma disease requires extremely high experience and professional knowledge of doctors. At the same time, it is difficult for physicians to efficiently complete the diagnosis of patients due to the current lack of well-developed osteosarcoma-assisted segmentation technology in hospitals [5]. Therefore, how to assist physicians in efficiently diagnosing osteosarcoma disease is a challenge that needs to be urgently solved. Although approximately 600 to 700 MRI images can be acquired from each osteosarcoma patient, only 10 to 20 of these MRI images can be used to diagnose osteosarcoma disease [6], and the process of manual screening by physicians is very inefficient. Developing countries are relatively economically backward, the problem of few medical resources and lack of medical personnel. In countries, there are many deficiencies in the stages of diagnosis and treatment of osteosarcoma. Tight medical resources and insufficient funding [7,8], coupled with the fact that medical equipment used to diagnose osteosarcoma is very expensive, means that most hospitals in these developing countries cannot afford the cost of the instruments [9]. It leads to increased pressure to diagnose osteosarcoma disease.

Diagnostic MRI images present the following disadvantages: MRI scans usually take a long time to acquire high-quality images; the MRI equipment itself is costly and requires specialized technicians to operate and maintain it, which increases operating costs; certain patients may not be able to undergo MRI scans, such as those with pacemakers and metal implants; MRI images are susceptible to patient movement, and even slight movement may result in blurring or artifacts, which may affect the accuracy of the diagnosis [10,11].

As cell segmentation techniques and deep learning algorithms continue to advance, precise instance segmentation of osteosarcoma cells can effectively assist physicians in making accurate and efficient diagnoses, thereby providing crucial support for treatment planning.

Cell segmentation is a critical task in computer vision, aiming at precise boundary localization and segmentation of cells in microscope images. With the development of deep learning, the processing of cell microscopy images has been greatly enriched in recent years. By segmenting osteosarcoma cell images, we can better understand their morphological features, cellular structure, and potential growth and metastasis mechanisms, providing an important reference basis for clinical diagnosis and treatment.

With the advancement and maturity of deep learning technology and convolutional neural networks, they play an increasingly important role in image segmentation, especially in medical images. Currently, medical image segmentation processing methods are roughly divided into two categories: the first category can be classified as traditional methods, including region-based methods, threshold-based methods, edge-based methods and clustering-based methods; the second category of methods can be summarized as methods using convolutional neural networks, which carry out feature extraction by convolution and then segmentation. The above methods only perform semantic segmentation of the image, while instance segmentation is a combination of target detection and semantic segmentation, which can both segment to get the edges of the objects and label the different individuals in the same kind of objects in the image.

2. Related work

The evolution of high-precision manufacturing systems and diagnostic imaging has significantly influenced quality improvement in engineering and medical fields. In manufacturing, Du et al. [12] introduced a Markov-based model to manage product sequence and bottlenecks, emphasizing the importance of handling sequential dependencies across stages to maintain quality. This concept mirrors the importance of upstream data preparation in osteosarcoma segmentation, where the accuracy of initial data processing can directly impact downstream segmentation results.

In a complementary effort, Wang et al. [13] incorporated variable stiffness structures (VSS) into machining models using a state-space framework. Their work on modeling elastic deformations aligns with the segmentation challenges encountered in medical imaging, where accurately capturing complex biological structures—such as osteosarcoma tumors—requires nuanced modeling. Similarly, in our segmentation approach, we employ multi-scale feature fusion to capture intricate variations within osteosarcoma images, analogous to the adaptive strategies used in VSS machining systems.

The insights from high-definition metrology (HDM) techniques further inform our work. Li et al. [14] used HDM to enhance surface texture evaluation, analogous to how we employ high-resolution fluorescence microscopy to achieve precise cell segmentation. Zhao et al. [15] leveraged dynamic modeling to optimize face milling parameters, enhancing process stability. Shao et al. [16,17] extended the application of HDM by developing models for predicting leakage channels in static sealing interfaces, focusing on spatial irregularities. This challenge resonates with the difficulties of accurately segmenting irregular tumor boundaries in osteosarcoma detection. Our work builds on these insights by introducing advanced segmentation techniques to ensure that irregular cell structures are reliably delineated for diagnostic accuracy.

In the context of osteosarcoma segmentation specifically, Anisuzzaman et al. [18] reviewed the state of deep learning for osteosarcoma detection, identifying key challenges and opportunities. Kayal et al. [19] evaluated nine algorithms, including traditional methods like Otsu thresholding and machine learning approaches such as logistic regression and support vector machines, to assess their effectiveness on osteosarcoma datasets. Their study underscores the need for more advanced models to overcome the limitations of traditional algorithms.

Loraksa et al. [20] found that image format impacts segmentation accuracy, with PNG images yielding optimal results across multiple CNN models. Zhang et al. [21] proposed a multiple supervised residual network for CT segmentation, demonstrating that deeper hierarchical features improve segmentation performance, outperforming traditional architectures like U-Net.

Building on these works, Shuai et al. [22] introduced the W-net++, a dual U-Net model that addresses spatial detail loss through multi-scale inputs, achieving superior segmentation of osteosarcoma lesions. Huang et al. [23] further advanced segmentation techniques with the MSFCN, an end-to-end network using edge supervision and multi-scale feature learning to enhance F1-scores in CT image analysis.

Jia et al. [4] focused on MRI segmentation with DecoupleSegNet, a lightweight model leveraging flow field learning to enhance pixel-level consistency, achieving 90.51 % IoU with fewer parameters. Their work demonstrates that optimized networks can maintain high accuracy while minimizing computational costs, aligning with our objective to design an efficient and adaptive segmentation framework.

Inspired by the multi-stage processing principles found in manufacturing [12,14,15], we develop a Twin Swin Transformer network to address the challenges of under-segmentation in osteosarcoma detection. This model integrates squeeze-and-excitation networks (SENet) [24], enabling adaptive channel learning and improving feature interactions.

3. Overall network structure

3.1. Mask R-CNN based on Swin Transformer backbone network

The network structure of Mask R-CNN [25] based on Swin Transformer backbone network is shown in Fig. 1.First, the input zipper images are fed into the Swin Transformer backbone network for feature

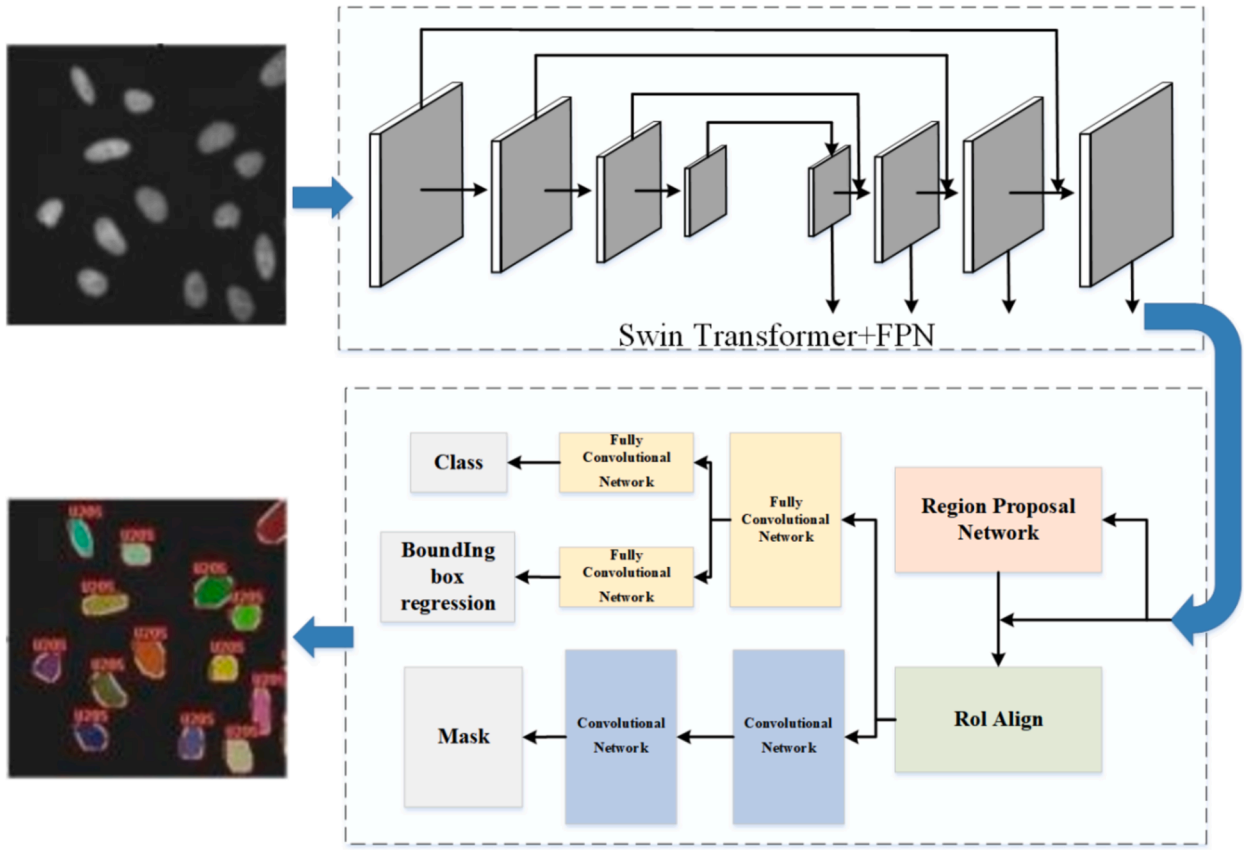

Fig. 1. Network structure diagram of Mask R-CNN model based on Swin Transformer backbone network.

extraction, obtaining multi-scale feature maps at different levels. Then, these feature maps are fused with the feature maps obtained from upsampling on the feature pyramid to generate the output feature maps of the feature pyramid structure. Next, centering on each point of the feature map, candidate frames of different scales and sizes are generated by the Region Proposal Network (RPN), and these candidate frames are scored and filtered to select the regions of interest with scores above a threshold. Meanwhile, candidate frames with high overlap are removed using Non-Maximum Suppression (NMS). Subsequently, Region of Interest Align (RoI Align) to obtain feature maps of consistent size. Finally, these aligned feature maps are fed separately into the classification and regression branches and the mask branch, where they are processed through multiple convolutional networks to generate the final category predictions, bounding box regression results, and masks, achieving zipper identification and segmentation.

The loss function of Mask R-CNN consists of three primary parts: classification loss, bounding box regression loss, and mask loss.

The total loss is calculated as follows:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

The classification loss is used to evaluate the accuracy of the model's category predictions for the candidate regions. The Cross-Entropy Loss function is utilized and can be expressed using the following formula:

$$L_{cls} = -\frac{1}{N_{cls}}\sum_{i=1}^{N} log[p_i^* p_i + (1-p_i^*)(1-p_i)] \tag{2}$$

where p_i^* is the probability that the predicted candidate box is a foreground, and p_i is the probability of the predicted candidate box.

The bounding box regression loss is used to evaluate the accuracy of the model's predictions for the position and size of the candidate region's bounding box. The Smooth L1 loss function is employed, and the formula is as follows:

$$L_{box} = \sum_{i\in\{x,y,w,h\}} smooth_{L1}(t_i - t_i^*) \tag{3}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 if|x| < 1 \\ |x| - 0.5 otherwise \end{cases} \tag{4}$$

where t_i and t_i^* represent the true and predicted bounding box parameters (including the center coordinates and dimensions), respectively, and $smooth_{L1}(x)$ is the Smooth L1 loss function.

The mask loss is used to evaluate the accuracy of the model's predictions for the target masks. A per-pixel binary cross-entropy loss function is employed, and the formula is as follows:

$$L_{mask} = -\frac{1}{m^2}\sum_{i,j} y_{i,j} \log(y_{i,j}^*) + (1-y_{i,j})\log(1-y_{i,j}^*) \tag{5}$$

where $y_{i,j}$ is the value of pixel (i,j) in the ground truth mask, $y_{i,j}^*$ is the value of pixel (i,j) in the predicted mask, and m is the width or height of the mask, m^2 indicates the total number of pixels in the mask.

3.2. Swin Transformer

The Swin Transformer efficiently models both local and global features by dividing the input image into fixed-size patches and applying a hierarchical structure and sliding window mechanism. Within each layer, window-based self-attention is performed, progressively reducing spatial dimensions while increasing the number of feature channels. This approach is highly effective for a wide range of computer vision tasks.

As shown in Fig. 2. First, the input image is divided into non-overlapping chunks of fixed size, each of which is spread and projected as a feature vector of fixed length. This converts the 2D image into a series of 1D vectors for subsequent Transformer coding. Next, these

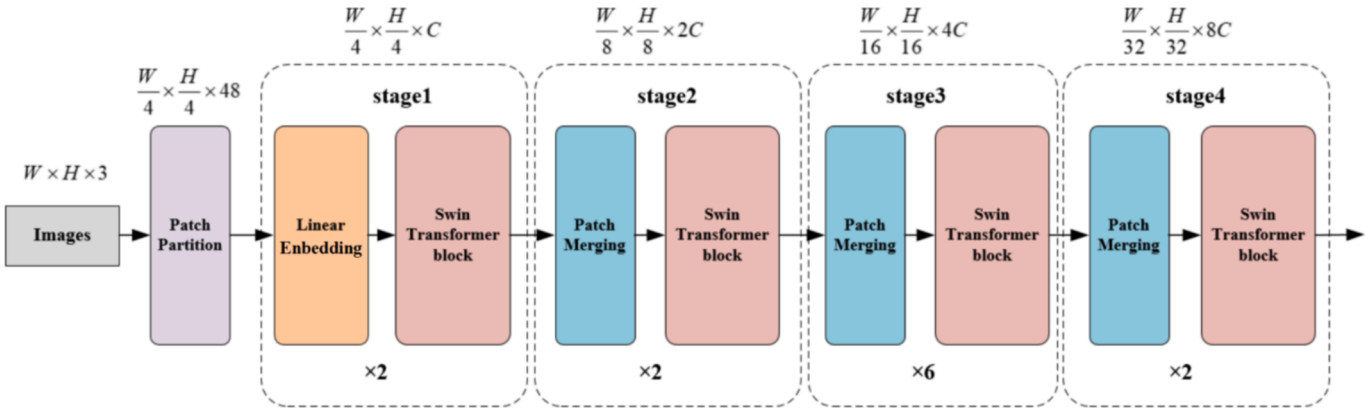

Fig. 2. Swin Transformer network structure diagram.

feature vectors are mapped to a higher dimensional feature space through a linear embedding layer, a step similar to the convolutional layer operation in traditional convolutional neural networks for increasing the dimensionality of the feature representation.Then, feature extraction is performed through multiple Swin Transformer blocks, while downsampling is performed between each stage through the Patch Merging module, which combines small adjacent blocks into larger blocks, reduces the size of the feature map, and increases the number of feature channels through linear layers, which is similar to the pooling operation in convolutional neural networks. After multi-layer Swin Transformer block and multiple stages of processing, the final feature representation with rich hierarchical information is obtained, and these features can be used for various downstream tasks.

The core module of Swin Transformer is the Swin Transformer Block, shown in Fig. 3, which consists of the Multi-head Self-Attention (MSA), the Multilayer Perceptron (MLP), and the Normalization Layer. Its key innovation is the sliding window mechanism.

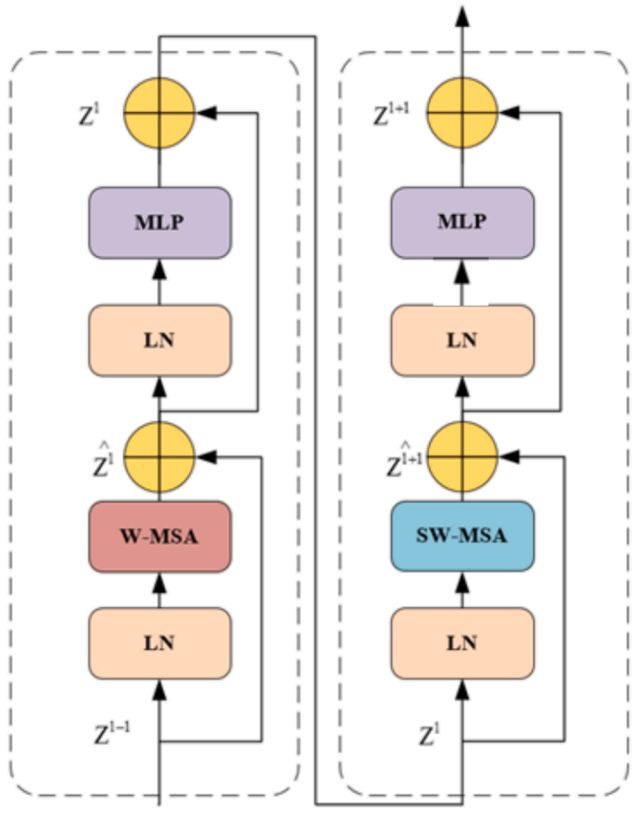

Fig. 3. Swin Transformer Block structure diagram.

Its network structure can be roughly divided into two: Window Multi-head Self Attention (W-MSA) network and Shifted Window Multi-head Self-Attention (SW-MSA) network. The specific process of the W-MSA network is as follows: firstly, the input feature map is divided into multiple windows, and multi-head self-attention computation is carried out within each window, which is then processed by hopping connection, normalization, and multilayer perceptual machine. The difference between the SW-MS network and the W-MSA network is that the divided windows are slid to form new windows before performing the MSA computation, and then the MSA computation is performed within these slid windows.

The window long self-attention is calculated as:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK\hat{T}/\sqrt{d}+B)\text{V} \tag{6}$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(head1,\ldots,headh)W^{O} \tag{7}$$

In the above equation, the projection parameter matrix satisfies the condition W^{O} is the linear transformation matrix of the output and $head_i$ denotes the output of the ith attention head.; B represents the relative positional offset; Q, K, and V are the features within the window mapped into the query (Q), key (K), and value (V) matrices by three linear transformations, respectively.

The first layer of Swin Transformer is a linear embedding layer, which first divides the input image into non-overlapping chunks of fixed size and flattens each chunk into a one-dimensional vector. A linear transformation is applied to each of the flattened chunk vectors to map them to a fixed-length feature vector. The linear transformation is usually implemented through a fully connected layer. A series of feature vectors are generated after all the chunks have been linearly transformed. These feature vectors are rearranged into a feature map.

In addition to the Swin Transformer Block, there is another important operation in Swin Transformer: patch merging. As shown in Fig. 4, it gradually reduces the spatial dimension between different levels and increases the number of feature channels to achieve hierarchical feature extraction. This process is similar to the pooling operation in convolutional neural networks, which can effectively reduce the computational complexity while retaining important feature information.

3.3. Module design of Twin Swin Transformer with multiscale feature fusion

The original Swin Transformer [26] performs self-attention computation within each window, reducing computational complexity but resulting in weak feature interactions between different windows and failing to capture global contextual information adequately. Furthermore, Swin Transformer is effective within local windows but has

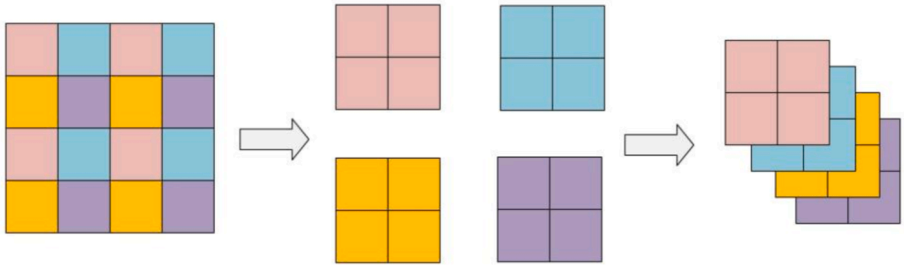

Fig. 4. Patch merging operation.

limitations in modeling complex non-linear relationships, especially when dealing with highly complex image features. Meanwhile, the Swin Transformer may be overfitted and undergeneralised on specific tasks or datasets.

The Twin Swin Transformer Block is able better to capture the complex non-linear relationships between input features, enabling the model to more accurately understand and process complex visual information, and improve the overall recognition and classification capabilities. By enhancing feature interactivity and capturing complex non-linear relationships, the Twin Swin Transformer Block helps to improve the model's generalization ability, enabling it to perform well on different datasets and tasks.

In addition, the Swin Transformer may lose some detail information during multi-level feature extraction, especially in deep networks where the detail information may be gradually weakened. To solve this problem, special Skip Connections are designed in this paper to enable the model to better retain and capture detail information, and thus perform better when dealing with high-resolution images and fine-grained features.

Module for Twin Swin Transformer based on Multi-scale feature fusion (MF Twin Swin) is shown in Fig. 5. Firstly, the original Swin Transformer Block is replaced by the twin Swin Transformer Block designed in this paper, and then a special skip connection and feature fusion module is set between each stage.

In this paper, the twin Swin Transformer Block is designed based on the original Swin Transformer Block, as shown in Fig. 6. It mainly consists of two parallel Swin Transformer Blocks, so it is called Twin Swin Transformer Block.The specific network structure is as follows:

The input feature vector of size (H × W) × C is divided into two feature vectors of the same size (H × W) × C/2 by equally spaced sampling, and these two feature vectors are input into two parallel Swin Transformer Blocks respectively, which output two feature vectors of size (H × W) × C/2. The two feature vectors are overlapped and merged into the (H × W) × C feature vector, and then the feature vectors are rearranged to obtain the H × W × C feature map. Next, further feature processing is performed on the feature map using a convolution kernel size of 1 × 1 to finally obtain an H × W × C feature map. To facilitate processing in the next module, the feature map is rearranged into (H × W) × C feature vectors and output.

The Twin Swin Transformer module is calculated as follows:

Suppose the input feature tensor is $X \in R^{H\times W\times C}$, This tensor is split into two sub-tensors in each Transformer Block:

$$X_1, X_2 = \text{Split}(X), X_1, X_2 \in R^{H\times W\times C/2} \tag{8}$$

Each sub-tensor is computed by a separate Transformer:

$$Y_1 = \text{SwinTransformer}(X_1), Y_2 = \text{SwinTransformer}(X_2) \tag{9}$$

Finally, the output features are passed through the feature fusion operation:

$$Y = \text{Concat}(Y_1, Y_2) \text{ and } Y = \text{Conv}_{1\times1}(Y) \tag{10}$$

The overall network flow is:

The network structure can be divided into four stages and a total of (2, 2, 6, 2) 12-layer network is used. In stage 1, an image dimension of H × W × 3 is input into the model, which is processed and computed by linear embedding layer to obtain a feature vector of (H/4 × H/4) × 96 in two dimensions, which is inputted into the two-layer twin Swin Transformer Block for feature extraction, and outputs a feature vector of the same size, which is then rearranged to form a feature map of H/4 × H/ 4 × 96 feature map as the output of the first stage of the feature pyramid.

In the second stage, the feature map is converted into a feature map of H/8 × H/8 × 192 by the Block Merge Module, which is feature fused with the feature vector output from the first stage linear embedding layer. This is done by rearranging the feature vectors output from the linear embedding layer of the first stage into a feature map of H/4 × H/

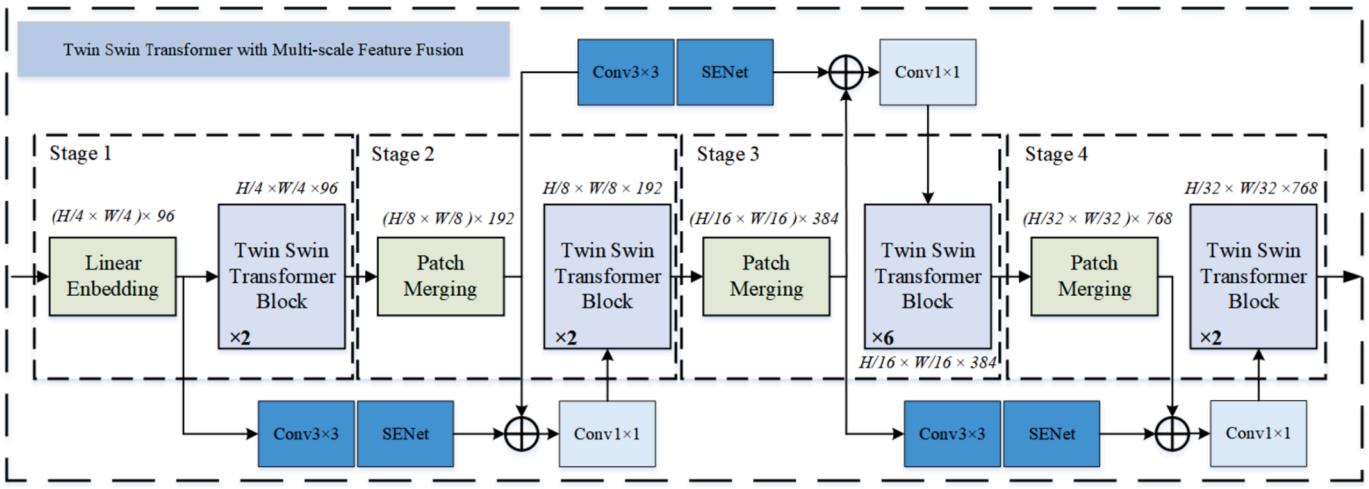

Fig. 5. Twin Swin Transformer module based on multi-scale fusion.

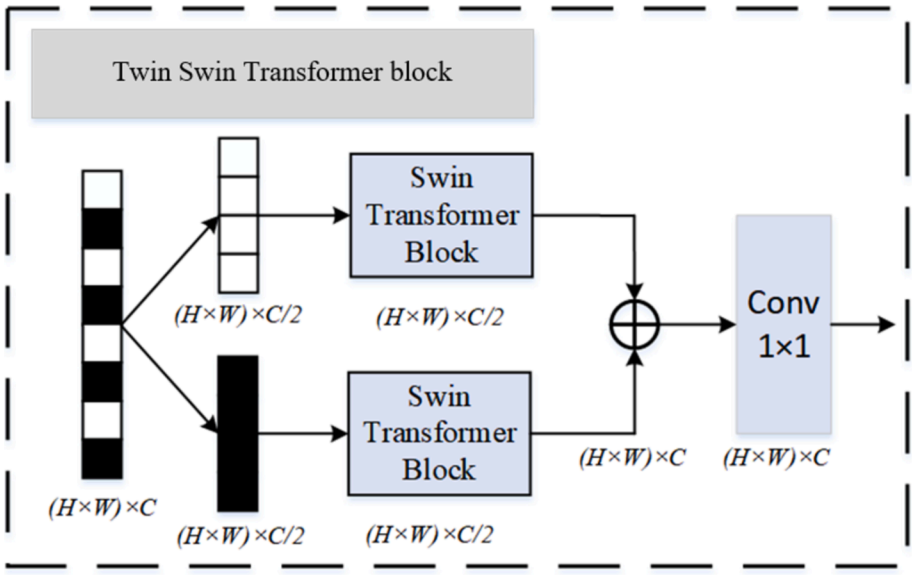

Fig. 6. Twin Swin Transformer module.

4 × 96, which is sequentially converted into a feature map of H/8 × H/8 × 192 by a convolutional layer with a convolutional kernel size of 3 × 3 and a SENet attention network, which is connected with the feature map output from the block merging module of the second stage in a top and bottom stacking, and is combined into a H/8 × H/8 × 384 feature map, which in turn is converted into a H/8 × H/8 × 192 feature map by a convolution kernel of size 1 × 1, which is rearranged into a (H/8 × H/8) × 192 feature vector input into the two-layer twinned Swin Transformer Block in the second stage to obtain a (H/8 × H/8) × 192 feature vectors, which are then rearranged into H/8 × H/8 × 192 feature maps as the feature map output of the second stage of the feature pyramid.

The network structure of the third and fourth stages is similar to that of the second stage. Specifically, the feature map of the second stage is converted into a feature map of H/16 × H/16 × 384 by the block merging module of the third stage, which is feature fused with the feature vectors output from the linear embedding layer of the second stage. The feature vector after performing feature fusion is input to the six-layer twin Swin Transformer Block in the third stage ultimately obtains the feature map of H/16 × H/16 × 384 by rearranging it, which is output as the feature map of the third stage of the feature pyramid. The feature map of the third stage is converted into the H/32 × H/32 × 768 feature map through the block merging module of the fourth stage, which is feature fused with the feature vector output from the linear embedding layer of the third stage. The feature vectors after performing feature fusion are input to the two-layer twin Swin Transformer Block in the fourth stage to finally get the H/32 × H/32 × 768 feature map by rearranging them, which is used as the feature map output of the fourth stage of the feature pyramid.

4. Results and discussion

In order to validate the segmentation effectiveness of the MF Twin Swin on osteosarcoma images, we compare it with three excellent and commonly used feature extraction networks: the ResNet50 [27], the GCNet [28], and the Swin Transformer [27]. ResNet50 improves the training efficiency of deep networks through residual connections, excelling at extracting local features from images. GCNet introduces a global context module, leveraging attention mechanisms to capture long-range dependencies and enhance global information understanding. Swin Transformer combines local windows with shifted window strategies to efficiently balance local and global feature extraction.

This section firstly describes the dataset of osteosarcoma images used for experiments and segmentation performance evaluation method. Subsequently, the findings of the experiments are described. Finally, qualitative and quantitative analyses of the osteosarcoma segmentation results obtained by different methods are done.

4.1. Training settings

The training hardware configuration for this paper is: CPU model is Intel(R) Xeon(R) W-2155 with 16G of RAM; operating system Ubuntu 22.04.3 LT; GPU model is NVIDIA GeForce RTX 2080 Ti.

The models in this paper were trained in the MMDetection framework, Python 3.9.17, CUDA 11.7; CuDNN 8.4, with 80 rounds of training, batch_size of 2, and no pre-training weights loaded.

MMDetection [29] is an open source, PyTorch-based target detection toolkit developed by the OpenMMLab team. It provides a large number of pre-trained models, a flexible configuration system, and rich data enhancement features, and is widely used in academic research and industrial applications. MMDetection supports a variety of target detection methods, including Faster R-CNN, Mask R-CNN, RetinaNet, etc., with high scalability and ease of use. With the MMDetection framework, model training, evaluation, and inference can be easily performed. Its modular design allows users to freely combine model components according to their needs, and supports distributed training and multiple optimisation algorithms, making efficient training on large-scale datasets possible. In the experiments in this paper, we conducted model training and evaluation using the flexible configuration system and efficient training process provided by MMDetection to verify the effectiveness of the proposed method. Therefore, MMDetection is adopted as the basic framework for training in this paper.

4.2. Data sets and evaluation indicators

The data set used for the experiments in this paper is BBBC039 [30]. This image set is a high-throughput chemical screen for the human osteosarcoma U2OS cell line. The image set consists of 200 osteosarcoma cell images, each of which is 520 × 696 pixels in size. In this experiment, 140 images of this image set are classified as the training set and the remaining 60 images are classified as the test set.

In order to quantitatively evaluate the recognition performance of the twin Swin Transformer on the osteosarcoma cell dataset, the mean Average Precision (mAP) is used in this paper.

The mAP is a standard metric for evaluating target detection and

instance segmentation models, which is widely used in academic research and industry. mAP is the average of the APs under multiple IoU thresholds, and a high mAP value usually indicates that the model has better detection accuracy and stability, which is an important measure of the model's merit. mAP [31] is specifically calculated as follows.

Firstly for each prediction frame and real frame, IoU is calculated. IoU is a measure of the similarity between the prediction results and real labels. The calculation formula is as follows:

$$IoU = \frac{IntersectionArea}{UnionArea} \tag{11}$$

Where Intersection Area denotes the area of the overlapping part of the prediction frame and the real frame, and Union Area denotes the total area of the prediction frame and the real frame, minus the area of the intersection area. The value of IoU ranges from 0 to 1.

Afterwards, True Positive (TP) and False Positive (FP) are determined based on the value of IoU. If the IoU of the prediction frame is greater than or equal to a given threshold (e.g., 0.5, 0.55, …, 0.95) and matches a true frame, it is considered a TP. otherwise, it is considered a FP.

Then the precision and recall are calculated, precision is used to indicate the proportion of positive samples in the prediction results, and recall is used to indicate the proportion of samples that are actually positive classes that are correctly recognised as positive classes by the model. The calculation formula is as follows:

$$Pre = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

Where TP denotes true cases, i.e., the number of samples that are actually positive and correctly predicted to be positive. FP denotes false positive cases, i.e., the number of samples that are actually negative but incorrectly predicted to be positive. FN denotes false negative cases, i.e., the number of samples that are actually positive but incorrectly predicted to be negative.

The precision-recall curve is first plotted through the sorted detection results. The area under the P-R curve is calculated to represent Average Precision (AP), which is calculated using interpolation in this paper:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{14}$$

Where P_n is the maximum precision corresponding to the nth recall and R_n is the nth recall.

Finally mAP is calculated by AP. the formula is as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{15}$$

where N is the number of categories.

The mAP in this paper is the mean value of the average precision (AP) calculated at multiple IoU thresholds (from 0.5 to 0.95 in steps of 0.05). mAP_50 is the average precision of the detection frame calculated at IoU = 0.50. Calculated in the same way as mAP, but with the IoU fixed at 0.50. mAP_75 is the mean accuracy of the detection frame calculated at IoU = 0.75. The calculation method is the same as that of mAP, but the IoU is fixed at 0.75. mAP_s and mAP_m represent the mAP values obtained when the area of the detected target is in the range of 0 to 322 and 322 to 622 pixel units, respectively.

4.3. Quantitative analysis

In order to quantitatively compare the effectiveness of the twin Swin Transformer for segmentation of osteosarcoma cells, we applied four deep learning methods to the BBBC039 osteosarcoma image set, and plotted the loss curves and the curves of the five evaluation metrics mentioned above based on the experimental results.

From Fig. 7, it can be seen that the convergence speed of the loss curves of the four methods basically the same, and the loss curve corresponding to the twin Swin Transformer method proposed in this paper is smoother, and the final loss value is the smallest. It shows that the model corresponding to the method proposed in this paper is more stable.

In order to further verify the effectiveness of the proposed twin Swin Transformer for osteosarcoma cell segmentation. When the loss function converges and the model tends to be stable, we selected the training results of the last 10 rounds and averaged the values of the four methods in each of the five evaluation metrics to obtain Table 1 as follows:

Combining Table 1 and Fig. 8, it can be seen that compared to the other three methods, the twin Swin Transformer proposed in this paper achieves the best results in all the five evaluation metrics of the index. There is an improvement over the best results obtained by the other three methods for each metric, specifically: segm_mAP improved by 0.02, segm_mAP_50 improved by 0.04, segm_mAP_75 improved by 0.04, segm_mAP_s improved by 0.02,segm_mAP_m improved by 0.06.

4.4. Qualitative analysis

To further validate the effectiveness of the Twin Swin Transformer with multi-scale feature fusion proposed in this paper for osteosarcoma cell segmentation. Fig. 9 shows the three representative images selected and the distribution of the three osteosarcoma cell images from top to bottom in the first column of each image can be described as uniformly distributed; cells of varying sizes with adhesions, and sparsely distributed cells.

As shown in Figs. 9 and 10, it can be seen that compared to Resnet, Gcnet, and Swin Transformer, the Twin Swin Transformer with multi-scale feature fusion method proposed in this paper achieves the best segmentation results on all three of the presented images Specifically, the method of ResNet50 has insufficient segmentation; there are many cases of segmenting multiple osteosarcoma cells into one cell, and the segmentation omission is the most serious, which will have the worst effect. The GcNet method is better than ResNet50, and more cases of segmentation omission exist. The segmentation effect of the Twin Swin Transformer with multi-scale feature fusion method is significantly improved compared to the ResNet50 and the GcNet methods; there is no segmentation of multiple osteosarcoma cells into a single cell, and

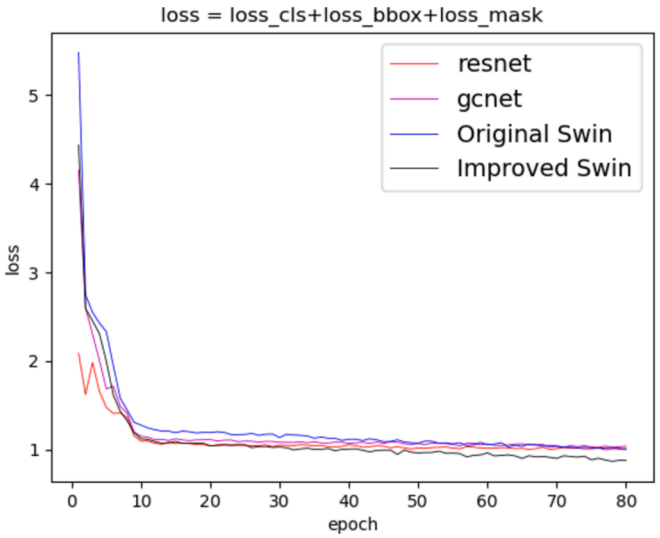

Fig. 7. Loss curve graphs for the four methods.

Table 1
Mean values of the four methods on the five evaluation indicators.

backbone	segm_mAP	segm_mAP_50	segm_mAP_75	segm_mAP_s	segm_mAP_m
Resnet50	0.54	0.65	0.60	0.56	0.48
Gcnet	0.51	0.62	0.57	0.55	0.39
Swin transformer	0.62	0.74	0.70	0.61	0.65
MF Twin Swin	0.64	0.78	0.74	0.63	0.71

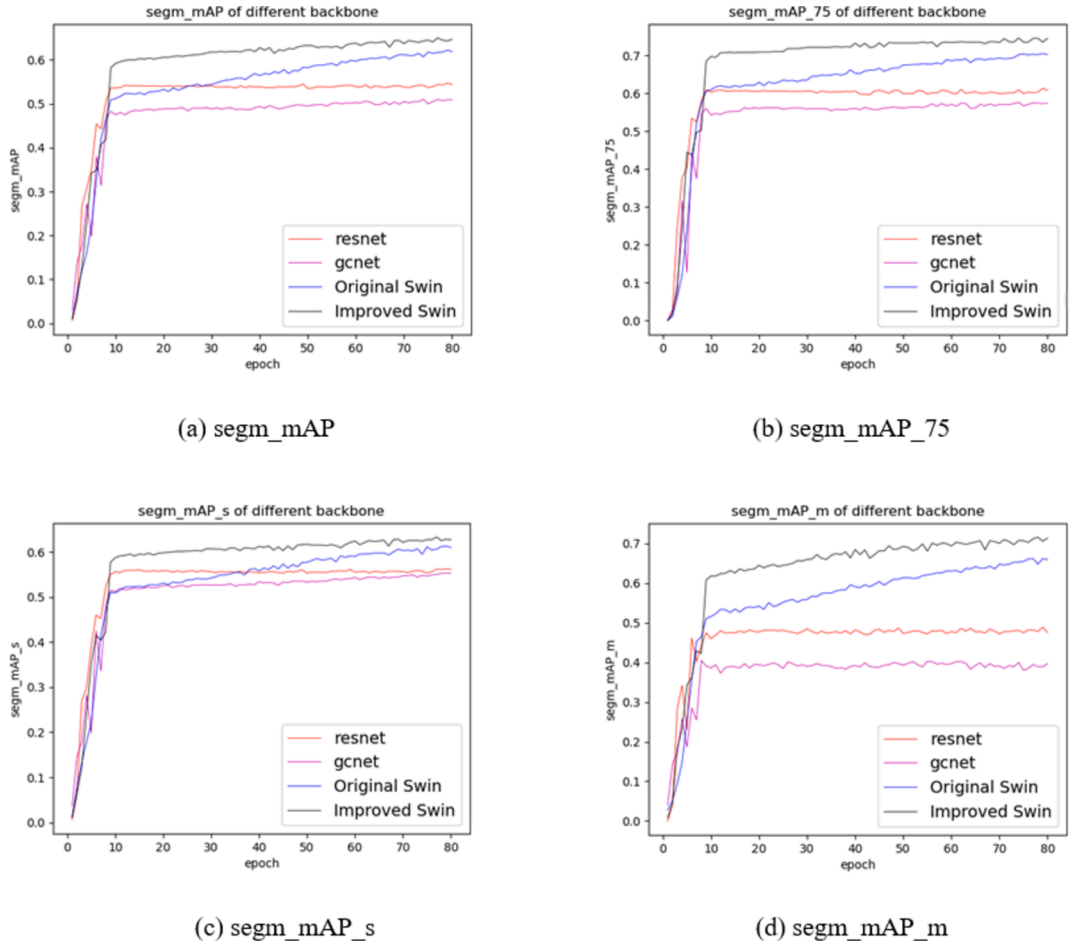

Fig. 8. Specific performance of the four methods on the four evaluation metrics.

segmentation omissions are substantially reduced. The Twin Swin Transformer with multi-scale feature fusion method has only a few cases of segmenting multiple osteosarcoma cells into a single cell, and the segmentation omission is the least among the four methods.

The method proposed in this paper focuses on replacing the Swin Transformer Block in the original Swin Transformer Network with a Twin Swin Transformer Block, which can enhance the feature interaction of the input features at different stages. In addition, SENet is added to the original network, which allows the network to focus on the most important feature channels. Moreover, Skip Connections are introduced between different Twin Swin Transformer Block stages to improve the model's ability to process detailed information. With these improvements, our method effectively segments osteosarcoma cells, thus assisting physicians in efficiently and accurately diagnosing bone cancer.

The paper centers on improving the original Swin Transformer Network, mainly emphasizing replacing the Swin Transformer Block with a Twin Swin Transformer Block to enhance the feature interaction of the input features. SENet is added to the original network, which allows the network to focus on the most important feature channels, thus improving the model's performance. Skip Connections are introduced between different Twin Swin Transformer Block stages to enhance the model's ability to process detailed information.

The experimental part deals with the evaluation setup, including the dataset, experimental metrics, and comparative analysis with existing methods. Specifically, the experiments are performed on a dataset of bone tumour cell images captured under Hoechst-stained fluorescence microscopy to demonstrate the proposed method's effectiveness in each segmentation metric.

Comparative analysis with existing methods ResNet50, GcNet, and Swin Transformer showed that the method performed well in terms of osteosarcoma cell segmentation effectiveness, especially for segm_mAP_75, segm_mAP_75, and cells with cell areas ranging from 322 to 622 pixel-unit scale sizes.

The method performs excellently in the bone tumor cell image segmentation task In conclusion, the method proposed in this paper dramatically improves the accuracy of the network in the osteosarcoma cell segmentation task based on the original Swin Transformer Network by replacing the Swin Transformer Block in the original network with a Twin Swin Transformer, adding SENet, and introducing Skip

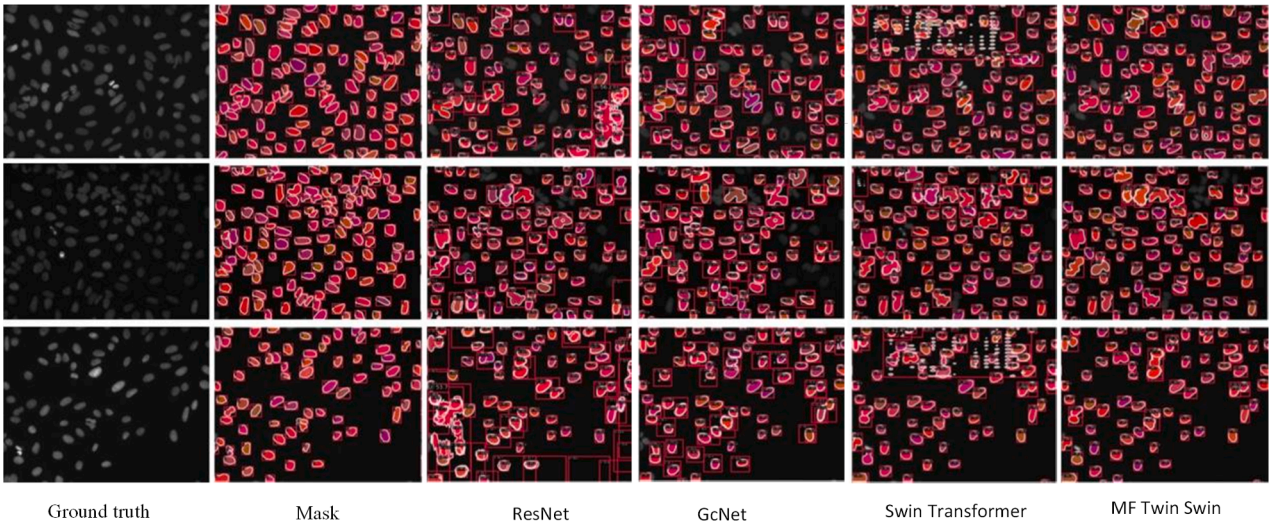

Fig. 9. Segmentation effect of each of the four methods on three images of osteosarcoma cells.

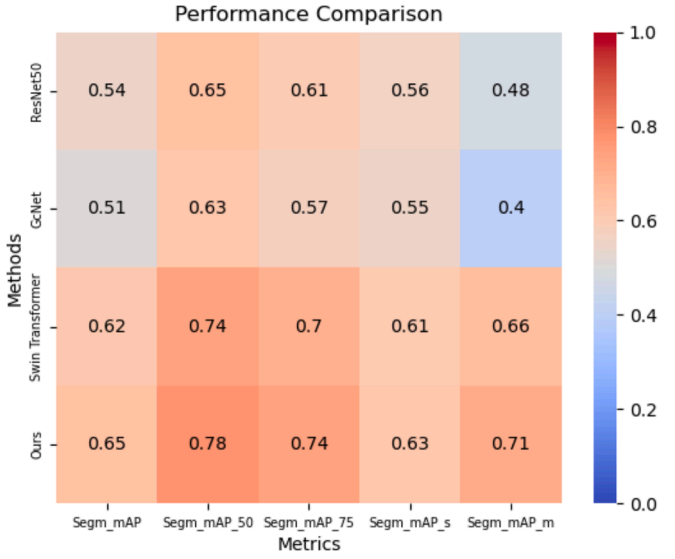

Fig. 10. Segmentation performance comparison.

Connections between different stages of the Twin Swin Transformer Block. The method performs excellently in the bone tumor cell image segmentation task and has a promising future in clinical bone cancer-assisted diagnosis.

Based on the above research findings, our future research directions include: 1) introducing automated hyperparameter search and adaptive learning rate strategies to optimize model efficiency; 2) employing pruning, quantization, and knowledge distillation to achieve lightweight design and mobile deployment; 3) enhancing the model's generalization ability across multi-center datasets through transfer learning and domain adaptation techniques; 4) exploring multimodal data integration to further improve the model's performance in recognizing complex lesions.

5. Conclusion

In this paper, we design the twin Swin Transformer Block, through which the interaction of input features can be enhanced to capture the complex nonlinear relationship between input features, which helps improve the model's generalisation ability. In addition, based on the Swin Transformer network, we design a Twin Swin Transformer network with muti-scale feature fusion by replacing the Swin Transformer Block in the original network with the designed Twin Swin Transformer Block, which enhances the feature interactions of the input features, and also Skip Connections is introduced to increase the ability to capture detail information. Moreover, SENet is introduced in the twin Swin Transformer network to adaptively adjust the importance of each channel to enhance the feature representation capability, thus improving the network's performance. The experimental results show that the twin Swin Transformer method with multi-scale fusion proposed in this paper achieves ideal segmentation results in the osteosarcoma cell segmentation task. Improving the segmentation accuracy of osteosarcoma cells enhances the precision and timeliness of diagnosis by better capturing the morphological and structural features of the tumor cells.

Future research directions focus on enhancing model accuracy and computational efficiency through deep feature extraction and efficient computing strategies. Efforts will also be directed towards optimizing parameter tuning and designing lightweight architectures to improve the model's applicability and scalability. Additionally, the research will explore the model's adaptability and generalization capabilities across multicenter datasets and various imaging devices, aiming to assess the potential feasibility and advantages of MF Twin Swin in practical clinical applications.

Ethics approval

This study utilized publicly available osteosarcoma datasets for research purposes. The dataset acquisition address is https://bbbc.broadinstitute.org/BBBC039. As such, ethics approval was not required for the use of these datasets, as they are openly accessible and do not involve direct human or animal subjects. The datasets were obtained from reputable sources and used in accordance with their respective terms of use and data sharing policies.

CRediT authorship contribution statement

Tingxi Wen: Supervision, Methodology. **Binbin Tong:** Writing – original draft, Software, Methodology, Conceptualization. **Yuqing Fu:** Software, Methodology. **Yunfeng Li:** Visualization, Software. **Mengde Ling:** Validation. **Xinwen Chen:** Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Science and Technology Program of Quanzhou (No. 2024QZC010R, No. 2024G11, No. 2024G16).

References

[1] R. Belayneh, M.S. Fourman, S. Bhogal, et al., Update on osteosarcoma[J], Curr. Oncol. Rep. 23 (2021) 1–8.
[2] C. Chen, L. Xie, T. Ren, et al., Immunotherapy for osteosarcoma: fundamental mechanism, rationale, and recent breakthroughs[J], Cancer Lett. 500 (2021) 1–10.
[3] F. Sadoughi, P.M. Dana, Z. Asemi, et al., DNA damage response and repair in osteosarcoma: defects, regulation and therapeutic implications[J], DNA Repair 102 (2021) 103105.
[4] J. Wu, Y. Guo, F. Gou, et al., A medical assistant segmentation method for MRI images of osteosarcoma based on DecoupleSegNet[J], Int. J. Intell. Syst. 37 (11) (2022) 8436–8461.
[5] M. Nasor, W. Obaid, Segmentation of osteosarcoma in MRI images by K-means clustering, Chan-Vese segmentation, and iterative Gaussian filtering[J], IET Image Proc. 15 (6) (2021) 1310–1318.
[6] R.A. Nabid, M.L. Rahman, M.F. Hossain, Classification of osteosarcoma tumor from histological image using sequential RCNN[C]//2020, in: 11th International Conference on Electrical and Computer Engineering (ICECE), 2020, pp. 363–366.
[7] J. Wu, F. Gou, Y. Tan, A staging auxiliary diagnosis model for nonsmall cell lung cancer based on the intelligent medical system[J], Comput. Math. Methods Med. 2021 (1) (2021) 6654946.
[8] J. Wu, Y. Tan, Z. Chen, et al., Decision based on big data research for non-small cell lung cancer in medical artificial system in developing country[J], Comput. Methods Programs Biomed. 159 (2018) 87–101.
[9] R. Cui, Z. Chen, J. Wu, et al., A multiprocessing scheme for PET image pre-screening, noise reduction, segmentation and lesion partitioning[J], IEEE J. Biomed. Health Inform. 25 (5) (2020) 1699–1711.
[10] R. Abdalla, M. Ahmed, MRI limitations: the main aspects and resolving techniques [J], Ind. J. Appl. Res. 10 (2020) 71–73.
[11] Z. Luo, W. Chen, X. Shen, et al., CT and MRI features of calvarium and skull base osteosarcoma (CSBO)[J], Br. J. Radiol. 93 (1105) (2020) 20190653.
[12] S. Du, R. Xu, L. Li, Modeling and analysis of multiproduct multistage manufacturing system for quality improvement[J], IEEE Trans. Syst., Man Cybernet.: Syst. 48 (5) (2016) 801–820.
[13] K. Wang, G. Li, S. Du, et al., State space modelling of variation propagation in multistage machining processes for variable stiffness structure workpieces[J], Int. J. Prod. Res. 59 (13) (2021) 4033–4052.
[14] G. Li, S. Du, B. Wang, et al., High definition metrology-based quality improvement of surface texture in face milling of workpieces with discontinuous surfaces[J], J. Manuf. Sci. Eng. 144 (3) (2022) 031001.
[15] G. Li, S. Du, D. Huang, et al., Dynamics modeling-based optimization of process parameters in face milling of workpieces with discontinuous surfaces[J], J. Manuf. Sci. Eng. 141 (10) (2019) 101009.
[16] Y. Shao, Y. Yin, S. Du, et al., A surface connectivity-based approach for leakage channel prediction in static sealing interface[J], J. Tribol. 141 (6) (2019) 062201.
[17] Y. Shao, Y. Yin, S. Du, et al., Leakage monitoring in static sealing interface based on three dimensional surface topography indicator[J], J. Manuf. Sci. Eng. 140 (10) (2018) 101003.
[18] D.M. Anisuzzaman, H. Barzekar, L. Tong, et al., A deep learning study on osteosarcoma detection from histological images[J], Biomed. Signal Process. Control 69 (2021) 102931.
[19] E. Baidya Kayal, D. Kandasamy, R. Sharma, et al., Segmentation of osteosarcoma tumor using diffusion weighted MRI: a comparative study using nine segmentation algorithms[J], SIViP 14 (2020) 727–735.
[20] C. Loraksa, S. Mongkolsomlit, N. Nimsuk, et al., Effectiveness of learning systems from common image file types to detect osteosarcoma based on convolutional neural networks (CNNs) models[J], J. Imag. 8 (1) (2021) 2.
[21] R. Zhang, L. Huang, W. Xia, et al., Multiple supervised residual network for osteosarcoma segmentation in CT images[J], Comput. Med. Imaging Graph. 63 (2018) 1–8.
[22] L. Shuai, X. Gao, J. Wang, Wnet++: a nested W-shaped network with multiscale input and adaptive deep supervision for osteosarcoma segmentation[C]//2021, in: IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), 2021, pp. 93–99.
[23] L. Huang, W. Xia, B. Zhang, et al., MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images[J], Comput. Methods Programs Biomed. 143 (2017) 67–74.
[24] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks[c]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018:) 7132–7141.
[25] X. Bi, J. Hu, B. Xiao, et al., Iemask r-cnn: Information-enhanced mask r-cnn[J], IEEE Trans. Big Data 9 (2) (2022) 688–700.
[26] Z. Liu, Y. Lin, Y. Cao, et al., Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[c]//proceedings of the IEEE/CVF International Conference on Computer Vision (2021:) 10012–10022.
[27] K. He, X. Zhang, S. Ren, et al., Deep Residual Learning for Image Recognition[c]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016:) 770–778.
[28] Y. Cao, J. Xu, S. Lin, et al., Gcnet: Non-Local Networks Meet Squeeze-Excitation Networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019:).
[29] F.J. Garcia-Espinosa, A.S. Montemayor, A. Cuesta-Infante, Automatic annotation for weakly supervised pedestrian detection[C]//International Work-Conference on the Interplay Between Natural and Artificial Computation, Springer International Publishing, Cham, 2022, pp. 308–317.
[30] V. Ljosa, K.L. Sokolnicki, A.E. Carpenter, Annotated high-throughput microscopy image sets for validation[J], Nat. Methods 9 (7) (2012) 637.
[31] Zhou Y, Yang X, Zhang G, et al. Mmrotate: A rotated object detection benchmark using pytorch[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 7331-7334.