International Journal of Imaging Systems and Technology

WILEY

RESEARCH ARTICLE

TS-Net: Trans-Scale Network for Medical Image Segmentation

HuiFang Wang[1] | YaTong Liu[1] | Jiongyao Ye[1] | Dawei Yang[2,3] | Yu Zhu[1]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China | [2]Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai, China | [3]Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai, China

Correspondence: Yu Zhu (zhuyu@ecust.edu.cn) | Dawei Yang (yang_dw@hotmail.com)

Received: 9 August 2024 | **Revised:** 26 February 2025 | **Accepted:** 3 March 2025

Funding: This work was supported by National Natural Science Foundation of China (62476088, 82170110) and Shanghai Municipal Science and Technology Major Project (ZD2021CY001) and Shanghai Municipal Key Clinical Specialty (shslczdzk02201). Science and Technology Commission of Shanghai Municipality (20DZ2254400, 20DZ2261200). Fujian Province Department of Science and Technology (2022D014).

Keywords: convolution modulation | deep supervision | edge loss | feature complementarity | medical image segmentation

ABSTRACT

Accurate medical image segmentation is crucial for clinical diagnosis and disease treatment. However, there are still great challenges for most existing methods to extract accurate features from medical images because of blurred boundaries and various appearances. To overcome the above limitations, we propose a novel medical image segmentation network named TS-Net that effectively combines the advantages of CNN and Transformer to enhance the feature extraction ability. Specifically, we design a Multi-scale Convolution Modulation (MCM) module to simplify the self-attention mechanism through a convolution modulation strategy that incorporates multi-scale large-kernel convolution into depth-separable convolution, effectively extracting the multi-scale global features and local features. Besides, we adopt the concept of feature complementarity to facilitate the interaction between high-level semantic features and low-level spatial features through the designed Scale Inter-active Attention (SIA) module. The proposed method is evaluated on four different types of medical image segmentation datasets, and the experimental results show its competence with other state-of-the-art methods. The method achieves an average Dice Similarity Coefficient (DSC) of 90.79% ± 1.01% on the public NIH dataset for pancreas segmentation, 76.62% ± 4.34% on the public MSD dataset for pancreatic cancer segmentation, 80.70% ± 6.40% on the private PROMM (Prostate Multi-parametric MRI) dataset for prostate cancer segmentation, and 91.42% ± 0.55% on the public Kvasir-SEG dataset for polyp segmentation. The experimental results across the four different segmentation tasks for medical images demonstrate the effectiveness of the Trans-Scale network.

1 | Introduction

Medical image segmentation plays an important role in medical image analysis, which can assist doctors in more efficient disease diagnosis [1, 2]. The main target of it is to accurately facilitate internal organs and extract lesions from the background pixels on diverse biomedical images, such as Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI) [3, 4]. Medical image segmentation is a challenging task because of various shapes and blurred boundaries [5, 6], as shown in Figure 1. At the same time, it requires a lot of time and effort to annotate medical images. Therefore, traditional manual-based medical image segmentation is tedious and limited.

In recent years, deep learning (DL) has provided state-of-the-art performance for various vision tasks such as

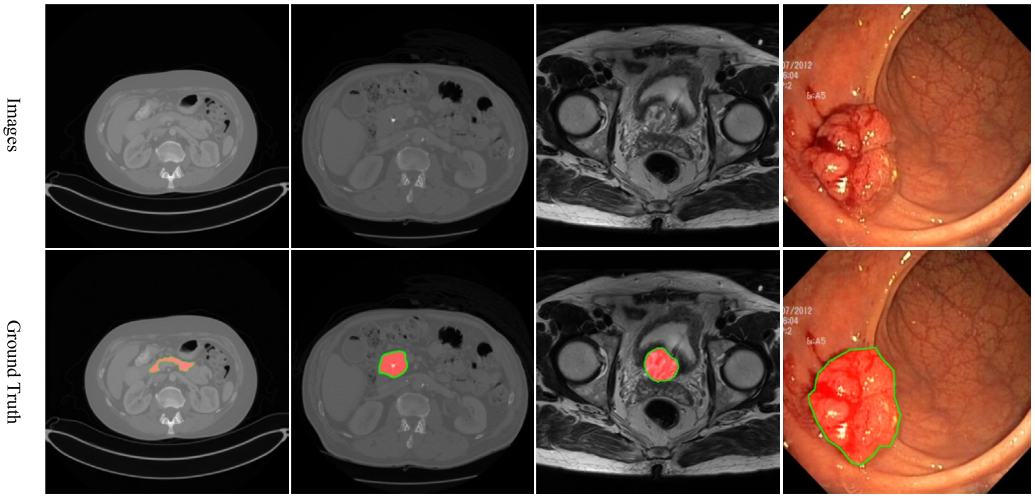

FIGURE 1 | Instances of medical images with their respective semantic segmentation annotations. The red area corresponds to the segmentation target, showing significant differences in morphological positions among different medical image segmentation samples.

image classification, segmentation, and recognition [7–9]. Convolutional neural networks (CNNs) have been used widely to extract complex features accurately through simple operations [10]. U-Net is one of the most popular medical image segmentation networks. Since the inception of U-shaped architecture, numerous researchers have been committed to advancing its capabilities to better suit the demands of medical image segmentation. For instance, Attention U-net [11] applied an attention gate (AG) mechanism to focus on structures of interest in medical images automatically, regardless of their size or shape. In addition, MA-UNet [12] introduced the multi-scale information based on a more lightweight Attention U-net. Although CNN models have remarkable representation capabilities, they still suffer from inevitable limitations such as lacking the ability to understand global contexts owing to the intrinsic locality of convolution operations.

To address the limitations of CNN, Transformer [13] integrates the self-attention mechanism and demonstrates significant success in natural language processing tasks. Thereafter, the Transformer has gradually been introduced to the field of computer vision [14–16]. The Vision Transformer [17] brought about a breakthrough in using Transformers for vision tasks, adopting patch-based input and multi-head self-attention. Building upon ViT's success, SegFormer [18] was proposed in 2021 by Xie et al. as a Transformer-based semantic segmentation approach, which surpassed the performance of other existing models. Nowadays, Transformer-based models have received increasing attention in the field of medical image segmentation. Medical Transformer [19] proposed gated axial attention to construct the main encoder block with a LoGo training strategy, successfully addressing the issue of insufficient medical sample data. PVT [20] leveraged a progressive shrinking pyramid structure in Transformer to overcome the limitation of scale invariance for pure Transformer models. TransUNet [21] made a pioneering contribution to the field of medical image segmentation by integrating Transformer into CNN. Specifically, TransUNet employed a combination of convolutional and attention mechanisms to extract global and fine-grained local context, thus leveraging the advantages of both Transformer and CNN. Through the effective connection of the fused channel-wise information with the decoder features, UCTransnet [22] reduced the ambiguity in the medical image segmentation process and narrowed the semantic gap, setting a new state-of-the-art standard in the field of medical image segmentation.

Different from the aforementioned methodologies, we propose a novel segmentation network called Trans-Scale that effectively leverages the deep –shallow layers interaction and the scale variation rules to obtain more accurate feature representations. Our approach embraces a two-stage segmentation framework, which progressively refines segmentation from coarse to fine levels. Initially, a pre-trained U-Net is utilized to extract the approximate contour of the target region for coarse detection. The fine-grained segmentation framework is built upon U-Net architecture, with the skip connection comprising the scale inter-active attention (SIA) module and the multi-scale convolution modulation (MCM) module. The SIA module employs the concept of feature complementarity, which utilizes deep semantics to supplement shallowly neglected information. By adopting subtraction operations to obtain the reverse weight map of the encoding features, the information that is easily overlooked in the encoding process can be added to the decoding process. It is worth noting that the MCM module dynamically adjusts the depth-wise separable convolution kernel size based on the scale of layer-specific features, facilitating efficient multi-scale convolutional modulation. The fine-grained segmentation network consists of five layers, where the fusion of encoder-derived feature maps and upsampled high-level semantic features from deep layers is accomplished through the SIA module in the preceding four layers. Subsequently, the MCM module leverages multi-scale convolutional modulation to further enhance the network's performance.

The main contributions of this work can be summarized as follows:

1. The proposed segmentation network Trans-Scale leverages the fusion of CNN and Transformer structures through the

designed MCM module, which employs the multi-scale convolutional modulation strategy to simplify the self-attention mechanism, effectively extracting the multi-scale global features and local features.

2. We employ a SIA module to focus on the information details lost from different layers through feature complementarity, which applies deep semantics to supplement shallowly neglected information. We further design the Edge loss function based on wavelet decomposition that refines the segmentation results by emphasizing the high-frequency texture features.
3. Experimental results on the NIH pancreas dataset, the MSD pancreas cancer dataset, the private PROMM (Prostate Multi-parametric MRI) prostate cancer dataset, and the Kvasir-SEG polyp dataset validate the effectiveness of the proposed Trans-Scale network compared to the state-of-the-art methods.

The paper is organized into six sections: Section 2 briefly reviews the related work. Section 3 describes the proposed Trans-Scale network and the edge loss function. In Section 4, we present the experiments with detailed analysis and results comparison. Finally, a discussion is provided in Section 5, and the conclusion is explained in Section 6.

2 | Related Work

Despite the persistent need for further improvement, deep learning methods for medical image segmentation have reached a relatively mature stage of development. There has been a notable emergence of exceptional deep-learning models for medical image segmentation in recent years.

Pancreatic cancer is a serious threat to human life and health due to its exceptionally high malignancy, necessitating the exploration of an efficient approach for accurately segmenting both the pancreas and pancreatic cancer. Primarily, in the realm of pancreas segmentation: Oktay et al. [11] designed a novel attention gate (AG) model to emphasize various target structures of the pancreas. Cai et al. [23] designed a novel CNN-RNN architecture to tackle inter-slice spatial non-smoothness during the pancreas segmentation process. To overcome the segmentation difficulties for uncertain regions, Zheng et al. [24] proposed an iterative workflow for progressively refining segmentation results. Li et al. [25] designed a multiscale attention mechanism to enhance semantic information through integrating scale variations. Wang et al. [26] proposed a dual-input FCN, which further improved the pancreas segmentation performance by the contrast-specific algorithm. To capture the pancreatic features flexibly, Huang et al. [27] combined a deformable convolution module with U-Net. Additionally, Liu et al. [28] presented an ensemble-based multiloss FCN for accurate feature representation. Except that, Li et al. [29] introduced the double adversarial networks with a pyramidal pooling module to achieve satisfactory results. Chen et al. [30] developed a fuzzy skip connection to facilitate the information transmission of variable pancreas targets between codecs. To efficiently capture the global features of the pancreas, Qiu et al. [31] integrated a residual Transformer block into U-Net in 2023.

Moreover, in the realm of pancreatic cancer segmentation: Wang et al. [32] presented an Inductive Attention Guidance Network that utilizes multi-instance learning to improve segmentation accuracy for pancreatic ductal adenocarcinoma. Chen et al. [33] proposed a model-driven approach based on spiral transformation, addressing the challenge of incorporating 3D contextual information into 2D models. Li et al. [34] proposed a position-guided deformable UNet to effectively tackle variations in pancreatic cancer segmentation. To mitigate the uncertainty caused by image registration in multi-modal MRI, Li et al. [35] designed a novel multi-scale adversarial network. Li et al. [36] proposed a dual-meta-learning method that leverages both common knowledge and salient information to enhance the pancreas cancer segmentation performance. Mahmoudi et al. [37] introduced a hybrid model that ingeniously ensembles the Attention U-Net and TAU-Net [38] for pancreatic cancer segmentation. Li et al. [39] Proposed a 3D FCN with three temperature-guided modules to effectively overcome local optima problems. Ju et al. [40] developed an approach based on spatial contextual cues and activated location offsets for precise pancreatic cancer segmentation. Considering that pancreatic cancer usually occupies a small region, Wang et al. [41] proposed a novel two-stage segmentation strategy for pancreatic cancer, combining a lightweight CNN for initial localization and an improved U-shaped network for fine segmentation. What's more, Liang et al. [42] developed a novel framework for automatic gross tumor volume segmentation, matching expert radiology oncologists' performance by integrating multimodal images and daily MRI scans. Li et al. [43] designed a CausegNet that focuses on extracting intrinsic structure features to reduce interference from background noise. Qiu et al. [44] proposed a cascaded segmentation framework based on a multi-scale U-Net to accurately locate pancreatic tumors of different sizes in 2024.

Prostate cancer is also a highly prevalent malignant disease. For prostate cancer segmentation, Zhang et al. [45] achieved competitive prostate cancer segmentation results by integrating channel and position attention mechanisms into the generator network of GAN [46]. Liu et al. [47] introduced a cascading pyramid convolution module and a double-input channel attention module to preserve small target features across different scales, resulting in dependable segmentation of prostate cancer (PCa) lesions. Song et al. [48] proposed the DMSA-V-Net, which is capable of learning comprehensive spatial structure features, strengthening the image understanding ability to effectively segment PCa lesions.

Although the above segmentation methods based on deep learning can automatically segment the organs or lesions area with a certain accuracy, the lack of medical data samples hinders the feature representation ability of the network. In order to better address the issues of inadequate feature extraction, we propose the Trans-Scale network to make full use of context information through the SIA module and MCM module. We further design the Edge loss function based on wavelet decomposition to overcome the problem of blurred edges and refine the prediction results.

3 | Method

In this section, we will delineate more details of the Trans-Scale network for medical image segmentation. We first

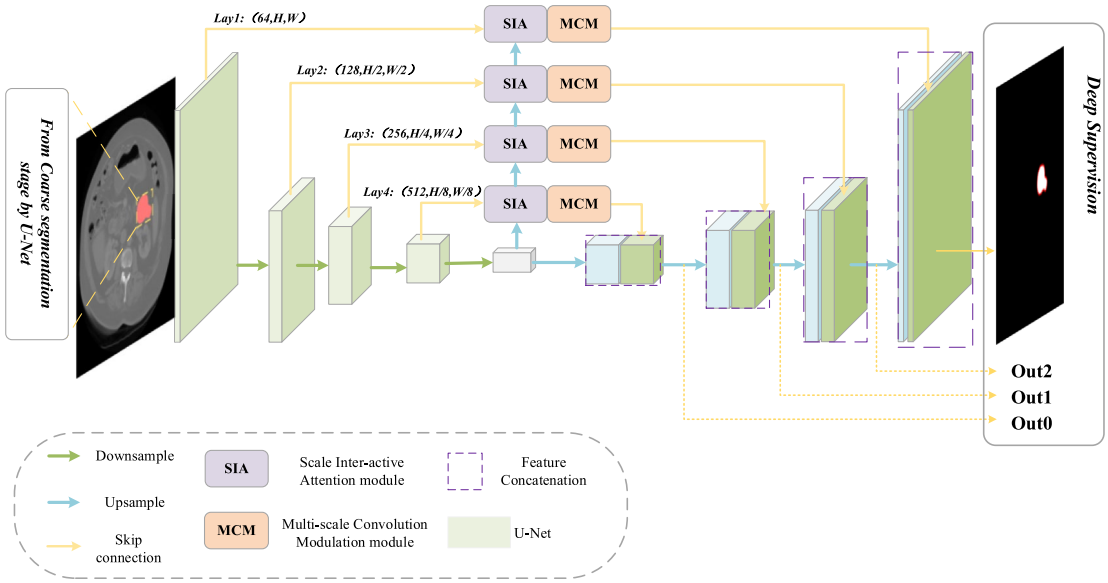

FIGURE 2 | Trans-Scale algorithm framework. A novel SIA (Scale Inter-active Attention) module and MCM (Multi-scale Convolution Modulation) module are designed in the skip connection part of the network to promote the information transmission between codec and codec. The Deep Supervision is conducted to enhance the robustness of the Trans-Scale network.

introduce the overall structure of the two-stage segmentation network. Then, Section 3.1 presents the SIA module based on a complementary concept, and Section 3.2 analyzes the MCM module using a convolution modulation strategy to achieve a more efficient self-attention mechanism in VIT. Finally, Section 3.3 elaborates on the edge loss function based on wavelet decomposition.

The overview of the proposed Trans-Scale segmentation network is shown in Figure 2. The model employs a coarse-to-fine two-stage segmentation framework. Specifically, we first use the pre-trained U-Net as the coarse segmentation network to detect the probable pancreas region and then feed it into the Trans-Scale network for fine segmentation.

The model framework adopts a symmetrical codec structure, and U-Net is used as the benchmark framework of the network. The network has a total of five layers. The encoder expands the receptive field through continuous downsampling operations, the channels' numbers are doubled layer by layer, from 64 to 512, and the corresponding encoded feature map size is halved per layer. The decoder uses upsampling to gradually restore the concatenated feature map to the original resolution. The skip connection mainly consists of the SIA module and the MCM module. In the first four layers, the feature maps obtained by the encoder and the up-sampled advanced semantic features from the fifth layer are integrated through the SIA module to complete the fusion of information between features of different scales. The SIA module makes up for the lack of attention to information by using high-level semantic information. Then we use the MCM module to further realize a more efficient self-attention mechanism through a convolutional modulation strategy and dynamically adjust the convolution kernel size based on the scale of layer-specific features, which skillfully complement the advantages of CNN and Transformer to enhance the segmentation performance. To improve network robustness, apply a deep supervision mechanism to upsample each decoder layer's output to match the ground truth size for supervised learning.

3.1 | Scale Inter-Active Attention Module

After U-Net encoding, diverse feature maps are generated at varying scales, capturing the semantic information of the original image across multiple layers. According to the framework of the traditional U-Net, the obtained feature maps will be used as the skip connection to directly perform the concatenation operation on the channel dimension with the upsampled high-level features. However, the direct concatenation ignores long-distance dependencies to a certain extent, thus affecting the enhancement of model performance. Inspired by the idea of feature complementarity, we propose the SIA module, which uses deep semantic information to supplement shallow-level easily overlooked detailed features. Meanwhile, it makes up for the attention of the proposed Trans-Scale Net to the region information that has not been paid attention to, so as to achieve more effective fusion between features.

The structure of the SIA module is presented in Figure 3. The SIA module acts on the feature maps output by the first four layers of the encoder. By performing an upsampling operation on the features of the next layer to match the scale of the features of this layer, the idea of complementary information between features is used to make up for the information details lost by the network, so as to realize the fusion between multi-scale features.

Specifically, we define the feature map of this layer as $F^i(i = 1, 2, 3, 4)$, and first perform upsampling and convolution operations on the feature map of the next layer G^{i+1} to obtain feature map G^i with the same size and dimension as F^i. Then the two feature maps are sent to the SIA module together. In the SIA module, we first use the Sigmoid [49] activation function to transform F^i to obtain a weight, which represents the

region of interest that the current feature F^i focuses on. And the inversion operation is implemented to obtain the reverse attention weight, which represents the regional information that the current feature F^i may ignore. Then the obtained result is multiplied element-wise with the feature G^i containing deep advanced semantic information to realize attention weighting. The operation fuses information of different scales by using the idea of feature complementarity. After using the complementary features of this layer to filter the deep advanced semantic feature G^i, the residual connection is used to retain the original features of this layer, and supplement the neglected information details. Finally, the multi-layer perceptron is used to further realize the fusion of multi-scale features. The SIA module uses information interaction between features of different scales to make up for the missing information in this layer, and realizes the selection of segmentation target features. The overall process of the SIA module can be expressed by Equation (1):

$$F^i_{fuse} = MLP\left[\left(1-\sigma\left(F^i\right)\right)\odot G^i + F^i\right] \tag{1}$$

Among them, σ represents the Sigmoid activation function that generates the weight, $\odot$ represents the element-wise product operation, and the feature F^i of this layer, the upsampled G^i and the output F^i_{fuse} have the same scale.

3.2 | Multi-Scale Convolution Modulation Module

After the fusion features F^i_{fuse} of four different scales are obtained through the SIA module, they are sent to the MCM module. The MCM module uses convolution modulation to implement a more efficient self-attention mechanism, thereby completing the fusion of information between features. The architecture of the MCM module is shown in Figure 4. Unlike the self-attention mechanism in VIT which generates attention weight A by multiplying Q and K^T, the attention weights A in the MCM module are obtained by $k \times k$ depthwise separable convolutions [50]. Then, the obtained result is multiplied element-by-element by the V obtained through the linear mapping (Conv 1×1) to achieve further attention fusion. The

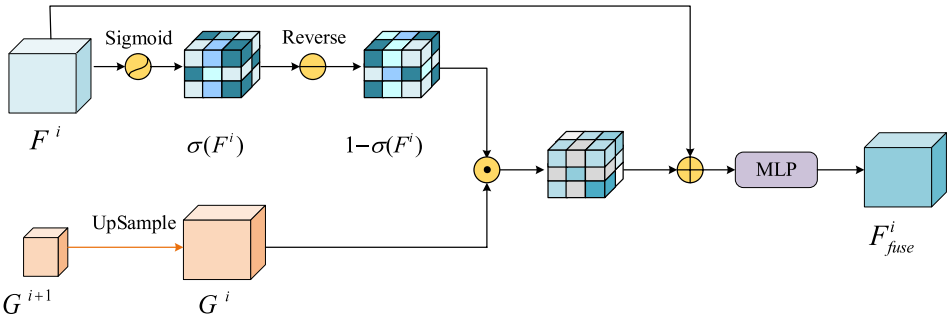

FIGURE 3 | The illustration of the SIA module. Taking the reverse of the encoder feature weights captures easily overlooked details.

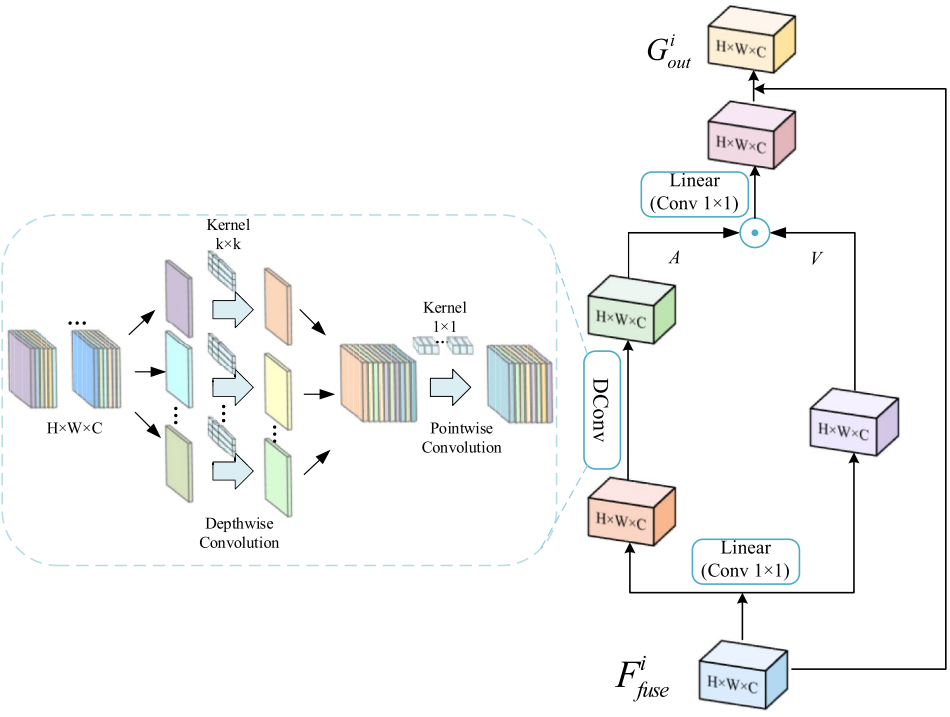

FIGURE 4 | The architecture of the Multi-scale Convolution Modulation. Utilizing depth-wise separable convolutions at different scales to weight the encoded features at each layer optimizes the network's scale-awareness capability.

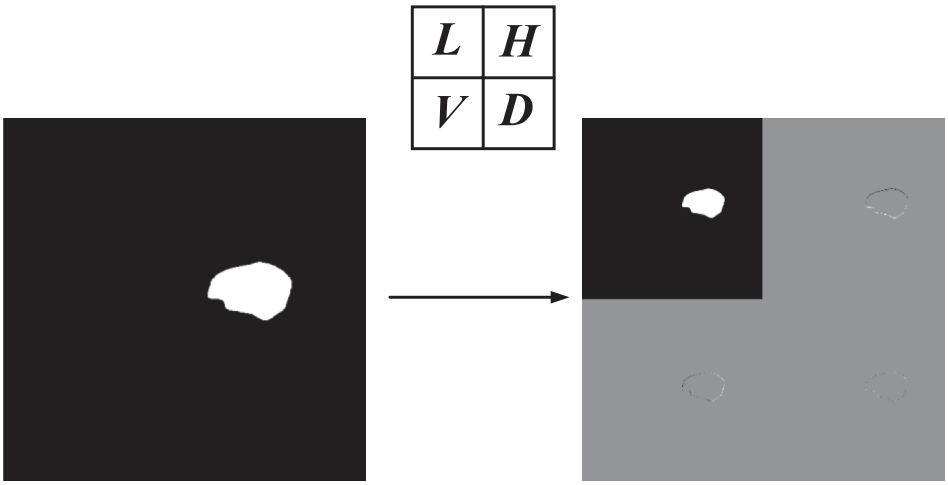

FIGURE 5 | The effect display of first-order wavelet decomposition. The high-frequency texture information is extracted by horizontal high frequency (H), vertical high frequency (V), and diagonal high frequency (D).

information across channels can be integrated through linear mapping. The specific process can be expressed by the following formula:

$$A = DConv_{k\times k}\left(W_A F^i_{fuse}\right) \tag{2}$$

$$V = W_V F^i_{fuse} \tag{3}$$

$$Attention = A \odot V \tag{4}$$

where $\odot$ represents element-wise product, W_A and W_V are weight matrices of two linear maps, and $DConv_{k\times k}$ represents $k \times k$ depthwise separable convolution. Multi-scale convolutional modulation is achieved by adjusting the size of depthwise separable convolutional kernels according to the scale of features in different layers. In the specific implementation process, considering that small-scale deep features have a large receptive field, we use small convolution kernels to match small-scale features, thus the size of the corresponding depth-separable convolution kernels $k \times k$ is sequentially set as 11 $\times$11, 9 $\times$9, 7 $\times$7, and 5 $\times$5, aligned with the number of network layers from top to bottom.

By setting depth-separable convolution kernels of different sizes, the multi-scale convolution modulation is realized, which further enhances the ability of the network to extract multi-scale features.

Finally, we send the four different-scale features generated by the MCM module and the underlying features generated by the encoder to the U-Net decoder and perform feature fusion layer by layer through the concatenation operation. We employ the deep supervision strategy after the decoder to further improve the robustness of the network. Specifically, the output feature maps from the end of each layer are represented as Out0, Out1, and Out2, respectively, up-sampled to the same size as the original image for final loss calculation.

3.3 | Edge Loss Function Based on Wavelet Decomposition

Owing to the slight size of the lesion area, blurred boundaries, and low contrast with surrounding tissues and organs, most segmentation networks pay too much attention to the redundant back area, which affects the segmentation performance. Based on this, we design an edge loss function based on wavelet decomposition to increase the attention to the edge of the target, so as to achieve a more refined segmentation result. Wavelet decomposition [51] has the ability of multi-resolution analysis that allows extracting high-frequency texture details, particularly at image edges. By introducing wavelet decomposition into the segmentation task, the attention to object edges can be improved.

Specifically, we implement wavelet decomposition on both prediction and ground truth to extract high-frequency texture information of object edges, where the high-frequency components include horizontal high frequency (H), vertical high frequency (V) and diagonal high frequency (D). Figure 5 demonstrates the effect of first-order wavelet decomposition on pancreatic cancer ground truth. As shown in Figure 5, indicates the result of using a low-pass filter to transform the image in the horizontal and vertical directions sequentially. indicates that the original image is transformed with the low-pass filter in the vertical directiofirstly, and thesentnd to the high-pass filter to perform the convolution operation in the horizontal direction to obtain the horizontal high-frequency information. Correspondingly, means that the low-pass filter is used to perfothe rm convolution operation in the horizontal direction, and then the high-pass filter is used to perform convolution in the vertical direction to obtain the vertical high-frequency information. represents the diagonal high-frequency information of the original image obtained by successively implementing the convolution on the image using a high-pass filter in the horizontal and vertical directions.

Finally, we apply L1 loss [52] on the two-dimensional plane for each direction component, and accumulate them to obtain

the final edge loss. The specific process can be expressed by Equation (5):

$$L_{edge}\left(\hat{Y},Y\right)=L_1\left(\hat{Y}^H,Y^H\right)+L_1\left(\hat{Y}^V,Y^V\right)+L_1\left(\hat{Y}^D,Y^D\right) \quad (5)$$

where $\hat{Y}$ represents the prediction, Y represents the ground truth, $L_1=\sum_{i=1}^{N}|\hat{y}_i-y_i|$, $\hat{y}_i$ and y_i denotes the the i-th pixel of prediction and the ground truth respectively, and N represents the total number of pixels.

4 | Experimental Results

4.1 | Datasets

We evaluate the performance of the proposed network on four different medical image segmentation datasets: (1) The public NIH(National Institutes of Health)pancreas segmentation dataset, which contains 82 contrast-enhanced abdominal 3D CT scans. The sizes of each CT volume vary from $512\times512\times181$ to $512\times512\times466$. The dataset is randomly divided into four subsets of 21, 21, 20, and 20 following the four-fold cross-validation, (2) The public Medical Segmentation Decathlon (MSD) Challenge pancreatic cancer segmentation dataset, which contains 281 abdominal enhanced CT scans with pancreatic cancer annotations. The sizes of each CT volume vary from $512\times512\times37$ to $512\times512\times751$. The dataset is divided into four parts of 70, 70, 71, and 71 according to the four-fold cross-validation, (3) The private PROMM (Prostate Multi-parametric MRI) prostate cancer segmentation dataset provided by the Shanghai Tongji hospital, which contains mpMRI sequences (ADC, T2W, and DWI) of 171 prostate cancer cases. Each sample contains 20–26 images. Following common settings for the prostate cancer segmentation task [53], we experiment with 5-fold cross-validation, and (4) The public Kvasir-SEG Polyp Segmentation Dataset consists of 1000 RGB images of gastrointestinal polyps with their respective ground truths. The pixel dimensions of the images range from 332×487 to 1920×1072. Following the official recommendation, the dataset is divided into 880 for training and 120 for testing.

The above datasets differ greatly in data distribution, imaging principle and segmentation target shape, among which the private prostate cancer segmentation dataset collected from hospital further strengthens the validation of model generalization performance.

4.2 | Evaluation Metrics

To evaluate the segmentation performance of our Trans-Scale network, we utilize popular metrics, including the Dice Similarity Coefficient (DSC), Precision, Recall, Average Symmetric Surface Distance (ASD) and 95% Hausdorff Distance (HD). These metrics can be defined as follows:

(1) DSC measures the similarity between the ground truth and the prediction, which is widely used for medical image segmentation tasks.

$$DSC=\frac{2\left\|\hat{Y}\cap Y\right\|}{\left\|\hat{Y}\right\|+\|Y\|} \quad (6)$$

(2) Precision measures the proportion of the correctly predicted foreground pixels to the total predicted foreground pixels.

$$Precision=\frac{\left\|\hat{Y}\cap Y\right\|}{\left\|\hat{Y}\right\|} \quad (7)$$

(3) Recall measures the proportion of the correctly predicted foreground pixels to the foreground pixels of ground truth.

$$Recall=\frac{\left\|\hat{Y}\cap Y\right\|}{\|Y\|} \quad (8)$$

(4) ASD is used to evaluate the accuracy of edge segmentation, which measures the average distances between the surface of the prediction and the ground truth.

$$ASD=\frac{1}{S\left(\hat{Y}\right)+S(Y)}\left(\sum_{\hat{y}\in S\left(\hat{Y}\right)}d(\hat{y},S(Y))+\sum_{y\in S(Y)}d\left(y,S\left(\hat{Y}\right)\right)\right) \quad (9)$$

(5) HD is used to evaluate the completeness of target boundary segmentation, which measures the edge contour distance of both the prediction and the ground truth.

$$HD=\max\left\{\max_{y\in S(Y)}\min_{\hat{y}\in S\left(\hat{Y}\right)}d\{y,\hat{y}\},\max_{\hat{y}\in S\left(\hat{Y}\right)}\min_{y\in S(Y)}d\{\hat{y},y\}\right\} \quad (10)$$

where $\hat{Y}$ and Y refer to the prediction and the ground truth respectively, $S\left(\hat{Y}\right)$ denotes the edge point set of the segmentation prediction, $S(Y)$ denotes the ground truth edge point set and $d\{y,\hat{y}\}$ denotes the Euclidean distance between pixel y and pixel $\hat{y}$.

4.3 | Implementation Details

Our framework is constructed using the PyTorch platform and trained on a NVIDIA GeForce RTX 3090 graphics card with 24GB memory. The input images are resized to 128×28 due to memory constraints. The model is optimized using stochastic gradient descent (SGD). The initial learning rate is set to $1\times1e{-}4$, with a momentum setting of 0.9.

The loss function used is the sum of the dice loss [54], the binary cross-entropy loss [55], and the edge loss function based on wavelet decomposition. Among them, the dice loss, which is commonly used in image segmentation tasks, can be defined as Equation (11):

$$L_{dice}=1-\frac{2\sum_{i=1}^{N}y_i\hat{y}_i}{\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}\hat{y}_i} \quad (11)$$

and the binary cross-entropy loss can be defined as Equation (12):

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{12}$$

so the final combined loss function is formulated as Equation (13):

$$L_{final\_loss} = L_{dice} + L_{bce} + L_{edge} \tag{13}$$

where N is the number of pixels, $\hat{y}_i$ is the predicted pixel result of the network, y_i is the value of the corresponding ground truth. The deep supervision strategy is employed to strengthen the robustness of the designed Trans-Scale Net. We refine the loss function during training by assigning proportional coefficients of 1, 0.6, 0.3, and 0.1 to each layer's output including prediction, Out2, Out1, and Out0, respectively, in accordance with their respective degrees of contribution for network performance. The specific training flow of the proposed TS-Net is given by Algorithm 1, the feature symbols correspond to the input and output of each module.

ALGORITHM 1 | Training of proposed TS-net model.

Input: The ROIs obtained from coarse segmentation

Output: Prediction,Out2,Out1,Out0

For every training epoch **do**:

  For every layer **do**:

    Encoder feature F_i ← U-Net based encoder(input)

    F^i_{fuse} ← SIA module (F_i, up(G_{i+1}))

    G^i_{out} ← MCM module(F^i_{fuse})

    Prediction,Out0,Out1,Out2 ← U-Net based decoder(G^i_{out}, up($F_{decoder\_i+1}$))

  End

  Loss ← Loss_function((Prediction, GT) ;(Out2,GT);(Out1,GT);(Out0,GT))

  Backpropagating and update

End

4.4 | Segmentation Results on NIH Dataset

To comprehensively evaluate the performance of the model, the proposed model Trans-Scale in this section will first be compared with other advanced pancreas segmentation methods. Moreover, based on the public NIH pancreas segmentation dataset, the results are compared and analyzed with the classic medical image segmentation models under the same experimental conditions.

4.4.1 | Comparison With State-Of-The-Art Pancreas Segmentation Methods

Table 1 exhibits the comparison of evaluation results on NIH pancreas segmentation datasets. Since the proposed network TS-Net utilizes the SIA module and MCM module to fully complete the interlayer interaction, the loss of details is reduced as much as possible. The TS-Net achieves the best pancreas segmentation performance among the eleven methods (Table 1) with DSC, Precision, and Recall (± standard deviation) values of 90.79 (±1.01)%, 90.62 (±2.37) % and 90.27 (±1.65) %, respectively. The ASD and HD are 0.56 (±0.06) mm and 1.93 (±0.05) mm, respectively, which also indicates that the similarity between the results of our method and manual delineations is high. Furthermore, it is noteworthy that our method exhibits the lowest standard deviation for each metric. The lowest standard deviation demonstrates our method's enhanced robustness and stability across varied CT scans compared to other approaches.

4.4.2 | Comparison With the Classic Medical Image Segmentation Models

Figure 6 presents that the proposed Trans-Scale network outperforms the classic segmentation models (MedT, PVT, TransUNet, and UCTransNet) on the pancreas segmentation task. From the boxplot component of the raincloud plot, we

TABLE 1 | The performance (evaluated by the DSC, Precision, Recall, ASD, and HD) of pancreas segmentation on the NIH dataset.

Method	DSC (%)	Precision (%)	Recall (%)	ASD (mm)	HD (mm)
Oktay et al. [11]	83.1 ± 3.8	82.5 ± 7.3	84 ± 5.3	—	—
Cai et al. [23]	83.3 ± 5.6	84.5 ± 6.2	82.8 ± 8.37	—	—
Li et al. [29]	83.31 ± 6.32	84.09 ± 8.65	83.30 ± 8.54	—	—
Liu et al. [28]	84.10 ± 4.91	83.60 ± 5.85	85.33 ± 8.24	—	—
Zheng et al. [24]	84.37	83.10	86.26	—	—
Li et al. [25]	86.10 ± 3.52	84.97 ± 6.18	86.43 ± 5.30	1.27 ± 0.43	4.40 ± 2.99
Qiu et al. [31]	86.25 ± 4.25	—	—	—	—
Wang et al. [26]	87.4 ± 6.8	89.5 ± 5.8	87.7 ± 7.9	2.89 ± 4.78	18.4 ± 28.19
Huang et al. [27]	87.25 ± 3.27	88.98	89.97	—	—
Chen et al. [30]	87.91 ± 2.65	90.43 ± 3.77	85.77 ± 4.61	0.73 ± 0.16	—
Ours	**90.79 ± 1.01**	**90.62 ± 2.37**	**90.27 ± 1.65**	**0.56 ± 0.06**	**1.93 ± 0.05**

Note: Optimal values are shown in bold. "—" indicates the data are not reported in the literature.

can observe that the proposed Trans-Scale network achieves the highest DSC score with minimal fluctuation range, which presents the superiority of our method. Additionally, a more densely packed distribution of data point colors and the peak of the half violin plot intuitively reflect the robustness of the Trans-Scale network.

To further demonstrate the advantages of the network, Table 2 compares the other indicators with the classic Transformer-based segmentation models on the NIH dataset. The proposed segmentation network Trans-Scale significantly outperforms other models with all p-values <0.05 (Dsc index, t-test). It further demonstrates that TS-Net effectively integrates the MCM module to promote multi-level feature interaction and the SIA module to compensate for lost details, thereby enhancing segmentation performance.

Figure 7 illustrates a visual comparative analysis of pancreas segmentation outcomes between our proposed network and the classic medical image segmentation models, utilizing the NIH dataset for evaluation. The ground truth is delineated within the original image in the leftmost column. From Figure 7, we can see that the segmentation results provided by our proposed network Trans-Scale exhibit a closer fit to the ground truth and obtain improved continuity compared to the segmentation results of other models. Specifically, as indicated in the first row of Figure 7, the Trans-Scale network delineates the edge of the pancreatic head and tail (highlighted within the yellow box) more completely compared to other models.

4.5 | Segmentation Results on MSD Dataset

Since the datasets and codes of some pancreatic cancer segmentation tasks are not publicly available, the comparison with other pancreatic cancer segmentation models is based on the pancreatic cancer segmentation data reported in the literature.

4.5.1 | Comparison With State-Of-The-Art Pancreatic Cancer Segmentation Methods

Table 3 shows the comparison of evaluation results on different pancreatic cancer segmentation datasets. Since the pancreatic

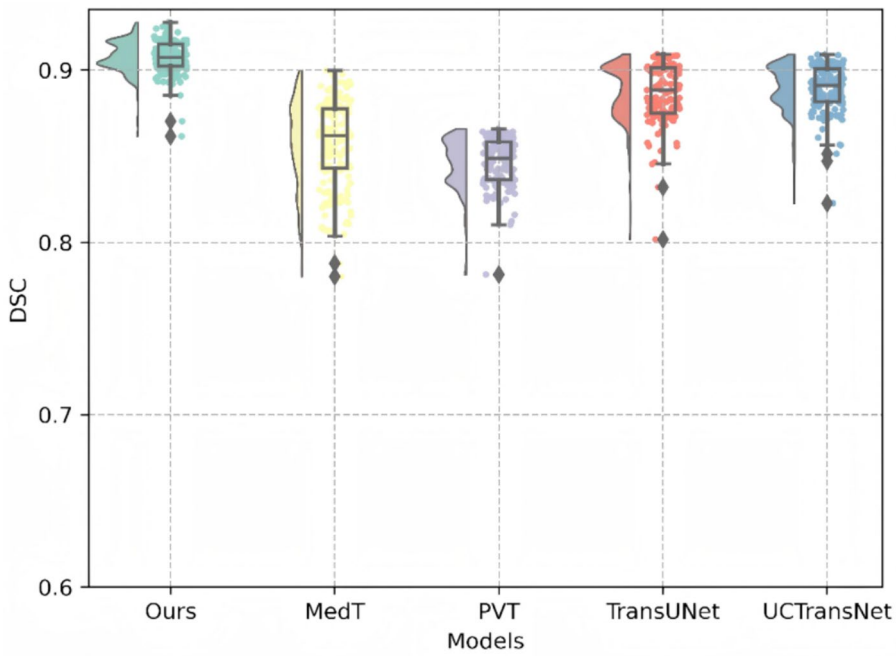

FIGURE 6 | Performance comparison (DSC) with different classic segmentation models in a raincloud plot for the NIH dataset. Most of the DSC scores of our proposed model are above 0.9.

TABLE 2 | The performance (evaluated by the DSC, Precision, Recall, ASD, and HD) of pancreas segmentation on the NIH dataset. The p-value is obtained by comparing the DSC of ours with other methods. Optimal values are shown in bold.

Method	DSC (%)	Precision (%)	Recall (%)	ASD (mm)	HD (mm)	p-value
MedT [19]	85.50 ± 7.91	86.45 ± 6.14	86.89 ± 7.02	0.87 ± 0.22	2.17 ± 0.13	$2.99 \times e^{-11}$
PVT [20]	84.82 ± 1.27	87.33 ± 2.39	85.38 ± 3.17	1.01 ± 0.07	2.29 ± 0.06	$3.54 \times e^{-13}$
TransUNet [21]	88.59 ± 2.48	86.57 ± 3.27	89.13 ± 3.85	0.66 ± 0.12	2.01 ± 0.13	$1.26 \times e^{-10}$
UCTransNet [22]	88.87 ± 1.29	87.72 ± 2.56	88.52 ± 3.21	0.62 ± 0.08	1.99 ± 0.06	$4.53 \times e^{-9}$
Ours	**90.79 ± 1.01**	**90.62 ± 2.37**	**90.27 ± 1.65**	**0.56 ± 0.06**	**1.93 ± 0.05**	—

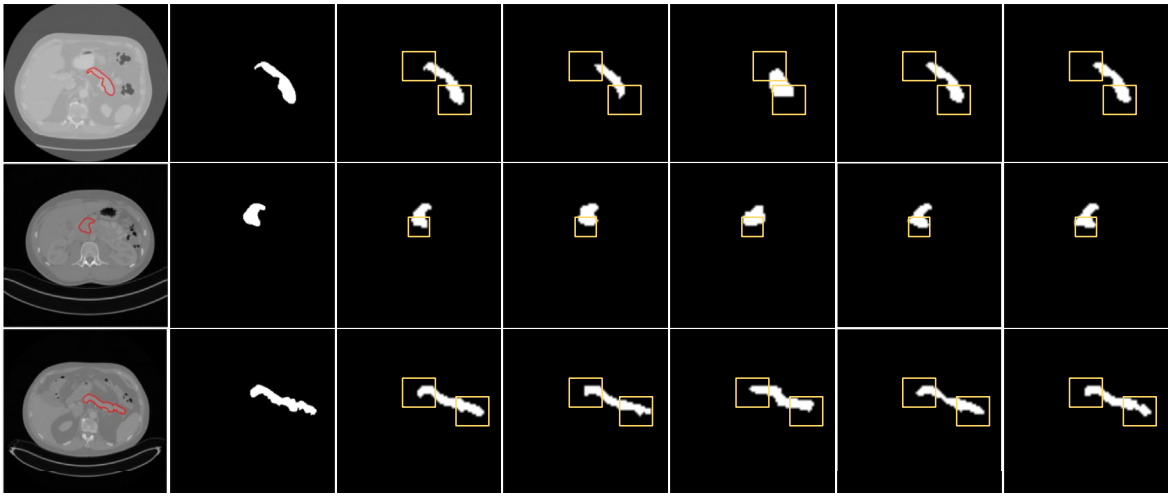

FIGURE 7 | Comparison of the visualization results for different classic medical image segmentation models on the NIH dataset. The ground truth is delineated within the original image at the leftmost column. Segmentation regions with clear distinctions between different models are highlighted with yellow boxes.

TABLE 3 | The performance (evaluated by the DSC, Precision and Recall, ASD, and HD) of pancreatic cancer segmentation on different datasets.

Method	Dataset	DSC (%)	Precision (%)	Recall (%)	ASD (mm)	HD (mm)
Mahmoudi et al. [37]	138 Cases	57.30 ± 15.00	57.80 ± 23.00	78.00 ± 9.00	—	3.73 ± 0.78
Wang et al. [32]	800 Cases	60.29 ± 21.60	—	—	—	—
Li et al. [36]	631 Cases	64.16	—	—	—	—
Li et al. [35]	327 Cases	65.60 ± 15.32	—	—	3.01 ± 4.16	—
Liang et al. [42]	40 Cases	73.00 ± 9.00	—	—	1.82 ± 0.84	8.11 ± 4.09
Li et al. [34]	Public(MSD) 281 Cases	50.12 ± 30.86	—	—	—	—
Li et al. [36]		57.53	—	—	6.64	14.78
Li et al. [39]		59.85	—	63.07	3.77	—
Wang et al. [41]		63.40 ± 23.67	—	—	—	—
Ours		**76.62 ± 4.34**	**87.27 ± 8.09**	**82.43 ± 7.77**	**1.17 ± 0.14**	**2.33 ± 0.10**

Note: Optimal values are shown in bold. "—" indicates the data are not reported in the literature.

cancer region is smaller than the pancreatic organ and has blurred edges, the pancreatic cancer segmentation index is still at a relatively low level. The upper part of the table is the indicator results on the private pancreatic cancer segmentation datasets, and the lower part is the indicator distribution on the public MSD dataset. The proposed network Trans-Scale achieves a DSC score of 76.62%, which is superior to the DSC indicators on other pancreatic cancer segmentation datasets and performs well on the MSD pancreatic cancer segmentation dataset. The precision of 87.27% and the recall of 82.43% indicate that the designed network Trans-Scale has a higher segmentation accuracy and a lower probability of misdiagnosis. The ASD distance of 1.17 mm and the HD distance of 2.33 mm prove that the segmentation result of the designed pancreatic cancer segmentation network Trans-Scale is closer to the ground truth on the edge. The smaller variance fluctuation proves that the model has higher robustness.

4.5.2 | Comparison With the Classic Medical Image Segmentation Models

Figure 8 illustrates the superior performance of our Trans-Scale network compared to classic Transformer-based segmentation models for pancreatic cancer segmentation. The diverse shape of pancreatic cancer, being smaller and more complex compared to the pancreas, poses a significant segmentation challenge, with model DSC predominantly ranging between 0.7 and 0.8.

The comparison of DSC score further validates the advantages of Trans-Scale network from two perspectives: Firstly, the higher DSC score substantiates the effectiveness of our proposed model in pancreatic cancer segmentation. It can be clearly seen that our network's DSC values are mainly distributed above 0.7 compared to other models. Secondly,

our network's predictions also exhibit minimal fluctuations, demonstrating its remarkable stability.

Furthermore, Table 4 shows that the designed segmentation network Trans-Scale achieves higher precision and recall on the MSD dataset compared with other methods. This demonstrates that TS-Net produces fewer false positives and false negatives in the segmentation results for pancreatic cancer.

Figure 9 reflects the comparison of the visualization results of pancreatic cancer segmentation between our proposed network and the classic segmentation models on the MSD dataset, where the purple is the ground truth, and the sky blue is the prediction of different networks. From Figure 9, we can see the following: on the one hand, for the pancreatic cancer region with irregular edges, the prediction results of the proposed Trans-Scale model are generally consistent with the ground truth in terms of shape and structure; on the other hand, the proposed model is closer to the ground truth in the target contour compared with other methods, showing the outperforming feature extraction ability of the model. Although the pancreatic cancer area to be segmented is smaller in size than the pancreas, and the redundant background in the CT image occupies a larger area to some extent, the prediction mitigates the occurrences of over-segmentation or under-segmentation on the target edge as much as possible when compared with other methods.

4.6 | Segmentation Results on Prostate PROMM Dataset

The prostate cancer segmentation experiment is conducted on the dataset provided by Shanghai Tongji Hospital. Since the PROMM dataset is private, we only compare the Trans-Scale network with the classic medical image segmentation models.

Figure 10 shows the DSC scores of the five-fold cross-validation of the Trans-Scale network on the PROMM prostate cancer segmentation dataset, which reflects that our network can balance the differences between different samples and achieve accurate prostate cancer segmentation. For the prostate segmentation task with a smaller data sample size, the proposed Trans-Scale network also demonstrates effective extraction of target features.

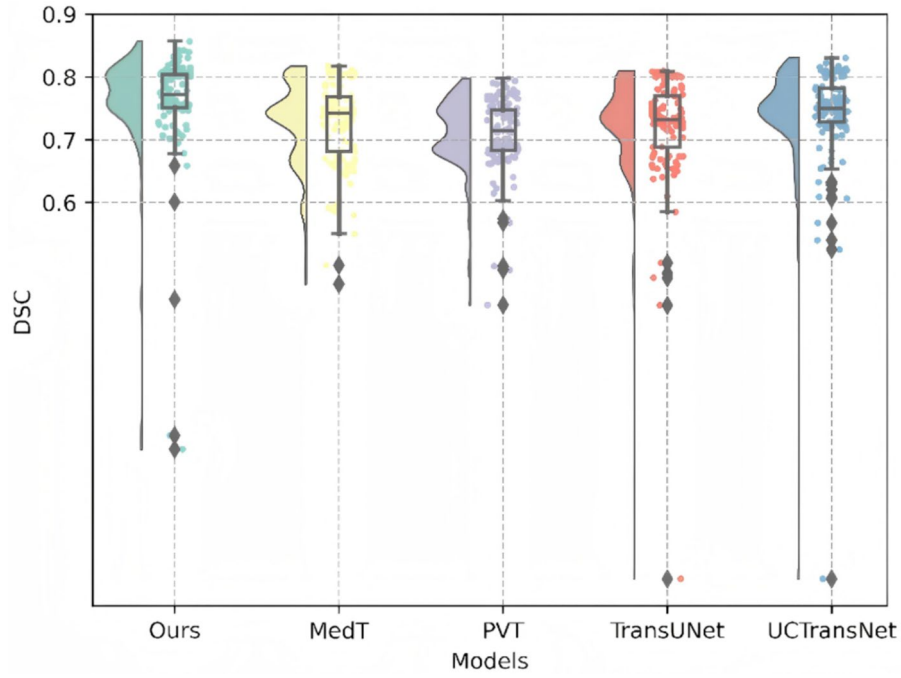

FIGURE 8 | Performance comparison (DSC) with different classic segmentation models in raincloud plot for MSD dataset. The average DSC achieved by our segmentation model exceeds that of other models.

TABLE 4 | The performance (evaluated by the DSC, Precision, Recall, ASD, and HD) of pancreatic cancer segmentation on the MSD dataset.

Method	DSC (%)	Precision (%)	Recall (%)	ASD (mm)	HD (mm)	*p*-value
MedT [19]	74.20 ± 7.65	82.82 ± 9.07	80.06 ± 10.24	1.29 ± 0.18	2.42 ± 0.12	$3.25 \times e^{-14}$
PVT [20]	71.47 ± 5.65	82.29 ± 8.90	79.95 ± 9.07	1.43 ± 0.15	2.60 ± 0.09	$5.12 \times e^{-15}$
TransUNet [21]	73.30 ± 8.26	84.95 ± 10.67	79.06 ± 15.57	1.37 ± 0.19	2.52 ± 0.16	$2.17 \times e^{-12}$
UCTransNet [22]	75.18 ± 5.13	85.23 ± 8.94	77.35 ± 11.29	1.20 ± 0.10	2.37 ± 0.12	$1.78 \times e^{-15}$
Ours	**76.62 ± 4.34**	**87.27 ± 8.09**	**82.43 ± 7.77**	**1.17 ± 0.14**	**2.33 ± 0.10**	—

Note: Optimal values are shown in bold.

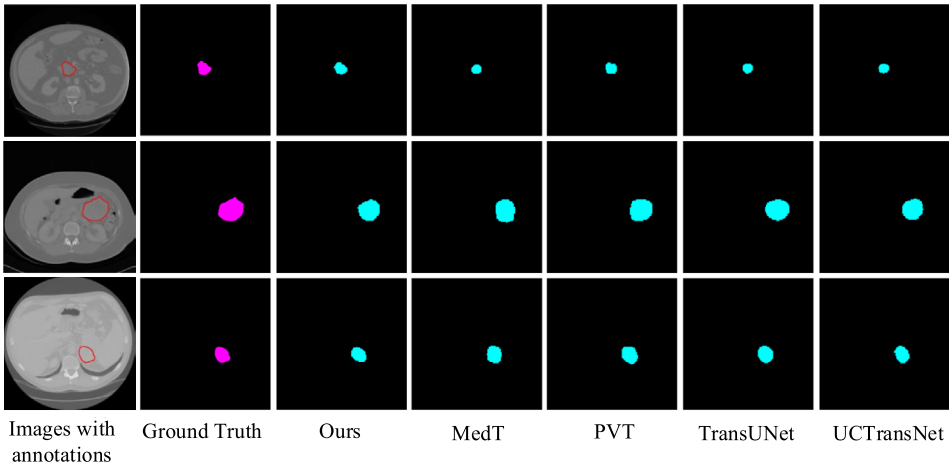

FIGURE 9 | Comparison of the visualization results for different classic medical image segmentation models on the MSD dataset. The ground truth is delineated within the original image in the leftmost column.

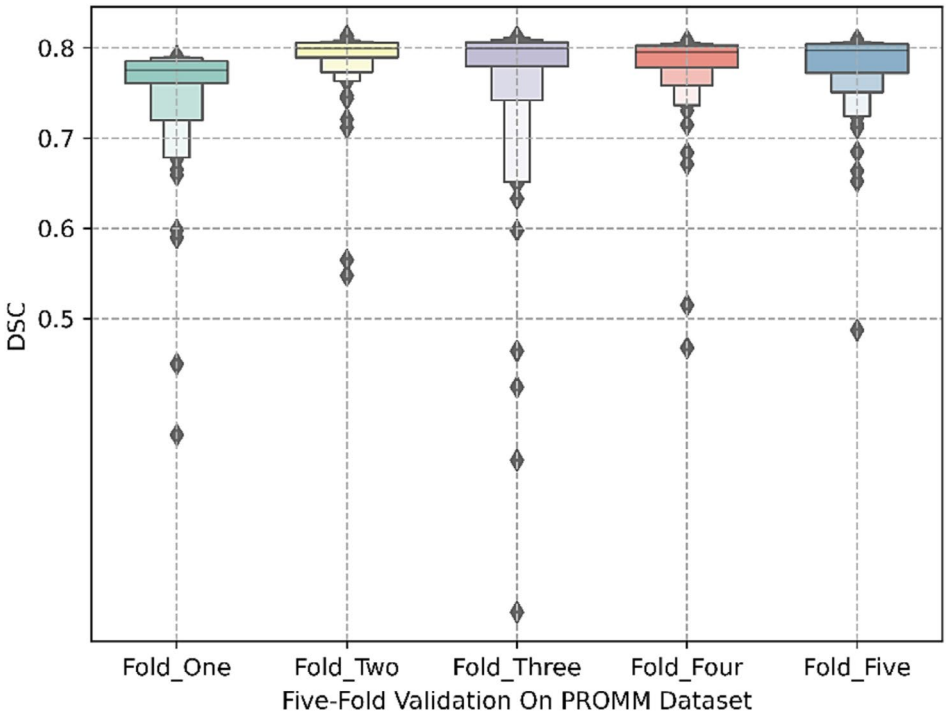

FIGURE 10 | Enhanced boxplot representation for five-fold cross-validation Results of the DSC Metric on the PROMM dataset. The consistency of the DSC distribution between different folds proves the robustness of our proposed model.

Table 5 shows the comparison of the results between our proposed network and the classic Transformer-based segmentation models on the PROMM dataset. It can be seen that the Trans-Scale network shows better segmentation indicators than other models. Although the prostate cancer area accounts for a larger proportion in the PROMM dataset, the effective number of slices containing the target area is very limited, which limits the performance of the PVT network and the MedT network to a certain extent. It causes the models to pay too much attention to the background region, thus affecting the prediction of the model. In addition, the following information can be seen from Table 5: on the one hand, the proposed network achieves the ASD distance of 1.50 mm and the HD distance of 2.23 mm, indicating that the distance between the prediction and the ground truth is closer. The competitive performance on this dataset further reflects the outperforming generalization performance of our proposed model; on the other hand, the proposed Trans-Scale Network might be disturbed by the redundant background to a certain extent so that it fluctuates more variance, reflecting that the network still has room for further enhancement.

Figure 11 shows the visualization results for prostate cancer segmentation on the PROMM dataset. The three columns on the left show the renderings of MRI in the three modalities of T2W, ADC and DWI. It can be seen that T2W shows the clearest MRI image, where the prostate cancer area and the surrounding background are very clear, and the amount of information

TABLE 5 | The performance (measured by the DSC, Precision, Recall, ASD, and HD) of prostate cancer segmentation on the PROMM dataset.

Method	DSC (%)	Precision (%)	Recall (%)	ASD (mm)	HD (mm)	*p*-value
MedT [19]	67.82 ± 15.87	69.26 ± 15.28	51.56 ± 15.31	2.94 ± 0.63	2.96 ± 0.25	$1.25 \times e^{-13}$
PVT [20]	58.50 ± 1.82	67.35 ± 6.59	53.73 ± 4.49	2.34 ± 0.30	2.88 ± 0.06	$2.76 \times e^{-11}$
TransUNet [21]	79.80 ± 1.18	84.82 ± 6.11	77.80 ± 5.00	1.69 ± 0.22	2.29 ± 0.04	$5.48 \times e^{-11}$
UCTransNet [22]	79.83 ± 1.09	87.32 ± 7.23	89.65 ± 4.18	1.54 ± 0.12	2.26 ± 0.04	$4.69 \times e^{-13}$
Ours	**80.70 ± 6.40**	**90.53 ± 7.02**	**92.76 ± 7.65**	**1.50 ± 0.29**	**2.23 ± 0.16**	—

Note: Optimal values are shown in bold.

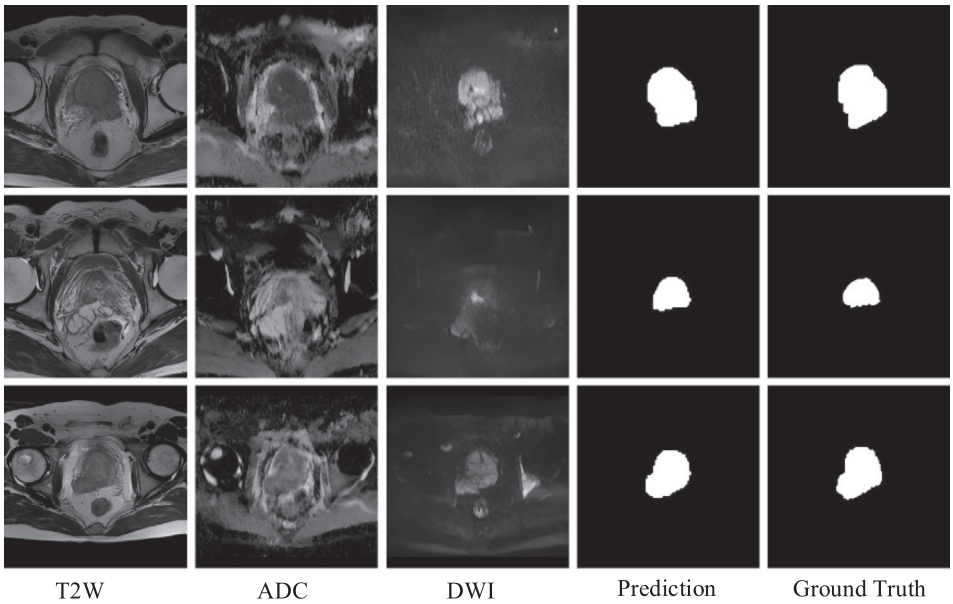

FIGURE 11 | The visualization of prediction results on the PROMM dataset.T2W, ADC, and DWI represent three different modalities in MRI images.

that can be provided is also the largest. Compared with the T2W image, the ADC is slightly blurred. The effect of the DWI image is the most blurred, and only the approximate prostate cancer area can be seen. However, the influence of the surrounding redundant background is eliminated in the DWI image to some extent, and the prediction of the network can be further promoted. Based on this, the network adopts a three-channel input composed of three different modalities and assists the prediction of the network by providing supplementary information of different modalities. From the two columns on the right, it can be seen that the prediction and the ground truth are very close in shape, reflecting the outperforming performance of the proposed Trans-Scale network.

4.7 | Segmentation Results on Kvasir-SEG Dataset

On the Kvasir-SEG polyp segmentation Dataset, TS-Net attained the highest dice score of 91.42% and the minimum distance metrics for HD and ASD, as illustrated in Table 6, demonstrating its strong generalization capabilities. Despite considerable disparities in polyp shapes, sizes, positional distributions, and color variations across different samples, TS-Net consistently delivers complete and precise predictions, as presented in Figure 12 (The red solid line represents the ground truth (GT), and the green solid line represents the predicted results). This reaffirms the model's ability to extract target features effectively, thus significantly reducing the likelihood of mis-segmentation.

5 | Discussion

5.1 | Comparison of Computational Efficiency

To assess the computational efficiency of TS-Net, we conducted an analysis on the MSD pancreatic cancer segmentation dataset, where we compared the parameter count, training time, and inference time of the TS-Net network with four other state-of-the-art methods for pancreatic cancer segmentation, as presented in Table 7.

On the one hand, it is evident that TS-Net has the fewest parameters compared to the other methods and has achieved the highest Dice score. On the other hand, TS-NET significantly reduced both training and inference times, showcasing a remarkable balance between precision and efficiency. This underscores the superior practical performance and applicability of TS-Net.

TABLE 6 | The performance (measured by the DSC, Precision, Recall, ASD, and HD) of polyp segmentation on the Kvasir-SEG dataset.

Method	DSC(%)	Precision(%)	Recall(%)	ASD(mm)	HD(mm)	*p*-value
MedT [19]	88.24 ± 6.23	89.93 ± 5.88	89.79 ± 5.76	6.34 ± 2.31	5.57 ± 2.15	$6.23 \times e^{-14}$
PVT [20]	87.67 ± 2.03	89.50 ± 1.93	89.63 ± 1.82	7.11 ± 1.94	6.50 ± 1.76	$3.11 \times e^{-15}$
TransUNet [21]	89.83 ± 1.82	91.33 ± 1.66	91.26 ± 1.73	5.02 ± 1.69	4.83 ± 1.31	$4.45 \times e^{-15}$
UCTransNet [22]	89.95 ± 1.36	91.15 ± 1.58	92.78 ± 1.62	3.79 ± 1.69	4.67 ± 1.25	$1.04 \times e^{-14}$
Ours	**91.42 ± 0.55**	**91.36 ± 0.43**	**93.64 ± 0.46**	**2.87 ± 0.12**	**3.05 ± 0.14**	—

Note: Optimal values are shown in bold.

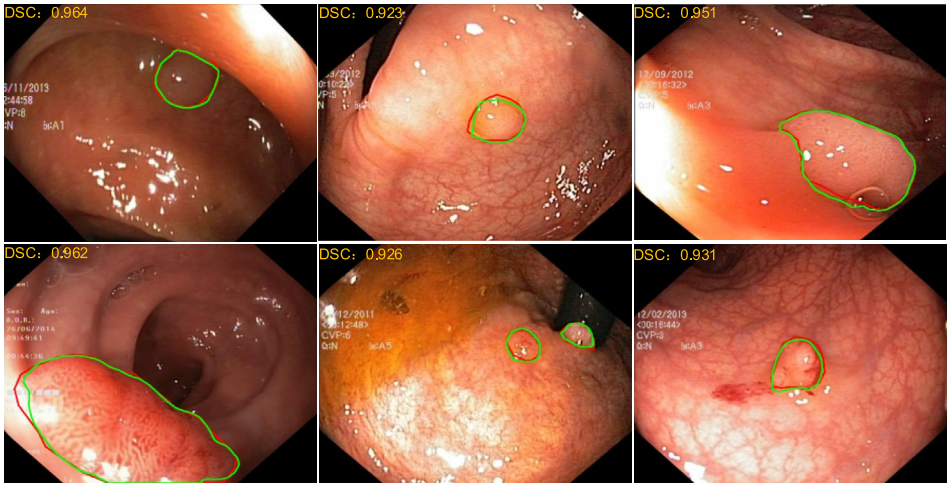

FIGURE 12 | The visualization of prediction results on the Kvasir-SEG polyp segmentation dataset. The red solid line represents the ground truth (GT), and the green solid line represents the predicted results. Polyp regions with different morphologies can be accurately segmented by our proposed model.

TABLE 7 | The comparison of computational efficiency on the public MSD dataset.

Method	Params (M)	Training time (h)	Inference time (min)	DSC (%)
Qiu et al. [44]	104.69	7–8	8–9	59.37 ± 5.78
Ju et al. [40]	45.17	5–6	7–8	63.32 ± 6.37
Chen et al. [33]	98.45	7–8	8–9	66.65 ± 15.21
Li et al. [43]	86.93	6–7	7–8	71.06 ± 6.85
Ours	**32.18**	**4–5**	**3–4**	**76.62 ± 4.34**

Note: Optimal values are shown in bold.

5.2 | Ablation Study

In order to validate the effectiveness of the fundamental modules within our proposed network, we implement ablation experiments on the Trans-Scale model using the MSD dataset. As indicated in Table 8, we individually removed each of the critical modules designed for the network and assessed the resulting network's performance based on the DSC metric.

Table 8 shows the results of the ablation experiment on the MSD dataset. The last row is the Trans-Scale network proposed. The data from the first row reflects that the model DSC index drops by 1.31% without SIA module, indicating the designed SIA module can effectively use different scale features to compensate for overlooked information by the network. The proposed MCM module has a 2.25% impact on the network. It demonstrates MCM module using convolution modulation can fully combine the advantages of CNN and Transformer to enhance the ability of extracting features. The Edge loss function based on wavelet transform has the greatest impact on the network. The network performance will be reduced by 3.43% after removing this module, indicating that in the pancreatic cancer segmentation task, the supervision of the target edge with the help of the loss function can effectively enhance the high-frequency texture details. As shown in Figure 13, where the red line represents the ground truth and the green line represents the predicted results, the final predictions in

the last column exhibit contours closer to the ground truth compared to those in the third column without utilizing the edge loss. Moreover, adding this loss function during training has an almost negligible impact on the training time, resulting in minimal computational overhead.

TABLE 8 | The ablation study of the Trans-Scale Network on the MSD dataset.

SIA	MCM	Dice loss	BCE loss	Edge loss	DSC(%)
—	√	√	√	√	75.31
√	—	√	√	√	74.37
√	√	√	√	—	73.19
√	√	√	—	—	72.98
√	√	—	√	—	72.02
√	√	√	√	√	**76.62**

Note: "—" indicates the exclusion of a module. The Optimal values is shown in bold.

5.3 | Limitations and Future Work

Despite the Trans-Scale Network demonstrating remarkable performance for various medical image segmentation tasks, there is still room for improvement. Specifically, the fine-segmentation stage heavily relies on the fast localization from the coarse stage. However, we observe that a small number of the ROIs (regions of interest) obtained from the coarse stage fail to encompass all target regions, thereby limiting the performance enhancement of these slices during the fine segmentation stage. As shown in Figure 14, the red solid line and green solid line represent the ground truth and the prediction, respectively. The segmentation network's attention is focused on the main target regions, leading to the omission of some topologically disconnected small target areas.

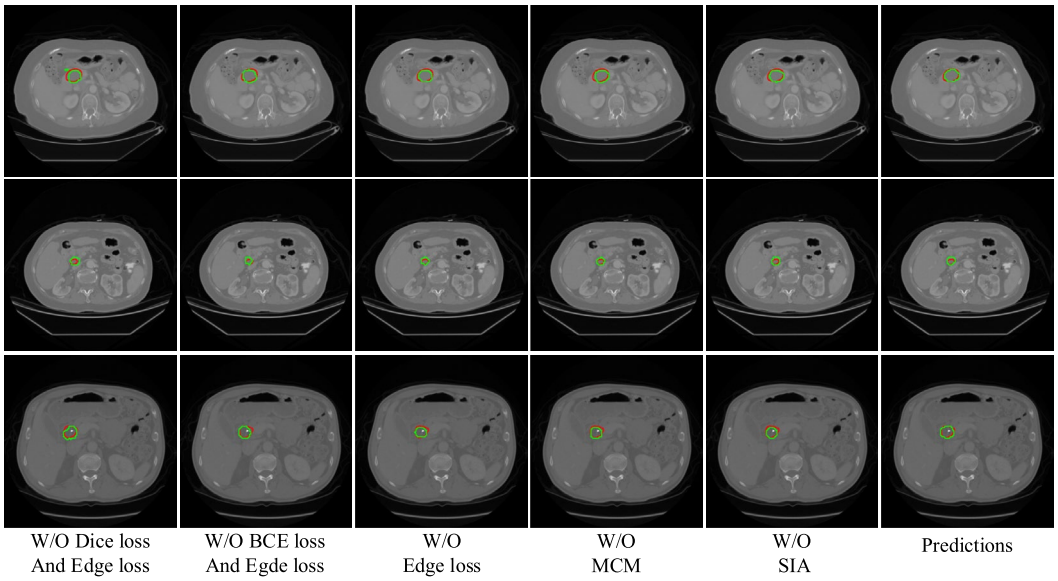

FIGURE 13 | The visualization of ablation experiments on the MSD dataset. The red solid line represents the ground truth (GT), and the green solid line represents the predicted results. W/O means without a module.

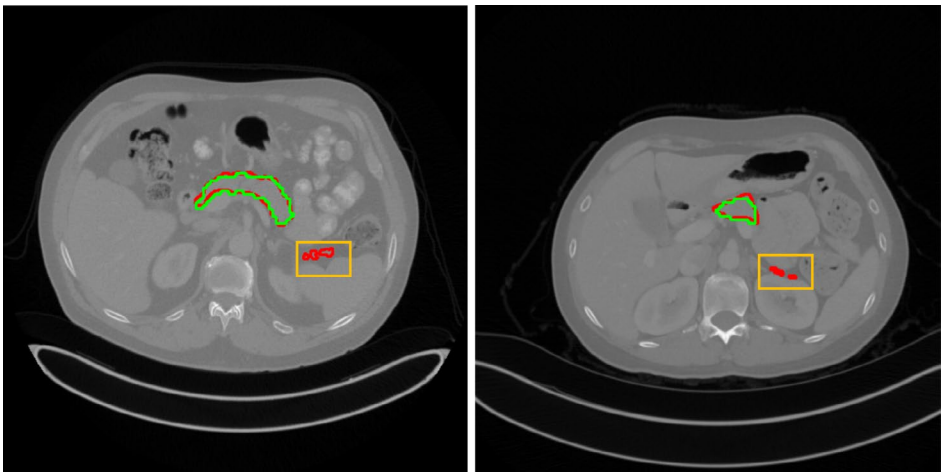

FIGURE 14 | The visualization of failure cases on the NIH pancreas segmentation dataset. The red solid line represents the ground truth (GT), and the green solid line represents the predicted results. The partially missed small target regions are highlighted with yellow boxes.

Therefore, we plan to explore how to extract more accurate ROIs from the coarse stage, such as designing an attention mechanism for small target regions or integrating coarse segmentation results into the fine segmentation stage to achieve end-to-end learning and iterative optimization in the future.

6 | Conclusion

This paper proposes a novel medical image segmentation network, Trans-Scale. Specifically, a SIA module, a MCM module, and an edge loss function based on wavelet decomposition are proposed. Among them, the SIA module compensates for the information details lost by the network, effectively facilitating the interaction between high-level semantic features and low-level spatial features. The proposed MCM module simplifies the self-attention mechanism utilizing convolution modulation that incorporates multi-scale large-kernel convolution into depth-separable convolution, enhancing the feature interaction ability of the network. For the problems of blurred edges and low contrast in medical images, the edge loss function based on wavelet decomposition is used to effectively supervise the high-frequency texture information and further improve the segmentation performance of the network.

Our method is evaluated on four different medical image datasets, and the experimental results demonstrate that the proposed Trans-Scale network achieves superior segmentation performance compared with other state-of-the-art methods. We also make a comparison with the classic medical image segmentation networks, which further proves the outperforming segmentation performance of the proposed network. Finally, the visualization of segmentation results proves the practicability of the network.

Author Contributions

HuiFang Wang: conceptualization, methodology, software, validation, writing – original draft, writing – editing. **Dawei Yang:** data curation, resources, funding acquisition. **Yu Zhu:** methodology, resources, supervision, project administration, funding acquisition, writing – review and editing. **YaTong Liu:** conceptualization, software. **Jiongyao Ye:** resources, writing – review.

Acknowledgments

The authors greatly appreciate the financial support of the National Natural Science Foundation of China (62476088, 82170110), Fujian Province Department of Science and Technology (2022D014), Science and Technology Commission of Shanghai Municipality (20DZ2254400, 20DZ2261200), Shanghai Municipal Science and Technology Major Project (ZD2021CY001) and Shanghai Municipal Key Clinical Specialty (shslczdzk02201).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Medical Segmentation Decathlon at http://medicaldecathlon.com/, reference number [56].

References

1. J. Li, W. Wang, C. Chen, et al., "TransBTSV2: Towards Better and More Efficient Volumetric Segmentation of Medical Images," (2022). arXiv:2201.12785.arXiv e-prints.

2. H. Ma, Y. Zou, and P. X. Liu, "MHSU-Net: A More Versatile Neural Network for Medical Image Segmentation," *Computer Methods and Programs in Biomedicine* 208 (2021): 106230.

3. B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "Transattunet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence* 8 (2023): 55–68.

4. X. Xie, W. Zhang, X. Pan, et al., "Canet: Context Aware Network With Dual-Stream Pyramid for Medical Image Segmentation," *Biomedical Signal Processing and Control* 81 (2023): 104437.

5. Z. Ren, Y. Zhang, and S. Wang, "A Hybrid Framework for Lung Cancer Classification," *Electronics* 11, no. 10 (2022): 1614.

6. Z. Ren, Q. Lan, Y. Zhang, and S. Wang, "Exploring Simple Triplet Representation Learning," *Computational and Structural Biotechnology Journal* 23 (2024): 1510–1521.

7. M. Baldeon-Calisto and S. K. Lai-Yuen, "AdaResU-Net: Multiobjective Adaptive Convolutional Neural Network for Medical Image Segmentation," *Neurocomputing* 392 (2020): 325–340.

8. T. Shan and J. Yan, "SCA-Net: A Spatial and Channel Attention Network for Medical Image Segmentation," *IEEE Access* 9 (2021): 160926–160937.

9. J. Zhuang, "LadderNet: Multi-Path Networks Based on U-Net for Medical Image Segmentation," (2018). arXiv preprint arXiv:181007810.

10. Z. Ren, Y. Zhang, and S. Wang, *Large Foundation Model for Cancer Segmentation*, vol. 23 (SAGE Publications Sage, 2024), 15330338241266205, https://doi.org/10.1177/15330338241266205.

11. O. Oktay, J. Schlemper, L. L. Folgoc, et al., "Attention u-Net: Learning Where to Look for the Pancreas," (2018). arXiv preprint arXiv:180403999.

12. Y. Cai and Y. Wang, "Ma-Unet: An Improved Version of Unet Based on Multi-Scale and Attention Mechanism for Medical Image Segmentation," in *Paper Presented at: Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)* (IEEE, 2022).

13. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," *Advances in Neural Information Processing Systems* 30 (2017).

14. S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical Image Segmentation Using Squeeze-And-Expansion Transformers," (2021). arXiv preprint arXiv:210509511.

15. G. Xu, X. Wu, X. Zhang, and X. He, "Levit-Unet: Make Faster Encoders With Transformer for Medical Image Segmentation," (2021). arXiv preprint arXiv:210708623.

16. X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-Unet: Axial Fusion Transformer Unet for Medical Image Segmentation," in *Paper Presented at: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (IEEE, 2022).

17. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale," (2020). arXiv preprint arXiv:201011929.

18. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation With Transformers," *Advances in Neural Information Processing Systems* 34 (2021): 12077–12090.

19. J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated Axial-Attention for Medical Image Segmentation," in *Paper Presented at: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,*

Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24 (Springer: 2021).

20. W. Wang, E. Xie, X. Li, et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions," in *Paper Presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021).

21. J. Chen, Y. Lu, Q. Yu, et al., "Transunet: Transformers Make Strong Encoders for Medical Image Segmentation," (2021). arXiv preprint arXiv:210204306.

22. H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the Skip Connections in u-Net From a Channel-Wise Perspective With Transformer," in *Paper Presented at: Proceedings of the AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2022).

23. J. Cai, L. Lu, F. Xing, and L. Yang, "Pancreas Segmentation in CT and MRI Images via Domain Specific Network Designing and Recurrent Neural Contextual Learning," (2018). arXiv preprint arXiv:180311303.

24. H. Zheng, Y. Chen, X. Yue, et al., "Deep Pancreas Segmentation With Uncertain Regions of Shadowed Sets," *Magnetic Resonance Imaging* 68 (2020): 45–52.

25. W. Li, S. Qin, F. Li, and L. Wang, "MAD-UNet: A Deep U-Shaped Network Combined With an Attention Mechanism for Pancreas Segmentation in CT Images," *Medical Physics* 48, no. 1 (2021): 329–341.

26. Y. Wang, G. Gong, D. Kong, et al., "Pancreas Segmentation Using a Dual-Input v-Mesh Network," *Medical Image Analysis* 69 (2021): 101958.

27. M. Huang, C. Huang, J. Yuan, and D. Kong, "A Semiautomated Deep Learning Approach for Pancreas Segmentation," *Journal of Healthcare Engineering* 2021 (2021): 3284493.

28. S. Liu, X. Yuan, R. Hu, et al., "Automatic Pancreas Segmentation via Coarse Location and Ensemble Learning," *IEEE Access* 8 (2019): 2906–2914.

29. M. Li, F. Lian, and S. Guo, "Automatic Pancreas Segmentation Using Double Adversarial Networks With Pyramidal Pooling Module," *IEEE Access* 9 (2021): 140965–140974.

30. Y. Chen, C. Xu, W. Ding, S. Sun, X. Yue, and H. Fujita, "Target-Aware U-Net With Fuzzy Skip Connections for Refined Pancreas Segmentation," *Applied Soft Computing* 131 (2022): 109818.

31. C. Qiu, Z. Liu, Y. Song, et al., "RTUNet: Residual Transformer UNet Specifically for Pancreas Segmentation," *Biomedical Signal Processing and Control* 79 (2023): 104173.

32. Y. Wang, P. Tang, Y. Zhou, W. Shen, E. K. Fishman, and A. L. Yuille, "Learning Inductive Attention Guidance for Partially Supervised Pancreatic Ductal Adenocarcinoma Prediction," *IEEE Transactions on Medical Imaging* 40, no. 10 (2021): 2723–2735.

33. X. Chen, Z. Chen, J. Li, Y.-D. Zhang, X. Lin, and X. Qian, "Model-Driven Deep Learning Method for Pancreatic Cancer Segmentation Based on Spiral-Transformation," *IEEE Transactions on Medical Imaging* 41, no. 1 (2021): 75–87.

34. Z. Li, H. Pan, Y. Zhu, and A. K. Qin, "PGD-UNet: A Position-Guided Deformable Network for Simultaneous Segmentation of Organs and Tumors," in *Paper Presented at: 2020 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2020).

35. J. Li, C. Feng, Q. Shen, X. Lin, and X. Qian, "Pancreatic Cancer Segmentation in Unregistered Multi-Parametric MRI With Adversarial Learning and Multi-Scale Supervision," *Neurocomputing* 467 (2022): 310–322.

36. J. Li, L. Qi, Q. Chen, Y.-D. Zhang, and X. Qian, "A Dual Meta-Learning Framework Based on Idle Data for Enhancing Segmentation of Pancreatic Cancer," *Medical Image Analysis* 78 (2022): 102342.

37. T. Mahmoudi, Z. M. Kouzahkanan, A. R. Radmard, et al., "Segmentation of Pancreatic Ductal Adenocarcinoma (PDAC) and Surrounding Vessels in CT Images Using Deep Convolutional Neural Networks and Texture Descriptors," *Scientific Reports* 12, no. 1 (2022): 3092.

38. K. M. Jeon, G. W. Lee, N. K. Kim, and H. K. Kim, "TAU-Net: Temporal Activation u-Net Shared With Nonnegative Matrix Factorization for Speech Enhancement in Unseen Noise Environments," *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021): 3400–3414.

39. Q. Li, X. Liu, Y. He, D. Li, and J. Xue, "Temperature Guided Network for 3D Joint Segmentation of the Pancreas and Tumors," *Neural Networks* 157 (2023): 387–403.

40. J. Ju, J. Li, Z. Chang, et al., "Incorporating Multi-Stage Spatial Visual Cues and Active Localization Offset for Pancreas Segmentation," *Pattern Recognition Letters* 170 (2023): 85–92.

41. W. Zhisheng, L. Qing, and D. Xuehai, "A Novel Coarse-To-Fine Segmentation Method for Pancreatic Cancer," in *Paper Presented at: 2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)* (IEEE, 2020).

42. Y. Liang, D. Schott, Y. Zhang, et al., "Auto-Segmentation of Pancreatic Tumor in Multi-Parametric MRI Using Deep Convolutional Neural Networks," *Radiotherapy and Oncology* 145 (2020): 193–200.

43. C. Li, Y. Mao, S. Liang, J. Li, Y. Wang, and Y. Guo, "Deep Causal Learning for Pancreatic Cancer Segmentation in CT Sequences," *Neural Networks* 175 (2024): 106294.

44. D. Qiu, J. Ju, S. Ren, et al., "A Deep Learning-Based Cascade Algorithm for Pancreatic Tumor Segmentation," *Frontiers in Oncology* 14 (2024): 1328146.

45. G. Zhang, W. Wang, D. Yang, et al., "A Bi-Attention Adversarial Network for Prostate Cancer Segmentation," *IEEE Access* 7 (2019): 131448–131458.

46. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144.

47. Y. Liu, Y. Zhu, W. Wang, B. Zheng, X. Qin, and P. Wang, "Multi-Scale Discriminative Network for Prostate Cancer Lesion Segmentation in Multiparametric MR Images," *Medical Physics* 49, no. 11 (2022): 7001–7015.

48. E. Song, J. Long, G. Ma, et al., "Prostate Lesion Segmentation Based on a 3D End-To-End Convolution Neural Network With Deep Multi-Scale Attention," *Magnetic Resonance Imaging* 99 (2023): 98–109.

49. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database," in *Paper Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009).

50. Q. Wei, X. Li, W. Yu, et al., "Learn to Segment Retinal Lesions and Beyond," in *Paper Presented at: 2020 25th International Conference on Pattern Recognition (ICPR)* (IEEE, 2021).

51. A. Turečková, T. Tureček, Z. Komínková Oplatková, and A. Rodríguez-Sánchez, "Improving CT Image Tumor Segmentation Through Deep Supervision and Attentional Gates," *Frontiers in Robotics and AI* 7 (2020): 106.

52. Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille, "A 3d Coarse-To-Fine Framework for Automatic Pancreas Segmentation," *arXiv Preprint arXiv:171200201* 2 (2017): 2.

53. F. Farheen, M. S. Shamil, N. Ibtehaz, and M. S. Rahman, "Segmentation of Lung Tumor From CT Images Using Deep Supervision," (2021). arXiv preprint arXiv:211109262.

54. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," (2016).

55. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27, no. 3 (1948): 379–423.

56. M. Antonelli, A. Reinke, S. Bakas, et al., "Medical Segmentation Decathlon," *Nature Communications* 13, no. 1 (2022): 4128, https://doi.org/10.1038/s41467-022-30695-9.