Neurocomputing 517 (2023) 279–293

Contents lists available at ScienceDirect

Neurocomputing

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

# TD-Net: Trans-Deformer network for automatic pancreas segmentation

Shunbo Dai [a], Yu Zhu [a,c,*], Xiaoben Jiang [a], Fuli Yu [a], Jiajun Lin [a], Dawei Yang [b,c,*]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[c] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, China

ARTICLE INFO

*Article history:*
Received 11 March 2022
Revised 13 September 2022
Accepted 24 October 2022
Available online 30 October 2022
Communicated by Zidong Wang

*Keywords:*
Pancreas segmentation
Deformable convolution
Vision transformer
Wavelet decomposition
Deep supervision

ABSTRACT

Accurate and efficient pancreas segmentation is the basis for subsequent diagnosis and qualitative treatment of pancreatic cancer. Segmenting the pancreas from abdominal CT images is a challenging task because the morphology of the pancreas varies greatly among different individuals and may be affected by problems such as the unbalanced category and blurred boundaries. This paper proposes a two-stage Trans-Deformer network to solve these problems of pancreas segmentation. To be specific, we first use 2D Unet for coarse segmentation to generate candidate regions of the pancreas. In the fine segmentation stage, we propose to integrate deformable convolution into Vision Transformer (VIT) for solving the deformation problem of the pancreas. For the problem of blurred boundaries caused by low contrast in the pancreas, a multi-input module based on wavelet decomposition is proposed to make our network pay more attention to high-frequency texture information. In addition, we propose using the Scale Inter-active Fusion (SIF) module to merge local features and global features. Our method was evaluated on the public NIH dataset including 82 abdominal contrast-enhanced CT volumes and the public MSD dataset including 281 abdominal contrast-enhanced CT volumes via fourfold cross-validation. We have achieved the average Dice Similarity Coefficient (DSC) values of 89.89 ± 1.82 % on the NIH dataset, and 91.22 ± 1.37 % on the MSD dataset, outperforming other exiting state-of-the-art pancreas segmentation methods.



## 1. Introduction

The pancreas is an important digestive organ of the human body, responsible for regulating the physiological functions of the whole body, and is susceptible to various diseases. Pancreatic cancer is a malignant tumor that has a high mortality rate [1] and a low survival rate after treatment. In the United States alone in 2021, about 48,000 people die of pancreatic cancer, and about 60,000 new patients are diagnosed with pancreatic cancer [2]. Regardless of the stage of disease, the five-year survival rate of patients is only about 10 % [2]. Fortunately, early diagnosis and timely treatment can delay the development of pancreatic cancer and even eliminate it [3]. Accurately segmenting the pancreas from CT images is helpful for timely monitoring of abnormal volume changes and abnormal growth of the pancreas, providing the possibility for the prevention, diagnosis and surgical treatment of pancreatic cancer. Because it is time-consuming and labor-intensive to manually outline the boundary of the pancreas layer by layer, automatically identifying and segmenting the pancreas using radiological images has become a research hotspot. Furthermore, automatic segmentation of the pancreas is an important prerequisite for medical image analysis and surgical diagnosis plans.

In recent years, with the rapid development of deep learning research and neural networks, the automatic segmentation of many organs and tissues has achieved good results, such as the heart [4], liver [5], spleen [6], lung [7], left and right kidneys [8,9] and so on. However, compared with other organs, the accuracy of pancreas segmentation is still relatively low. Automatic segmentation of the pancreas remains a challenging task due to the very limited volume of the pancreas in abdominal CT scans. The main difficulties come from the following aspects: 1) the pancreas only occupies a small part of the entire CT image, as shown in Fig. 1. There is a serious imbalance between the target and the background, which makes it easy for the network to pay attention to the non-target background area, resulting in misclassification; 2) the pancreas is irregular in shape and easily deformed, and the shape, size, and position of the pancreas in the abdomen of different patients are different greatly; 3) in the CT image, the contrast between the pancreas and its surrounding tissues is weak,

* Corresponding authors at: School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.
*E-mail addresses:* zhuyu@ecust.edu.cn (Y. Zhu), yang_dw@hotmail.com (D. Yang).

https://doi.org/10.1016/j.neucom.2022.10.060
0925-2312/

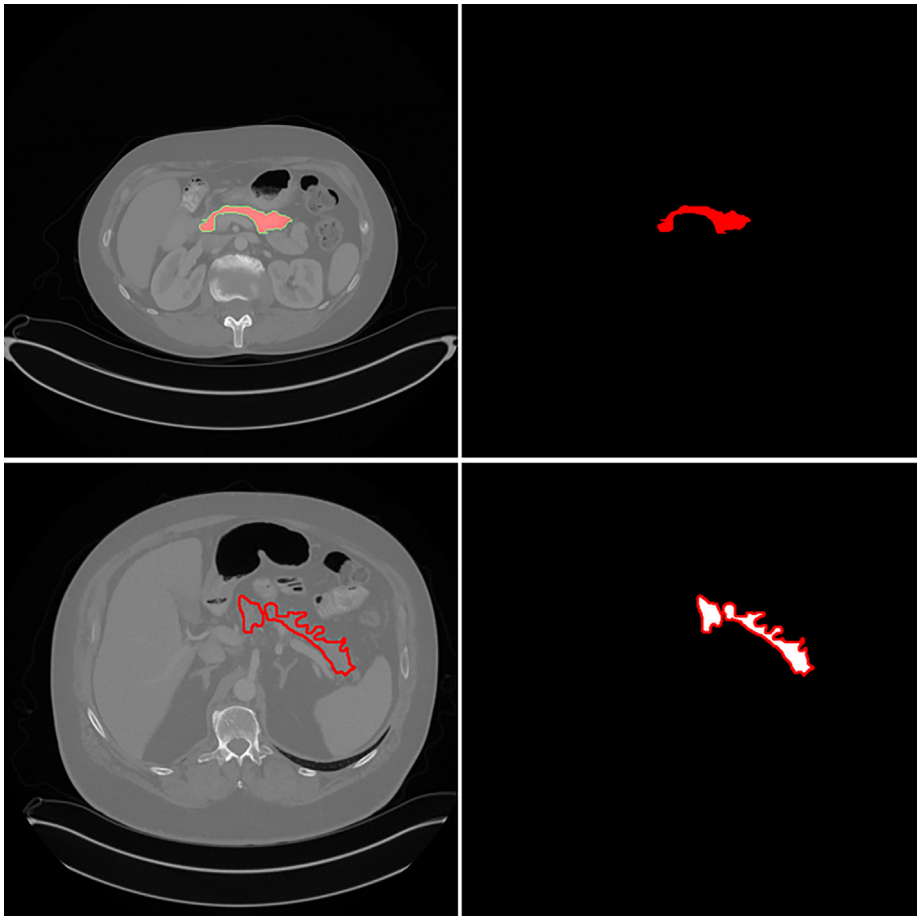

**Fig. 1.** Typical slice display of the CT scan in the abdomen with the pancreas area highlighted.

which is caused by the similar range of voxel intensities, so it is difficult to distinguish the pancreas from these tissues. The borders of the pancreas are blurred and even difficult to be seen.

These characteristics greatly increase the difficulty of segmentation and may lead to the problem of blurred boundaries in the pancreas segmentation task of purely traditional image segmentation algorithms, such as clustering [10], region growing [11] and wavelet decomposition [12]. The characteristics also cause the phenomenon that traditional convolutional segmentation networks will pay more attention to the partial area of the pancreas, while ignoring rich global context information, which limits the further improvement of the segmentation accuracy to a certain extent. Recently, the global attention mechanism based on Transformer [13] has been able to effectively solve the above problems and fully integrate local features and global features, which greatly improves the performance of the network and the accuracy of segmentation. The original Transformer was used in the field of natural language processing (NLP). In recent years, with the development of deep learning research, Vision Transformer (VIT) [14] successfully realized the application of Transformer in the field of computer vision and achieved remarkable achievements. A series of VIT-based segmentation models quickly occupied the field of medical segmentation [15–17], which further promoted the development of the field of medical image segmentation. Nevertheless, these existing VIT-based networks have not addressed the problems in pancreas segmentation well, mainly due to the small size and the deformation of the pancreas.

Recently, Dai et al. [18] proposed a deformable convolution to solve the problem that the size of the receptive field in the standard convolution cannot perfectly adapt to the geometric deformation of the target. Specifically, the convolution kernel of the deformable convolution is variable, so the corresponding receptive field can change adaptively according to the change of the target shape, which perfectly matches the situation where the pancreas is deformed in size and shape in different patients. Due to the fixed receptive field, it is difficult to further improve the performance of standard convolution on the task of pancreas segmentation. Meanwhile, deformable convolution has been applied to the pancreas segmentation task [19,20]. Huang et al. [19] proposed DUNet for pancreas segmentation by combining deformable convolution and Unet. DUNet can flexibly capture pancreatic features and improve the geometric modeling ability of UNet, finally achieving the Dice coefficient of 87.25 ± 3.27 % on the NIH dataset. Wang et al. [20] proposed a dual-input v-mesh fully convolutional network (FCN) to segment the pancreas. The contrast of the pancreas was increased by complementing the image processed by a contrast-specific graph-based visual saliency (GBVS) algorithm. By fusing the spatial transformation and fusion (SF) model with multi-branch residual deformable convolutional layers, a Dice coefficient of 87.40 ± 6.80 % was finally achieved on the NIH dataset. However, deformable convolution inevitably can only focus on local features and cannot combine with global features.

Considering all of these, we propose to integrate deformable convolution into the VIT architecture, which solves the problem of deformation of the pancreas, and perfectly integrates local features and global features. Our Trans-Deformer network can adaptively focus on the corresponding area according to the shape of the target. Meanwhile, we propose a multi-input module based

on two-dimensional wavelet decomposition to deal with the problem of blurry boundaries caused by low contrast in the pancreas. The strategy enables our network to pay more attention to the marginal high-frequency information of the pancreas. In response to the problem of categories imbalance between the background and target caused by the small size of the pancreas, the binary cross-entropy and the dice loss are used to alleviate this problem. In addition, deep supervision is added to enhance the robustness of the network. These designs improve the performance of the network on the task of pancreas segmentation.

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

1) We put forward the Trans-deformer module. By subtly fusing deformable convolution into VIT, it solves the deformation of the pancreas in the pancreas segmentation task. And the proposed Trans-deformer module can be quickly transferred to other segmentation tasks.
2) The Scale Inter-active Fusion (SIF) module is designed to integrate local features and global features through attention interaction.
3) We come up with a multi-input module based on two-dimensional wavelet decomposition, which makes our network pay more attention to the high-frequency texture information of the target edge. It solves the problem of blurred boundaries caused by low contrast in the pancreas.
4) We propose the Trans-Deformer network that achieves accurate segmentation of the pancreas and outperforms state-of-the-art methods on publicly available pancreas datasets, achieving the average DSC scores of 89.89 ± 1.82 % on the NIH dataset, and 91.22 ± 1.37 % on the MSD dataset.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 elaborates on the primary principles and the network architecture of our Trans-Deformer network. Experiments with detailed implementation, evaluation criteria and results are explained in Section 4. Section 5 is the discussion and the conclusion is drawn in Section 6.

## 2. Related work

We have introduced the overall research background and current situation in Section 1, and clarified the significance and difficulties of pancreas segmentation research. In the following subsections, we will review the research process of pancreas segmentation in more detail, and introduce the evolution and the latest development direction of Vision Transformer (VIT).

### 2.1. Pancreas segmentation

With the rapid development and progress of deep learning, more and more methods have begun to deal with the challenging task of segmenting the pancreas from abdominal CT scans. In the existing pancreas segmentation methods, if they are divided according to the type of input, they can be roughly divided into two categories: input according to 2D slice and input according to 3D block, thus giving birth to 2D network and 3D network. 2D segmentation networks, such as fully convolutional neural networks (FCN) [21] and U-Net [22], have laid an important foundation for the field of medical image segmentation [23]. Li et al. [24] proposed an automatic pancreas segmentation model using double adversarial networks with a pyramidal pooling module, achieving the Dice coefficient of 83.31 ± 6.32 %. Li et al. [25] proposed the multiscale attention dense residual U-shaped network (MAD-UNet) to solve the problems of intraclass inconsistency and interclass indistinction in the segmentation of the pancreas, achieving the Dice coefficient of 86.10 ± 3.52 % on the NIH dataset. Li et al. [26] proposed three strategies: skip network, residual network and multi-scale cross-domain information fusion to solve the problems in pancreas segmentation, and finally achieved an 87.57 ± 3.26 % Dice coefficient.

Because the 2D network ignores the continuity of the pancreas in the three-dimensional space to a certain extent, it limits the further improvement of the network segmentation performance. 3D-based segmentation networks, such as V-Net [27] and 3D U-Net [28], can directly extract features from three-dimensional spatial information, thereby avoiding the bottleneck of 2D segmentation networks. Zhu et al. [29] proposed a 3D-based coarse-to-fine cascaded pancreas segmentation network and finally achieved the Dice coefficient of 84.59 ± 4.86 % on the NIH dataset. Zhang et al. [30] proposed a segmentation framework that combines multi-atlas registration and the 3D level set. By using the 3D level set to refine predicted probability maps, it effectively compensated for the defect of incomplete pancreatic edge prediction and achieved an 84.47 ± 4.36 % Dice coefficient on the NIH dataset.

However, due to the high requirements of the 3D network on GPU memory, the CT scan image input to the 3D network is usually cut into small pieces or down-sampled to a smaller size, which limits its spatial context learning to a certain extent.

Based on this, the emergence of the 2.5D network has attracted widespread attention. The 2.5D network makes up for the lack of spatial context information in the 2D network, and it reduces the computational cost compared to the 3D network. The final segmentation result is obtained by feeding three axial 2D slices of the CT image or one axial adjacent slice into the network, and finally performing fusion. Since the 2.5D network uses 2D convolution to implicitly extract spatial features, it will be affected by the topology to a certain extent. The features from input slices of different channels are finally mixed and output within one channel dimension. The lack of corresponding label matches will cause features to be confused when fused, making it difficult to distinguish them from each other [31]. Zhou et al. [32] proposed a fixed-point FCN model for segmenting the pancreas from abdominal CT images. The results of segmentation in three directions were voted and fused by the majority, finally obtaining an 82.37 % Dice coefficient. Li et al. [33] proposed a model-driven stack-based fully convolutional network with a sliding window fusion algorithm to capture local spatial context features between slices for pancreas segmentation, achieving the Dice coefficient of 85.7 ± 4.1 %. Li et al. [31] designed a 2.5D network for generating light-weight 3D voxels by stacking three adjacent slices into three input channels to balance the use of contextual information for pancreas segmentation, achieving the Dice coefficient of 86.49 ± 1.44 % on the NIH dataset.

Of course, pancreas segmentation methods can also be divided according to the number of experimental stages. They can be divided into a one-stage network that is directly segmented and a two-stage network that is positioned first and then segmented [34–37]. Zhang et al. [34] proposed a lightweight deep convolutional neural network that utilized the Scale Transferable Feature Fusion Module (STFFM) and the Prior Propagation Module (PPM) for coarse-to-fine pancreatic segmentation, achieving the Dice coefficient of 84. 9 % on the NIH dataset. Chen et al. [35] proposed a dual-view feature learning network based on attention mechanism and multi-scale supervision, and finally achieved the Dice coefficient of 85.19 ± 4.73 % on the NIH dataset through a two-stage TVMS-Net with first localization and then segmentation. Li et al. [36] proposed a probabilistic map guided bidirectional recursive UNet (PBR-UNET), which performed coarse segmentation through probability maps, and finally achieved the Dice coefficient of 85.35 ± 4.13 % through bidirectional recursion on the NIH data-

set. Hu et al. [37] proposed a saliency perception model based on geodesic distance. By transforming the probability map into a saliency map through the saliency transformation, and introducing a saliency perception module that combined the saliency map with the image context information in the fine segmentation stage, they obtained the Dice coefficient of 85.49 ± 4.77 % on the NIH dataset. Statistics have found that in terms of pancreas segmentation, the two-stage network is generally performing better [35] than the direct segmentation network. The phenomenon is mainly produced by the unbalanced category caused by the small size, and the blurred boundaries caused by the low contrast in the pancreas.

Although the above methods have made significant achievements in the field of pancreas segmentation, the traditional convolutional neural networks based on standard convolution have fixed the receptive field, which ignores the rich global context information to a certain extent, making it difficult to further optimize the network performance.

### 2.2. Technology evolution based on VIT

The Transformer architecture was first proposed by Vaswani et al. [13] to improve the performance of machine translation. Due to the perfect combination of local information and global information, it has been widely used in the field of natural language processing (NLP). In recent years, a large number of experiments have been carried out to transfer the advantages of Transformer to the field of computer vision. Detection Transformer (DETR) [38] used the Transformer network to realize the end-to-end target detection task for the first time. Vision Transformer (VIT) [14] laid the foundation for the application of Transformer in the field of computer vision by cutting the image into individual patches and transforming them into sequences, then cleverly integrating them into the Transformer architecture. Segmentation Transformer (SETR) [39] used VIT as the encoder of the network and CNN as the decoder of the network to complete the prediction of the semantic map.

In the field of medical image segmentation, the rapid development of Transformer-based network architecture has further promoted the advancement of the field of medical image segmentation. UCTransNet [40] integrated the Transformer self-attention mechanism into the channel dimension for the first time, which made up for the gap in semantics and resolution between low-level and high-level features through effective feature fusion and multi-scale channel cross attention. Unlike VIT, which usually has a low-resolution output and high computational and storage costs, PVT [41] could be trained on dense partitions of the image to obtain a high-resolution output. And it could gradually downscale the pyramid feature map to reduce large feature map calculation. Valanarasu et al. [42] proposed a self-attention mechanism network system called MedT based on Transformer. By using local–global training strategies, it was superior to traditional neural convolutional networks in segmentation tasks. Chen et al. [43] proposed TransUNet, which integrated Unet and VIT, to improve the performance of the synapse multi-organ CT dataset in multi-organ segmentation tasks. Ji et al. [44] proposed the MCTrans network combining rich feature information and semantic structure through the self-attention mechanism and cross-attention mechanism, which achieved better performance than Attention Unet [45]. Hatamizadeh et al. [6] proposed a UNETR network using VIT as an encoder to improve the performance of seg-

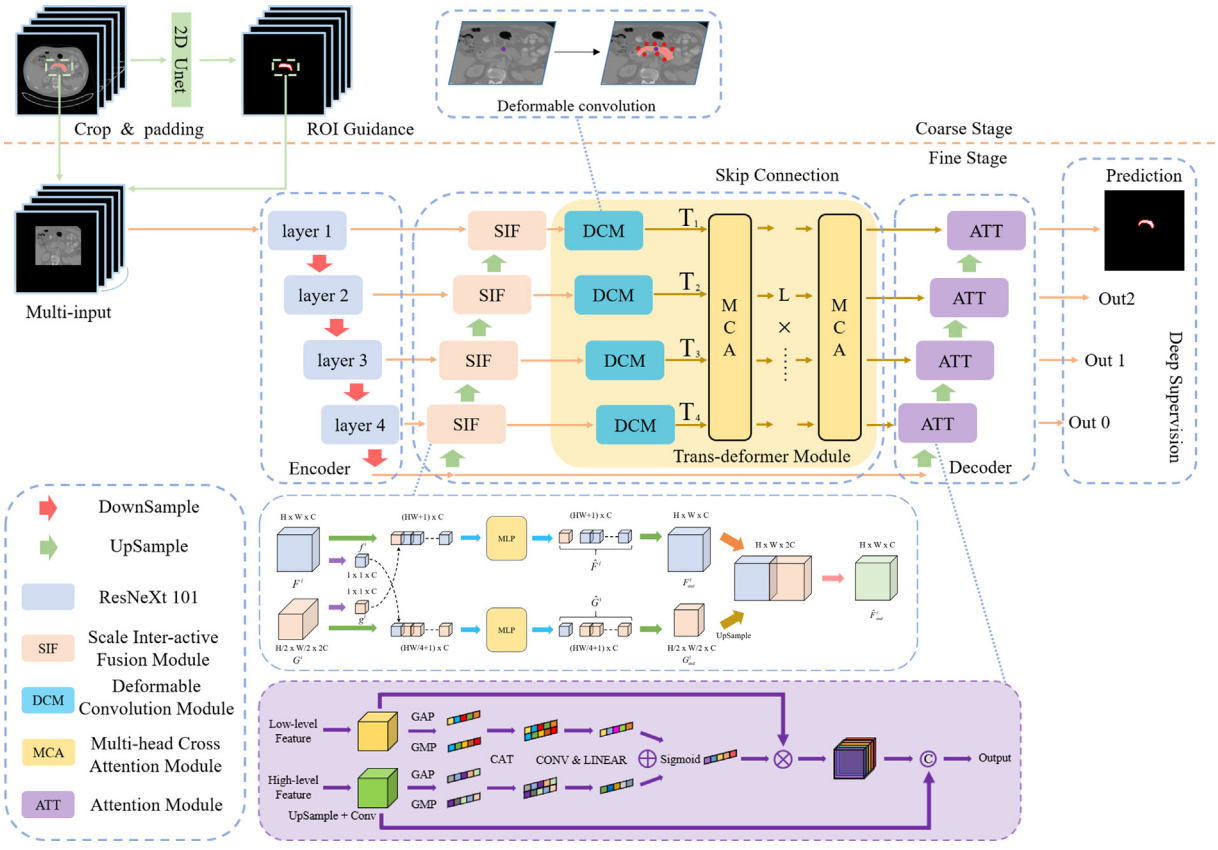


**Fig. 2.** The overview of the proposed Trans-Deformer framework.

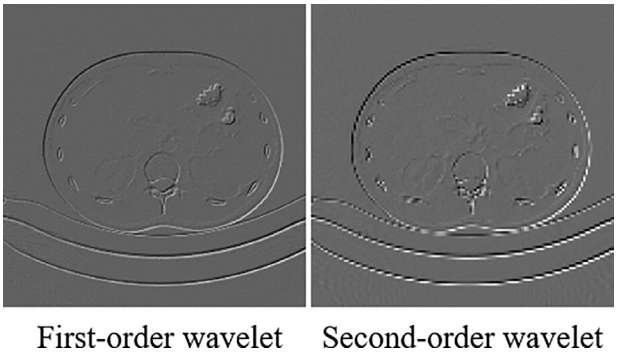


**Fig. 3.** The first-order (left) and second-order (right) wavelet decomposition effects of an abdominal CT scan slice.

menting large-volume targets such as brain tumors and spleen on the MSD dataset.

These VIT-based networks have achieved good results on most medical segmentation tasks, but they have not achieved good index improvement in the segmentation of small organs such as the pancreas, which are easily deformable and small in size, and also have fuzzy edges.

## 3. Method

In this section, we describe the proposed Trans-Deformer network in more detail. We will first introduce the two-stage network implementation process, and then present three innovative modules we put forward. Section 3.1 clarifies the multi-input module based on two-dimensional wavelet decomposition. Section 3.2 elaborates on the SIF module that integrates local features and global features and Section 3.3 explains the Trans-deformer module that fuses deformable convolution into VIT.

As shown in Fig. 2, the pancreas segmentation is implemented in a coarse-to-fine framework. First, we train a 2D Unet [22] network for coarse segmentation. Of course, it can be replaced with any other 2D segmentation network. Then, based on the pancreatic probability map obtained by rough segmentation, we crop the original image and fill it to the original size to obtain the focused pancreas area, and send it to the Trans-Deformer network for fine segmentation. Inspired by the Unet network, the proposed Trans-Deformer network shaped like the letter W consists of an encoder based on ResNeXt 101 [46], a decoder based on the channel cross attention module [47], and a skip connection. The skip connection includes a SIF module that integrates local features and global features, and a Trans-deformer module based on multi-head cross attention (MCA) and deformable convolution.

### 3.1. Multi-input module based on two-dimensional wavelet decomposition

As we mentioned above, the low contrast and the blurred boundaries seriously affect the performance of the existing network on the task of segmenting small organs such as the pancreas. Based on this, we propose to introduce two-dimensional wavelet decomposition into the network. To be specific, we perform discrete wavelet transform [48] on the original image in the preprocessing stage to generate high-frequency features containing the edge information as illustrated in Fig. 3. The operation helps the network learn the edge of the pancreas by providing more texture information to alleviate the problem of blurred boundaries.

Through two-dimensional wavelet decomposition, the original image is split into low-frequency information and high-frequency information. The low-frequency information contains the essential characteristics of the image, mainly the areas where the bright or gray value changes slowly in the image. The high-frequency information including diagonal, vertical and horizontal directions, mainly highlights the edge texture of the image, which is a supplement to the image details. Notably, the high-frequency information is exactly what is needed in the task of pancreas segmentation.

After successively applying the first-order and second-order discrete wavelet transform to the original image, we take out the high-frequency components and resize them to the original image size, as a supplement to the edge information of the pancreas. Then, to supplement inter-slice context information, two wavelet components and the original image with the adjacent CT images before and after the image form five inputs, which are cropped based on the result of the coarse segmentation and padded with zeros to the size of the original image. Finally, generated inputs as presented in Fig. 4 are sent to the network. The above operations force the network to pay more attention to the edge of the

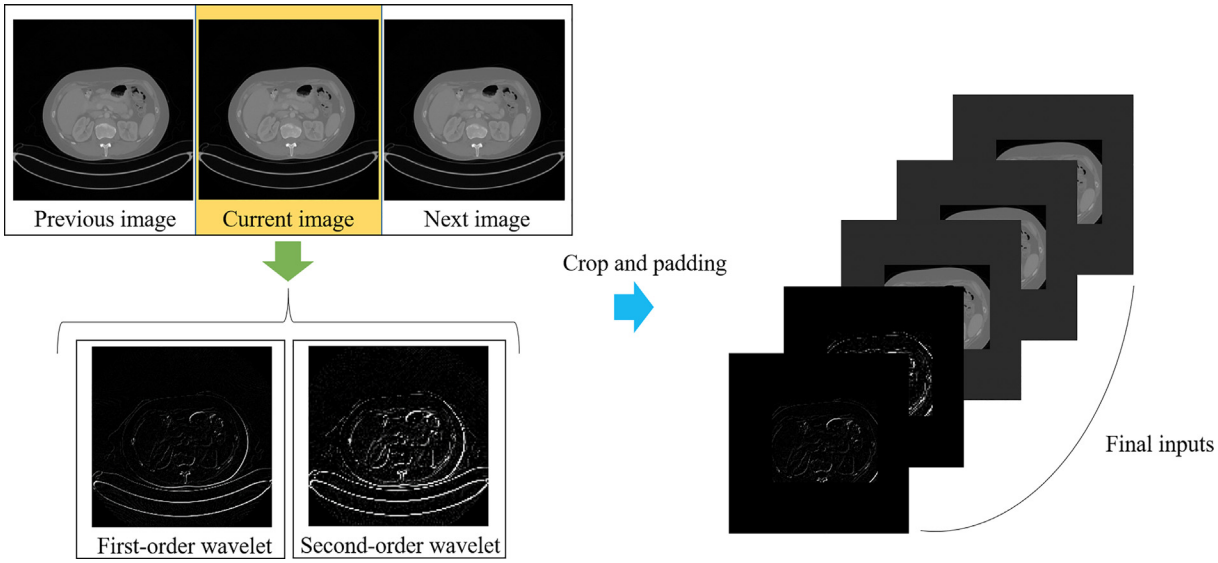


**Fig. 4.** The five inputs of the network.

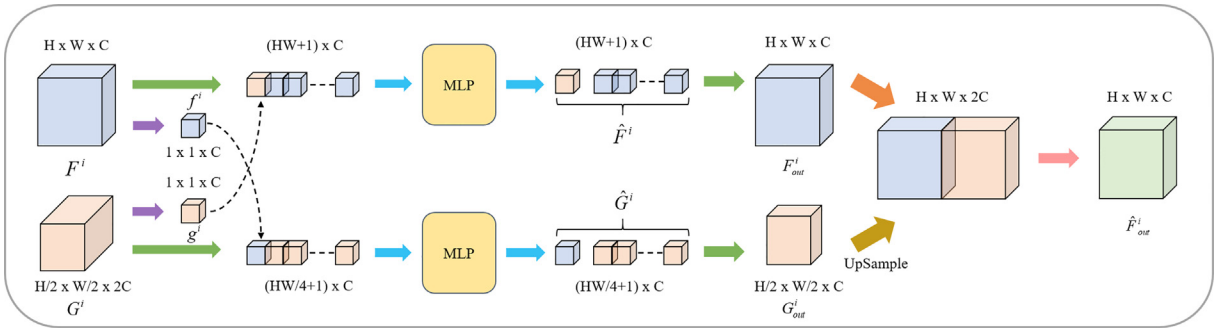


**Fig. 5.** The illustration of the SIF module.

pancreas by providing more accurate edge information and richer texture information.

### 3.2. Scale Inter-active fusion module

After encoding by ResNeXt 101, we get feature maps of different scales, which respectively represent the information at different scales of layers corresponding to the original image. According to the architecture of the traditional Unet network, in addition to the feature map at the bottom layer, the feature maps of other layers will be directly concatenated as a skip connection to the upsampled high-level feature maps. Nevertheless, this straightforward concatenation is difficult to capture the inner relations among feature maps at different scales. And the long-range dependence between global context information and local structural features is ignored easily. Inspired by the self-attention mechanism in Transformer, we propose a novel Scale Inter-active Fusion (SIF) module, which can combine local information with global information, and improve the performance of the network by merging feature maps of different scales.

As shown in Fig. 5, the proposed SIF module can integrate the features between the low-level and high-level. In the following, we select the low-level large-scale feature for detailed analysis, and the other branch will perform the same operation.

To be specific, we define the low-level large-scale feature as $F^i \in \mathbb{R}^{C\times H\times W}$ (primary branch), where $i(i=1,2,3,4)$ represents the layer where the SIF module is located, and define the high-level small-scale feature as $G^i \in \mathbb{R}^{2C\times(H/2)\times(W/2)}$ (complementary branch). The SIF module is always at the same layer as the large-scale feature. Then we reshape $F^i \in \mathbb{R}^{C\times H\times W}$ into $F^i = \left[f_1^i, f_2^i, \cdots, f_{H\times W}^i\right] \in \mathbb{R}^{C\times(H\times W)}$ and compress $G^i$ as the following:

$$g^i = Flatten\left(AvgPool\left(G^i\right)\right) \in \mathbb{R}^{C\times 1}$$

where AvgPool is a two-dimensional average pooling operation, followed by flattening. At this time, $g^i$ represents the global feature that contains all the information of $G^i$, and the channel has been halved for subsequent feature fusion with local information $F^i$. Further, $F^i$ is concatenated with $g^i$ into a sequence of length $H\times W+1$, which will be sent to the multilayer perceptron for attention fusion:

$$\begin{aligned}\hat{F}^i &= MLP\left(\left[g^i, f_1^i, f_2^i, \cdots, f_{H\times W}^i\right]\right)\\ &= \left[\hat{f}_0^i, \hat{f}_1^i, \cdots, \hat{f}_{H\times W}^i\right] \in \mathbb{R}^{C\times(H\times W+1)}\end{aligned}$$

After nonlinear mapping, the fused feature $\hat{F}^i$ is obtained. It not only contains the local features of the low-level, but also integrates the high-level global features of different scales. We get the mixed feature $F_{out}^i = \left[\hat{f}_1^i, \cdots, \hat{f}_{H\times W}^i\right] \in \mathbb{R}^{C\times(H\times W)}$ after removing the head component. Then we reshape it to the original size as the final result $F_{out}^i \in \mathbb{R}^{C\times H\times W}$ of the primary branch. In the same way, the same transformation is performed on the small-scale feature to get $G_{out}^i \in \mathbb{R}^{C\times(H/2)\times(W/2)}$. Before concatenating the two different scale features, we implement an up-sampling operation on $G_{out}^i$ to make the scale match $F_{out}^i$. Finally, through the convolution operation, the feature map that combines the low-level local information and the high-level global information is converted to $\hat{F}_{out}^i \in \mathbb{R}^{C\times H\times W}$, to replace the original feature $F^i$ of the low-level. The SIF module can capture the relations among feature maps of different scales through multilayer perceptron. Meanwhile, the local information of the low-level and the global information of the high-level is interacted through nonlinear mapping to make the network robust.

### 3.3. Trans-deformer module based on deformable convolution

After obtaining the four fused feature maps of different scales through the SIF module, we perform tokenization by reshaping the feature maps into sequences of flattened 2D patches to generate tokens. Different from the way that the original VIT generates tokens based on standard convolution, we use deformable convolution to generate tokens. The main difference between deformable convolution and standard convolution is the addition of a deformable offset field, which contains the learnable offset for each position in the feature map. The added deformable offset field enhances the network's ability to extract features, enabling the network to adaptively match the shape of the pancreas. By fusing deformable convolution into VIT, the network can flexibly capture the morphological differences of the pancreas to achieve high-precision segmentation, finally achieving the effect of solving the deformation of the pancreas.

The overall structure of the proposed Trans-deformer module is presented in Fig. 6. After using deformable convolution to generate tokens of four different scales, we will send them to $L$-layer multi-head cross attention module. The resulting tokens from four layers will generate four different $Q$, while $K$ and $V$ will be generated by the four concatenated tokens $T_{cat}$. Specifically, the resulting tokens $T_i(i=1,2,3,4)$ will generate $Q_i(i=1,2,3,4)$, $K$, $V$ as follows:

$$Q_i = T_i W_Q,\ K = T_{cat} W_K,\ V = T_{cat} W_V$$

where $W_Q \in \mathbb{R}^{C_i\times d}$, $W_k \in \mathbb{R}^{C_{cat}\times d}$, $W_v \in \mathbb{R}^{C_{cat}\times d}$ are three weight matrices, $C_i$ is the channel dimension of layer $i(i=1,2,3,4)$, $C_{cat}$ is the concatenation of the four-layer channel dimension, and $d$ is the number of patches. In our experiment, $C_1$ to $C_4$ are set to 64, 128, 256 and 512 respectively. With $Q_i \in \mathbb{R}^{C_i\times d}$, $K \in \mathbb{R}^{C_{cat}\times d}$, $V \in \mathbb{R}^{C_{cat}\times d}$,

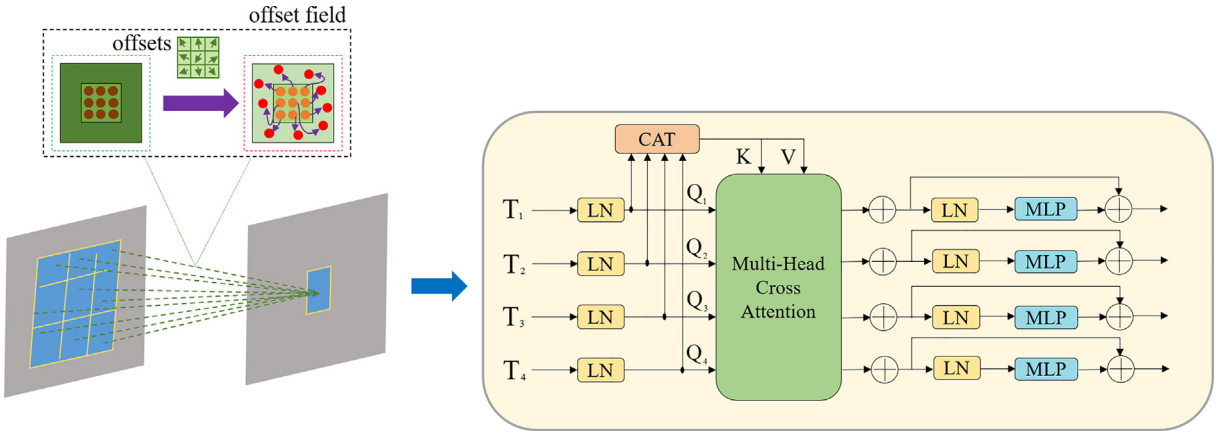


**Fig. 6.** The structure of the proposed Trans-deformer module. The offset field at the top left shows the evolution from standard convolution to deformable convolution. Among them, the sampling points of standard convolution and deformable convolution are respectively marked in orange and red. The purple curve from the orange point to the red point represents the learnable offset of the deformable convolution. The offset shift is shown above the purple arrow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the cross attention (CA) mechanism is calculated by the following formula:

$$\mathrm{CA}_i(Q,K,V) = \mathrm{softmax}\left(\frac{Q_i K^T}{\sqrt{C_{cat}}}\right)V$$

In the experiment, the number of heads of the multi-head cross attention (MCA) module is set to 4, that is, the $MCA_i$ is obtained by averaging four $CA_i$. The size of the patch is set to 16. Then, applying a multilayer perceptron (MLP) and residual operator, the output of the Trans-deformer module is obtained as follows:

$$O_i = MCA_i + MLP(Q_i + MCA_i)$$

we omit layer normalization (LN) in the above formula for simplicity. The process in Equation (5) is repeated $L$ times to build an $L$-layer multi-head cross attention module, where $L$ is set to 4 in our implementation. Different from the self-attention of the original VIT, our attention fusion is carried out in the channel dimension, which can effectively interact with the information among different layers.

Finally, we upsample the feature maps of four different scales obtained by the Trans-deformer module and the bottom-level feature map generated by the encoder, to complete the final fusion through the attention module layer by layer. The internal structure of the attention module is shown in the purple area at the bottom of Fig. 2. To be specific, we first upsample the feature map of the high-level and halve the channel through a convolution operation to match the low-level feature map from the skip connection. Then we perform global average pooling and global maximum pooling on the two branches respectively to capture the comprehensive correlation of features among channels, after that concatenate the results of each branch and pass them through the multilayer perceptron. Before the final fusion, the resulting feature maps of the two branches will be summed to perform sigmoid activation for suppressing irrelevant features. At last, the generated channel cross attention acts directly on the original skip connection of the low-level as a coefficient, then concatenates with the upsampling feature map of the high-level to complete the feature fusion of the low-level.

Meanwhile, to speed up the network convergence and increase the robustness of the network, we supplement the deep supervision module after the decoder. Specifically, the fusion feature map obtained at the end of each layer is directly upsampled to the original image size, which is denoted as Out0, Out1, and Out2 respectively, and output together with the final prediction. The three additional outputs will participate in the final loss calculation along with the prediction, further optimizing the performance of the network by supervising high-level features.

## 4. Experimental results

### 4.1. Datasets

We evaluate the performance of the proposed network on two publicly available datasets: 1) 82 abdominal contrast-enhanced CT scans from the National Institutes of Health (NIH) Clinical Center pancreas segmentation dataset [49], which is the most widely used publicly dataset for the pancreas segmentation task. Each CT volume is 512 × 512 × D, where D ∈ [181,466] is the number of slices along the transverse plane. The slice thickness varies from 1.5 mm to 2.5 mm along with the depth of the CT scan. We follow the principle of fourfold cross-validation and randomly divide the dataset into four fixed subsets 21, 21, 20 and 20. In one cross-validation, we train on three of the subsets and test on the remaining subset, repeating four times and averaging; 2) 281 abdominal contrast-enhanced CT scans with labeled pancreas and pancreatic tumor from the Medical Segmentation Decathlon (MSD) challenge pancreas segmentation dataset [50], where each CT volume is 512 × 512 × D, and D ∈ [37,751] is the number of slices in CT scan. Following previous studies [36] on the MSD dataset, we combine the pancreas and pancreatic tumor into a single entity as the segmentation target. We divide it into four subsets containing 70, 70, 70 and 71 CT volumes respectively. The rest of the operations are consistent with the NIH dataset. The 2D visualization of two datasets is shown in Fig. 7.

### 4.2. Evaluation metrics

To evaluate the proposed network segmentation performance, we use five different evaluation metrics, namely, the Dice similarity coefficient (DSC), precision, recall, average symmetric surface distance (ASD) and 95 % Hausdorff distance (HD). We define $S(X)$ and $S(Y)$ represent the edge point set of the prediction result and the ground true respectively. And $d\{x,y\}$ represents the Euclidean distance between voxel $x$ and voxel $y$. The details of the evaluation metrics are as follows:

1) DSC: the Dice similarity coefficient is the most common index for evaluating segmentation results in the field of medical

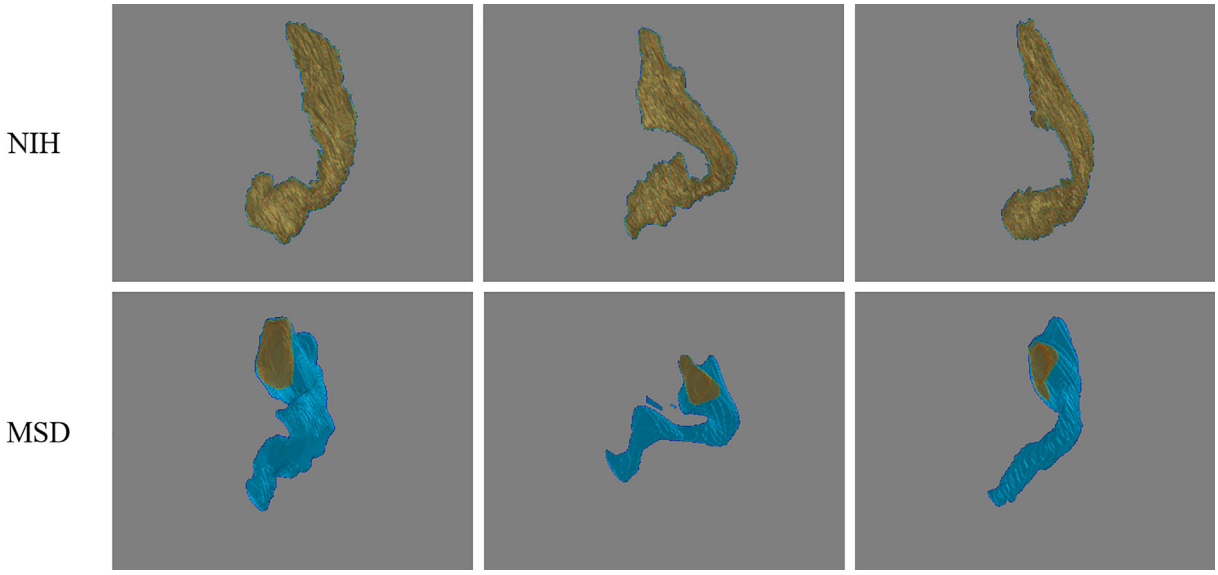


**Fig. 7.** 2D visualization of two datasets. Among the MSD dataset, the blue area represents the pancreas, and the grass green area represents the pancreatic tumor. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

image segmentation, which mainly measures the similarity between the prediction and the ground truth.

2) Precision: the precision is the percentage of positives correctly predicted among all positives predicted in the prediction.

3) Recall: the recall is the percentage of positives correctly predicted among all positives in the ground truth.

4) ASD: the average symmetric surface distance is the average distance between the prediction boundary and the ground truth boundary, reflecting the accuracy of edge segmentation.

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

5) HD: the 95 % Hausdorff distance can evaluate the degree of pancreatic boundary segmentation. The smaller the value of HD, the more complete the pancreatic boundary segmentation.

$$HD = \max \left\{ \max_{y \in S(Y)} \min_{x \in S(X)} d\{y, x\}, \max_{x \in S(X)} \min_{y \in S(Y)} d\{x, y\} \right\}$$

### 4.3. Implementation details

We implemented our framework based on the PyTorch platform on the Ubuntu system equipped with an NVIDIA GeForce RTX 3090 graphics card of 24 GB memory. Due to memory limitations, the images that entered the network were resized to 128 × 128. For the encoder, we loaded the weights of pre-trained ResNeXt 101 on ImageNet [51].

In our experiments, we used the sum of binary cross-entropy and dice loss as the final loss function. The expression of binary cross-entropy usually used in binary classification is shown as follows:

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

and the dice loss, which is effective for categories imbalance [52], can be defined as:

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

so the loss function consists of the following:

$$L_{loss} = L_{bce} + L_{dice}$$

where $\hat{y}_i$ represents the predicted value of the network, $y_i$ is the value of corresponding ground truth, and $N$ is the number of pixels. We adopted the strategy of deep supervision to enhance the robustness of the network. Considering that shallow low-level features

**Table 1**
The results (measured by the DSC, Precision, Recall and Testing time) of pancreas segmentation on the NIH dataset. "-" denotes that the corresponding results are not provided in the literature. Optimal results (described by mean ± std) are shown in bold.

| Method | DSC(%) | Precision(%) | Recall(%) | Testing time |
|---|---|---|---|---|
| M. Li et al. [24] | 83.31 ± 6.32 | 84.09 ± 8.65 | 83.30 ± 8.54 | – |
| Y. Zhang et al. [30] | 84.47 ± 4.36 | – | – | 3–5 min |
| D. Zhang et al. [34] | 84.90 | – | – | 4–5 min |
| H. Chen et al. [35] | 85.19 ± 4.73 | 86.09 ± 5.93 | 84.58 ± 8.09 | 3–4 min |
| J. Li et al. [36] | 85.35 ± 4.13 | 83.45 ± 7.19 | 82.76 ± 8.21 | – |
| P. Hu et al. [37] | 85.49 ± 4.77 | – | – | – |
| H. Li et al. [33] | 85.70 ± 4.10 | 87.40 ± 5.20 | 84.80 ± 7.50 | – |
| W. Li et al. [25] | 86.10 ± 3.52 | – | 86.43 ± 5.30 | – |
| J. Li et al. [31] | 86.49 ± 1.44 | – | – | 14–15 min |
| M. Huang et al. [19] | 87.25 ± 3.27 | 88.98 | 89.97 | 10–11 min |
| Y. Wang et al. [20] | 87.40 ± 6.80 | – | 87.70 ± 7.90 | – |
| F. Li et al. [26] | 87.57 ± 3.26 | 86.63 ± 3.70 | 89.55 ± 4.03 | 10–11 min |
| Ours | **89.89 ± 1.82** | **89.59 ± 1.75** | **91.13 ± 1.48** | **3–4 min** |

**Table 2**
The results (measured by the ASD and HD) of pancreas segmentation on the NIH dataset. "-" denotes that the corresponding results are not provided in the literature. Optimal results (described by mean ± std) are shown in bold.

| Method | ASD(mm) | HD(mm) |
|---|---|---|
| Y. Wang et al. [20] | 2.89 ± 4.78 | 18.41 ± 28.19 |
| W. Li et al. [25] | 1.27 ± 0.43 | 4.40 ± 2.99 |
| J. Li et al. [36] | 1.10 ± 0.40 | 3.68 ± 2.30 |
| Ours | **0.78 ± 0.08** | **2.09 ± 0.07** |

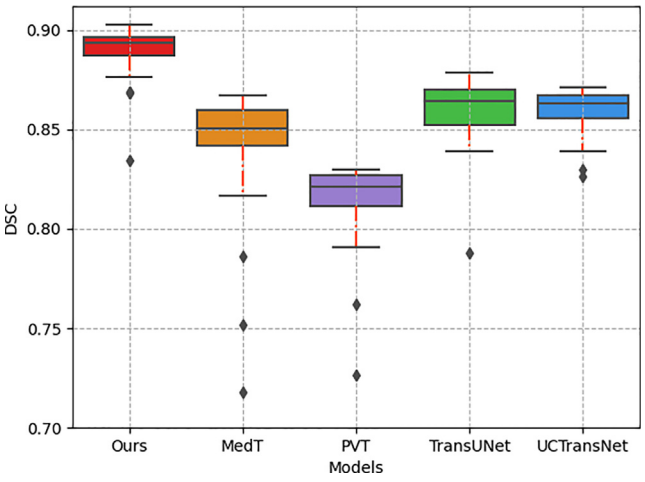


**Fig. 8.** The DSC comparison of the proposed network with mainstream VIT-based medical image segmentation networks on the NIH dataset.

have a greater impact on network performance [53–55], we set the dice loss scale coefficients of prediction, Out2, Out1 and Out0 as 1, 0.6, 0.3 and 0.1 respectively in training.

For data preprocessing, we empirically truncated CT intensity values to the range [−100,240] HU and normalized them to the range [0,1]. Data augmentation operations included random rotations with angles that were integer multiples of 90 degrees, random scaling with scaling factors between [0.8,1.2], and random

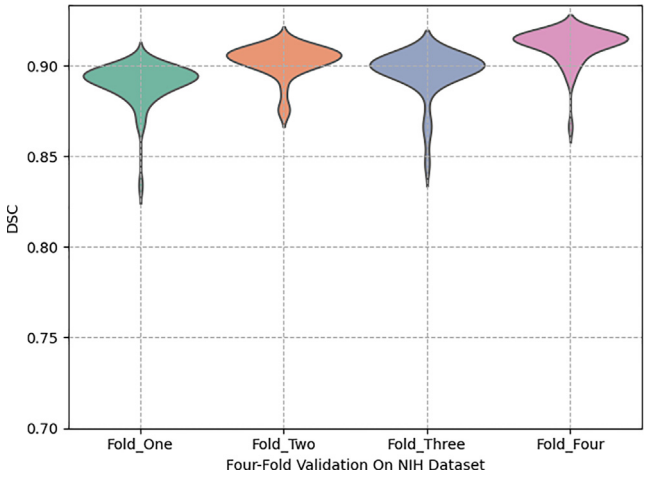


**Fig. 10.** The DSC distribution of the proposed network on the NIH dataset. The four colors represent different folds.

flipping to prevent overfitting of the model. In the final output prediction stage, we binarized the prediction with a bound of 0.5.

For model training, we chose stochastic gradient descent as the optimizer of our network, the initial learning rate was set to 1 × 1e-4, and the momentum was set to 0.9. We set the epoch number to 30 and the batch size to 8. The training time for an epoch was about 8 min.

### 4.4. Segmentation results on NIH dataset

To evaluate the advantages of our model from different perspectives, we first compare networks that perform well on the NIH dataset. In addition, we compare with the current mainstream medical segmentation networks based on VIT. Finally, we visualize the segmentation results of the proposed Trans-Deformer network.

#### 4.4.1. Comparison with state-of-the-art methods

Table 1 shows the comparison of our network with the current state-of-the-art networks [19,20,24–26,30,31,33–37] on the NIH dataset using the fourfold cross-validation.

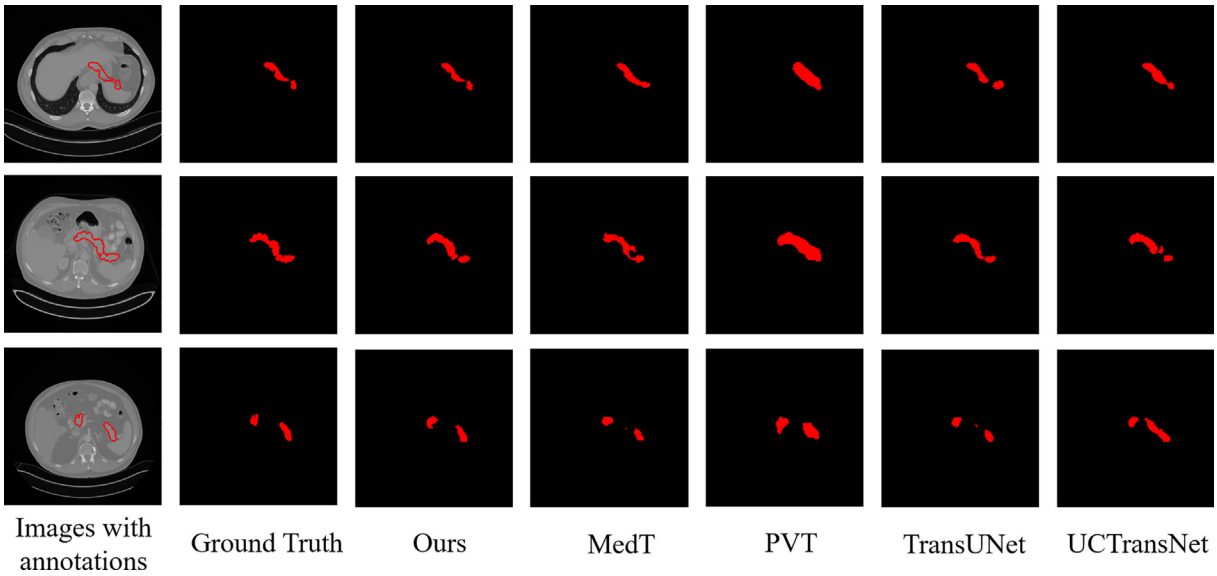


**Fig. 9.** Comparison of segmentation results with different VIT-based mainstream medical image segmentation networks on the NIH dataset. The leftmost column is an overlay that outlines the ground truth in the original image.

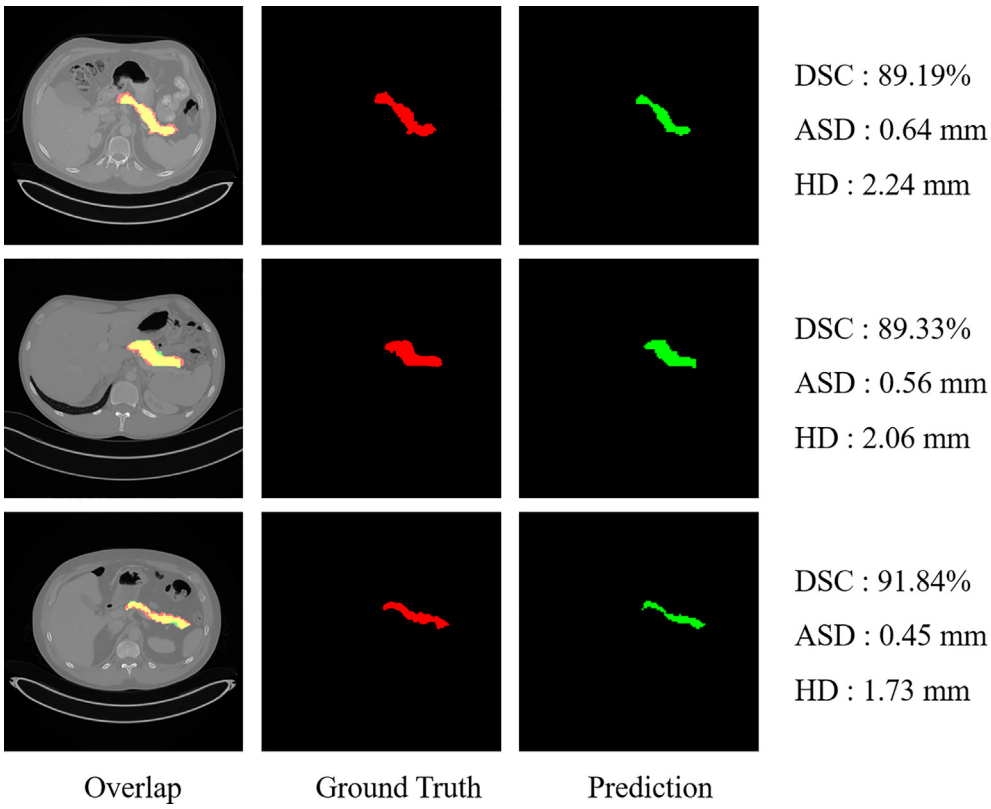


**Fig. 11.** The presentation of segmentation results on the NIH dataset. Green represents the prediction of the network, red represents the ground truth, and yellow represents the overlapping area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
The results (measured by the DSC, Precision, Recall and Testing time) of pancreas segmentation on the MSD dataset. "-" denotes that the corresponding results are not provided in the literature. Optimal results (described by mean ± std) are shown in bold.

| Method | DSC(%) | Precision(%) | Recall(%) | Testing time |
|---|---|---|---|---|
| H. Chen et al. [35] | 76.60 ± 7.30 | 87.70 ± 8.30 | 69.20 ± 12.80 | – |
| Y. Zhang et al. [30] | 82.74 | – | – | – |
| D. Zhang et al. [34] | 85.56 | – | – | 16–17 min |
| J. Li et al. [36] | 85.65 | – | – | – |
| W. Li et al. [25] | 88.52 ± 3.77 | – | **91.86 ± 5.06** | – |
| Ours | **91.22 ± 1.37** | **93.22 ± 2.79** | 91.35 ± 1.63 | **5–6 min** |

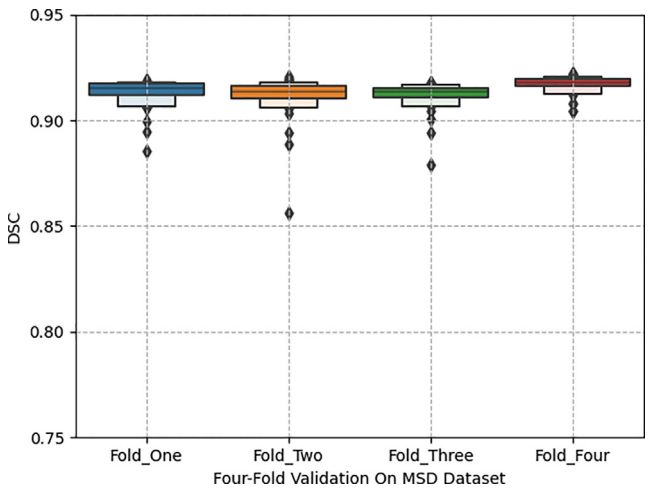


**Fig. 12.** The enhanced boxplot representation of fourfold cross-validation results on the MSD dataset.

The comparison shows that the proposed Trans-Deformer network outperforms state-of-the-art methods, achieving the best DSC of 89.89 %. The standard deviation of 1.82 illustrates that our method is robust for different cases on the NIH dataset. While maintaining better precision and recall, the testing time consumption is relatively low, indicating the potential of the proposed network for the task of pancreas segmentation. The high segmentation accuracy and robustness demonstrate the superiority of our method.

**Table 4**
The results (measured by the ASD and HD) of pancreas segmentation on the MSD dataset. "-" denotes that the corresponding results are not provided in the literature. Optimal results (described by mean ± std) are shown in bold.

| Method | ASD(mm) | HD(mm) |
|---|---|---|
| H. Chen et al. [35] | – | 14.70 ± 14.93 |
| W. Li et al. [25] | 0.95 ± 0.53 | 3.78 ± 4.00 |
| Ours | **0.61 ± 0.11** | **1.97 ± 0.09** |

**Table 5**
The ablation experiments of the proposed network on the NIH dataset. "–" denotes without a module.

| Wavelet decomposition | SIF | Trans-deformer | DSC(%) |
|---|---|---|---|
| – | √ | √ | 88.54 |
| √ | – | √ | 87.38 |
| √ | √ | – | 86.71 |
| √ | √ | √ | 89.89 |

To more intuitively evaluate the advantages of the proposed network in segmenting pancreatic margins, we present the evaluation results based on distance in Table 2.

The results show that our network has better performance on pancreatic margins. The ASD distance of 0.78 mm and the HD distance of 2.09 mm illustrate that the distance between the prediction of the proposed network and the ground truth is shorter, which is consistent with our competitive similarity-based metrics, further indicating that the proposed Trans-Deformer network is effective. The smaller variance demonstrates that the network is robust.

#### *4.4.2. Comparison with the mainstream VIT-based medical image segmentation networks*

To ensure fairness, we test and compare the current mainstream VIT-based medical image segmentation networks on the NIH dataset, which contains MedT, PVT, TransUNet, UCTransNet, and all configurations are consistent with the proposed network.

Fig. 8 illustrates that the proposed Trans-Deformer network outperforms the mainstream VIT-based segmentation networks on the pancreas segmentation task. The results confirm the advantages of our network from two aspects: on the one hand, the prediction of our network only fluctuates in a small range, which justifies that our method has good stability and robustness; and on the other hand, the higher DSC demonstrates the effectiveness of our method.

Fig. 9 visualizes the segmentation effect of our network and the VIT-based mainstream medical image segmentation networks. The proposed network segmentation results are closer in shape to the ground truth, which demonstrates that our network can more effectively mitigate the effects of being distracted by irrelevant background regions. And accurate segmentation can also be achieved for the edges and discontinuities of the pancreas such as shown in the first row of Fig. 9.

**Table 6**
The generalization experiments of the proposed network on the NIH dataset and the MSD dataset. The results are shown in DSC(%).

| Test \ Train | NIH | MSD | NIH + MSD |
|---|---|---|---|
| NIH | 89.89 % | 88.13 % | 90.87 % |
| MSD | 89.17 % | 91.22 % | 90.84 % |

#### *4.4.3. Visualization of results*

From Fig. 10 we can see that our network achieves outperforming performance on the NIH dataset, and the distribution of DSC values at different folds is quite close, indicating that the proposed network has high robustness and can mitigate the effects of sample changes.

Fig. 11 shows the segmentation results of the proposed Trans-Deformer network on the NIH dataset. The presentation of precise segmentation results and quantitative metrics indicate that our network's prediction is very close to the ground truth, implying that our network can effectively capture differences in pancreas shape and size between individuals. The competitive DSC justifies that the deep supervision strategies and the adopted loss function can deal with the problem of category imbalance, making our network focuses more on pancreas regions than redundant background regions. The ASD and HD metrics illustrate that the proposed Trans-Deformer network can finely segment pancreas contours and clearly outline the edge of the pancreas to a certain extent, thus solving the problem of deformation of the pancreas.

### *4.5. Segmentation results on MSD dataset*

Table 3 shows the comparison of the proposed network with current methods [25,30,34–36] that perform well on the MSD dataset.

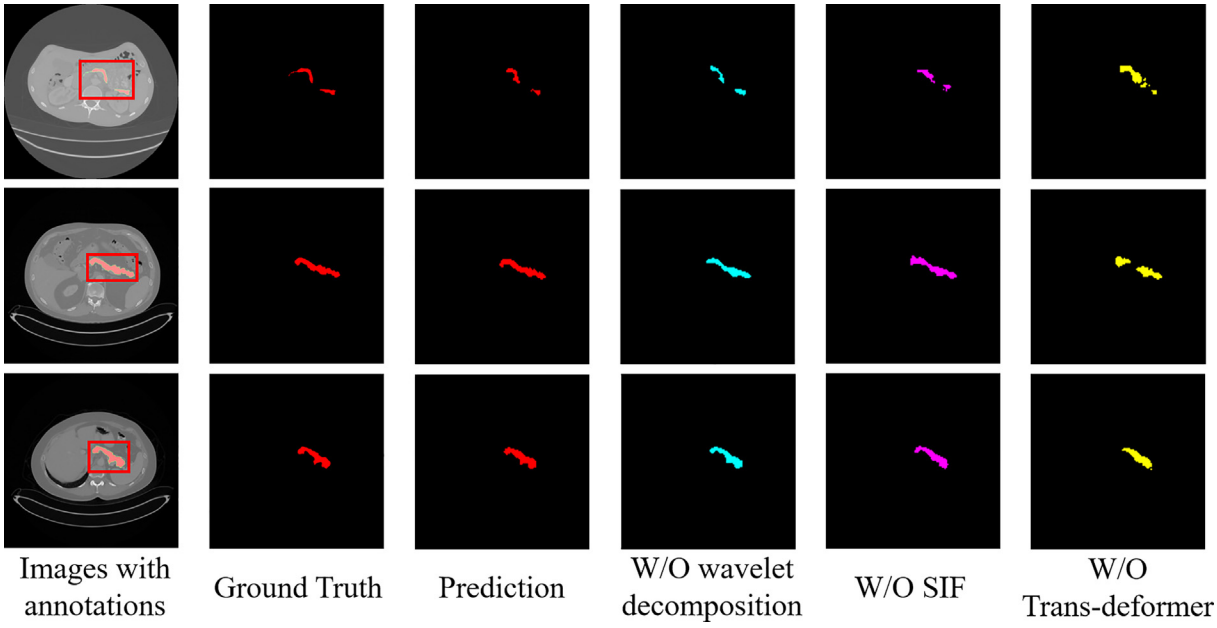


**Fig. 13.** The demonstration of segmentation results of ablation experiments on the NIH dataset. The leftmost column shows the overlay of the original image and the ground truth, and W/O means without a module.

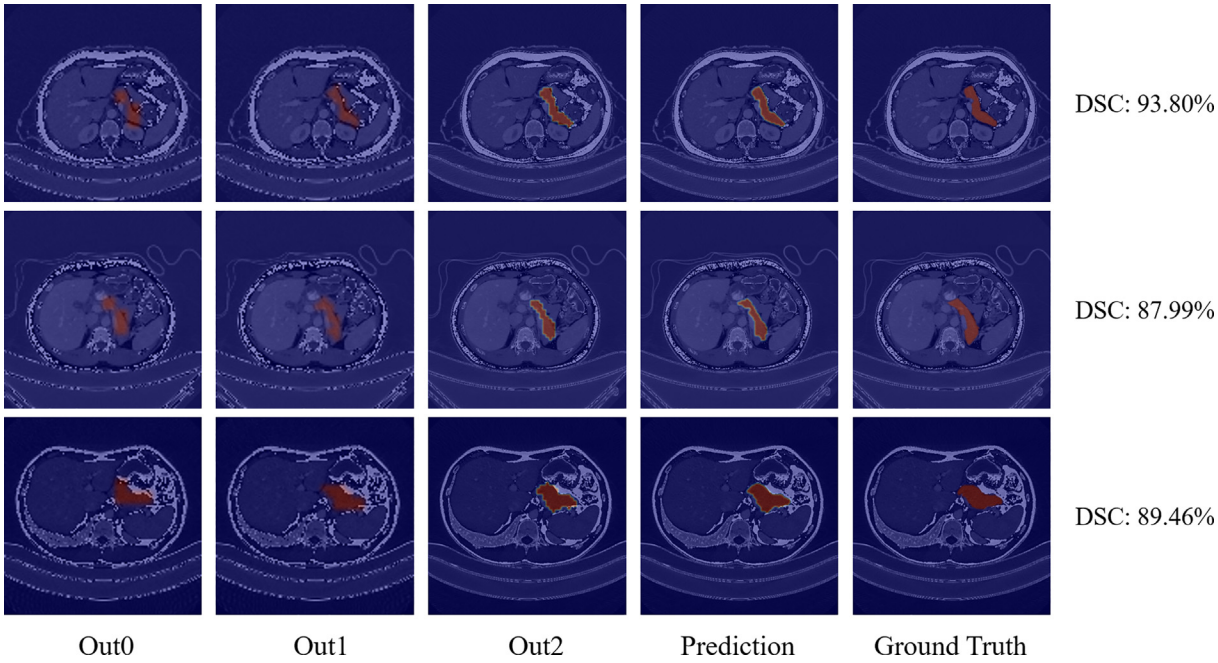


**Fig. 14.** The visualization of the output of different layers in the deep supervision strategy. Among them, Out0, Out1, Out2 and prediction represent the different layer outputs respectively of the proposed Trans-Deformer network. The last column shows that the ground truth is directly mapped to the original image.

From Table 3 we can see that our network achieves a superior DSC of 91.22 % when compared to state-of-the-art methods on the MSD dataset. Nevertheless, it should be noted that our recall is 91.35 %, 0.51 % lower than Li et al. [25], which means that the network performance still has potential for further improvements. Despite this, the standard deviation of the recall of 1.63 is lower, and higher precision is achieved with a shorter testing time, indicating that the network has better performance and higher robustness.

Fig. 12 illustrates the DSC for the fourfold cross-validation of the proposed Trans-Deformer network on the MSD dataset, which reflects our network's ability to achieve accurate segmentation of the pancreas across different datasets.

Table 4 presents the distance evaluation between the prediction result and the ground truth on the MSD dataset. The ASD distance of 0.61 mm and the HD distance of 1.97 mm indicate that our network achieves finer segmentation of pancreatic margins.

## 5. Discussion

### 5.1. Ablation study

To demonstrate that the core modules in the proposed network are effective, we conduct ablation experiments of the proposed Trans-Deformer network on the NIH dataset. As shown in Table 5, we drop the proposed three innovative modules respectively in the proposed network and measure the segmentation DSC metric of the remaining network.

Table 5 shows that the proposed Trans-deformer module has the greatest impact on network performance. After removing it, the DSC score of the network will decrease by 3.18 %. The second most influential is the SIF module, without it, the network DSC will reduce by 2.51 %. The wavelet decomposition strategy also has a 1.35 % impact on network performance. To more intuitively display the impact of the proposed innovative modules on the network, we visualize the segmentation performance of the network after discarding a submodule alone in Fig. 13. Fig. 13 presents the following information: firstly, the network achieves the relatively complete pancreas segmentation with the Trans-deformer module, thus solving the problem of pancreas deformation to a certain extent; secondly, the SIF module further refines the segmentation results by fusing local features and global features; at last, the details of the edge of the pancreas can be further improved with the wavelet decomposition module, and finally accurate pancreas segmentation can be achieved. The visualization of segmentation results further confirms the effectiveness of each innovation module, which is consistent with the data in Table 5.

### 5.2. Model generalization on different datasets

To further demonstrate the advantages of the proposed Trans-Deformer network, we performed generalization verification on two datasets of NIH and MSD. Specifically, the following four sets of experiments were conducted: 1) trained on the NIH dataset and tested on the MSD dataset; 2) trained on the MSD dataset and tested on the NIH dataset; 3) trained on an equal mix of NIH and MSD datasets, and tested on the remainder of the NIH dataset; 4) trained on an equal mix of NIH and MSD datasets, and tested on the remainder of the MSD dataset. During the implementation of the experiment, to be consistent with the previous experimental settings, we set the ratio of the training set to the test set to 3:1, and set the total case number of samples in each experiment to 80. The mixed dataset consisted of 30 cases randomly selected from the NIH and MSD datasets respectively to form a mixture of 60 cases. During the test, 30 cases were randomly selected from the remaining cases after the training set was taken out. The experimental results are shown in Table 6.

From the data in Table 6, we can see that the proposed network achieves superior performance and small DSC fluctuations on two datasets, indicating that the proposed Trans-Deformer network has strong generalization ability on the pancreas segmentation task thanks to our innovative modules. Furthermore, the following

points can be seen: 1) When the test set is fixed, the DSC score trained on the same data set as the test set is better than training on another dataset, indicating that there is a certain degree of difference between the two datasets. 2) When the training set is fixed, the fluctuation of 0.72 % in the second column is much lower than the fluctuation of 3.09 % in the third column, implying that the NIH dataset contains a wider variety of samples. This might be because although the MSD dataset contains more samples, the average total number of slices per sample and the number of valid slices that contain the pancreas are less than in the NIH dataset.

### 5.3. Visualization of deep supervision strategy

To highlight the effectiveness of the deep supervision strategy, we present its advantages from a visual perspective.

Fig. 14 shows the class activation map (CAM) effect output by different layers after the deep supervision strategy is adopted. We can see that in the underlying feature map such as Out0, the proposed network has been able to focus on the pancreas area through the constraints of the deep supervision strategy, which further explains the reason why the proposed Trans-Deformer network finally achieves high-precision segmentation of the pancreas.

### 5.4. Limitations and future work

Although the proposed Trans-Deformer network has achieved competitive performance in the pancreas segmentation task, it is clear that the network still has room for improvement. Our network currently segments the pancreas in a fully supervised manner, which requires pixel-level annotated data to train the network. Unfortunately, in the actual medical image segmentation scene, it is laborious to perform pixel-level annotation on the target area, and it is usually difficult to obtain high-quality segmentation annotations. In contrast, it is more efficient to only perform image-level annotation on samples. Therefore, in the follow-up research, we will consider optimizing the proposed network in a weakly supervised manner by combining it with the classification task.

## 6. Conclusion

This paper proposes the Trans-Deformer network for pancreas segmentation. In dealing with the challenging task of segmenting the pancreas, the proposed network effectively solves the problems of pancreas deformation, the unbalanced category caused by the small size, and the blurred boundaries caused by the low contrast, and further improves segmentation metrics. Specifically, we propose a Trans-deformer module combining deformable convolution with VIT, which enables the generated tokens to change adaptively according to the shape of the pancreas, solving the problem of pancreas deformation by improving segmentation accuracy. Meanwhile, the proposed SIF module can perfectly fuse local features and global features, and enable the network to have a clearer expression, ensuring the intrinsic connection between the low-level and the high-level. The module based on two-dimension wavelet decomposition helps the network to pay more attention to the edge of the pancreas by providing high-frequency texture information, which solves the problem of the blurred boundaries of the pancreas. In addition, the adopted deep supervision strategy accelerates the convergence of the network to a certain extent, and improves the robustness and generalization of the network.

Our method was evaluated on the publicly available NIH dataset and MSD dataset. The results demonstrated that the proposed Trans-Deformer network achieved comparable performance, not only outperforming other state-of-the-art methods on both datasets, but also surpassing the mainstream VIT-based medical image segmentation networks, which proved that the proposed network is effective. Moreover, the proposed Trans-Deformer network can be flexibly integrated into any other segmentation network and can be quickly adapted to other segmentation tasks.

## Data availability

Data will be made available on request.

## 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 in part by the Science and Technology Commission of Shanghai Municipality (20DZ2254400, 21DZ2200600, 20DZ2261200), National Scientific Foundation of China (82170110), Fujian Province Department of Science and Technology (2022D014).

## References

[1] P. Ghaneh, E. Costello, J.P. Neoptolemos, Biology and management of pancreatic cancer, Gut 56 (8) (2007) 1134–1152.
[2] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2021, CA: Cancer J. Clin. 71 (1) (2021) 7–33.
[3] Y. Ning, Z. Han, L. Zhong, C. Zhang, Automated pancreas segmentation using recurrent adversarial learning, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 927–934.
[4] X. Liao, Y. Qian, Y. Chen, X. Xiong, Q. Wang, P.-A. Heng, MMTLNet: Multi-Modality Transfer Learning Network with adversarial training for 3D whole heart segmentation, Comput. Med. Imaging Graphics 85 (2020).
[5] W. Tang, D. Zou, S. Yang, J. Shi, J. Dan, G. Song, A two-stage approach for automatic liver segmentation with Faster R-CNN and DeepLab, Neural Comput. Appl. (2020) 1–10.
[6] A. Hatamizadeh et al., Unetr: Transformers for 3d medical image segmentation, in: ProceedIngs of the IEEE/CVF WInter Conference on Applications of Computer Vision, 2022, pp. 574–584.
[7] M. Kim, B.-D. Lee, Automatic lung segmentation on chest X-rays using self-attention deep neural network, Sensors 21 (2) (2021) 369.
[8] N. Heller et al., The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge, Med. Image Anal. 67 (2021).
[9] N. Bouteldja et al., Deep learning–based segmentation and quantification in experimental kidney histopathology, J. Am. Soc. Nephrol. 32 (1) (2021) 52–68.
[10] A.Z. Arifin, A. Asano, Image segmentation by histogram thresholding using hierarchical cluster analysis, Pattern Recogn. Lett. 27 (13) (2006) 1515–1521.
[11] R. Pohle, K.D. Toennies, Segmentation of medical images using adaptive region growing, in: Medical Imaging 2001: Image Processing, International Society for Optics and Photonics, 2001, pp. 1337–1346.
[12] J. Gao, B. Wang, Z. Wang, Y. Wang, F. Kong, A wavelet transform-based image segmentation method, Optik 208 (2020).
[13] A. Vaswani et al., Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[15] G.-P. Ji et al., "Progressively Normalized Self-Attention Network for Video Polyp Segmentation," arXiv preprint arXiv:2105.08468, 2021.
[16] Y. Gao, M. Zhou, D.N. Metaxas, UTNet: a hybrid transformer architecture for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 61–71.
[17] Y. Zhang et al., "A Multi-Branch Hybrid Transformer Networkfor Corneal Endothelial Cell Segmentation," arXiv preprint arXiv:2106.07557, 2021.
[18] J. Dai et al., Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
[19] M. Huang, C. Huang, J. Yuan, D. Kong, A Semiautomated Deep Learning Approach for Pancreas Segmentation, J. Healthcare Eng. 2021 (2021).
[20] Y. Wang et al., Pancreas segmentation using a dual-input v-mesh network, Med. Image Anal. 69 (2021).
[21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical

image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[23] P.F. Christ et al., Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 415–423.
[24] M. Li, F. Lian, S. Guo, Automatic Pancreas Segmentation Using Double Adversarial Networks With Pyramidal Pooling Module, IEEE Access 9 (2021) 140965–140974.
[25] W. Li, S. Qin, F. Li, L. Wang, MAD-UNet: A deep U-shaped network combined with an attention mechanism for pancreas segmentation in CT images, Med. Phys. 48 (1) (2021) 329–341.
[26] F. Li, W. Li, Y. Shu, S. Qin, B. Xiao, Z. Zhan, Multiscale receptive field based on residual network for pancreas segmentation in CT images, Biomed. Signal Process. Control 57 (2020).
[27] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE, 2016, pp. 565–571.
[28] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 424–432.
[29] Z. Zhu, Y. Xia, W. Shen, E. Fishman, A. Yuille, A 3D coarse-to-fine framework for volumetric medical image segmentation, in: 2018 International conference on 3D vision (3DV), IEEE, 2018, pp. 682–690.
[30] Y. Zhang et al., A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set, Med. Image Anal. 68 (2021).
[31] J. Li et al., A 2.5 D semantic segmentation of the pancreas using attention guided dual context embedded U-Net, Neurocomputing 480 (2022) 14–26.
[32] Y. Zhou, L. Xie, W. Shen, Y. Wang, E.K. Fishman, A.L. Yuille, A fixed-point model for pancreas segmentation in abdominal CT scans, in: International conference on medical image computing and computer-assisted intervention, Springer, 2017, pp. 693–701.
[33] H. Li, J. Li, X. Lin, and X. Qian, "A Model-Driven Stack-Based Fully Convolutional Network for Pancreas Segmentation," in: 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), 2020: IEEE, pp. 288-293.
[34] D. Zhang, J. Zhang, Q. Zhang, J. Han, S. Zhang, J. Han, Automatic pancreas segmentation based on lightweight DCNN modules and spatial prior propagation, Pattern Recogn. 114 (2021).
[35] H. Chen, Y. Liu, Z. Shi, Y. Lyu, Pancreas segmentation by two-view feature learning and multi-scale supervision, Biomed. Signal Process. Control 74 (2022).
[36] J. Li, X. Lin, H. Che, H. Li, X. Qian, Pancreas segmentation with probabilistic map guided bi-directional recurrent UNet, Phys. Med. Biol. 66 (11) (2021).
[37] P. Hu et al., Automatic Pancreas Segmentation in CT Images With Distance-Based Saliency-Aware DenseASPP Network, IEEE J. Biomed. Health. Inf. 25 (5) (2020) 1601–1611.
[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[39] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881-6890.
[40] 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," arXiv preprint arXiv:2109.04335, 2021.
[41] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," arXiv preprint arXiv:2102.12122, 2021.
[42] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," arXiv preprint arXiv:2102.10662, 2021.
[43] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
[44] Y. Ji et al., Multi-Compound Transformer for Accurate Biomedical Image Segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 326–336.
[45] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.
[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492-1500.
[47] L. Chen *et al.*, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5659-5667.
[48] M.J. Shensa, The discrete wavelet transform: wedding the a trous and Mallat algorithms, IEEE Trans. Signal Process. 40 (10) (1992) 2464–2482.
[49] H.R. Roth et al., Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2015, pp. 556–564.
[50] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv preprint arXiv:1902.09063, 2019.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in: 2009 IEEE conference on computer vision and pattern recognition, 2009: Ieee, pp. 248-255.
[52] Q. Wei et al., "Learn to segment retinal lesions and beyond," in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021: IEEE, pp. 7403-7410.
[53] A. Tureckova, T. Turecek, Z. Kominkova Oplatkova, A.J. Rodriguez-Sanchez, Improving CT Image Tumor Segmentation Through Deep Supervision and Attentional Gates, Front. Robotics AI 7 (2020) 106.
[54] 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:1712.00201, vol. 2, 2017.
[55] F. Farheen, M. Shamil, N. Ibtehaz, and M. S. Rahman, "Segmentation of Lung Tumor from CT Images using Deep Supervision," arXiv preprint arXiv:2111.09262, 2021.

**Shunbo Dai** was born in Xiangyang, Hubei Province, China in 1998. He received a B.S. degree in communication engineering from Shanghai Normal University in 2020. He is currently pursuing an M.S. degree at East China University of Science and Technology. His research interests include medical image processing in deep learning and computer vision.

**Yu Zhu** *Member IEEE* received a Ph.D. degree from Nanjing University of Science and Technology, China, in 1999. She is currently a professor in the department of electronics and communication engineering of East China University of Science and Technology. Her research interests include image processing, computer vision, multimedia communication and deep learning, especially, for the medical auxiliary diagnosis by artificial intelligence technology. She has published more than 90 papers in journals and conferences.

**Xiaoben Jiang** is pursuing a Ph.D. degree at East China University of Science and Technology. His current research interests include digital image processing and computer vision. His experience includes the denoising method on chest X-ray images and CT images, and detection of COVID-19 cases from denoised CXR images. He has published in journals in the crossing field of medical science and computer vision, and has been involved in publicly and privately funded projects.

**Fuli Yu** was born in Taizhou, Zhejiang, China in 1998. She received a B.S. degree in information engineering from East China University of Science and Technology, Shanghai, in 2019. She was a recipient of the Outstanding Graduate title of Shanghai universities, class of 2015. Since 2019, she has continued her master's degree at ECUST. Her research interests include image processing, artificial intelligence and its applications.

**Jiajun Lin** obtained his Ph.D. degree from TSINGHUA University, Beijing. He is a professor at the School of Information Science and Engineering, East China University of Science and Technology. His research interests include Intelligent Information Processing and Security of Industry Control Systems.

**Dawei Yang** is dedicated to the early diagnosis of lung cancer and relevant studies, with special interests in the management of pulmonary nodules and validation of diagnostic biomarker panels based on MIOT, CORE and radiomics artificial intelligence (AI) platform. He is a member of the IASLC Prevention, Screening and Early Detection Committee. Since 2011, he has published 16 SCI research articles and 9 as the first author, including which on Am J Resp Crit Care (2013), Can Lett (2015, 2020) and Cancer (2015 and 2018), etc. As a presenter for oral or poster presentations in ATS, WCLC, APSR, ISRD couple times. He is one of the peer reviewers for international journals, such as J Cell Mol Med, J Transl Med, etc.