Applied Intelligence (2024) 54:12272–12292
https://doi.org/10.1007/s10489-024-05780-9

MDHT-Net: Multi-scale Deformable U-Net with Cos-spatial and Channel Hybrid Transformer for pancreas segmentation

HuiFang Wang[1] · DaWei Yang[2,3] · Yu Zhu[1] · YaTong Liu[1] · JiaJun Lin[1]

Accepted: 16 August 2024 / Published online: 11 September 2024

Abstract

Accurate pancreas segmentation is essential for the diagnosis of pancreas disease, while it is still challenging due to the variable structure and small size of the pancreas. In this paper, we propose a Multi-scale Deformable U-Net with Cos-spatial and Channel Hybrid Transformer (MDHT-Net) for pancreas segmentation. To mitigate the ambiguity between the codec stages, the Cos-spatial and Channel Hybrid Transformer (CCHT) module is designed as a novel skip connection, enhancing the network's ability to perceive spatial information and reveal the inter-channel relationships within different layers' features. Furthermore, the CCHT efficiently aggregates multi-stage contextual information by improving the self-attention mechanism in two different manners, overcoming the limitation of computational complexity. In addition, to comprehensively understand deep semantic information, the Multi-scale Feature Adaptive-extraction (MFA) module is proposed to dynamically enhance the network's receptive field by integrating the pancreas characteristics of scale variations. The experimental results present that our proposed MDHT-Net achieves superior performance compared to other existing state-of-the-art methods on two public pancreas datasets, with the mean Dice coefficient of $91.07 \pm 1.19\%$ for NIH and $91.52 \pm 0.66\%$ for MSD, respectively. Given the effectiveness and advantages of our proposed MDHT-Net, it is expected to be a potential tool to assist clinicians in detecting pancreas disease and making reasonable treatment plans.

Keywords Pancreas segmentation · Deformable convolution · Transformer · Multi-scale information

1 Introduction

Pancreatic cancer is one of the most common and deadliest malignancies, with a five-year survival rate of less than 5% [1–3]. Accurate pancreas segmentation technology can effectively assist clinicians in the clinical analysis and diagnosis of pancreatic cancer using computer assistance. In recent years, with the rapid development of deep learning, the automatic segmentation of other organs such as the lungs, kidneys, and liver has achieved relatively high accuracy [4–6]. However, accurately segmenting the pancreas from CT images still presents many challenges. Compared to other organs, the pancreas exhibits high variability in size, shape, and position among different patients, and it typically occupies only a small portion of the entire CT image volume, as shown in Fig. 1. Furthermore, the pancreas has weak contrast with surrounding tissues, and its boundaries are often blurry, which can lead to interference from the background regions in CT images for deep neural networks, resulting in low segmentation accuracy [7–9]. Therefore, the pancreas is considered one of the most complex organs to segment, making the development of a robust and accurate automatic pancreatic segmentation model of profound significance.

In 2015, Ronneberger proposed U-Net [10] for medical image segmentation and attained outstanding performance, subsequently, many scholars designed a series of U-Net variants such as R2U-Net [11], U-Net++ [12], TransUNet [13] and so on. Nowadays U-shaped networks have been the mainstream framework for medical image segmentation [14, 15].

HuiFang Wang and DaWei Yang contributed equally to this work.

✉ Yu Zhu
zhuyu@ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

[2] Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China

[3] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai, China

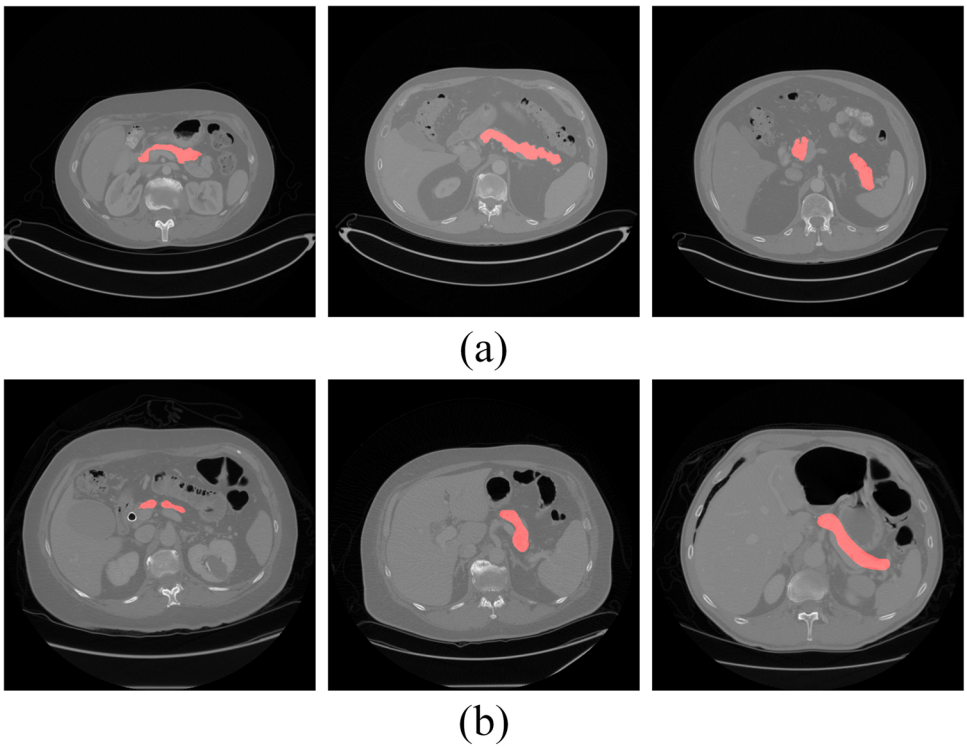


Fig. 1 CT slices from the two public pancreas datasets: (a) Some CT examples from the NIH pancreas dataset. (b) Some CT examples from the MSD pancreas dataset. Three slices of different cases are randomly chosen from each dataset

For pancreas segmentation: Oktay et al. [16] proposed a novel Attention Gate (AG) model that can be integrated into U-Net to highlight the important salient features of pancreas. Li et al. [17] proposed three cross-domain information fusion strategies to solve the problem that U-Net can't distinguish the regions between pancreas and background in low-contrast CT images. Li et al. [18] introduced the Multiscale Attention Dense residual U-shaped network (MAD-UNet) to address the problems of intraclass inconsistency. Additionally, to improve the consistency of adjacent CT slices, Li et al. [19] designed a stack-based U-Net architecture to fuse the two-dimensional and local three-dimensional context information. Although the aforementioned methods improved the information propagation process in U-shaped architectures by incorporating inter-slice relationships or adding attention mechanisms, the lack of specific analysis of pancreatic features limited the segmentation performance.

Specially, considering the highly variable appearances of the pancreas across different patients and slices, Huang et al. [20] integrated the deformable convolution into U-Net to flexibly capture the pancreas locations of various shapes in 2021, the excellent segmentation results demonstrate deformable convolution is more suitable for pancreas segmentation compared with traditional standard convolution. The deformable convolution is illustrated in Fig. 2, which utilizes an additional parallel standard convolution layer to learn the offsets for each sampling point of the input image, and then the obtained offsets are added to original sampling positions, thus the deformable convolution can adaptively adjust predefined receptive field according to the target.

Since Transformer [21] gained tremendous success in NLP(Natural Language Processing), Dosovitskiy et al. [22] first introduced Vision Transformer(ViT) for image classi-

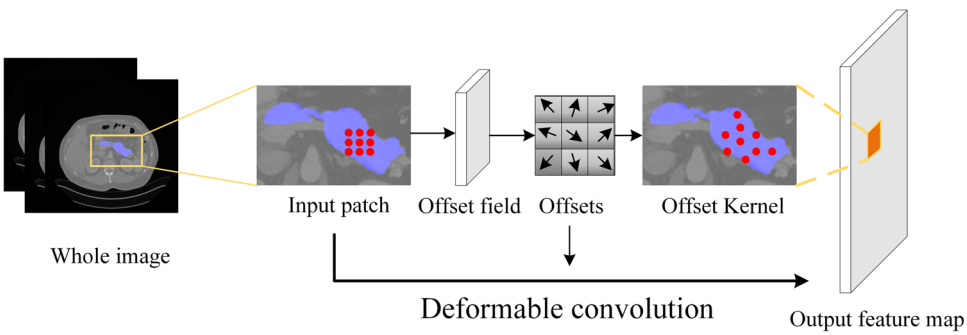


Fig. 2 Illustration of 3×3 deformable convolution. The offset field is learned from the input feature maps by applying a standard convolutional layer

fication and accomplished a immense break through self-attention mechanism. After that, Transformer has been widely applied in computer vision in recent years. In the field of pancreas segmentation: Chen et al. [23] deployed a channel-wise transformer into 3D U-Net as the skip connection, coordinating global features to assist the network learning. Cheng et al. [24] sequentially utilized the DenseA-SPP module and Transformer in the center of U-Net, realizing the construction of long-range dependency. Qu et al. [25] proposed a transformer-guided progressive fusion network, effectively combining the CNN branch and the transformer branch to enhance the feature representation of the pancreas.

Motivated by the aforementioned methodologies, a novel network called MDHT-Net for pancreas segmentation is proposed in this paper. The network employs a coarse-to-fine segmentation strategy, the pre-trained U-Net is utilized as the coarse segmentation network for the fast location of pancreas regions, and then fed the focused region into the MDHT-Net for fine segmentation. The proposed MDHT-Net consists of five layers. To balance the network's capability to learn various pancreatic features and the demand for computational resources, only the third and fourth layers corresponding to the encoder-decoder in the U-shaped framework are replaced with dual deformable convolution blocks. The Cos-spatial and Channel Hybrid Transformer (CCHT) is designed as a skip connection, improving the self-attention mechanism to efficiently extract global features from both spatial and channel dimensions with linear complexity. Furthermore, the Multi-scale Feature Adaptive-extraction (MFA) module is proposed to dynamically optimize the network's receptive field, facilitating the transmission of multi-scale information between codecs.

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

(1) A novel network MDHT-Net is proposed for pancreas segmentation, which skillfully integrates the deformable U-shaped framework with a designed hybrid Transformer, thus automatically perceiving position variations in pancreas feature contours and effectively fusing local and global contextual information. It makes the network adaptively process the complex and varied structure of the pancreas.
(2) A Cos-spatial and Channel Hybrid Transformer (CCHT) module is proposed by introducing the Cosine Spatial attention mechanism (CSA) and Efficient Channel attention mechanism (ECA). The CSA mechanism explores the spatial dependency for multi-layer features and realizes linear computational complexity through the cosine decomposition theorem. The ECA mechanism further establishes inner-relationships between different channels for global features extraction. The CCHT module significantly alleviates the ambiguity between codecs, thereby enhancing the accuracy and robustness of the decoder's output.
(3) A Multi-scale Feature Adaptive-extraction (MFA) module is developed by the multi-branch atrous convolution structure with a scale attention mechanism, comprehensively calibrates and optimizes the deep-level semantics through the regularities of scale variations, thus reducing the probability of mis-segmentation.
(4) Experimental results on the NIH pancreas dataset(Mean DSC: 91.07±1.19%) and the MSD pancreas datasset(Mean DSC: 91.52 ± 0.66%) show that our network outperforms existing state-of-the-art methods, demonstrating the effectiveness of our proposed MDHT-Net.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3 describes the architecture of the proposed MDHT-Net in detail. Section 4 presents the experimental settings and results. Finally, discussions and conclusions are provided in Section 5 and Section 6, respectively.

2 Related work

2.1 Pancreas segmentation

Traditional pancreas segmentation methods mainly include multi-atlas techniques, region growing algorithms, and statistic shape models [26, 27]. Karasawa et al. [28] proposed a multi-atlas pancreas segmentation approach based on vessel structure around the pancreas, selecting atlases with high pancreatic similarity to the unlabeled CT volume. Tam et al. [29] applied region-growing to label pancreas region, returning the segmented result which has the same characteristics as the seed point. Hammon et al. [30] incorporated spatial relationships across the pancreas, surrounding organs, and vessels to acquire a constrained statistical shape model for pancreas segmentation. However, these methods are tedious with limited performance due to the manual intervention. Most of the traditional methods achieved low Dice coefficient(<75%) for pancreas segmentation [19].

Due to the rapid development of deep neural networks, the method based on deep learning has gradually become the mainstream method in the field of pancreas segmentation [31, 32], as Table 1 represented. The deep-learning-based methods for pancreas segmentation can be divided into the single-stage methods and the two-stage methods. Zheng et al. [33] proposed a two-dimensional deep learning-based method to describe the uncertain regions at pancreatic MRI images in the process of iterative segmentation, gradually correcting segmentation results and obtaining an 84.37% Dice coefficient on the NIH dataset. To make full use of the local context during the segmentation process, Li et al.

Table 1 The previous advanced pancreas segmentation methods

Authors	Pros.	Adv.	Disad.	Results
Single-Stage				
Zheng et al. [33]	Describing uncertain regions based on shadowed sets.(2020)	Solving the problem of topological fracture for pancreas segmentation.	Lack of analysis of application performance.	84.37% Dice on the NIH Dataset.
Li et al. [34]	A probabilistic-map-guided bi-directional recurrent UNet.(2021)	Avoiding losing precise context during the segmentation process.	Lack of generalization validation on other tasks.	85.35% Dice on the NIH Dataset.
Chen et al. [35]	A fuzzy skip connection module. (2022)	Reducing the redundant information of non-target regions.	Lack of visual presentation of case indicators.	87.91% Dice on the NIH Dataset.
Two-Stage				
Dogan et al. [36]	A two-phase approach using Mask-RCNN and 3D U-Net.(2021)	The computational costs and complexity are low.	Lack of generalization validation on other tasks.	86.15% Dice on the NIH Dataset.
Qiu et al. [37]	A cascaded multi-scale feature calibration UNet.(2023)	Excellent inference time.	Lack of visual presentation of case indicators.	86.30% Dice on the NIH Dataset.
Xia et al. [38]	A multipath fusion network based on 2.5D.(2023)	Incorporating Z-axis information during the segmentation process.	Lack of analysis of application performance.	87.01% Dice on the NIH Dataset.
Yao et al. [39]	Different connections for multi-layer interaction.(2023)	Fighting the vanishing gradient problem by the long connection.	Lack of visual presentation of ablation experiment.	87.87% Dice on the NIH Dataset.

[34] introduced a bi-directional recurrent scheme to optimize the network and earned an 85.35% Dice coefficient on the NIH dataset. Considering the small and changeable structural characteristics of the pancreas, Chen et al. [35] designed a fuzzy skip connection module to transform the low-level features into high-level semantic features, achieving a relatively impressive Dice coefficient of 87.91% on the NIH dataset in 2022.

Compared with the aforementioned single-stage segmentation methods, the two-stage methods primarily leverage the coarse localization results from the first stage to assist the fine segmentation in the second stage network. Dogan et al. [36] adopted Mask R-CNN to detect the pancreatic candidate region roughly, then obtained the precise segmentation result through 3D U-Net. To further improve the segmentation framework from coarse to fine, Qiu et al. [37] designed a Cascaded Multi-scale Feature Calibration UNet (CMFCUNet) for pancreas segmentation, attaining a Dice coefficient of 86.30% on the NIH dataset. Xia et al. [38] incorporated Z-axis information with Multi-path Transformer fusion network(MTr-Net), efficiently addressing the boundary deformation of the pancreas. To overcome the semantic gap problem between the codec in fine-segmentation stage, Yao et al. [39] designed different connections for multi-layer interaction and got a Dice score of 87.87% on the NIH dataset.

Different from the previous advanced pancreas segmentation methods, the proposed MDHT-NET simultaneously considers the changeable structural characteristics of the pancreas and the importance of information interaction during the segmentation process. By incorporating deformable convolutions in deep layers, the network adeptly learns the spatial position of the pancreas. Furthermore, it aggregates various attention mechanisms to facilitate effective information interaction between the codec stages, resulting in decoded outputs that are closer to the ground truth and achieve an excellent Dice coefficient of $91.07 \pm 1.19\%$ on the NIH dataset.

2.2 Attention meachnisam

Attention mechanism has been widely applied in computer vision tasks in recent years, which can effectively extract important information from the region of interest and suppress irrelevant information [40–42]. The attention mechanism usually can be divided into spatial attention and channel attention [43, 44]. Most of the researchers choose to combine these two kinds of attention for better feature representation. For instance: Huang et al. [45] proposed a Discriminative Feature Attention Network for pancreas segmentation, utilizing a Bottleneck Attention Module (BAM) to learn spatial and channel-wise attention, and successfully obtain discriminative hierarchical features. Chen et al. [46] designed an effective Residual Multi-Scale Dilated Attention (RMSA) module to capture comprehensive inter-channel relationships and multi-scale spatial features, enhancing the segmentation of pancreas. Yan et al. [47] incorporated a hybrid attention module into U-Net, acquiring a more robust feature representation for pancreas segmentation.

Different from the above methods, we successfully integrate global information and local information in both spatial and channel dimensions through the CSA mechanism and ECA mechanism with linear computational complexity, significantly enhancing the overall network's feature extraction capabilities.

2.3 ViT for medical image segmentation

Transformer was originally proposed by Vaswani et al. [21] for machine translation, which can model long-range dependencies through the self-attention mechanism and achieve state-of-the-art performance in the NLP domain. Inspired by the remarkable performance of Transformer in NLP domain, many researchers attempted to apply Transformer in computer vision. Specifically, the Vision Transformer [22] (ViT) proposed by Dosovitskiy et al. presents that the pure Transformer-based methods can also perform very well in computer vision. ViT divides an input image into a series of fixed-size patches, followed by fed into transformer blocks for image classification tasks.

To overcome the quadratic computational complexity of ViT, swim-transformer [48] introduced the window-based attention mechanism and enhanced local feature interaction through shifted windows. PVT [49] handled the segmentation task of high-resolution input images through a progressive pyramid architecture, thus acquiring better performance with lower computation costs. Additionally, Transformer is also popular in the field of medical image segmentation. Swin-Unet [50] constructed a U-shaped network based on the swim-transformer block for medical image segmentation, taking full advantage of the Transformer's global awareness capacity. To address the problem of insufficient medical data, MedT [51] integrated an additional control mechanism into the self-attention module and further improved the performance through a Local-Global training strategy. TransUNet [13] adopted the CNN-Transformer hybrid structure for medical segmentation primary, which effectively leveraged both CNN's ability for local feature extraction and the Transformer's advantage for global information interaction. UCTransNet [52] redesigned the skip connection of the origin U-Net through the multi-scale channel-wise cross-attention mechanism, significantly optimizing the combination of U-Net and Transformer.

3 Method

In this part, an overview of the proposed MDHT-Net is presented in Section 3.1. The key modules of MDHT-Net: the Cos-spatial and Channel Hybrid Transformer (CCHT) module and the Multi-scale Feature Adaptive-extraction (MFA) module are introduced in Section 3.2 and Section 3.3 respectively. Finally, the loss function is described in Section 3.4.

3.1 Overview

The overall architecture of the proposed MDHT-Net is illustrated in Fig. 3. The coarse location of pancreas region is accomplished by the pre-trained U-Net from the first stage, followed by the application of the designed MDHT-Net for fine segmentation based on the focused pancreas region. MDHT-Net follows an asymmetric encoder-decoder framework, it mainly consists of four parts: feature encoder, CCHT module, MFA module, and feature decoder.

Considering the augmented computational resources required by deformable convolutions to facilitate the network in learning the complex morphological characteristics of the pancreas, we select to substitute the third and fourth layers of the associated encoder and decoder with 3×3 deformable convolutional blocks. Simultaneously, the initial two layers continue to employ standard 3×3 convolutions for preliminary feature extraction. To effectively fuse the multi-scale spatial-wise and channel-wise features from different encoder levels with sufficient receptive fields, the CCHT module is designed to connect the encoding path and the decoding path. To further optimize the receptive field of the network and adaptively acquire changes in scale, the MFA module is designed to replace the center part in the origin U-Net structure. During the encoding phase, the input is downsampled four times, with each stage halving the spatial dimensions and doubling the number of channels. In the corresponding decoding phase, the original input dimensions $H \times W \times C$ are restored through four stages of upsampling. To better constrain the network convergence, the deep supervision mechanism is employed, where the output of each layer is adjusted to the same dimensions as the ground truth through a 1×1 convolution followed by an upsampling operation.

3.2 Cos-spatial and Channel Hybrid Transformer

In order to extract and fuse the effective feature expressions from different layers, we design a Cos-spatial and

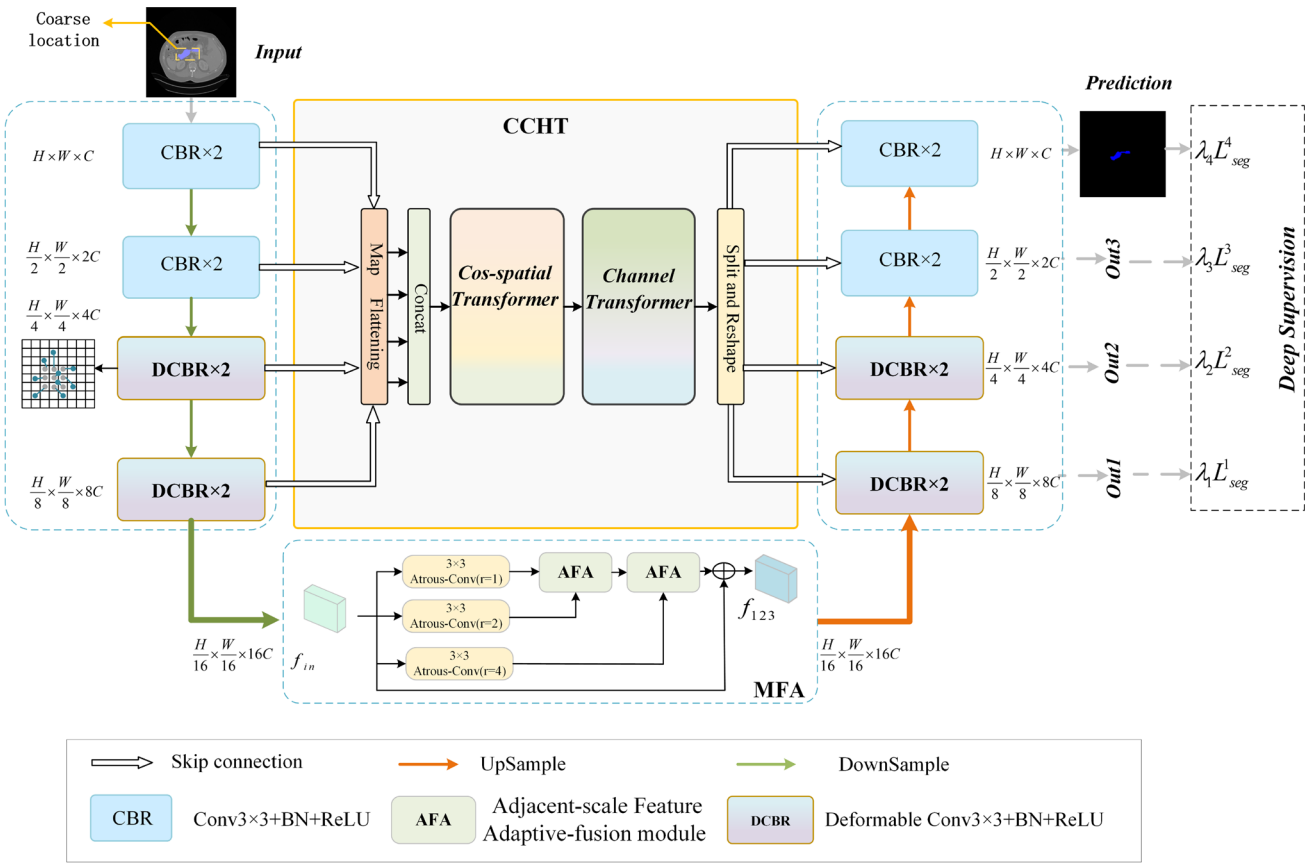


Fig. 3 Network architecture of the proposed MDHT-Net

Channel Hybrid Transformer (CCHT) as the skip connection between codecs, as depicted in Fig. 4(a). Initially, we perform a reshape operation on the multi-scale feature maps originating from the encoder: $X_1 \in R^{H\times W\times C}$, $X_2 \in R^{\frac{H}{2}\times\frac{W}{2}\times 2C}$, $X_3 \in R^{\frac{H}{4}\times\frac{W}{4}\times 4C}$, $X_4 \in R^{\frac{H}{8}\times\frac{W}{8}\times 8C}$. These feature maps are adjusted to have uniform channel dimensions C and flattened into 1D sequences, then concatenate along the token dimension to form $X_s \in R^{N\times C}$. Subsequently, within the Cos-spatial transformer, we employ Cosine Spatial Attention(CSA) mechanism to fuse and interconnect spatial dependency within the multi-stage feature maps, simultaneously enhancing local information representation while extracting global context information. Furthermore, within the channel transformer, we utilize the Efficient Channel Attention (ECA) mechanism to explore the intrinsic relationships among these multi-stage feature maps in the channel dimension.

Self-attention mechanism has proven to be effective in compensating for the global information overlooked by CNN networks, but its persistent bottleneck in quadratic computational complexity. Considering the extensive lengths of multi-stage feature sequences, in order to strike an ideal balance between computational efficiency and performance, we adopt the Cosine Spatial Attention mechanism (CSA) to comprehensively capture global spatial context and local details for inputs with linear computational complexity.

As shown in Fig. 4(b), the input sequence X_s is first mapped to $Q \in R^{N\times d}$, $K \in R^{N\times d}$,$V \in R^{N\times d}$ respectively, where d represents the embedding dimension in transformer. To avoid the aggregation of negatively correlated contextual information, the CSA mechanism employs the ReLU function as the linear mapping kernel function to enforce non-negative attributes. By applying the associative property, the left multiplication of Q and K can be transformed

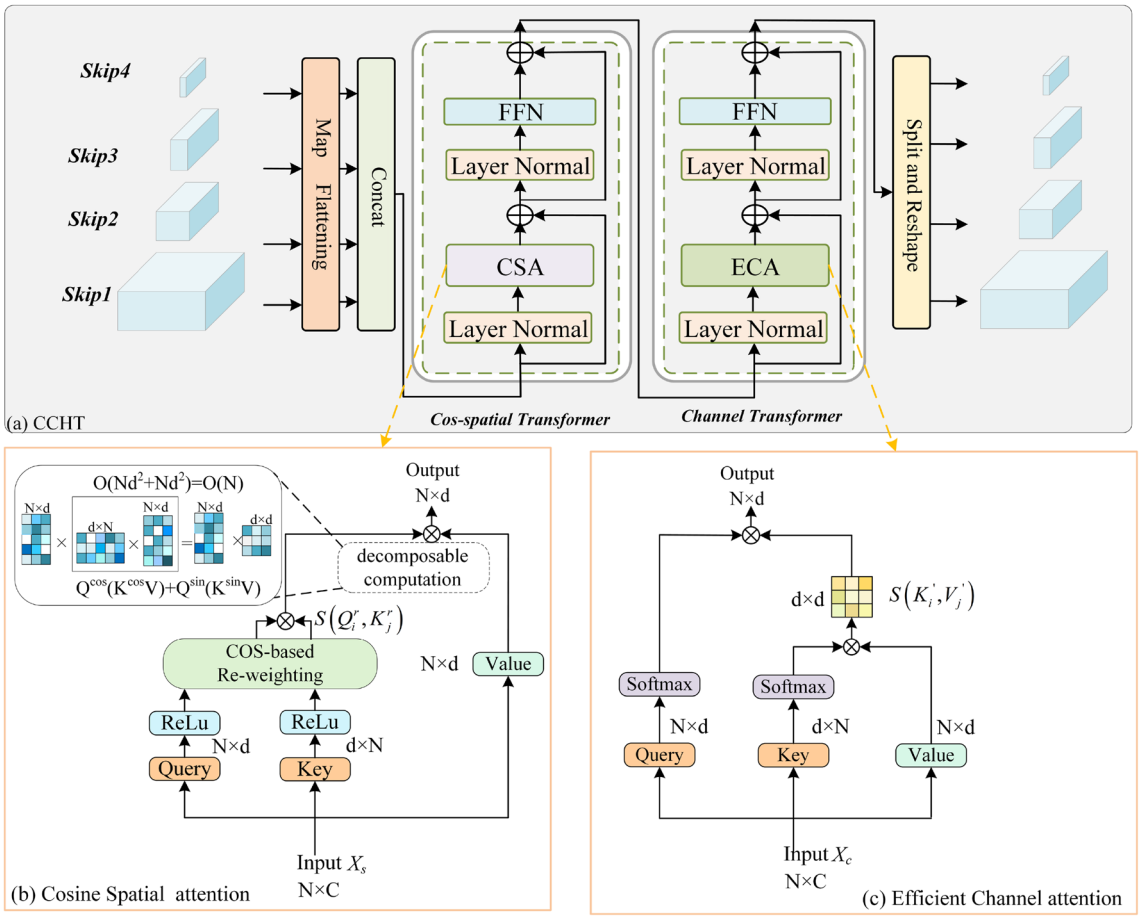


Fig. 4 The illustration of the Cos-spatial and Channel Hybrid Transformer (CCHT) module. (a) The architecture of CCHT module. (b) The Cosine Spatial Attention mechanism is employed to extract global spatial context information and reduce the quadratic computational complexity of self-attention mechanism by the Cosine decomposition theorem. (c) The Efficient Channel Attention mechanism is applied to extract the dependency between any two channels with linear computational complexity

into the right multiplication of K and V to reduce the computational complexity from $O(N^2)$ to $O(N)$, which can be formulated as:

$$O_i = \frac{\sum_j (ReLU(Q_i)ReLU(K_j))^T V_j}{\sum_j ReLU(Q_i)ReLU(K_j)^T} = \frac{ReLU(Q_i)\sum_j (ReLU(K_j))^T V_j}{ReLU(Q_i)\sum_j ReLU(K_j)^T} \tag{1}$$

Furthermore, the cosine-based re-weighting mechanism is introduced to focus on the distribution of the attention matrix, thereby expediting model convergence and enhancing training stability. The formula is as follows:

$$S(Q_i^r, K_j^r) = Q_i^r K_j^{rT} \cos(\frac{\pi}{2} \times \frac{i-j}{N}) \tag{2}$$

Where Q_i^r represents $ReLU(Q_i)$,K_j^r represents $ReLU(K_j)$, and $S(Q_i^r, K_j^r)$ is the obtained spatial attention map. N refers to the length of the sequence, and $i, j \in (1, N)$denote the position of each token. A smaller value of $i - j$ means a closer proximity between tokens, leading to an increase in attention weights. Conversely, the opposite holds true. Therefore, it effectively penalizes the weights associated with tokens that are more distantly positioned while enhancing the significance of local contextual information within shorter token distances.

According to the Ptolemy's theorem, the equation above can be decomposed into:

$$Q_i^r K_j^{rT} \cos(\frac{\pi}{2} \times \frac{i-j}{N}) = (Q_i^r \cos(\frac{\pi i}{2N}))(K_j^r \cos(\frac{\pi j}{2N}))^T + (Q_i^r \sin(\frac{\pi i}{2N}))(K_j^r \sin(\frac{\pi j}{2N}))^T \tag{3}$$

From Eq. 5, the CSA mechanism improves the non-decomposable non-linear softmax operation in self-attention into an efficient decomposable linear operation with reweighting mechanism:

$$\begin{cases} Q_i^{cos} = (Q_i^r \cos(\frac{\pi i}{2N})), K_j^{cos} = (K_j^r \cos(\frac{\pi j}{2N})) \\ Q_i^{sin} = (Q_i^r \sin(\frac{\pi i}{2N})), K_j^{sin} = (K_j^r \sin(\frac{\pi j}{2N})) \end{cases} \tag{4}$$

$$CRA = Q^{cos}(K^{cos}V) + Q^{sin}(K^{sin}V) \tag{5}$$

After obtaining spatial information from multi-stage features, it is essential to further explore the inter-channel relationships. We employ the Efficient Channel Attention mechanism (ECA), which also possesses linear computational complexity, to extract the dependency between any two channels within the multi-stage features. As shown in Fig. 4 (c), the input sequences $X_c \in R^{N\times C}$ is mapped into $Q \in R^{N\times d}$, $K \in R^{N\times d}$ and $V \in R^{N\times d}$ at first, and then the channel attention matric $S(K_i^{'}, V_j^{'})$ can be acquired by aggregating the K^T and V. This expression can be formulated as:

$$S(K_i^{'}, V_j^{'}) = softmax(\frac{K}{\sqrt{N}})^T \times V \tag{6}$$

From the above equation, it can be observed that the vector $k_i^T \in R^{1\times N}$ of the i th channel in K^T build relationship with the other channels by weighting with $v_j \in R^{N\times 1}$. Finally, the channel attention output, which is obtained by applying the channel attention matric $S(K_i^{'}, V_j^{'})$ to weight Q, can be formally defined as:

$$CHA = softmax(\frac{Q}{\sqrt{N}}) \times S(K_i^{'}, V_j^{'}) \tag{7}$$

After the concatenated sequences pass through the Cos-spatial transformer and Channel transformer, we then split them according to the corresponding length of each layer's input from the encoder and transform them into 2D feature maps to input the decoder. The CCHT module ingeniously integrates skip connections from different encoder levels, facilitating a seamless fusion of multi-size information, thus effectively mitigating the issue of semantic inconsistency in information transmission across codecs. Furthermore, the CCHT module introduces an innovative approach to skip connections, uniquely emphasizing the combination of spatial and channel factors, resulting in more comprehensive and robust feature representations.

3.3 Multi-scale Feature Adaptive-extraction module

To further facilitate the multi-scale context information transmission and improve the receptive fields, the MFA (Multi-scale Feature Adaptive-extraction) module is designed to cascade the bottom of the encoder and decoder. As illustrated in Fig. 5, the proposed MFA module consists of a multi-branch parallel atrous convolution structure with the AFA (Adjacent-scale Feature Adaptive-fusion) module. Firstly, We perform parallel atrous convolutions on the input feature maps obtained from the last layer of the encoder, using dilation rates of 1, 2, and 4, respectively. To minimize the introduction of additional parameters, we employ a shared-weight strategy for the convolutional kernels across these three branches. After that, the feature maps derived from adjacent branches following atrous convolution are subjected to scale information interaction through the AFA module. In the AFA module, it can be seen that the input feature maps f_1, f_2 are concatenated firstly, and then the concatenated features f_{12} are reflected into scale attention map through 1×1 convolution and softmax operation from two different branches.

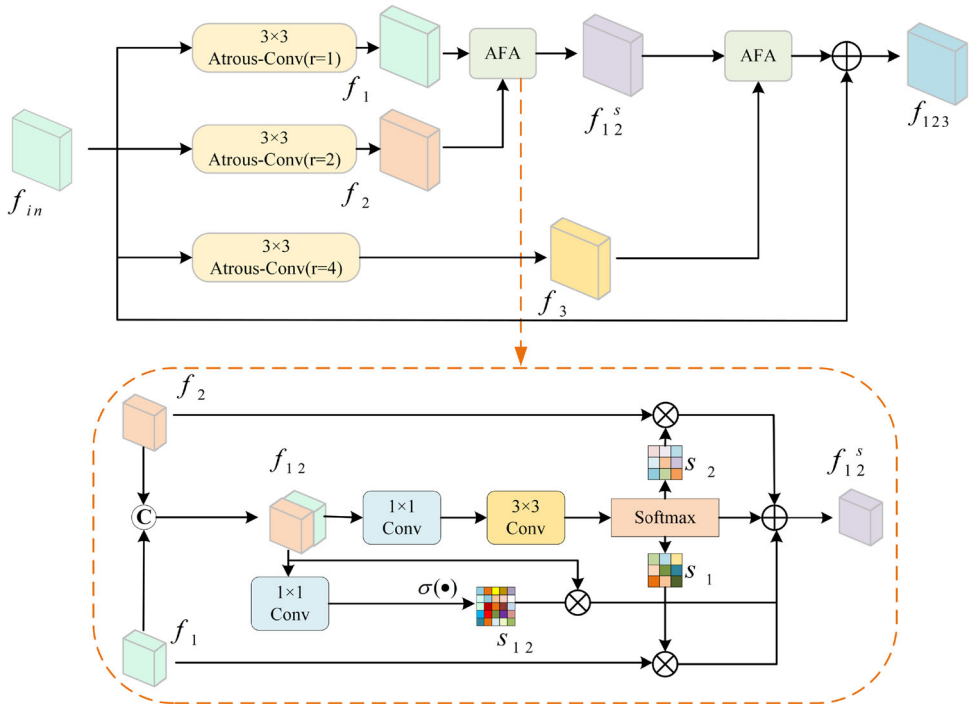


Fig. 5 The architecture of Multi-scale Feature Adaptive-extraction module

For one branch, the concatenated feature f_{12} is compressed into a scale attention martic S_{12} using 1×1 convolution, which is subsequently rescaled to the $(0, 1)$ range by applying a sigmoid operation. For the other branch, f_{12} is also mapped as two scale attention matrices(s_1, s_2) after convolution and softmax operations. The final fused map f_{12}^s, spanning adjacent scales, is derived as the weighted sum of f_1, f_2 and f_{12}, each multiplied by their corresponding scale attention matrices:

$$f_{12}^s = s_1 \otimes f_{12} + s_1 \otimes f_1 + s_2 \otimes f_2 \tag{8}$$

The f_{12}^s is further fused with f_3 flowing the similar approach described above. Finally, the aggregated muti-scale maps f_{123} can be represented as:

$$f_{123} = CBR(f_{in} + f_{23}^s) \tag{9}$$

Where $CBR(\cdot)$ denotes 1×1 convolution,batch normal and ReLU activation, f_{in} denotes the input feature maps, f_{23}^s is the fused map of the f_{12}^s and f_3. Thus, multi-scale information can be adaptively extracted by the MFA module, aiding the network in flexibly perceiving scale variations.

3.4 Loss function

In this paper, the hybrid loss consists of binary cross-entropy loss and dice loss is employed to optimize network training. Binary cross-entropy loss is commonly used for binary classification tasks and dice loss is effective for the class imbalanced problems in segmentation tasks, which can be formulated as:

$$\begin{cases} L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_l + (1-y_i)\log 1 - \hat{y}_l\right] \\ L_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_l}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_l} \end{cases} \tag{10}$$

where $\hat{y}_l$ represents the predicted value of the network, y_i is the value of corresponding ground truth, and N is the number of pixels. Therefore, the hybrid loss can be defined as:

$$L_{seg} = L_{bce} + L_{dice} \tag{11}$$

To fully optimal training of parameters across different network depths and enhance overall robustness, the deep supervision strategy is adopted to refine the hybrid loss. Finally, the total loss function can be defined as:

$$L_{total} = \sum_{i=1}^{4} \lambda_i L_{seg}^i \tag{12}$$

Where $i = (1, 2, 3, 4)$ represents the ith layer prediction of MDHT-Net, the proportional coefficients λ_1 λ_2 λ_3 and λ_4

are set as 0.1, 0.3, 0.6 and 1 respectively according to the impact of each layer.

4 Experimental results

4.1 Datasets

The performance of the proposed segmentation network MDHT-Net is validated with the NIH and MSD pancreas datasets.

NIH_Pancreas: The NIH_Pancreas dataset consists of 82 abdominal contrast-enhanced CT scans from the National Institutes of Health (NIH) Clinical Center. The sizes of the scans vary from $512 \times 512 \times 181$ to $512 \times 512 \times 466$ with a thickness between 1.5-2.5 mm.

MSD_Pancreas: The MSD_Pancreas dataset consists of 281 abdominal contrast-enhanced CT scans with labeled pancreas and pancreatic tumors from the Medical Segmentation Decathlon (MSD) challenge. The sizes of the scans vary from $512 \times 512 \times 37$ to $512 \times 512 \times 751$ with a thickness of 2.5 mm. Following previous studies on the MSD dataset, we treat the pancreas and pancreatic tumor as a whole for segmentation.

A fourfold cross-validation approach is employed for both datasets. We divided the original data into four pieces (NIH: #1-#20, #21-#40, #41-#61, and #62-#82; MSD: #1-#70, #71-#140, #141-#210, and #211-#281). We chose one piece as a test set each time and the remaining three pieces as a training set, repeating four times and averaging.

4.2 Implementation details

The proposed network is implemented on the PyTorch framework with one NVIDIA GeForce RTX 3090 graphics card of 24GB memory. During data preprocessing, all CT images intensity values are truncated into range [-100,240] HU firstly, and then normalized to be the range of [0,1]. Different data augmentation techniques including random rotations (90°, 270°), random flipping, and random scaling are adopted to reduce overfitting.

In the training phase, the input images are resized to 128×128 for both NIH and MSD datasets. Stochastic gradient descent (SGD) is applied as the optimizer with an initial learning rate $1 \times$ 1e-4 and the momentum 0.9. The batch size is set to 8 and the proposed network MDHT-Net is trained for 40 epochs.

4.3 Evaluation metrics

To evaluate the performance of the proposed MDHT-Net, we quantitatively analyze the segmentation results with 5 evaluation metrics, including Dice Similarity Coefficient (DSC), Sensitivity, Specificity, Average Symmetric Surface Distance (ASD) and Hausdorff Distance (HD). These metrics can be expressed as:

$$DSC = \frac{2 \times |f_{pre} \cap f_{GT}|}{(|f_{pre}| + |f_{GT}|)} \tag{13}$$

Where f_{pre} represents the prediction result of the model and f_{GT} denotes the real mask information. DSC is one of the most common indexes to evaluate the effectiveness of image segmentation method. It calculates the spatial overlap between the segmentation and ground truth. The closer its value is to 1, the more similar it is.

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

Where True Positive (TP) indicates that the prediction is positive and the ground truth is positive. False Positive (FP) indicates that the prediction is positive but the ground truth is negative. False Negative (FN) represents prediction is negative, but ground truth is positive. Sensitivity is used to measure the proportion of positive samples correctly recognized by the model out of the total number of actual positives. Specificity is used to measure the proportion of negative samples correctly recognized by the model and the total number of actual negatives. The higher the sensitivity and specificity, the lower the model's recognition error rates for true foreground and true background, respectively.

$$HD = max\left\{ \underset{x \in f_{GT}}{max} \underset{y \in f_{pre}}{\min} d\{x, y\}, \underset{x \in f_{GT}}{max} \underset{y \in f_{pre}}{\min} d\{y, x\} \right\} \tag{16}$$

Where x and y denote the voxels of the ground truth and the prediction results, respectively, and $d\{x, y\}$ represents the Euclidean distance between x and y. The Hausdorff distance is used to assess whether the edge of the pancreas is completely segmented. The smaller the Hausdorff value is, the more complete the pancreatic margin segmentation.

$$ASD = \frac{1}{2}\left\{ \underset{x \in f_{GT}}{mean} \underset{y \in f_{pre}}{\min} d\{x, y\}, \underset{x \in f_{GT}}{mean} \underset{y \in f_{pre}}{\min} d\{y, x\} \right\} \tag{17}$$

The average symmetric surface distance is used to evaluate the accuracy of edge segmentation. The smaller the ASD value is, the more accurate the pancreatic margin segmentation.

4.4 Segmentation results on NIH dataset

4.4.1 Comparison to the state-of-the-art

The effectiveness of MDHT-Net is first evaluated using NIH pancreas segmentation dataset. From Table 2, it can be seen that the proposed MDHT-Net achieves state-of-the-art performance in most of the indicators compared to previous advanced pancreas segmentation methods. It is worth noting that the mean dice coefficient of MDHT-Net is 91.07 ± 1.19%, exhibiting a statistically significant 3.18% improvement in segmentation performance compared to the second-ranked results [35]. In terms of Sensitivity and Specificity, the results from MDHT-Net also increased by 0.5% and 0.83% respectively compared with the second-best method [38], resulting in a decreased error rate for pancreas segmentation. Simultaneously, the minimal Floating Point Operations (FLOPs) metric implies that the model has a lower computational complexity. The MDHT-Net not only achieves a reduction of approximately 4.49G Flops compared to suboptimal results [37] but also remarkably outperforms it in terms of the DSC metric. It demonstrates that the MDHT-Net effectively strikes an optimal balance between computational cost and performance improvement. What's more, the MDHT-Net saves over half of the testing time compared to the majority of methods during the segmentation process. The shortest testing time implies that MDHT holds promising prospects for clinical application.

4.4.2 Comparison to the mainstream segmentation models

In order to further confirm the superiority of the proposed MDHT-Net, we also compare our segmentation network with the mainstream segmentation models, including MedT, PVT, SwinUNet, TransUNet, and UCTransNet. For the sake of fairness, all these models are executed within the same experimental environment. As for the DSC shown in Fig. 6, the proposed MDHT-Net not only outperforms all other models but also presents a more concentrated data distribution.

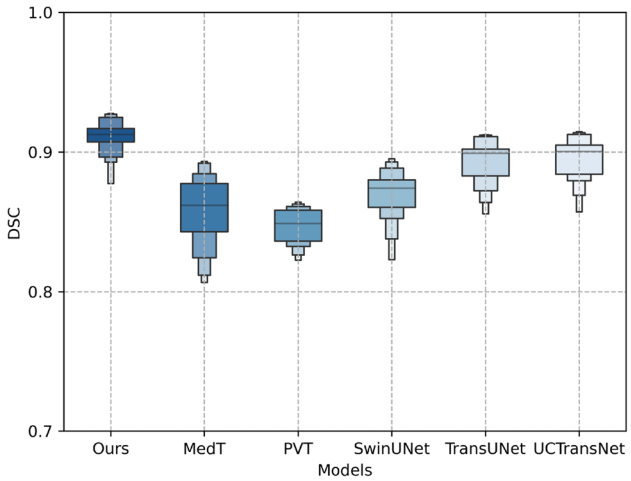


Fig. 6 Boxplots of DSC for different methods on the NIH dataset

Table 3 shows the performance comparison using the distance-based indicators of ASD and HD, it is clear that the proposed MDHT-Net obtains superior performance with the smallest values for ASD and HD, which is consistent with the presentation of the DSC index.

Figure 7 presents the visual comparison of different mainstream medical image segmentation models on the NIH dataset, intuitively reflects that our proposed MDHT-Net effectively mitigates over-segmentation and under-

Table 2 The results (measured by the DSC, Sensitivity,Specificity,Flops, and Testing time) of pancreas segmentation on NIH datasets

Method	Year	DSC(%)↑	Sensitivity(%)↑	Specificity(%)↑	Flops(G)↓	Testing time↓
Zheng et al. [33]	2020	84.35±7.69	86.23±6.35	85.01±6.04	66.25	7-8min
Li et al. [19]	2020	85.68±3.21	84.37±7.45	89.44±6.31	37.32	5-6min
Li et al. [34]	2021	85.33±4.11	82.74±8.20	89.62±7.25	55.62	7-8min
Li et al. [18]	2021	86.30±4.52	84.93±5.15	86.41±5.30	32.21	5-6min
Huang et al. [20]	2021	87.23±6.70	89.95±7.53	90.25±7.28	86.45	10-11min
Chen et al. [46]	2022	85.29±4.75	84.55±8.29	89.03±7.01	30.27	5-6min
Chen et al. [35]	2022	87.89±2.45	85.72±4.41	87.71±4.02	53.24	7-8min
Qiu et al. [37]	2023	86.28±5.01	86.95±5.02	88.56±5.35	29.34	**1-2min**
Xia etal. [38]	2023	87.02±3.26	91.08±4.85	90.99±3.79	36.27	5-6min
Yao et al. [39]	2023	87.85±2.74	88.32±3.61	90.53±4.06	59.83	7-8min
Ours	2024	**91.07 ± 1.19**	**91.58 ± 1.31**	**91.82 ± 1.24**	**24.85**	**1-2min**

↑ means the higher the better and ↓ represents the opposite. Optimal results (described by mean ± std) are shown in bold

Table 3 The results (measured by the ASD and HD) of pancreas segmentation on the NIH dataset

Method	ASD(mm)↓	HD(mm)↓
MedT [51]	0.88±0.14	2.20±0.10
PVT [49]	1.01±0.11	2.30±0.09
SwinUNet [50]	0.81±0.11	2.16±0.10
TransUNet [13]	0.66±0.10	2.00±0.09
UCTransNet [52]	0.64±0.08	1.99±0.09
Ours	**0.55 ± 0.08**	**1.93 ± 0.08**

↓ means the lower the better. Optimal results (described by mean ± std) are shown in bold

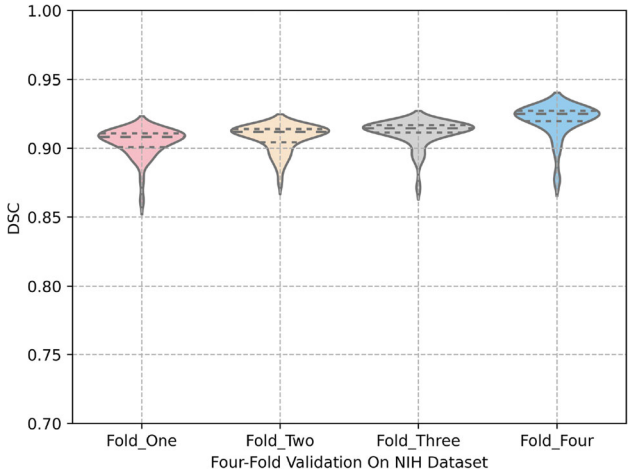


Fig. 8 Violinplots of DSC for four-fold cross-validation on NIH dataset

segmentation issues in the segmentation results, which proves that the MDHT-Net can accurately extract the global context information through CCHT module. Moreover, MDHT-Net's advanced capability to interact and fuse multi-scale feature information further enhances the model's accuracy, resulting in a more complete segmentation of the pancreatic structure compared to other segmentation networks.

4.4.3 Visualization of results

From the DSC distribution of four-fold cross-validation on the NIH dataset in Fig. 8, it can be seen that across different folds are remarkably consistent, demonstrating the high robustness of the MDHT-Net network in mitigating the impact of sample variations.

From the visualization of some segmentation results on the NIH dataset in Fig. 9, it is evident that our method maintains a high level of concordance with the manually labeled ground truth annotations, demonstrating its remarkable segmentation accuracy. Notably, despite the variances in pancreas shape and spatial distribution across the three cases (case 39#, case 56#, case 72#) within the CT images, our method consistently and precisely delineates the pancreas.

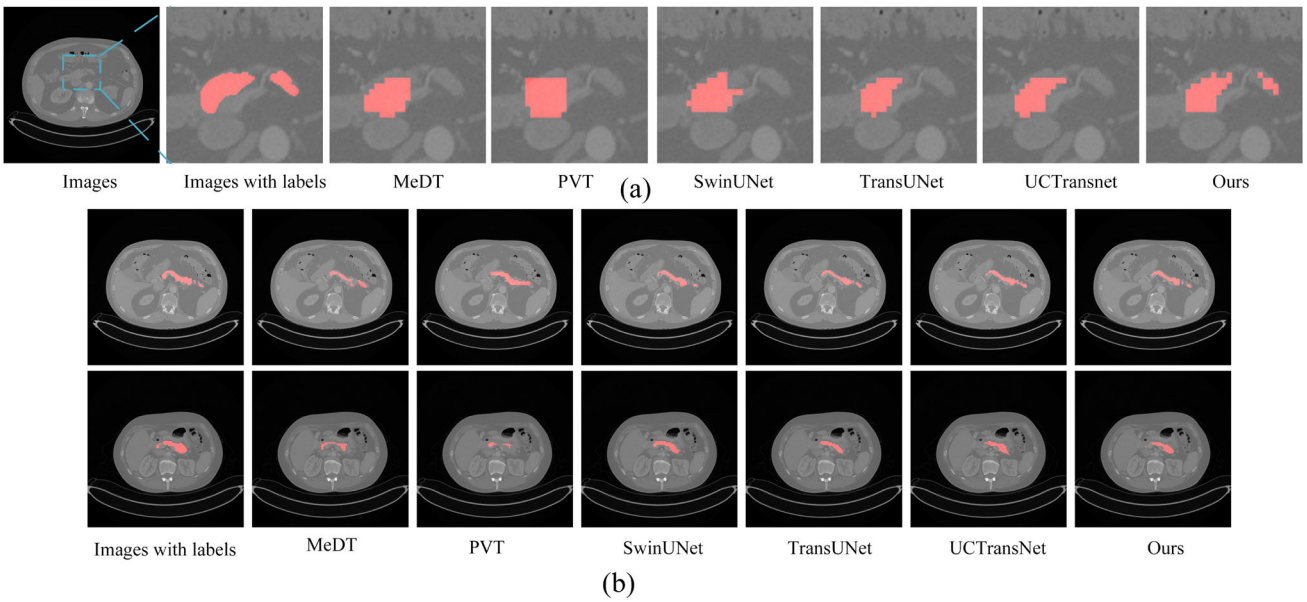


Fig. 7 The visual comparison of different mainstream medical image segmentation models on the NIH dataset. The red regions represent the GT and the predicted results by different models. (a) Local clip-focused view of the pancreas region. (b) Global view of pancreas segmentation in CT slice

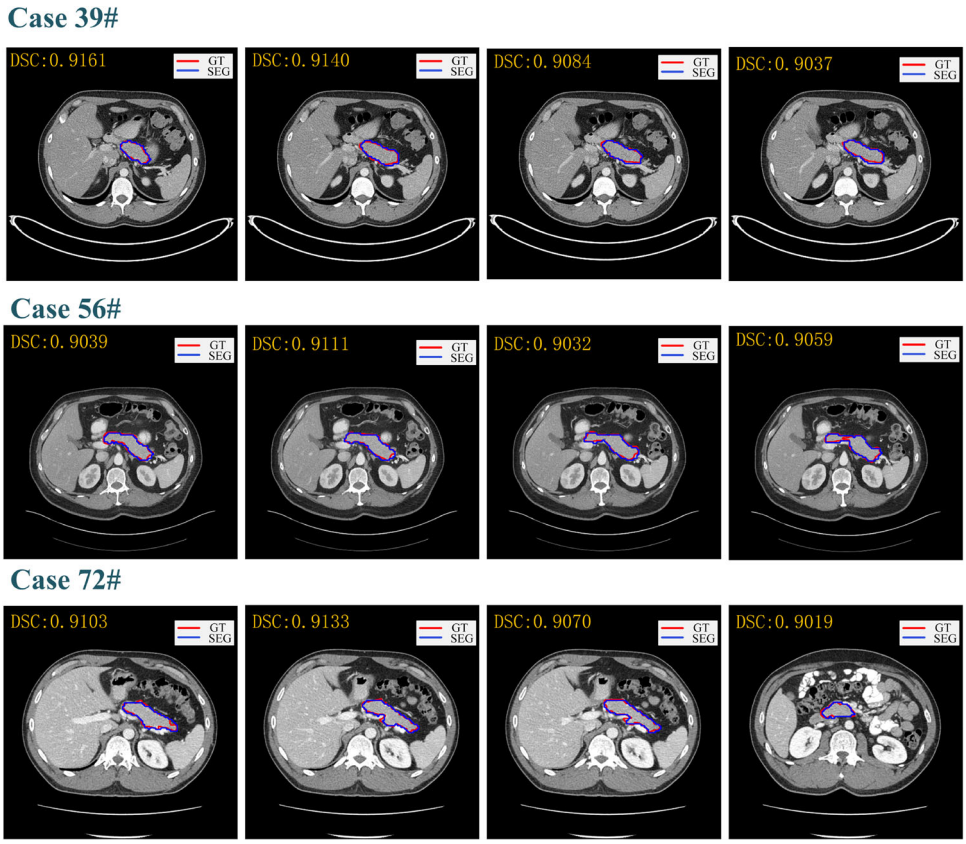


Fig. 9 The visualization of segmentation results for different cases in the NIH dataset. The red solid line denotes the ground truth, and the blue solid line denotes the prediction results

4.5 Segmentation results on MSD dataset

4.5.1 Comparison to the state-of-the-art

To further verify the effectiveness of MDHT-Net, four-fold cross-validation experiments, as well as a series of comparative experiments, are conducted on the public pancreas MSD dataset, maintaining the same experimental environment as NIH pancreatic segmentation. As shown in Table 4, the proposed MDHT-Net continues to demonstrate significant advantages across the four metrics when compared to previous state-of-the-art methods.The DSC shows a impressive enhancement of 3.02% compared to the second-ranked results [34]. Simultaneously, the highest Sensitivity and Specificity indicate a high degree of similarity between the results of our method and manual

Table 4 The results (measured by the DSC,Sensitivity, Specificity, and Testing time) of pancreas segmentation on MSD datasets

Method	Year	DSC(%)↑	Sensitivity(%)↑	Specificity(%)↑	Testing time↓
Li et al. [18]	2021	88.50±2.78	91.02±4.92	89.59±3.71	8-9 min
Chen et al. [46]	2022	76.55±8.30	69.34±11.82	70.53±10.76	8-9min
Chen et al. [35]	2022	86.76±4.56	82.1±8.71	80.11±8.05	8-9min
Qiu et al. [37]	2023	85.56±4.68	84.56±6.37	86.72±5.52	**3-4min**
Xia et al. [38]	2023	87.79±4.34	89.84±8.02	89.51±8.17	8-9min
Ours	2024	**91.52 ± 0.66**	**91.23 ± 1.23**	**92.15 ± 1.08**	**3-4min**

↑ means the higher the better and ↓ represents the opposite. Optimal results (described by mean ± std) are shown in bold

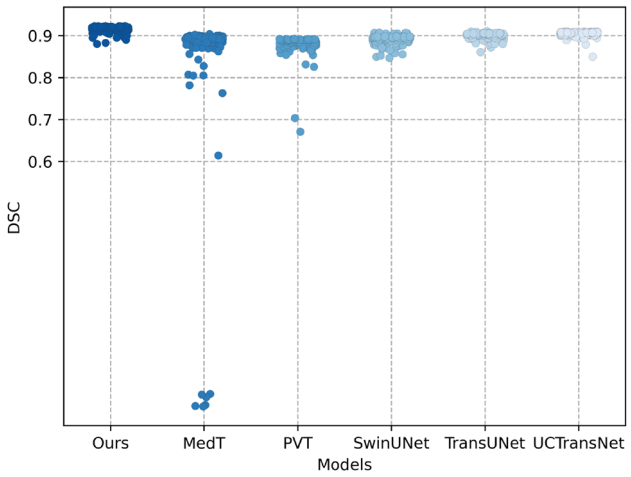


Fig. 10 Scatterplots of DSC for different methods on the MSD dataset

delineations. It demonstrates that MDHT-Net effectively reduces the probabilities of target foreground omission and background misclassification during the segmentation process. Additionally, occupying a shorter testing time compared with previous advanced segmentation methods on the MSD dataset further validates the potential applicability of MDHT-Net.

4.5.2 Comparison to the mainstream segmentation models

As depicted in Fig. 10, MDHT-Net still maintains a slight advantage over other mainstream segmentation models in terms of the DSC metric, obtaining consistently distributed results with no outliers, which highlights the model's robustness.

The comparison of distance-based indicators is provided in Table 5. The lowest values for ASD (0.52 ± 0.07 mm) and HD (1.89 ± 0.05 mm) substantiate the accuracy and completeness of MDHT-Net's edge segmentation performance.

Similarly, visual assessments of the segmentation results achieved by different models on the MSD dataset are presented in Fig. 11. Overall, the MDHT-Net is observed to retain pancreatic shape features that closely match the ground truth. Particularly, for the case of challenging structures within the pancreas, as exemplified in the third row in Fig. 11, MDHT-Net exhibited superior performance when compared to the slightly less favorable results obtained by TransUNet and UCTransNet.

4.5.3 Visualization of results

As shown in Fig. 12, the results of the four-fold cross-validation experiments on the MSD dataset consistently maintained mean DSC values above 90%, further substantiating the model's superiority and robustness. From Fig. 13, it is evident that MDHT-Net maintains outstanding segmentation performance, even in scenarios with significant structural variations within the same case. For instance, in the case of 16, although there is a topological disconnection in the fourth image, the overall structural integrity is still reasonably well-preserved in the segmentation results.

5 Discussion

5.1 Ablation study

To assess the individual impact of each component on the performance enhancement of the proposed MDHT-Net, the ablation experiments are conducted on the NIH dataset for analysis. Specifically, the Deformable U-shape network is chosen as the baseline, Baseline+CCHT indicates the addition of the CCHT module based on the Baseline, while Baseline+ MFA denotes the addition of the MFA module based on the Baseline.

It is obvious that the DSC score is Significantly improved with the addition of CCHT module, by 3.16% compared with the Baseline. This reflects that the CCHT module brings significant performance improvements by alleviating ambiguities between the codec, further validating its ability to extract global information based on the linear complexity transformer-architecture. As shown in Fig. 14, the overall contour features of Baseline+CCHT's visual results are obviously more consistent with Ground Truth compared to the Baseline. What's more, a notable 1.14% decrease in DSC variance compared to the Baseline has been achieved, demonstrating that the CCHT module markedly enhances network robustness and effectively promotes the interaction of multi-layer information.

With the integration of the MFA module into the baseline, the overall segmentation performance has been improved.

Table 5 The results (measured by the ASD and HD) of pancreas segmentation on the MSD dataset

Method	ASD(mm)↓	HD(mm)↓
MedT [51]	0.75±0.12	2.02±0.10
PVT [49]	0.82±0.10	2.10±0.06
SwinUNet [50]	0.78±0.07	2.05±0.08
TransUNet [13]	0.65±0.08	2.00±0.07
UCTransNet [52]	0.62±0.08	1.98±0.06
Ours	**0.52 ± 0.07**	**1.89 ± 0.05**

↓means the lower the better. Optimal results (described by mean ± std) are shown in bold

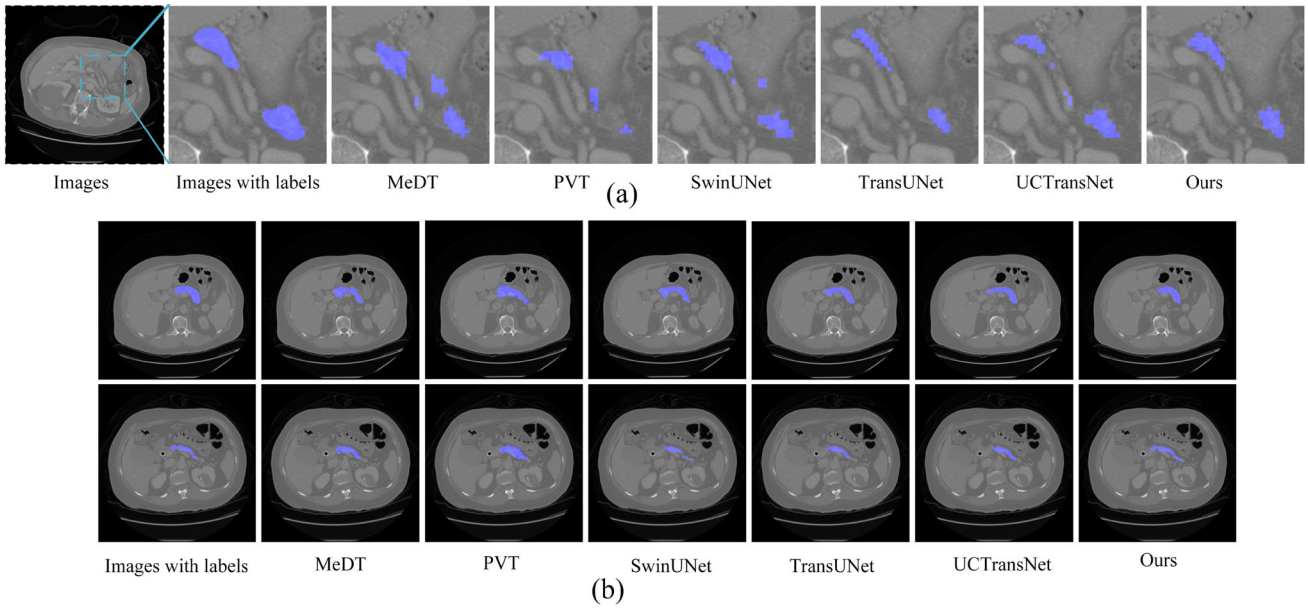


Fig. 11 The visual comparison of different mainstream medical image segmentation models on the MSD dataset. The blue regions represent the GT and the predicted results by different models. (a) Local clip-focused view of the pancreas region. (b) Global view of pancreas segmentation in CT slice

This demonstrates that, within the framework of the baseline utilizing direct concatenation-based skip connections, the MFA module can effectively leverage the scale variation regularities to dynamically optimize the network's receptive field, thereby extracting more accurate features of the pancreas.

In summary, combined with MFA and CCHT modules, the proposed MDHT-Net achieves the best performance across the five metrics. Simultaneously, as illustrated in Fig. 14, the proposed MDHT-Net's visual results are closest to the Ground Truth compared to other variations of the network (Table 6).

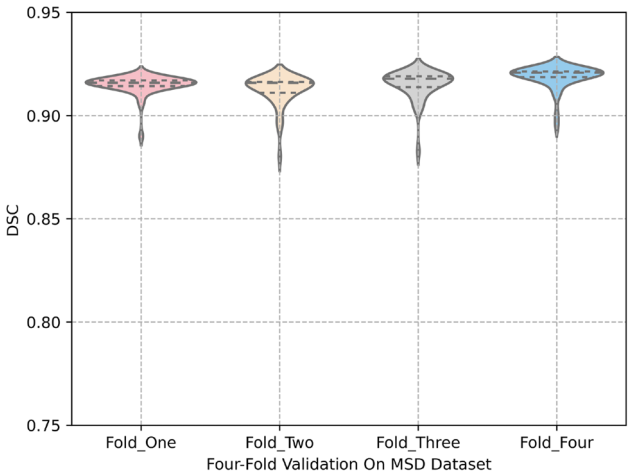


Fig. 12 Violinplots of DSC for four-fold cross-validation on MSD dataset

5.2 Generalization ablity

To verify the generalization performance of the proposed model, we conducted experiments on the public skin cancer dataset ISIC 2018. Following the official dataset partitioning, 2,594 images are used for training, with 100 images for validation and 1000 images for testing. Finally,the MDHT-NET achieved an excellent Dice coefficient of $91.54 \pm 1.02\%$ on the test set. The remarkable segmentation accuracy achieved in the task of skin cancer demonstrates that the pancreas segmentation network, MDHT-NET, can be successfully applied to other medical image segmentation tasks, showcasing its efficient generalization ability. We look forward to further exploring the potential of applying this network to a wider range of medical image segmentation tasks in future research endeavors.

Despite significant differences in data format, distribution, and target morphology compared to pancreas segmentation dataset, MDHT-NET still achieved superior segmentation performance. As illustrated in Fig. 15, the red lines represent the Ground Truth (GT) of skin cancer, while the blue lines represent the predictions. For melanomas of varying shapes and sizes in different cases, MDHT-NET's segmen-

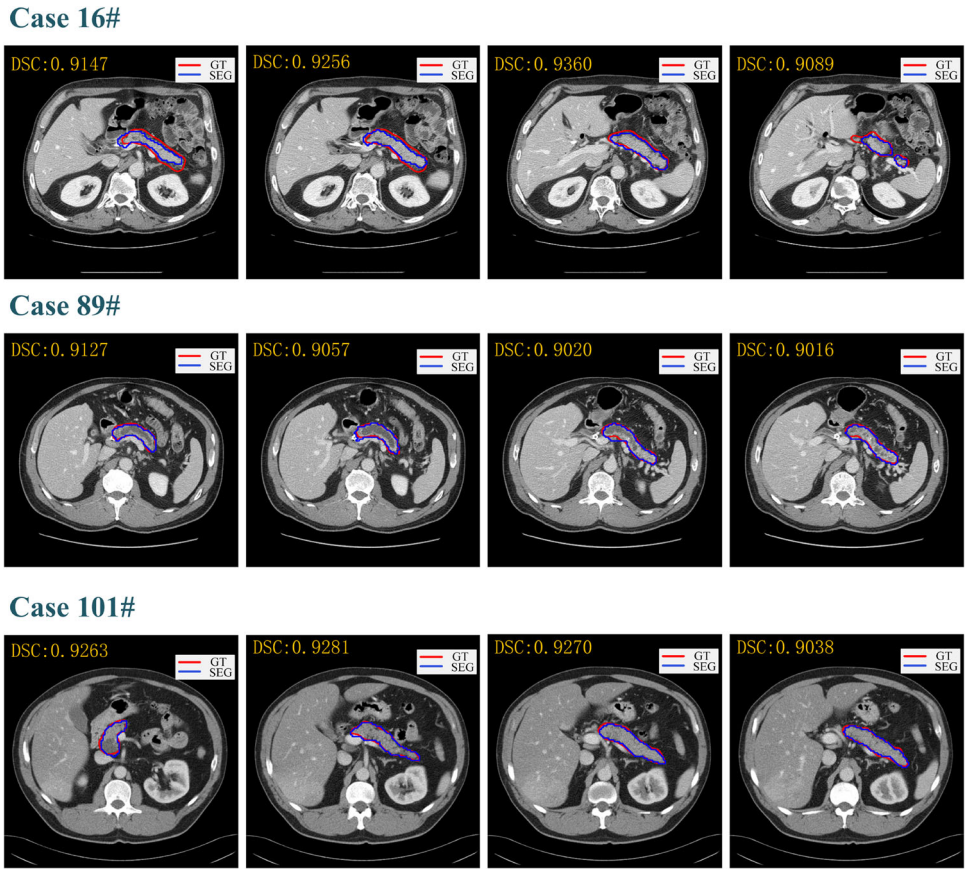


Fig. 13 The visualization of segmentation results for different cases in the MSD dataset. The red solid line denotes the ground truth, and the blue solid line denotes the prediction results

tation results closely match the GT. Even for samples with significant color and contour fluctuations, as shown in the first and fifth rows of Fig. 15, respectively, MDHT-NET can accurately capture their structural features. As depicted in Table 7, compared to the suboptimal pure transformer-based network(SwinUNet), the Dice coefficient and Sensitivity of MDHT-NET are increased by 0.82% and 0.92% respectively for skin cancer segmentation. It demonstrates MDHT-NET effectively combines local spatial positional information extracted by deformable convolution with global information obtained from various attention mechanisms.

5.3 Limitations and future work

Although our proposed method has achieved competitive performance, there are still some limitations.

On one hand, for some CT slices with blurred backgrounds, the pancreatic boundaries tend to adhere to surrounding soft tissues and organs, leading to mis-segmentation by the network. In Fig. 16(a), the arrow indicates the area where the target pancreas in the CT scan exhibits low contrast compared to the surrounding background. Consequently, the predicted pancreas shape shows deviations influenced by neighboring organs. In the future, we will employ the following strategies to enhance the segmentation quality of blurry edge samples: Introducing edge operators in the network to enhance edge features in the feature maps; Proposing a novel edge loss function to supervise the pancreas edge features.

On the other hand, the proposed MDHT-Net is a two-stage segmentation method, where the fine segmentation results of MDHT-NET rely to some extent on the accuracy of the coarse segmentation network in locating the pancreatic region. For some scattered small pancreatic segmentation targets, the network tends to miss these areas, resulting in under-segmentation, as shown in Fig. 16(b). To tackle this issue, we plan to design an enhancement module that integrates a coarse segmentation network and a fine segmentation network, fostering mutual learning between them. We aim to construct a novel coarse-to-fine segmentation framework that enables end-to-end training, mitigating the problem of missed segmentation by the coarse segmentation network for small, scattered pancreas targets.

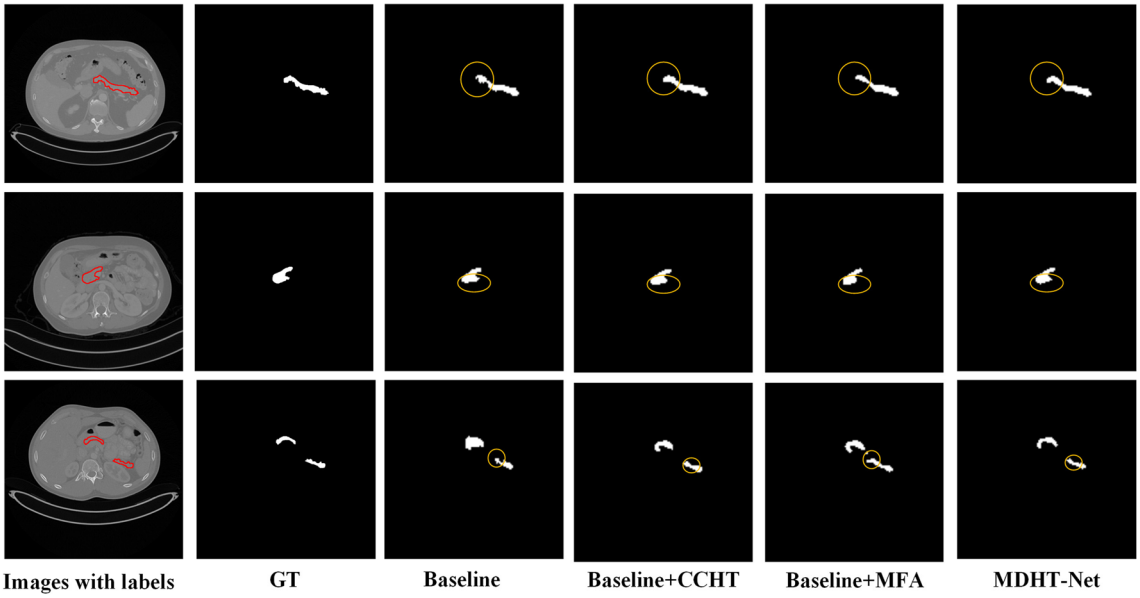


Fig. 14 The visualization of ablation experimental results on the NIH dataset. The leftmost column indicates that GT is outlined by solid red lines in the original image

Table 6 The ablation experimental results of the proposed network on the NIH dataset

Method	DSC(%)↑	Sensitivity(%)↑	Specificity(%)↑	ASD(mm)↓	HD(mm)↓
Baseline(Deformable U-Net)	87.66±2.37	88.30±2.24	89.35±2.01	0.79±0.15	2.13±0.13
Baseline+CCHT	90.82±1.23	90.70±1.53	91.64±1.50	0.57±0.08	1.94±0.08
Baseline+ MFA	89.18±1.89	90.12±1.82	91.05±1.77	0.68±0.12	2.02±0.13
MDHT-Net	**91.07 ± 1.19**	**91.58 ± 1.31**	**91.82 ± 1.24**	**0.55 ± 0.08**	**1.93 ± 0.08**

↑means the higher the better and ↓ represents the opposite. Optimal results (described by mean ± std) are shown in bold

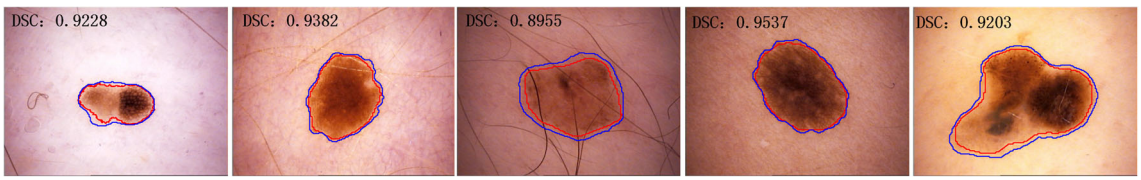
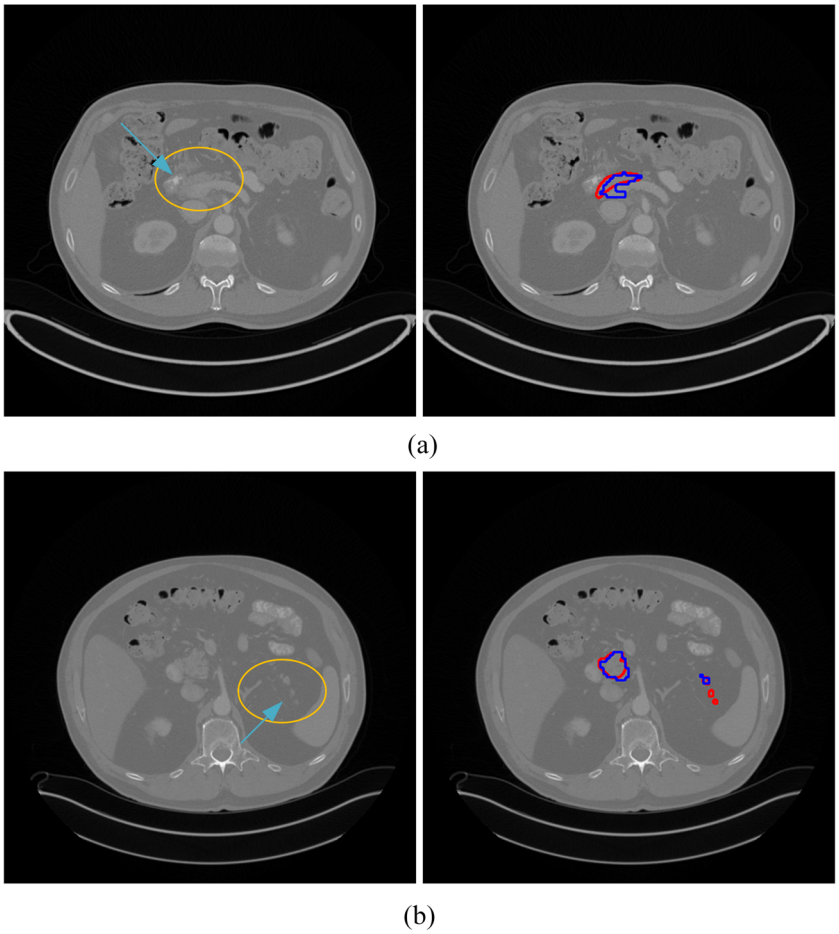


Fig. 15 The visualization of segmentation results on the ISIC 2018 dataset. The red line represents GT and the blue line represents prediction

Table 7 The results(measured by the DSC, Sensitivity, Specialty, ASD and HD) of skin cancer segmentation on the ISIC2018 dataset

Method	DSC(%)↑	Sensitivity(%)↑	Specificity(%)↑	ASD(mm)↓	HD(mm)↓
MedT	87.89±4.89	88.23±4.76	96.35±5.03	15.25±7.65	31.35±10.52
PVT	87.62±4.53	88.16±4.82	96.62±4.79	17.31±7.09	32.46±10.01
SwinUNet	90.72±3.65	92.12±3.58	96.80±3.93	10.82±5.31	28.55±6.88
TransUNet	89.42±3.01	90.01±4.15	95.96±4.01	13.15±5.05	31.02±6.72
UCTransNet	89.53±2.45	90.07±2.09	**96.81±3.13**	12.87±4.91	30.14±4.99
Ours	**91.54 ± 1.02**	**93.04 ± 1.16**	**96.81 ± 1.47**	**10.05 ± 2.78**	**25.30 ± 3.05**

↑ means the higher the better and ↓ represents the opposite. Optimal results (described by mean ± std) are shown in bold



(a)

(b)

Fig. 16 The visualization of segmentation failure cases on the NIH dataset. The leftmost column shows the original CT images.The red line represents GT and the blue line represents prediction in the second column. (a) An segmentation case of pancreatic blurred boundaries. (b) An segmentation case of dispersed pancreatic small regions

6 Conclusion

In this paper, a novel pancreas segmentation network is proposed, namely MDHT-Net, which skillfully integrates the spatial attention mechanism, channel attention mechanism and scale attention mechanism to fully extract context information and promote multi-layer information interaction. The dual deformable convolution blocks are utilized in the third and fourth layers of MDHT-Net, flexibly capturing the changeable features of the pancreas while also avoiding the dependence on computing resources. The Cos-spatial and Channel Hybrid Transformer is introduced to establish long-term dependency by improving the self-attention mechanism, and comprehensively extracting global features from both spatial and channel dimensions. Simultaneously, the Multi-scale Feature Adaptive-extraction module is designed to optimize the receptive field of the network and adaptively extract the multi-scale information. Extensive experiments on the NIH dataset and MSD dataset are performed, which demonstrate that our proposed MDHT-Net can achieve state-of-the-art performance when compared to previous advanced models.

Acknowledgements This work was supported in part by the Science and Technology Commission of Shanghai Municipality (20DZ2254400, 20DZ2261200), National Scientific Foundation of China (82170110), Fujian Province Department of Science and Technology (2022D014), Shanghai Municipal Science and Technology Major Project (ZD2021C Y001) and Shanghai Municipal Key Clinical Specialty (shslczdzk02201).

Availability of data The data is available from http://medicaldecathlon.com/.

Declarations

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.

Ethics approval This work does not involve experimental procedures with human subjects or animals.

References

1. Ilic M, Ilic I (2016) Epidemiology of pancreatic cancer. World J Gastroenterol 22(44):9694

2. Pelosi E, Castelli G, Testa U (2017) Pancreatic cancer: molecular characterization, clonal evolution and cancer stem cells. Biomedicines 5(4):65
3. Wang X, Wu X, Zhang Z et al (2018) Monensin inhibits cell proliferation and tumor growth of chemo-resistant pancreatic cancer cells by targeting the egfr signaling pathway. Sci Rep 8(1):17914
4. Ju J, Li J, Chang Z et al (2023) Incorporating multi-stage spatial visual cues and active localization offset for pancreas segmentation. Pattern Recognit Lett 170:85–92
5. Paithane PM, Kakarwal S (2022) Automatic pancreas segmentation using a novel modified semantic deep learning bottom-up approach. Int J Intell Syst Appl Eng 10(1):98–104
6. Dai S, Zhu Y, Jiang X et al (2023) Td-net: Trans-deformer network for automatic pancreas segmentation. Neurocomputing 517:279–293
7. Ma J, Lin F, Wesarg S et al (2018) A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, Springer, pp 480–487
8. Roth HR, Lu L, Lay N et al (2018) Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med Image Anal 45:94–107
9. Yu Q, Xie L, Wang Y et al (2018) Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8280–8289
10. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical image computing and computer. Springer, pp 234–241
11. Alom M, Hasan M, Yakopcic C et al (2018) Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv:1802.06955 10
12. Zhou Z, Siddiquee MMR, Tajbakhsh N et al (2019) Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39(6):1856–1867
13. Chen J, Lu Y, Yu Q et al (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306
14. Ruan J, Xiang S, Xie M et al (2022) Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp 1150–1156
15. Zhang X, Cao X, Wang J et al (2023) G-unext: a lightweight mlp-based network for reducing semantic gap in medical image segmentation. Multimed Syst 29(6):3431–3446
16. Oktay O, Schlemper J, Folgoc LL et al (2018) Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999
17. Li F, Li W, Shu Y et al (2020) Multiscale receptive field based on residual network for pancreas segmentation in ct images. Biomed Signal Process Control 57:101828
18. Li W, Qin S, Li F et al (2021) Mad-unet: A deep u-shaped network combined with an attention mechanism for pancreas segmentation in ct images. Med Phys 48(1):329–341
19. Li H, Li J, Lin X et al (2020) A model-driven stack-based fully convolutional network for pancreas segmentation. 2020 5th International Conference on Communication. Image and Signal Processing (CCISP), IEEE, pp 288–293
20. Huang M, Huang C, Yuan J et al (2021) A semiautomated deep learning approach for pancreas segmentation. J Healthc Eng 2021
21. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Advances in neural information processing systems 30
22. Dosovitskiy A, Beyer L, Kolesnikov A et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929
23. Chen L, Wan L (2023) Ctunet: automatic pancreas segmentation using a channel-wise transformer and 3d u-net. Vis Comput 39(11):5229–5243
24. Fei C, Luo J (2022) Dtunet: A transformer-based unet combined with denseaspp module for pancreas segmentation. 2022 15th International Congress on Image and Signal Processing. BioMedical Engineering and Informatics (CISP-BMEI), IEEE, pp 1–7
25. Qu T, Li X, Wang X et al (2023) Transformer guided progressive fusion network for 3d pancreas and pancreatic mass segmentation. Med Image Anal 86:102801
26. Wang Z (2016) A new approach for segmentation and quantification of cells or nanoparticles. IEEE Trans Ind Inform 12(3):962–971
27. van Donkelaar S, Daamen L, Andel P et al (2022) Superpixel-based context restoration for self-supervised pancreas segmentation from ct scans. In: 34rd Benelux conference on artificial intelligence and the 31th Belgian Dutch conference on machine learning
28. Karasawa K, Oda M, Kitasaka T et al (2017) Multi-atlas pancreas segmentation: atlas selection based on vessel structure. Med Image Anal 39:18–28
29. Tam TD, Binh NT (2015) Efficient pancreas segmentation in computed tomography based on region-growing. In: Nature of computation and communication: international conference, ICTCC 2014, Ho Chi Minh City, Vietnam, November 24-25, 2014, Revised Selected Papers 1, Springer, pp 332–340
30. Hammon M, Cavallaro A, Erdt M et al (2013) Model-based pancreas segmentation in portal venous phase contrast-enhanced ct images. J Digit Imaging 26:1082–1090
31. Azad R, Bozorgpour A, Asadi-Aghbolaghi M et al (2021) Deep frequency re-calibration u-net for medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3274–3283
32. Ma H, Zou Y, Liu PX (2021) Mhsu-net: A more versatile neural network for medical image segmentation. Comput Methods Prog Biomed 208:106230
33. Zheng H, Chen Y, Yue X et al (2020) Deep pancreas segmentation with uncertain regions of shadowed sets. Magn Reson Imaging 68:45–52
34. Li J, Lin X, Che H et al (2021) Pancreas segmentation with probabilistic map guided bi-directional recurrent unet. Phys Med Biol 66(11):115010
35. Chen Y, Xu C, Ding W et al (2022) Target-aware u-net with fuzzy skip connections for refined pancreas segmentation. Appl Soft Comput 131:109818
36. Dogan RO, Dogan H, Bayrak C et al (2021) A two-phase approach using mask r-cnn and 3d u-net for high-accuracy automatic segmentation of pancreas in ct imaging. Comput Methods Prog Biomed 207:106141
37. Qiu C, Song Y, Liu Z et al (2023) Cmfcunet: cascaded multi-scale feature calibration unet for pancreas segmentation. Multimed Syst 29(2):871–886
38. Xia F, Peng Y, Wang J et al (2023) Mtr-net: A multipath fusion network based on 2.5 d for medical image segmentation. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp 2896–2903
39. Yao X, Qiu C, Song Y et al (2023) Pancreas segmentation optimization based on coarse-to-fine scheme. Intell Autom Soft Comput 37(3)
40. Gao C, Ye H, Cao F et al (2021) Multiscale fused network with additive channel-spatial attention for image segmentation. Knowl-Based Syst 214:106754

41. Jiang X, Zhu Y, Liu Y et al (2023) Mc-dc: an mlp-cnn based dual-path complementary network for medical image segmentation. Comput Methods Prog Biomed 242:107846
42. Zhan B, Song E, Liu H (2023) Fsa-net: Rethinking the attention mechanisms in medical image segmentation from releasing global suppressed information. Comput Biol Med 161:106932
43. Gu R, Wang G, Song T et al (2020) Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans Med Imaging 40(2):699–711
44. Abed A, Akrout B, Amous I (2024) Convolutional neural network for head segmentation and counting in crowded retail environment using top-view depth images. Arab J Sci Eng 49(3):3735–3749
45. Mx Huang, Yj Wang, Cf Huang et al (2022) Learning a discriminative feature attention network for pancreas ct segmentation. Appl Math- J Chin Univ 37(1):73–90
46. Chen H, Liu Y, Shi Z et al (2022) Pancreas segmentation by two-view feature learning and multi-scale supervision. Biomed Signal Process Control 74:103519
47. Yan Y, Zhang D (2021) Multi-scale u-like network with attention mechanism for automatic pancreas segmentation. PLoS One 16(5):e0252287
48. Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
49. Wang W, Xie E, Li X et al (2021) Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 568–578
50. Cao H, Wang Y, Chen J et al (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer, pp 205–218
51. Valanarasu JMJ, Oza P, Hacihaliloglu I et al (2021) Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp 36–46
52. Wang H, Cao P, Wang J et al (2022) Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence, pp 2441–2449

HuiFang Wang was born in Anhui Province, China in 2000. She received a B.S. degree in communication engineering from Anhui University of Science and Technology in 2022. She is pursuing an M.S. degree at East China University of Science and Technology. Her research interests include medical image processing in deep learning and computer vision.

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.

Yu Zhu received the PhD degree in optical engineering 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 artificial intelligence, image processing, computer vision, multimedia communication and deep learning. She has published more than 150 papers in journals and conferences.

YaTong Liu is a doctoral candidate at the School of Information Science and Engineering, East China University of Science and Technology. She is an AI algorithm engineer specializing in medical image processing, deep learning algorithms, intelligent analysis of multi-source medical images, lesion segmentation and detection, and pattern recognition. She has published in journals at the intersection of medical and computer vision, and has been involved in publicly and privately funded projects.

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.