Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

# MAFE-Net: A Multi-Level Attention Feature Extraction Network for Pancreas Segmentation

Jiawei Chen[1], Wenjie Chen[1], Zhipeng Zhu[1], and Qi Ye* [1]

[1]School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China.

**Abstract.** Accurate and automatic segmentation of the pancreas from abdominal computed tomography (CT) scans is crucial for diagnosing and treating pancreatic diseases. However, the pancreas is a tiny target abdominal organ with high anatomical variability and low tissue contrast in CT scans, making segmentation tasks challenging. To address this challenge, we propose a multilevel attention feature extraction network to segment the pancreas in abdominal CT images. Specifically, a multi-field attention convolution module (MFAC) and a connection feature fusion module (CFF) are added to the encoding and decoding structure to improve the extraction of pancreatic features. To further enhance the segmentation network's extraction of pancreatic edge features, we propose a decoding feature recall module (DFC), which can be migrated to other encoding and decoding structures and pruned to capture pancreatic edge information better. We compared the performance of our method with that of the most advanced method on the NIH pancreatic segmentation dataset and the challenging pancreatic cancer CT image dataset collected by the Zhujiang Hospital of Southern Medical University. The experimental results show that the DSC of our method on NIH dataset and pancreatic cancer dataset is 84.69% and 78.18% respectively, which is superior to the existing methods.

**Keywords:**
Pancreas segmentation,
Multi-field attention convolution,
Connection feature fusion,
Decoding feature recall.

**Article Info.:**
Volume: 3
Number: 4
Pages: 445 - 463
Date: December/2024
doi.org/10.4208/jml.231101

**Article History:**
Received: 01/11/2023
Accepted: 22/09/2024

**Communicated by:**
Chenglong Bao

## 1 Introduction

In recent years, pancreatic cancer has become an increasingly serious public health problem worldwide. Its early detection is very difficult, and the choice of treatment options is also very limited [26]. By 2030, pancreatic cancer will become the second leading cause of cancer deaths worldwide [12]. Pancreatic cancer is still the tumor with the lowest survival rate at present, the five-year survival rate is between 5% and 15% [1]. Computed tomography images are the main means for doctors to obtain information about pancreas and pancreatic cancer. We can segment the pancreas through CT images to better assist doctors in diagnosing and treating pancreatic cancer.

The pancreas is a retroperitoneal organ. Due to the influence of surrounding organs such as stomach and duodenum, as well as the invasive growth characteristics of pancreatic cancer, it is difficult to obtain the typical features of pancreatic cancer in the early stage by CT and magnetic resonance imaging (MRI) [18]. Therefore, pancreatic segmentation is one of the most representative tasks in the field of medical image segmentation, for the following reasons:

*Corresponding author. yeqi@m.scnu.edu.cn

- The pancreas is closely related to surrounding organs and has a high degree of similarity to surrounding tissues and organs, and its edge position is difficult to distinguish.
- Targets vary greatly in shape, size, and location [5]. The proportion of the pancreas in the entire abdomen is less than 0.5%. In Fig. 1.1, we can better understand information about the pancreas.

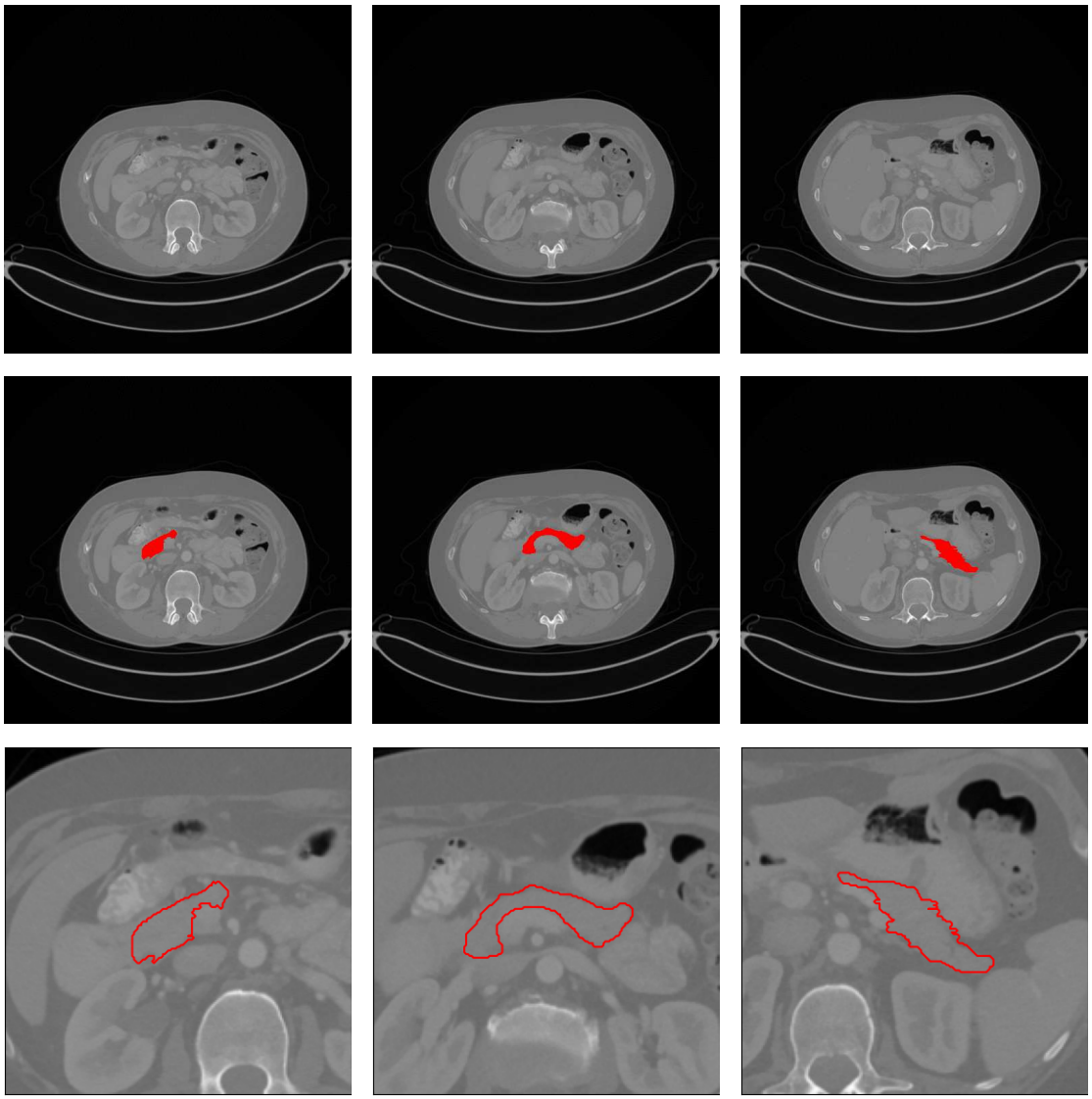

Figure 1.1: Here are three CT images from one case. The first row is the original CT image. The second row is a CT section of the pancreas labeled in red. The third row is a CT image cut based on the position of the pancreas, with the red color indicating the edge of the pancreas.

In Fig. 1.1, we can observe that the proportion of the pancreas in the entire CT image is very small. In the second row of Fig. 1.1, we compare the position and shape of the pancreas in three sections, and we observe significant differences in the distribution, shape, and size of the pancreas in different CT images. In the third row of Fig. 1.1, we can find that the edge of the pancreas is highly similar to surrounding related tissues or organs and is difficult to distinguish. The uncertainty of the shape, size, and location of the pancreas poses a significant challenge to pancreatic segmentation.

With the continuous development of deep learning, more and more researchers are applying it to medical image segmentation. Compared with the segmentation of lung [19], liver [25,28] and other large organs, the segmentation of pancreas and pancreatic cancer is more difficult. Most existing deep learning segmentation networks are based on a coding-decoding network structure called UNet [13] to extract features. However, networks learn through a unified sensory field, ignoring other features of the same data. The global loss caused by the output probability map and real labels makes the network not fully learn the edge features of the pancreas and pancreatic cancer, and the boundary segmentation of the pancreas and pancreatic cancer is not accurate. Therefore, we have introduced the multi-field attention convolution module to expand the receptive field of convolution, allowing the network to learn more features; add the connection feature fusion modules to the encoding and decoding layers. For ordinary feature stitching, the CFF can give different weights to each channel, making the network pay more attention to the required features; add the decoding feature callback module to the decoder. During training, the DFC outputs decoders at each layer. By calculating the loss functions of different decoding layers and performing weighted summation, it can enhance the learning of the network for related weak learning regions, increase the generalization ability of the network, and improve the accuracy of segmentation; during prediction, the DFC is pruned to reduce the model size and improve prediction speed.

This paper proposes the multilevel attention feature extraction network (MAFE-Net). This article adds three attention modules based on UNet: the MFAC module in the codec, the CFF module in the encoding and decoding connection, and the DFC module in the decoder. Specifically, based on UNet, the convolution in UNet is replaced with the MFAC module; adding the CFF module to the semantic connection of the same layer encoding and decoding; and The DFC module is added to the decoder. This paper trains and tests the model on the NIH-CT-82 dataset [14]. The results show that this model is superior to existing segmentation models. Our contributions are as follows:

- We proposed the MAFE-Net network structure, adding the MCF modules to the network, expanding the receptive field of each layer of the network, and allowing the network to learn more features.
- This paper proposes the DFC module that calculates the loss of output features at each decoding layer to improve learning of features such as pancreatic edges. During testing, pruning can be performed to reduce model parameters.
- In feature fusion, to give different weights to each channel and reduce the differences in encoding and decoding semantic features at each layer, we propose the CFF module that uses attention to select key features.

## 2 Related work

Pancreas segmentation is mainly divided into traditional image feature segmentation methods and depth learning-based segmentation methods.

**Traditional image feature segmentation methods.** Several traditional methods have been pivotal in the early stages of pancreas segmentation. Shan *et al.* [16] proposed a method leveraging the Otsu threshold to obtain an approximate contour of the pancreas, followed by morphological operations to separate the pancreas from surrounding tissues. This threshold and morphology-based segmentation method laid the groundwork for more refined approaches. Similarly, Tam *et al.* [21] utilized region growth technology to label pancreatic regions and achieve segmentation results, emphasizing the importance of region-based segmentation methods. Furthermore, Shimizu *et al.* [17] introduced a sophisticated method combining abdominal standardization, atlas-guided segmentation, and the expectation-maximization (EM) algorithm. They integrated the active contour model (Chan-Vese model) into DenseUNet to form a deep active contour network (DACN), significantly advancing medical image segmentation.

**Deep learning-based segmentation methods.** The advent of deep learning has significantly enhanced pancreas segmentation techniques. Zhou *et al.* [29] developed a two-stage method where a rough network segments the pancreatic boundary, followed by a fine network for detailed segmentation, demonstrating the potential of deep learning in refining segmentation processes. Asaturyan *et al.* [2] employed a 3D energy minimization algorithm and a loss function based on the Hausdorff metric and sinusoidal components to refine segmentation, highlighting the utility of advanced mathematical models in improving accuracy. Roth *et al.* [15] presented an automated system for segmenting pancreas from 3D CT volumes, employing a two-stage cascaded approach for localization and segmentation, thereby illustrating the effectiveness of automated deep learning systems. Cai *et al.* [3] proposed an RNN to address spatial inconsistency in segmentation across adjacent image slices, refining CNN outputs to improve shape smoothness, which showcases the integration of RNNs in enhancing spatial coherence. Additionally, several researchers have focused on 3D networks for pancreatic segmentation. Wang *et al.* [24] introduced a multimodal fusion and calibration network for tumor segmentation using 3D PET-CT images, and Zheng *et al.* [27] developed a scalable transformation network (ECTN) within a cascaded two-stage framework for precise pancreatic segmentation. Despite the high accuracy of 3D networks, their parameter count increases geometrically compared to 2D networks, posing challenges in computational efficiency.

**Advances in UNet-based architectures.** The UNet architecture, introduced by Ronneberger *et al.* [13] at the MICCAI conference, has inspired numerous enhancements due to its symmetrical structure and excellent segmentation effects. Oktay *et al.* [11] incorporated attention gates into the UNet network, proposing Attention UNet to enhance the learning of task-related features. This integration marked a significant step in improving segmentation accuracy through focused learning mechanisms. Zhou *et al.* [30] redesigned the skip connections in the UNet network and added dense blocks and convolutional layers, resulting in UNet++, which improves segmentation accuracy through enhanced

structural design. Lu *et al.* [10] introduced the CBAM attention module into the UNet network, replacing the convolution module with a ring residual module to effectively utilize spatial context information, thereby improving segmentation outcomes. Additionally, Hu *et al.* [9] proposed Squeeze and Excitation (SE) blocks, which recalibrate channel feature responses by modeling interdependencies between channels, further advancing the UNet architecture. Szegedy *et al.* [20] introduced the Inception architecture, integrating multi-scale feature information for improved classification and detection, which has influenced subsequent designs in segmentation networks.

**Recent innovations.** Recent studies have introduced novel architectures and training strategies to further advance pancreas segmentation. Valanarasu *et al.* [23] proposed a gated axial-attention model that extends existing architectures by incorporating a control mechanism in the self-attention module. They also introduced a Local-Global (LoGo) training strategy for enhanced performance on medical images, showcasing innovative approaches to model training and attention mechanisms. Guo *et al.* [6] proposed SegNeXt, demonstrating that convolutional attention is more efficient and effective than the self-attention mechanism in Transformers. They designed a convolutional attention network that employs inexpensive convolutional operations, highlighting the shift towards more efficient and effective context encoding methods. Hatamizadeh *et al.* [7, 22] redefined the 3D brain tumor semantic segmentation task as a sequence-to-sequence prediction problem with Swin UNETR. This model utilizes a hierarchical Swin Transformer encoder and an FCNN-based decoder connected via skip connections at multiple resolutions, illustrating the application of advanced transformer architectures in medical image segmentation.

# 3 Methods

## 3.1 Overview

We describe the details of MAFE-Net in this section. Fig. 3.1 illustrates the overall architecture of MAFE-Net. MAFE-Net mainly includes an encoder, decoder, and decoding feature recall module. The encoder and decoder are composed of convolutional and MFAC modules. The MFAC module is shown in Fig. 3.2. As shown in Fig. 3.5, the CFF module is added to the feature connection between the encoder and decoder to eliminate the problems caused by irrelevant and noisy responses in the skip connection. As shown in Fig. 3.3, the DFC module is added to the decoder to weighted sum the loss after feature extraction of the output of each decoding layer, which can capture weak learning feature regions and improve the segmentation of gray value abnormal regions.

## 3.2 Multi-field attention convolution module

Suppose convolutions with the same core size are used in encoding and decoding. In that case, the receptive field of the network will be limited, which will result in the network being unable to capture the semantic information of the larger receptive field, thereby losing some feature information. Inspired by [9], we first convolution or maximize pooling

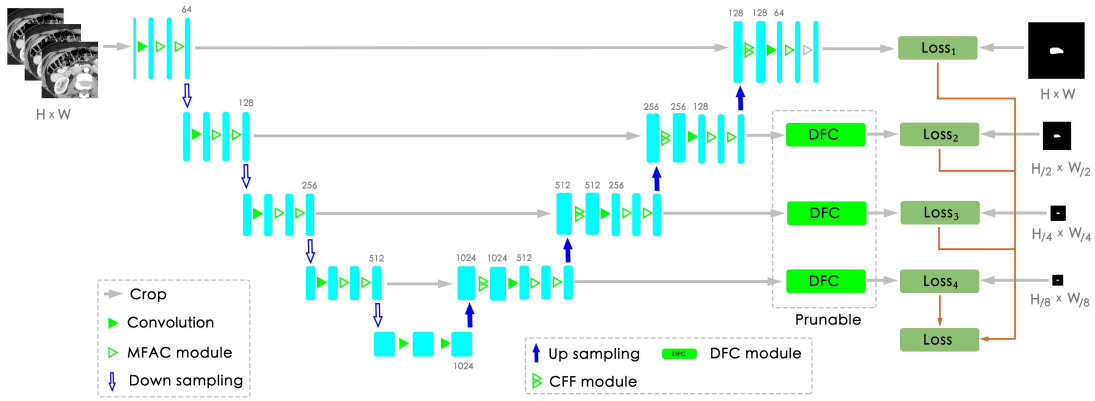


Figure 3.1: Multilevel attention feature extraction network. MAFE-Net has the encoding and decoding structure, including the MAFC and CFF modules in the encoder and decoder. The DFC modules are added to each decoding layer for output, and each decoding layer constitutes a loss function to improve the decoder's learning of pancreatic-related features.

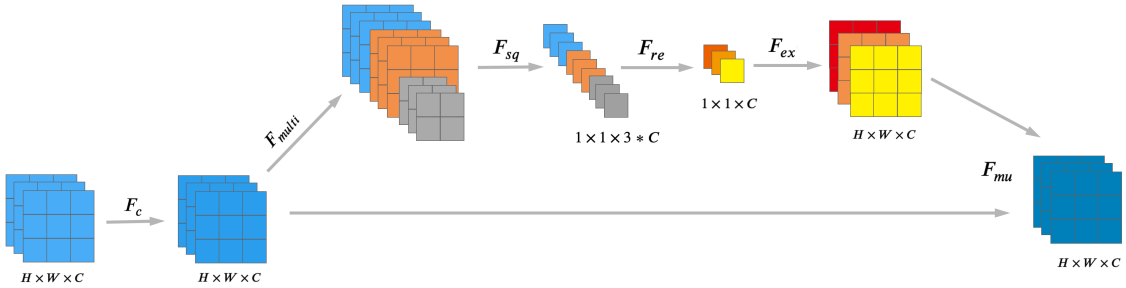


Figure 3.2: The overall architecture of MFAC. The MFAC module is nested in the encoder and decoder, obtaining different receptive field feature information through convolution of multiple core sizes. According to the weight values of various receptive field feature information, the feature information is given corresponding weights, and the corresponding feature information is output.

of multiple cores with different sizes for semantic information $X$, then obtain the corresponding weight $W$ for $X$ by squeezing and expanding, and then output $W$ and $X$ by dot multiplication. The MFAC module can better give corresponding weights to different receptive fields in encoding and decoding so that the network can mine more information related to the segmented region.

In Fig. 3.2, if the input semantic information is $X$, first pass through two layers of convolution

$$X_1 = F_c(X) = W_1\big(W_2(X)\big), \tag{3.1}$$

where $W_1$ is the $1 \times 1$ convolution and $W_2$ is the $3 \times 3$ convolution.

Input $X_1$ into multiple convolutional cores of different sizes and maximize pooling, and concat the output

$$X_2 = F_{multi}(X_1) = connect\big(H_1(X_1), H_2(X_1), H_3(X_1)\big), \tag{3.2}$$

where $H_2$ is the $3 \times 3$ convolution, $H_1$ is the $1 \times 1$ convolution, $H_3$ is the maximum pool size of the core, and *connect* is the number of channels for connecting.

Then input $X_2$ to the global pooling and full connection layers for extrusion operations

$$X_3 = F_{re}\big(F_{sq}(X_2)\big) = L_2\big(L_1\big(Q(X_2)\big)\big), \tag{3.3}$$

where $Q$ is global pooling, and $L_1$ and $L_2$ are full connectivity layers.

Finally, perform expansion and dot multiplication operations. $X_3$ first passes through the sigmod function, then performs expansion, and finally performs dot multiplication with $X_1$

$$X_4 = F_{mu}\big(X_1, F_{ex}(X_3)\big) = X_1 * \big(EX\big(Sigmod(X_3)\big)\big), \tag{3.4}$$

where *Sigmod* is the *Sigmod* function and $EX$ is a matrix expansion operation.

Then the semantic information $X$ passes through the MFAC module to obtain $X_4$. $X_4$ is semantic information with multiple different receptive fields, which allows the network to learn more features of the pancreas.

### 3.3 Decoding feature recall module

After the medical image is encoded, although the network has extracted rich feature information, and before decoding each layer, the CFF module connects the feature information output by the same layer coding, there is still the possibility of missing the feature information of small target objects. The deeper the network goes, the more advanced semantic information can be extracted, and relatively simple information will start to be omitted. As the number of network layers increases, the network undergoes a degradation phenomenon. For this issue, [8] proposes the deep residual learning framework to solve the problem of network degradation. The residual network structure block superimposes the upper layer's output onto the lower layer's input to solve the network degradation problem. In order to solve the problem of decoder degradation, this paper proposes the decoding feature recall module. The network structure diagram of DFC is shown in Fig. 3.3. The DFC module consists of the convolution with the core of $3 \times 3$, the MFAC module, and the *Sigmod* function.

We add DFC modules to coding layers other than the first during training. The feature information output by the $i$ $(i > 1)$ layer decoding layer $Y_i$ will be input to the $i-1$ coding layer, and $Y_i$ will also be input to the DFC module. At this point, the layer $i$ decoding and DFC module can act as the dividing network $U_i$, which ultimately outputs the result $Y_{1i}$, which can be reduced by $i-1$ times the original image $(H \times W)$ to form the loss function. In this way, the output of each layer of decoding layer will be calculated into a loss function and back-propagated through the loss function, so that the network will pay attention to the output of the $i$ layer decoding layer, thereby avoiding network degradation issues.

During the test, we can see from Fig. 3.1 that although each decoding layer has output, they only calculate the loss function, and the actual output is only the output decoded by layer 1 for the predicted image. We can prune the DFC module to reduce the model size and improve the prediction speed when predicting.

DFC modules can be applied to networks with different encoding and decoding structures for the prunability of DFC modules. During training, it can be connected to the

decoding layer output. During testing, this portion of the network can be pruned without increasing the network parameters. To verify this conjecture, we apply it directly to the UNet network, as shown in Fig. 3.4. During training, we added a DFC module to the decoding layer I of the UNet network. During testing, pruning is performed directly so that the parameters of the UNet network do not change. The addition of DFC modules to the network during training can enhance the network's learning of related weak learning areas and increase the network's generalization ability. Under the same data, the dice coefficient (DSC) of UNet+DFC is 0.76% higher than that of UNet.

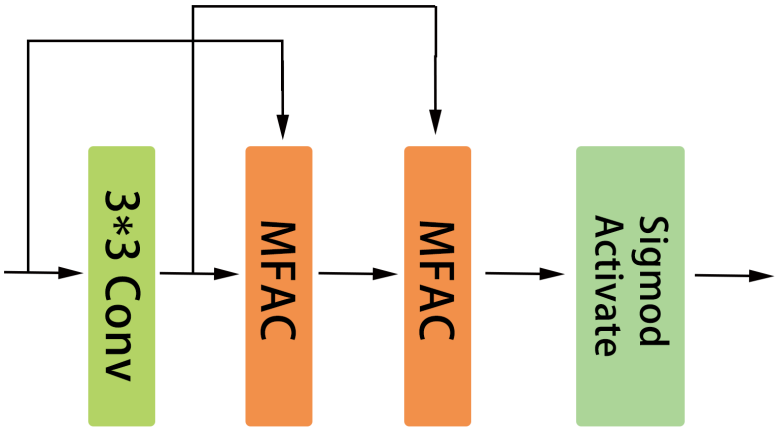


Figure 3.3: The overall architecture of DFC. The DFC module consists of convolution, MFAC module, and sigmod function. This module can be added to networks with different encoding and decoding structures to recall relevant decoding features.

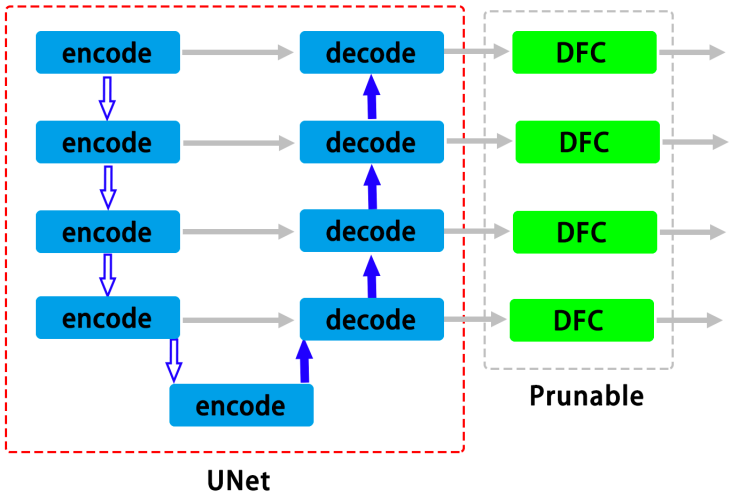


Figure 3.4: UNet with DFC module. The red dotted line shows the UNet network structure and the gray dotted line shows the DFC module. Encode and decode represent the decoding and encoding layers of the UNet network, respectively.

### 3.4 Connection feature fusion module

In a segmented network of encoding and decoding, the semantic information of encoding and decoding at the same layer differs significantly. If simply connecting two semantic information with significant differences, there will be a precipitous information gap inside the new semantic information. That makes the network pay attention to semantic gaps, which affects its attention. However, if the semantic information of the encoder is not connected, the network will directly lose the semantic information of the encoder, which is not conducive to network feature extraction. To preserve the semantic information generated by the encoder and reduce the semantic gap between the encoder and decoder, this paper designs a module to connect feature fusion. CFF can generate a weight matrix of the semantic information connected by the encoder and decoder, significantly reducing the semantic differences at the connection. It can form more continuous new semantic information while retaining the semantic information of the encoder and decoder.

The structure diagram of the CFF module is shown in Fig. 3.5. Assuming that the semantic information $X_1$ and $X_2$ of the same layer codec and decoder are to be connected. After inputting $X_1$ and $X_2$ into the CFF module, a new weight $W$ is generated, and then

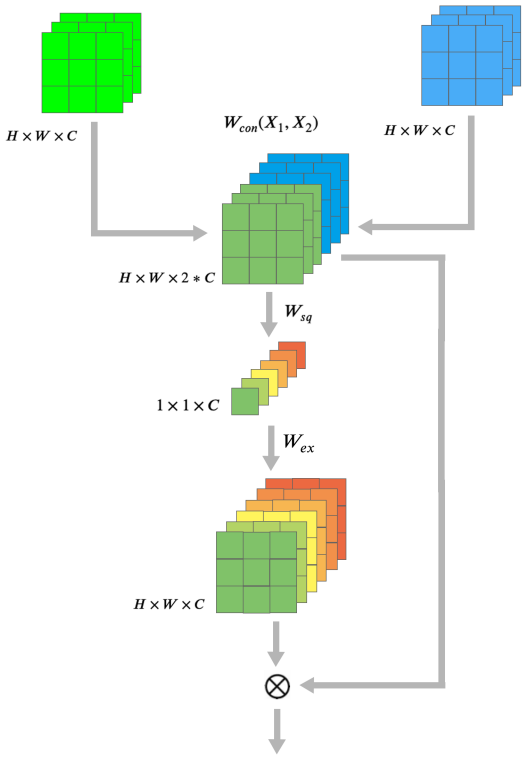


Figure 3.5: The overall architecture of CFF. The CFF module is located at the encoder and decoder connection. $X_1$ and $X_2$ are from the encoding and decoding layers, respectively. According to the weight value of each channel, the corresponding weight is given to the feature information to enhance useful features and suppress features that are not currently useful.

point multiplication is performed to form new semantic information. The detailed process of the CFF module is described below.

$X_1$ and $X_2$ first pass the convolution with a core of $1 \times 1$ before connecting on the channel

$$Y_1 = W_{con}(X_1, X_2) = connect\big(K_1(X_1), K_2(X_2)\big), \tag{3.5}$$

where in $K_1$ and $K_2$ represent convolutions with a core of $1 \times 1$ , and *connect* represents the number of channels for splicing operations.

After compressing $Y_1$, the corresponding weight for each channel number is obtained

$$Y_2 = W_{sq}(Y_1) = ReLU\Big(H_3\big(D_1\big(D_2(Y_1)\big)\big)\Big), \tag{3.6}$$

where $Re$ is the ReLu function, $H_3$ is global pooling, and $D_1$ and $D_2$ are convolutions with a core of $3 \times 3$.

Finally, $Y_2$ is expanded to obtain the weight $W$, then dot multiplied with $Y_1$ to obtain new semantic information $X_3$

$$X_3 = W * Y_1 = W_{ex}(Y_2) * Y_1 = EX\big(Sigmod(Y_2)\big) * Y_1, \tag{3.7}$$

where $Sigmod$ is the $Sigmod$ function and $EX$ is a matrix expansion operation.

The semantic information $X_1$ and $X_2$ encoded and decoded at the same layer are connected through the CFF module. This effectively reduces the semantic gap between $X_1$ and $X_2$. Weighting based on different feature information can highlight some features and enhance their expression.

### 3.5 Loss function of training

In Section 3.3, we introduce the DFC module. We add the DFC module to each layer of decoding to output the probability graph $Y_{1i}$ ($0 < i < 5$, where $i$ is the number of decoding layers). $Y_{1i}$ and the real tag $P$ perform loss calculations. In Fig. 3.1, we can see that the output size of layer 1 decoding is the same as the original image size. Assuming that the original image size $H \times W$, the output size of layer 2 decoding is half the original image size $H/2 \times W/2$, and the output size of layer $i$ decoding is $H/2^{i-1} \times W/2^{i-1}$. Therefore, when calculating the loss of decoding at layer $i$, we need first to reduce the length and width of the real label $P$ to half the original size by $i-1$ and then reduce the size of the label $P$ to obtain $P_{i-1}$. The loss between probability plots $Y_{1i}$ and $P$ is calculated as follows.

The first layer of decoding is directly calculated using $Diceloss$

$$Diceloss = \frac{1 - 2\sum_{h=1}^{H}\sum_{w=1}^{W} C_{1_{h,w}} P_{h,w}}{\sum_{h=1}^{H}\sum_{w=1}^{W} C_{1_{h,w}} + \sum_{h=1}^{H}, \sum_{w=1}^{W} P_{h,w}}, \tag{3.8}$$

where $H$ and $W$ represent the height and width of the image, $C_{1_{h,w}}$ represent the values of probability graph $Y_{11}$ at $(h, w)$, and $P_{1_{h,w}}$ represent the values of the real label $P$ at $(h, w)$.

To make the network pay more attention to the edge characteristics of the target, we use edge loss to calculate the decoding loss function at layer $i$. Layer $i$ decoding loss calculation

$$K_{h,w} = \begin{cases} |C_{i_{h,w}} - P_{i_{h,w}}|, & |C_{i_{h,w}} - P_{i_{h,w}}| < a, \\ 0, & \text{otherwise}, \end{cases} \tag{3.9}$$

$$Edgeloss_i = \frac{1}{m} \sum_{h=1}^{H^{'}} \sum_{w=1}^{W^{'}} K_{h,w}, \tag{3.10}$$

where $H^{'}$ and $W^{'}$ represent the scaled length and width of the real label $P$, and $C_{i_{h,w}}$ represent the values of the probability graph $Y_{1i}$ at $(h, w)$. $P_{i_{h,w}}$ represents the value of the real label $P_i$ in $(h, w)$, and $m$ represents the number of non zero elements in the $K$ matrix, $0 < a < 1$.

By calculating the Diceloss decoded at layer 1 and the loss decoded at layer $i$, the total network loss is obtained by weighted summation

$$loss = Diceloss + \sum_{i=2}^{4} b_i * Edgeloss_i, \tag{3.11}$$

where $0 < b_i < 0.5$.

# 4 Experiment

## 4.1 Introduction and preprocessing of data sets

**Datasets.** To validate the segmentation method proposed in this paper, numerical experiments were conducted using two pancreatic datasets: the publicly available NIH Pancreas Dataset [14,15] and a private dataset provided by the Zhujiang Hospital of Southern Medical University.

The NIH Pancreas Dataset [1] was collected by the National Institutes of Health in the United States. The NIH dataset includes 82 cases, each containing CT data with a resolution of $512 \times 512 \times L$, where $L \in [181, 466]$ represents the number of slices along the body's long axis. The slice thickness ranges from 1.5 mm to 2.5 mm. A medical student manually annotated the pancreas organ in this dataset and then verified it by an experienced radiologist.

The private dataset used in this paper was provided by the Zhujiang Hospital of Southern Medical University(ZJH) and consists of venous phase CT images of pancreatic cancer cases, totaling 45 cases. The image data for each case is $512 \times 512 \times L$, where $L$ represents the number of slices, which varies among different patients. The labels in this dataset include both the pancreas and pancreatic cancer. Medical students manually annotated all labels and subsequently checked and finalized them by professional chief surgeons.

**Data preprocessing.** CT images use grayscale to reflect X-ray absorption by organs and tissues, which varies among different organs. Due to the wide grayscale range of CT slices, window width and level adjustments are required during image reading. This study fo-

[1] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

cuses on segmenting the pancreas and pancreatic cancer, retaining detailed abdominal organs to facilitate segmentation.

We set the CT value range to $[-100HU, 240HU]$ and map these values to grayscale pixel values. Fig. 4.1 shows CT slices after adjusting window width and level, with clearer edges and reduced interference from surrounding organs, aiding feature learning and improving segmentation accuracy.

CT images are 3D, formed by stacking slices, with size and resolution determined by pixel size and spacing. Spacing, the distance between adjacent pixels, varies across cases. For consistent training and prediction, all images were resampled to a $(1, 1, 1)$ spacing.

Original CT slices alone do not sufficiently teach the network model about pancreatic positional variations, limiting generalization. Therefore, data augmentation was used, increasing training samples fivefold by enlarging, shrinking, rotating, translating, and flipping the original data.

Fig. 4.1 shows the small proportion of the pancreas in the image and the blank areas around the CT image. Training on the whole image would cause the model to learn irrelevant parameters and use excessive memory. Thus, the CT scan size was adjusted to $256 \times 256$ based on the pancreas's approximate range, ensuring each slice contains the complete pancreatic region.

Figure 4.1: The first image is the original CT slice. The second image shows the CT slice with adjusted CT values in the range of $[-100HU, 240HU]$. The third image displays the ground truth labels for the slice.

## 4.2 Evaluating metrics

The Dice similarity coefficient (DSC) [4] is used as measurements for experiment results. Dice is used to evaluate the similarity between the predicted sample $Y_{11}$ and the real label $P$

$$DSC = \frac{2\|Y_{11} \cap P\|}{\|Y_{11}\| + \|P\|}. \tag{4.1}$$

## 4.3 Implementation details

Experiments were conducted using the PyTorch framework on a server with one GeForce RTX 3080 Ti GPU (11 GB RAM).

The network model used the RMSprop optimizer with a learning rate of 1e-4, batch size of 4, and 50 epochs.

Input image size was set to 256 × 256. In Table 4.1, we used the controlled variable method to select the optimal image size. We only compared 512 × 512 and 256 × 256 in the experiments. Further reduction in cropping, such as 128 × 128, would result in some slices not fully encapsulating the pancreas, which is detrimental to practical applications. Additionally, due to computer memory limitations, we could only conduct experiments with an epoch count of 2 for images of size 512 × 512. As seen in the table, the segmentation performance is significantly better with an image size of 256 × 256. Therefore, this study selects an image size of 256 × 256 for the experiments.

For the NIH dataset, four-fold cross-validation (4-CV) was performed by dividing the dataset into four parts (20, 20, 21, and 21 cases). Three parts were used for training and one for testing, with results averaged over four iterations to estimate accuracy.

For the ZJH dataset, 30 cases were randomly selected for training and 15 for testing to estimate accuracy.

Table 4.1: Image size selection.

| Size | Epochs=2 | Epochs=4 | Epochs=6 |
|---|---|---|---|
| 256*256 | 77.73 | 84.69 | 83.35 |
| 512*512 | 73.57 | * | * |

## 4.4 Comparison with the state-of-the-art methods

In this subsection, we compare our method with state-of-the-art methods, such as Medical Transformer, SegNeXt, and Swin UNETR methods. The experimental results are shown in the following Table 4.2.

**Performance-complexity trade-off.** MAFE-Net demonstrates exceptional performance on both the NIH and ZJH datasets while maintaining high efficiency in model complexity. It achieves the top Dice Similarity Coefficient (DSC) of 84.69% on the NIH dataset for

Table 4.2: Performance of each methods on pancreas segmentation and pancreatic cancer segmentation tasks in NIH and ZJH datasets.

| Model | Params. (M) | Dataset & Task | | |
|---|---|---|---|---|
| | | NIH | ZJH | |
| | | Pancreatic segmentation | Pancreatic segmentation | Pancreatic cancer segmentation |
| Medical Transformer | 1.8 | 75.62 | 67.2 | 51.03 |
| SegNeXt | 27.6 | 83.41 | **82.79** | 56.91 |
| U-Net | 39.4 | 80.97 | 74.53 | 67.91 |
| Swin UNETR | 149.1 | 82.84 | 78.91 | 56.75 |
| **MAFE-Net (Ours)** | 64.3 | **84.69** | 78.18 | **71.36** |

pancreatic segmentation, and outperforms other models with a DSC of 71.36% on the ZJH dataset for pancreatic cancer segmentation. Although MAFE-Net does not have the lowest parameter count, it achieves high performance, showing a better balance between performance and complexity.

**Comparison with state-of-the-art in pancreatic segmentation.** On the NIH dataset, MAFE-Net leads all models with a DSC of 84.69%, highlighting its high precision and reliability in the task of pancreatic segmentation. Furthermore, on the ZJH dataset, MAFE-Net follows closely with a DSC of 78.18%, just behind SegNeXt's 82.79%, indicating its consistent high performance across different datasets.

**Comparison with state-of-the-art in pancreatic cancer segmentation.** In the particularly important task of pancreatic cancer segmentation, MAFE-Net proves its capability to handle complex medical imaging tasks with a DSC of 71.36% on the ZJH dataset. In contrast, SegNeXt, which performed best on the ZJH dataset for pancreatic segmentation with a DSC of 82.79%, achieved only a DSC of 56.91% for pancreatic cancer segmentation. This performance advantage not only reflects MAFE-Net's excellence in segmentation accuracy but also shows the effectiveness of its algorithm in dealing with the challenging tissue of pancreatic cancer.

**Visual Comparison.** In Fig. 4.2, we present three CT images of pancreases with different shapes and locations from the NIH dataset and segmentation results from MAFE-Net and U-Net. As shown in Fig. 4.2, comparing the segmentation results of U-Net and MAFE-Net, both networks perform similarly well in regions where the pancreatic boundary is distinct. However, U-Net's results deviate significantly from the ground truth in areas where the boundary is blurry or in contact with other organs or tissues. U-Net tends to segment only the clearly defined internal parts of the pancreas accurately, while MAFE-Net excels in delineating the boundary regions. Additionally, in cases where the pancreas is disjointed in a single CT slice, our proposed network can accurately segment both parts. This demonstrates that the multi-level attention mechanism enables the network to learn richer features, resulting in more precise segmentation outcomes.

Fig. 4.3 shows the segmentation results of two pancreas and two pancreatic cancer cases from the ZJH dataset. It is evident that, whether segmenting the pancreas or pancreatic cancer, our proposed MAFE-Net model achieves more accurate segmentation.

In summary, MAFE-Net's performance across all tasks highlights its outstanding capabilities and demonstrates the efficiency of its model design. MAFE-Net, with its high DSC and reasonable parameter count, has proven its potential as an efficient and accurate segmentation model in pancreatic or pancreatic cancer segmentation tasks. Particularly in pancreatic cancer segmentation, MAFE-Net's performance significantly surpasses other models. This ability to balance performance and complexity gives MAFE-Net significant application prospects in medical image segmentation.

## 4.5 Ablation study

Ablation experiments were conducted on the NIH dataset to evaluate the impact of different modules on the network's performance. Specifically, the experiments assessed the

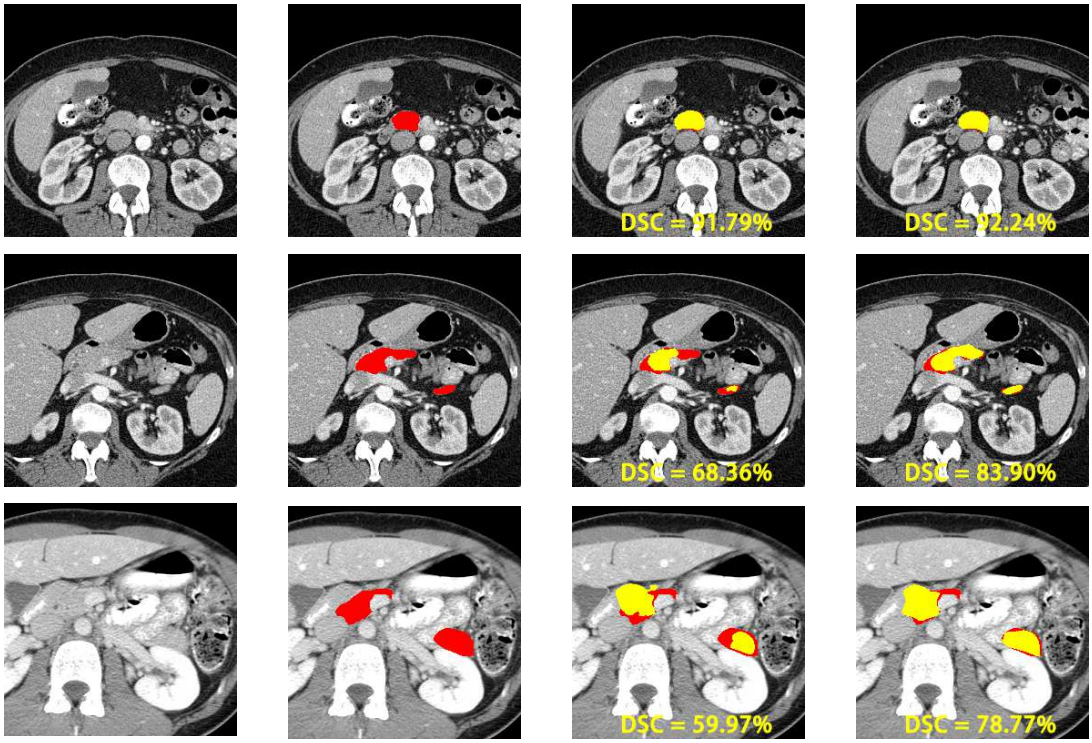


Figure 4.2: Segmentation of pancreas in NIH dataset. Red represents the real label and yellow represents the network prediction label.

MFAC, CFF, and DFC modules combined with a weighted loss function structure. The segmentation results, with and without these modules, were compared to validate the positive contributions of the proposed modules and structures to the pancreas segmentation task.

As shown in Table 4.3, the baseline model (UNet) achieved a DSC of 80.97% on the NIH dataset. By progressively adding the MFAC, CFF, and DFC modules, the DSC scores increased to 83.16%, 83.66%, and 84.69%, respectively. The MFAC module demonstrated the most significant improvement, increasing the DSC by 2.19%. When all three modules were

Table 4.3: Ablation study of the effectiveness of each module. Experiments on the NIH dataset. "✓" indicates that the module is added to baseline for experiment.Max DSC and Min DSC are taken after 20 epochs.

| MFAC | CFF | DFC | DSC(%) | Max DSC(%) | Min DSC(%) |
|---|---|---|---|---|---|
| | | | 80.97 | 81.48 | 79.26 |
| ✓ | | | 83.16 | 83.60 | 81.43 |
| ✓ | ✓ | | 83.66 | 84.44 | 82.73 |
| ✓ | ✓ | ✓ | 84.69 | 85.49 | 83.81 |

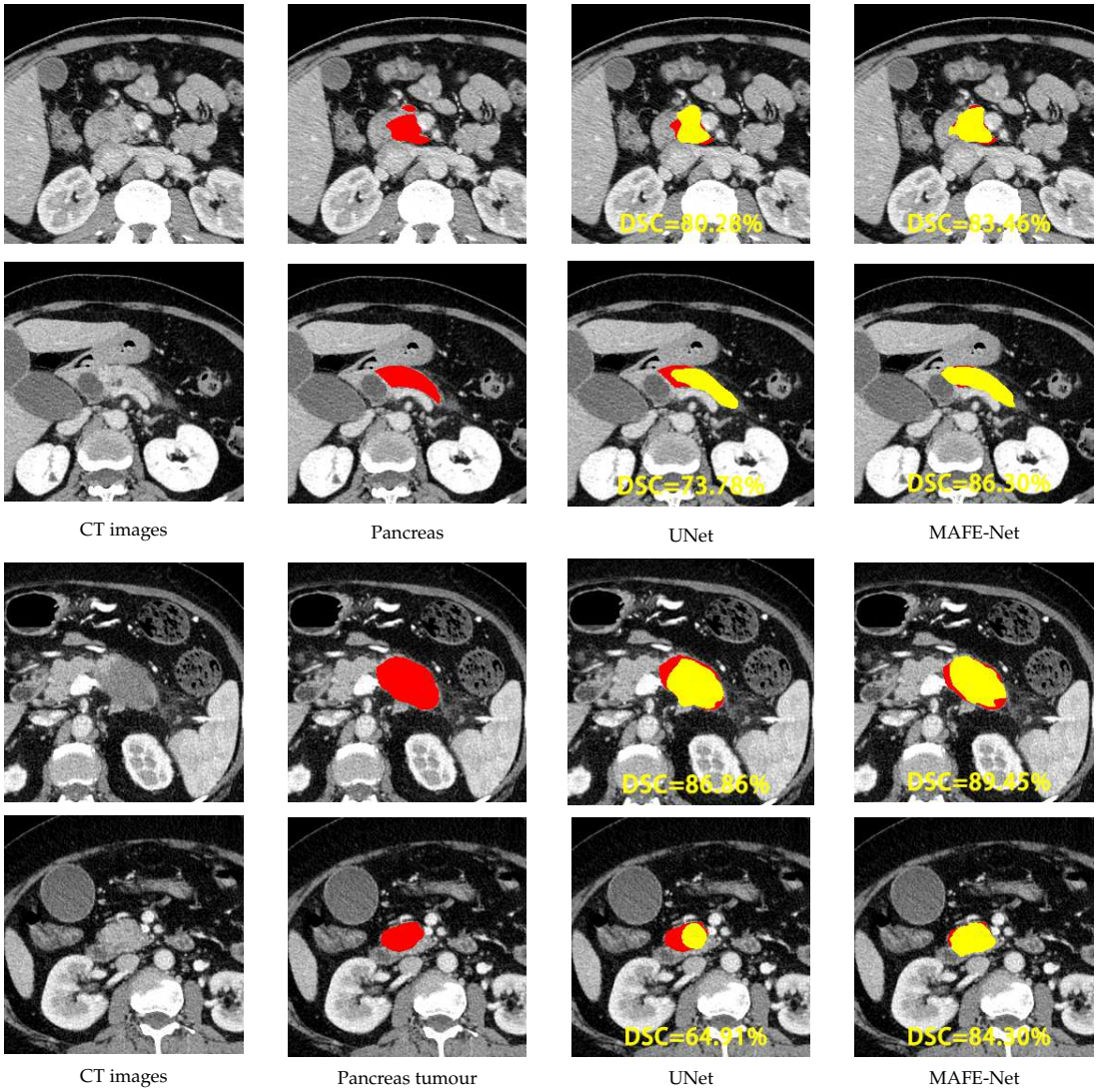


Figure 4.3: Segmentation of pancreas and pancreatic cancer in ZJH dataset. Red represents the real label and yellow represents the network prediction label.

used together, forming the proposed MAFE-Net, the performance peaked, with a 3.72% improvement over the baseline.

Due to insufficient model training in the initial stage, we selected the prediction results after 20 epochs to analyze the maximum and minimum values of DSC. It can be seen from Table 4.3 that the maximum and minimum values of dsc in the model after the addition of each module are significantly increased, indicating that the added modules have a positive effect on the segmentation effect.

The results in Fig. 4.4 illustrate the stability and effectiveness of each module in enhancing model performance.

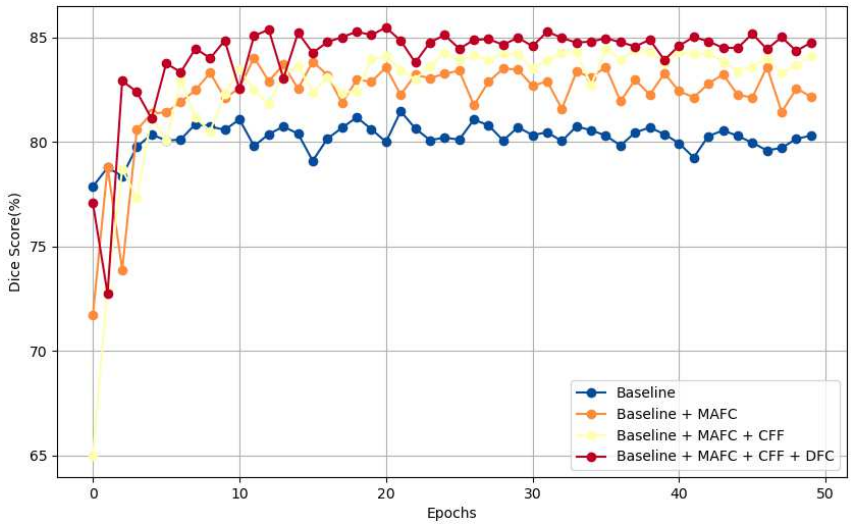


Figure 4.4: Training process of each model with different modules.

# 5 Discussion

Accurate segmentation of the pancreas and pancreatic cancer can better assist doctors in diagnosing and treating pancreatic cancer. This study aims to design a multilevel attention feature extraction segmentation network. However, the differences in shape, size, and position of the pancreas make automatic pancreatic segmentation challenging. Single network structure and Receptive field extract feature information is limited, especially at the edge of the pancreas (Table 4.2 shows that the DSC of UNet is only 80.97%). The image segmentation models with different Receptive fields effectively improve accuracy. Feature fusion also plays a crucial role in the accuracy of segmentation. This study introduces the convolution of different Receptive fields into the SE module [9] to construct the MAFC module. The MAFC module is used for different receptive fields to obtain different characteristic information to help the network pay attention to and learn useful characteristic information. This study proposes a CFF module to fuse the feature information of encoding and decoding, narrow the semantic gap between encoding and decoding, and preserve the practical features of encoding and decoding. Many encoding and decoding structure networks have only one layer of output, which can lead to the network paying too much attention to the output of the first layer, thereby ignoring the importance of other layer networks for pancreatic feature extraction. Therefore, this study connects the DFC module after the decoding layer to calculate the loss of decoding layer features. The DFC module can effectively improve the extraction of decoding layer features, making the network pay more attention to pancreatic edge features or features that the network itself learns weakly. The DFC module also has transferability. As shown in Table 4.3, the network of encoding and decoding structures added to the DFC module will improve its segmentation accuracy. In addition, based on the position and function of DFC module connections, DFC can prune without increasing network parameters during testing.

# 6 Conclusion

This study uses a new and effective deep learning-based segmentation network, MAFE-Net, to segment 2D pancreatic CT images. The MFAC module is proposed to obtain more feature information during the encoding and decoding. The CFF module proposes the fusion of encoded and decoded feature information to improve the effectiveness of feature extraction. In addition, adding transferable and prunable DFC modules after the decoding layer promotes the network's recall of feature loss in the decoding layer, making the network pay more attention to the pancreatic edge region. Finally, the MAFE-Net is verified on the NIH pancreas dataset and the Zhujiang Hospital of Southern Medical University dataset. The experimental results show that the proposed MAFE Net exceeds other most advanced methods, proving the proposed method's effectiveness.

# Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grand Nos. 12071157, 12026602).

# References

[1] C. Allemani et al., Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries, *Lancet*, 391(10125):1023–1075, 2018.

[2] H. Asaturyan, E. L. Thomas, J. Fitzpatrick, J. D. Bell, and B. Villarini, Advancing pancreas segmentation in multi-protocol MRI volumes using Hausdorff-sine loss function, in: *Machine Learning in Medical Imaging: 10th International Workshop*, Springer, 27–35, 2019.

[3] J. Cai, L. Lu, F. Xing, and L. Yang, Pancreas segmentation in CT and MRI via task-specific network design and recurrent neural contextual learning, in: *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, Springer, 3–21, 2019.

[4] L. R. Dice, Measures of the amount of ecologic association between species, *Ecology*, 26(3):297–302, 1945.

[5] A. Farag, L. Lu, H. R. Roth, J. Liu, E. Turkbey, and R. M. Summers, A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling, *IEEE Trans. Image Process.*, 26(1):386–399, 2016.

[6] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, Segnext: Rethinking convolutional attention design for semantic segmentation, *Adv. Neural Inf. Process. Syst.*, Curran Associates, Vol. 35, 1140–1156, 2022.

[7] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, in: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer, 272–284, 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, 2016.

[9] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 7132–7141, 2018.

[10] L. Lu, L. Jian, J. Luo, and B. Xiao, Pancreatic segmentation via ringed residual U-Net, *IEEE Access*, 7:172871–172878, 2019.

[11] O. Oktay et al., Attention U-Net: Learning where to look for the pancreas, *arXiv:1804.03999*, 2018.

[12] L. Rahib, B. D. Smith, R. Aizenberg, A. B. Rosenzweig, J. M. Fleshman, and L. M. Matrisian, Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States, *Cancer Res.*, 74(11):2913–2921, 2014.

[13] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, Vol. 9351, 234–241, 2015.

[14] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, Vol. 9349, 556–564, 2015.

[15] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation, *Med. Image Anal.*, 45:94–107, 2018.

[16] X. Shan, C. Du, Y. Chen, A. Nandi, X. Gong, C. Ma, and P. Yang, Threshold algorithm for pancreas segmentation in Dixon water magnetic resonance images, in: *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, 2367–2371, 2017.

[17] A. Shimizu, R. Ohno, T. Ikegami, H. Kobatake, S. Nawano, and D. Smutek, Segmentation of multiple organs in non-contrast 3D abdominal CT images, *Int. J. Comput. Assist. Radiol. Surg.*, 2:135–142, 2007.

[18] A. D. Singhi, E. J. Koay, S. T. Chari, and A. Maitra, Early detection of pancreatic cancer: Opportunities and challenges, *Gastroenterology*, 156(7):2024–2040, 2019.

[19] I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken, Computer analysis of computed tomography scans of the lung: A survey, *IEEE Trans. Med. Imaging*, 25(4):385–405, 2006.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1–9, 2015.

[21] T. D. Tam and N. T. Binh, Efficient pancreas segmentation in computed tomography based on region-growing, in: *Nature of Computation and Communication: International Conference*, Springer, 332–340, 2015.

[22] Y. Tang, D. Yang, W. Li, H. R. Roth, B. A. Landman, D. Xu, V. Nath, and A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3D medical image analysis, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 20698–20708, 2021.

[23] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Springer, 36–46, 2021.

[24] F. Wang, C. Cheng, W. Cao, Z. Wu, H. Wang, W. Wei, Z. Yan, and Z. Liu, MFCNet: A multi-modal fusion and calibration networks for 3D pancreas tumor segmentation on PET-CT images, *Comput. Biol. Medicine*, 155:106657, 2023.

[25] J. Wang, Z. Xu, Z.-F. Pang, Z. Huo, and J. Luo, Tumor detection for whole slide image of liver based on patch-based convolutional neural network, *Multimed. Tools Appl.*, 80:17429–17440, 2021.

[26] Y. Zhang, J. Wu, Y. Liu, Y. Chen, W. Chen, E. X. Wu, C. Li, and X. Tang, A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set, *Med. Image Anal.*, 68:101884, 2021.

[27] Y. Zheng and J. Luo, Extension-contraction transformation network for pancreas segmentation in abdominal CT scans, *Comput. Biol. Medicine*, 152:106410, 2023.

[28] Z. Zheng, X. Zhang, S. Zheng, H. Xu, and Y. Shi, Semi-automatic liver segmentation in CT images through intensity separation and region growing, *Procedia Comput. Sci.*, 131:220–225, 2018.

[29] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, A fixed-point model for pancreas segmentation in abdominal CT scans, in: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Springer, Vol. 10433, 693–701, 2017.

[30] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, Unet++: A nested U-Net architecture for medical image segmentation, in: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA*, Springer, 3–11, 2018.