Computer Methods and Programs in Biomedicine 233 (2023) 107493

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

MESTrans: Multi-scale embedding spatial transformer for medical image segmentation

Yatong Liu[a], Yu Zhu[a,b,*], Ying Xin[c], Yanan Zhang[c], Dawei Yang[b,d,*], Tao Xu[c,*]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai 200237, China
[c] Department of Pulmonary and Critical Care Medicine, the Affiliated Hospital of Qingdao University, Qingdao, Shandong 266000, China
[d] Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China

ARTICLE INFO

Article history:
Received 11 December 2022
Revised 15 March 2023
Accepted 16 March 2023

Keywords:
Computer-aided diagnosis
COVID-19
Medical image segmentation
Transformer

ABSTRACT

Background and objective: Transformers profiting from global information modeling derived from the self-attention mechanism have recently achieved remarkable performance in computer vision. In this study, a novel transformer-based medical image segmentation network called the multi-scale embedding spatial transformer (MESTrans) was proposed for medical image segmentation.

Methods: First, a dataset called COVID-DS36 was created from 4369 computed tomography (CT) images of 36 patients from a partner hospital, of which 18 had COVID-19 and 18 did not. Subsequently, a novel medical image segmentation network was proposed, which introduced a self-attention mechanism to improve the inherent limitation of convolutional neural networks (CNNs) and was capable of adaptively extracting discriminative information in both global and local content. Specifically, based on U-Net, a multi-scale embedding block (MEB) and multi-layer spatial attention transformer (SATrans) structure were designed, which can dynamically adjust the receptive field in accordance with the input content. The spatial relationship between multi-level and multi-scale image patches was modeled, and the global context information was captured effectively. To make the network concentrate on the salient feature region, a feature fusion module (FFM) was established, which performed global learning and soft selection between shallow and deep features, adaptively combining the encoder and decoder features. Four datasets comprising CT images, magnetic resonance (MR) images, and H&E-stained slide images were used to assess the performance of the proposed network.

Results: Experiments were performed using four different types of medical image datasets. For the COVID-DS36 dataset, our method achieved a Dice similarity coefficient (DSC) of 81.23%. For the GlaS dataset, 89.95% DSC and 82.39% intersection over union (IoU) were obtained. On the Synapse dataset, the average DSC was 77.48% and the average Hausdorff distance (HD) was 31.69 mm. For the I2CVB dataset, 92.3% DSC and 85.8% IoU were obtained.

Conclusions: The experimental results demonstrate that the proposed model has an excellent generalization ability and outperforms other state-of-the-art methods. It is expected to be a potent tool to assist clinicians in auxiliary diagnosis and to promote the development of medical intelligence technology.

1. Introduction

Convolutional neural networks (CNNs) have achieved state-of-the-art performance in many medical segmentation tasks [1,2], demonstrating the importance of convolutional operations for image modeling and understanding. The effectiveness of convolution operations can be attributed to many key factors, such as parameter (weight) sharing, local (sparse) connections, and translation invariance [3]. These properties are largely inspired by biological vision neuroscience [4], which give CNNs a strong inductive bias. Despite the significant success of CNN-based methods, they still have shortcomings in capturing global contextual information [5]. Because existing works obtain global information by generating a very large receptive field, this requires continuous down-sampling to make the stacked convolutional layers deep enough. However, deep networks can cause problems, such as local information loss

[*] Corresponding authors at: School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China (Y. Zhu); Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China (D. Yang); Department of Pulmonary and Critical Care Medicine, the Affiliated Hospital of Qingdao University, Qingdao, Shandong 266000, China (T. Xue).

E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), yang_dw@hotmail.com (D. Yang), xutao1008@163.com (T. Xu).

https://doi.org/10.1016/j.cmpb.2023.107493
0169-2607/© 2023 Elsevier B.V. All rights reserved.

and training difficulties on tiny datasets [6]. Some studies [7] have utilized non-local self-attention mechanisms in the model global context, however, because their computational complexity typically increases quadratically with the size of the space, they may only be appropriate for low-resolution feature maps. Other studies [8] have utilized atrous convolution to obtain long-distance information, which help segment large targets but not tiny targets.

Originally used for sequence-to-sequence predictive modeling in natural language processing (NLP) tasks [9,10], Transformer has recently attracted considerable interest in computer vision. The self-attention mechanism in Transformer can dynamically adjust the receptive field according to the input content; therefore, it is better than the convolution operation in modeling long-range dependencies. However, the number of pixels in a typical image is much larger than the number of data units (such as words), making it difficult to apply standard attention models to images. Despite many attempts [11,12], no dramatic change in NLP has yet occurred. The proposal of Vision Transformer (ViT) is an important step in applying Transformer in the field of computer vision [13]. The main contribution of this work is to use 2D image patches (rather than pixels) with location information as input, embed the image patches in a shared space, and use a self-attention module to learn the relationship between these embeddings. Although Transformer is good at modeling the global context, it has limitations in capturing fine-grained details, and the training of ViT requires huge datasets. Therefore, recent studies have attempted to link CNNs and Transformer to combine their advantages [14].

This paper proposes a novel transformer-based medical image segmentation network called multi-scale embedding spatial transformer (MESTrans). The introduction of the self-attention mechanism can improve the inherent limitation of the CNN, which is that its effective receptive field is smaller than the theoretical receptive field, making it challenging to cover the full image in practical experiments [15]. Based on U-Net, a multi-scale embedding block (MEB) and multi-layer spatial attention transformer (SATrans) structure were added to focus on the connections between multi-scale image patches and model long-distance global relationships. Meanwhile, a feature fusion module (FFM) was constructed, which combines the encoder features with the decoder features and adaptively focuses on important information through network training. The main innovations of this study are as follows:

(1) A novel medical image segmentation network, MESTrans, is proposed, which introduces a self-attention mechanism to improve the inherent limitations of the CNN. The proposed MEB, multi-layer SATrans, and FFM all play important roles in the network, which enhances its overall segmentation performance.

(2) An MEB and multi-layer SATrans based on multi-layer spatial attention are proposed, which can dynamically adjust the receptive field according to the input content, thereby enhancing the ability of the network to extract global features adaptively. This enables the modeling of the spatial relationship between multi-level and multi-scale image patches, effectively capturing global context information.

(3) An FFM is proposed, which performs global learning and soft selection between shallow and deep features, combining them adaptively to focus the network's attention on salient feature regions.

(4) The proposed network was validated using four different types of medical image datasets. Our method achieved 81.23% Dice similarity coefficient (DSC) on the COVID-DS36 dataset, 89.95% DSC and 82.39% intersection over union (IoU) on the GlaS dataset, 77.48% average DSC and 31.69 mm average Hausdorff distance (HD) on the Synapse dataset, and 92.3% DSC on the I2CVB dataset, all of these reaching exceptional results.

2. Related works

2.1. Medical image segmentation algorithms

Many classic segmentation networks have emerged in the field of image segmentation. The emergence of the fully convolutional network (FCN) [16] significantly advanced the development of segmentation networks. It modifies the last fully connected layer of the classification network to a convolutional layer and introduces an end-to-end fully convolutional mechanism to achieve pixel-level segmentation. In 2015, the classic encoder-decoder network U-Net [17] set off a wave in the field of medical image segmentation, and many improved networks based on U-Net have emerged since then. Attention U-Net [18] added a channel attention mechanism based on U-Net to enhance global feature representation. UNet++ [19] concatenated the first four layers of U-Net, allowing the network to learn the importance of different depth features. V-Net [20] provided a 3D image segmentation method and introduced a new objective function to deal with the extreme imbalance between the foreground and background. DeepLab v1 [21] combined a deep convolutional network with a probabilistic graphical model and proposed an atrous convolution algorithm to expand the receptive field and obtain more contextual information. Furthermore, it used fully convolutional conditional random fields [22] to improve the model's ability to capture details. DeepLabv2 [23] added the atrous spatial pyramid pooling (ASPP) module based on v1. ASPP used atrous convolution with different atrous rates to capture multi-scale global information of the image. The pyramid scene parsing network PSPNet [24] contained a hierarchical global prior structure called the pyramid pooling module, which could obtain contextual information at different scales and sub-regions, combined with four different pyramid scale features. Several studies have demonstrated the significance of extracting multi-scale information for medical image segmentation tasks. For instance, Shi et al. [25] proposed a lightweight network based on multi-scale input and feature fusion, which achieved good segmentation results for cardiac magnetic resonance images. Attention mechanisms have also been an active topic of research. To enhance the network's ability to perceive lesion boundaries, Fan et al. [26] used a set of implicitly recurrent reverse attention (RA) modules and explicit edge-attention guidance to establish the relationship between areas and boundary cues. Chaitanya et al. [27] proposed a strategy to extend the contrastive learning framework to segment 3D medical images under semi-supervised conditions with limited labeling. The new contrast strategy exploited domain-specific cues, namely similarity in the structure of 3D medical images. At the same time, this paper proposed a local version of contrast loss to learn local region-unique representations that are useful for pixelwise segmentation.

2.2. Transformer-based segmentation network

Xie et al. [28] proposed a framework that effectively combines convolution and Transformer. In this framework, a CNN was used to extract feature representations, and an efficient deformable Transformer structure was constructed to model the long-range dependencies of the feature maps. This module focused on a small part of the key areas so that the amount of calculation and space complexity were greatly reduced. Chen et al. [29] proposed a novel medical image segmentation network that combined the advantages of the Transformer and U-Net. The transformer first encoded the output feature map of the convolutional network as a learnable embedding sequence to extract global context information.

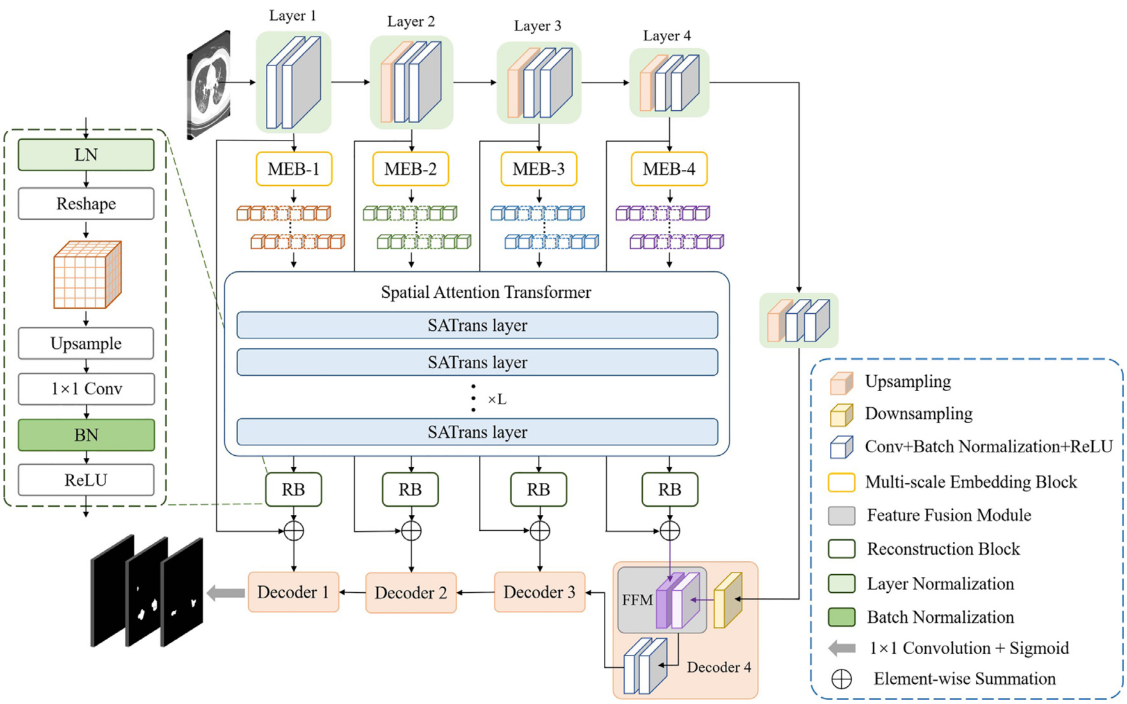

Fig. 1. The structure of the proposed MESTrans.

The decoder then up-sampled the encoded features and finally combined them with the encoder's feature map to achieve accurate segmentation. Cao et al. [30] proposed Swin-Unet, which is an encoder-decoder structure and composed of Swin Transformer [31] modules, making full use of global dependencies. Valanarasu et al. [32] proposed a Medical Transformer (MedT) network, which combined long and short-range dependencies by having a shallow global branch to extract features from the entire image and a deep local branch to process image patches. Tang et al. [33] proposed Swin UNETR for 3D medical image segmentation. The Swin Transformer modules comprised the encoder, which was pre-trained on three auxiliary tasks designed to solve the issue of a lack of medical annotation data, and the Transformer requires a large amount of data for training. In addition, based on the Swin Transformer, Du et al. [34] combined the two designed modules, the dense multiplicative connection module and local pyramid attention module, to propose SwinPA-Net. They cascaded multi-scale semantic feature information through dense multiplicative feature fusion to minimize the interference of shallow background noise and improve feature expression.

3. Method

In this section, we first introduce the overall architecture of the network, then describe the specific modules, and finally introduce the loss function used.

3.1. Network architecture

The proposed segmentation network MESTrans is mainly composed of a UNet encoder, MEB, multi-layer SATrans, and FFM. The overall structure is shown in Fig. 1, which is a codec structure. The encoder is based on UNet and the decoder employs an FFM to guide the learning of deep and shallow features. The network has four scale layers. The MEB at each layer transforms the feature map provided by the encoder into a multi-scale embedding vector, which is then supplied to the SATrans module. The SATrans is composed of L SATrans layers. Through the multi-head spatial attention mechanism, it constructs the spatial dependencies between image patches at different levels and captures multi-scale global information. The reconstruction block re-transforms the SATrans-generated vectors into feature maps. In the decoder part, the decoder block of each layer contains an FFM that performs global learning and soft selection between shallow and deep features. Finally, the prediction result of the original image resolution is output through a segmentation head.

3.2. Multi-scale embedding block

In the medical image segmentation task, we hope that the network can better perceive the target area; however, because the large variability in target size, it is difficult to locate the target area accurately when the target size is small. The main reason for this is that small target features are difficult to retain during continuous downsampling. To retain as many small target features as possible, we designed four MEBs before sending the feature map to the SATrans. For medical image segmentation tasks, preserving small target features by fusing multi-scale local and global information is more reliable. Therefore, we leverage large kernel convolutions in the MEB to enhance the feature extraction ability of the network for small targets. Large kernel convolutions can also expand the receptive field, reduce the loss of spatial location information, and enhance the spatial positioning ability of the network.

Without a loss of generality, Fig. 2 shows the structure of the MEB. The input of the MEB is the feature map $F^i \in \mathbb{R}^{H^i \times W^i \times C^i}$, $i = 1, 2, 3, 4$ output by the encoder. Taking the MEB of the i th layer as an example, a window of a certain size is first applied to the feature map to perform the window split operation:

$$P_j^i = Split\left(F^i\right) \in \mathbb{R}^{h^i \times w^i \times C^i}, j = 1, 2, \cdots N \tag{1}$$

where N is the number of image patches and $Split(\cdot)$ represents the operation of the window split on the feature map. The center point of each window is the same, but the scale is different, and the step size is kept unchanged, resulting in the same number of image patches being obtained via sampling. Then, multi-scale fea-

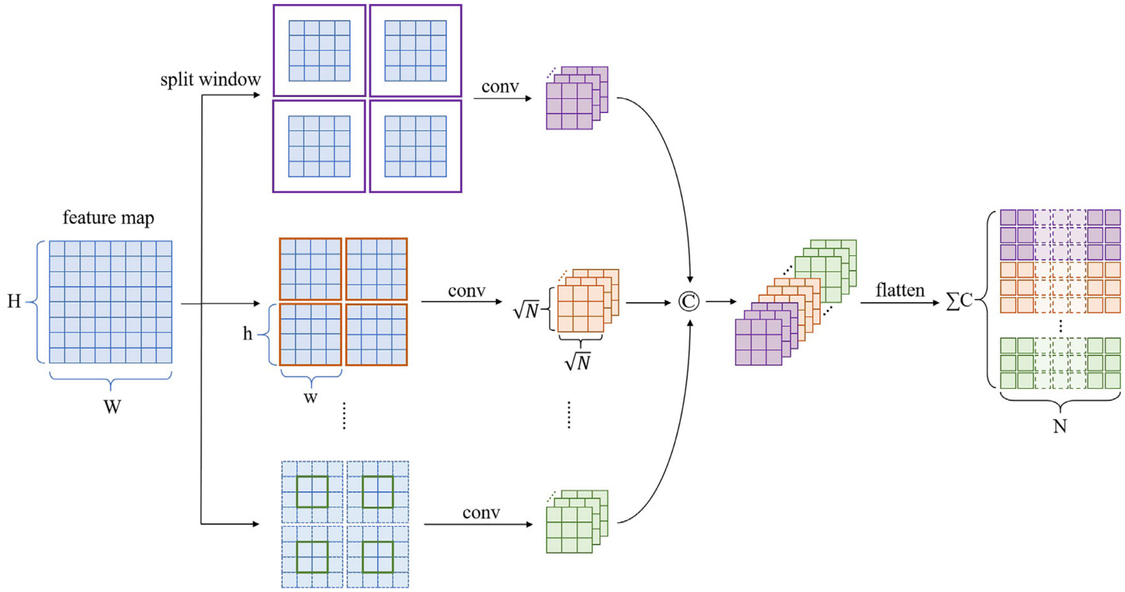

Fig. 2. Architecture of the multi-scale embedding block. The purple, orange, and green rectangles in the figure represent the size of the convolution kernels at different scales. Each layer of the multi-scale embedding block applies $\boldsymbol{M^i}$ convolution kernels of different scales.

Table 1
Parameters of multi-layer embedding. The first column is the number of layers, the second column is the scale of the input feature map of each layer, and the third column is the size of the convolution kernels used by each layer. The first to fourth layers use 8, 4, 2, and 1 kernels, respectively. The fourth column is the step size, the fifth column is the size of the sampled feature map, and the last column is the embedding size.

Layer	Scale	Convolution kernels	Step	Output	Embedding size
1	224 × 224 × 64	128,64,32,16,8,4,2,1	32	7 × 7 × 64	49 × 512
2	112 × 112 × 128	64,32,16,8	16	7 × 7 × 128	49 × 512
3	56 × 56 × 256	8,4	8	7 × 7 × 256	49 × 512
4	28 × 28 × 512	4	4	7 × 7 × 512	49 × 512

ture extraction is applied to all image patches:

$$V_o^i = S_{Conv\left(F^i, P^i\right)} \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times C^i}, o = 1, 2, \cdots M^i \tag{2}$$

where M^i denotes the number of convolution kernels used in the i th layer. The first to fourth layers use eight, four, two, and one kernels, respectively. $S\_Conv(A, B)$ represents the convolution operation on feature map A with a fixed window, and the window size is consistent with the size of B. Finally, the extracted features are concatenated in the channel dimension and flattened in the spatial dimension:

$$E^i = Flatten\left[Concat\left(V_1^i, V_2^i, \cdots V_{M^i}^i\right)\right] \in \mathbb{R}^{N\times\Sigma C} \tag{3}$$

where $\Sigma C = M^i \times C^i$ represents the length of the embedding vector of the final output. Table 1 lists the detailed parameters of the MEB for each layer.

3.3. Spatial attention transformer

In the medical image segmentation task, we expect the segmentation network to adaptively perceive the target region. Owing to the large variability between the target types of medical images, the boundary between the target and the surrounding tissue is blurred, and the difference is small. Thus, it is difficult for the network to identify the target location accurately. To overcome these difficulties, we introduce an attention mechanism.

We designed the SATrans between the encoder and decoder. The SATrans is improved based on the standard Transformer model, which consists of L SATrans layers. For the segmentation model, SATrans has two advantages: first, compared to conventional convolution, the self-attention mechanism can dynamically adjust the receptive field in accordance with the input content, and the ability of the network to adaptively extract global features will be enhanced; second, compared to the conventional skip connection, SATrans can combine multi-scale global information to guide each layer to output shallow features with higher discriminability.

Taking the first layer as an example, as shown in Fig. 3, the input is the four-layer embedding vector $T_i \in \mathbb{R}^{N\times d}, (i\ = 1, 2, 3, 4)$ output by the MEB, where i represents the number of layers, N represents the number of image patches, and d is the number of channels. As listed in Table 1, $N = 49$, and $d = 512$. The specific operation of the SATrans layer is as follows: First, the query vector (Q), key value (K), and value (V) are obtained through the weight matrix, which can be expressed by formula (4):

$$Q_i = T_i W_{Q_i}, K = T_\Sigma W_K, V = T_\Sigma W_V \tag{4}$$

where $W_{Q_i}, W_K, W_V \in R^{d\times d}$, $T_\Sigma = concat(T_1, T_2, T_3, T_4)$, and $T_\Sigma \in R^{4N\times d}$. Then, Q, K, and V perform a self-attention operation. A similarity matrix is generated using Q, K, and a weight on V to obtain the correlation between multi-level and multi-scale spatial regions. This can be expressed by formula (5):

$$A_i = soft\max\left(\frac{Q_i K^T}{\sqrt{d}}\right)V \tag{5}$$

The multi-head spatial attention mechanism refers to the use of multiple groups of W_{Q_i}, W_K, and W_V weight matrices to generate multiple groups of Q_i, K, and V, which can be expressed by formula

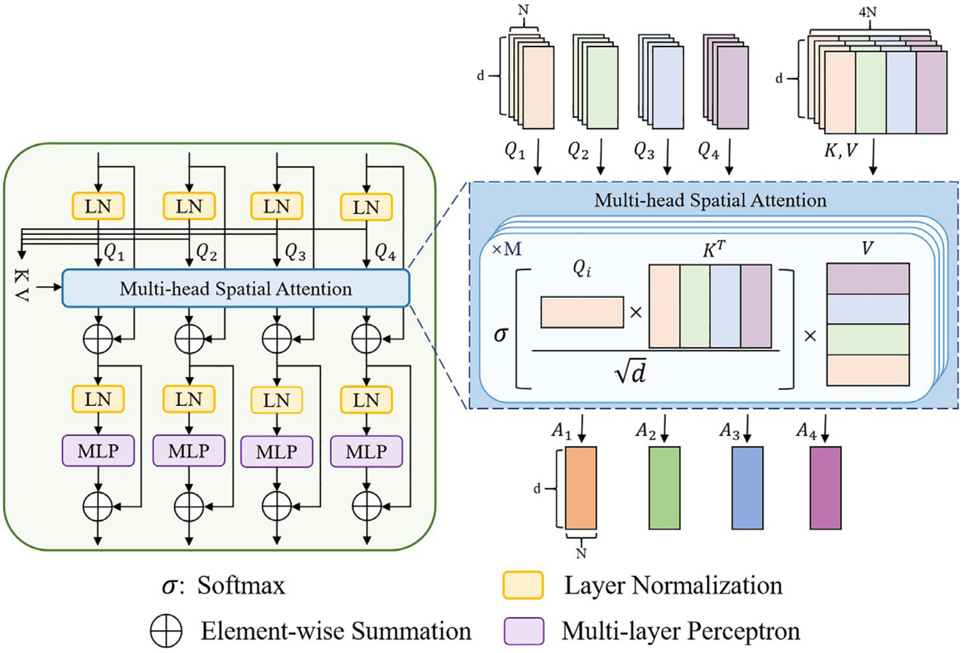

Fig. 3. Architecture of the spatial attention transformer layer.

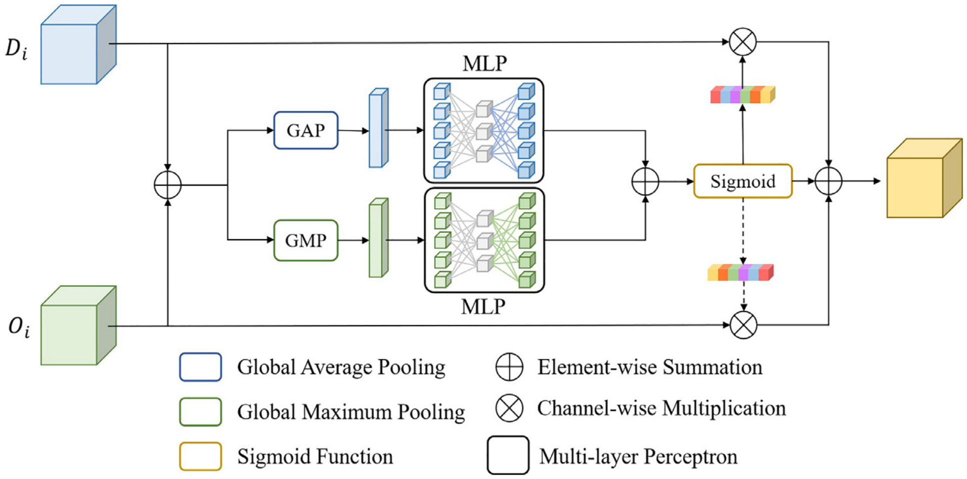

Fig. 4. Architecture of the feature fusion module.

(6):

$$MSA_i = \left(A_i^1 + A_i^2 +, \cdots, +A_i^M\right)/M \tag{6}$$

where M denotes the number of heads. Finally, add the multi-layer perceptron (MLP), normalization layer (LN), and residual connection structure, which can be expressed by formulas (7) and (8):

$$O_i^{'} = MSA(LN(Q_i)) + Q_i \tag{7}$$

$$O_i = MLP\left(LN\left(O_i^{'}\right)\right) + O_i^{'} \tag{8}$$

These two formulas are repeated L times to form an L-layer Transformer structure, and to re-convert the final output vector O_i into a feature map.

3.4. Feature fusion module

The overall medical image segmentation network has a codec structure. The shallow feature output by the encoder has more accurate target location information, but insufficient semantic features, whereas the deep features captured by the decoder have strong semantic information. Therefore, we propose an FFM that performs global learning and soft selection between shallow and deep features, combining them adaptively to focus the network's attention on salient feature regions.

The FFM structure is shown in Fig. 4. First, the feature map O_i output by the reconstruction block and decoder feature map D_i are added, and then the channel attention masks $F_{avg} \in \mathbb{R}^{C\times\ 1\ \times\ 1}$ and $F_{max} \in \mathbb{R}^{C\times\ 1\ \times\ 1}$ are obtained by global average pooling (GAP) and global maximum pooling (GMP), respectively. Then, perform the MLP operation and add them up to obtain CA_i. This can be expressed by formula (9):

$$CA_i = MLP(GAP(O_i + D_i)) + MLP(GMP(O_i + D_i)) \tag{9}$$

Finally, the sigmoid normalization operation is performed on CA_i to obtain S_i, and S_i is used to weigh the initial feature maps O_i and D_i, which is expressed by formula (10):

$$Z_i = D_i S_i + O_i(1 - S_i) \tag{10}$$

By assigning different weight values to O_i and D_i, the network can perform a soft selection, which enhances the adaptability of the network.

Table 2
Multi-class lesion segmentation comparison on the COVID-DS36 dataset. (mean±standard deviation of the Dice similarity coefficient, sensitivity, and specificity).

Methods	Metrics	GGO (%)	Interstitial Infiltrates (%)	Consolidation (%)
Attention U-Net [18]	DSC	73.47±0.92	75.70±0.93	80.18±0.74
	Sen.	70.26±0.88	77.56±0.55	81.26±0.31
	Spec.	99.93±0.01	99.57±0.01	99.87±0.01
UNet-CBAM [40]	DSC	74.69±1.07	74.40±1.00	81.21±1.06
	Sen.	76.56±1.47	77.68±0.73	81.99±0.92
	Spec.	99.88±0.01	99.56±0.04	99.88±0.01
UNet++ [19]	DSC	69.55±2.50	69.94±2.27	78.80±1.23
	Sen.	65.91±2.87	71.10±2.00	76.96±0.42
	Spec.	99.91±0.01	99.56±0.08	99.88±0.02
PAE-Net [41]	DSC	77.04±1.30	79.00±1.14	82.41±0.39
	Sen.	78.83±0.68	82.23±0.57	83.69±1.02
	Spec.	99.89±0.02	99.68±0.04	99.87±0.01
MedT [32]	DSC	65.21±2.07	60.45±1.39	75.18±1.09
	Sen.	68.97±1.96	65.56±1.28	74.96±0.71
	Spec.	99.70±0.02	99.09±0.04	99.80±0.01
TransUNet [29]	DSC	77.45±0.96	78.83±0.53	80.13±0.64
	Sen.	79.00±1.07	80.13±0.50	80.38±0.63
	Spec.	99.86±0.01	99.60±0.01	99.86±0.01
Swin-Unet [30]	DSC	78.14±0.88	76.78±1.42	80.68±1.19
	Sen.	80.65±0.78	79.74±0.82	81.61±1.26
	Spec.	99.86±0.01	99.50±0.06	99.87±0.01
MESTrans (Ours)	DSC	81.23±0.60	86.27±0.37	83.23±0.21
	Sen.	81.04±0.98	86.89±0.51	83.69±0.35
	Spec.	99.91±0.01	99.74±0.01	99.85±0.01

3.5. Loss function

The loss functions used are the dice loss and cross-entropy loss, which can be expressed by formula (11):

$$L = L_{dice} + L_{ce} \tag{11}$$

The dice loss is described by formula (12):

$$L_{dice}(X, Y) = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{12}$$

where X represents the prediction result and Y represents the true label. Dice loss is used to calculate the degree of overlap between the experimental predictions and true labels.

The cross-entropy loss function is expressed by formula (13):

$$L_{ce} = \sum_{i=1}^{N} y^i \log \hat{y}^i + \left(1 - y^i\right) \log\left(1 - \hat{y}^i\right) \tag{13}$$

where N is the number of samples, y is the true label, and $\hat{y}$ is the predicted result. It can be observed that when $y = 1$, the closer the prediction result is to 1, the smaller the loss function; when $y = 0$, the closer the prediction result is to 0, the smaller the loss function.

3.6. Dataset

This study conducted experiments using four datasets.

(1) The dataset COVID-DS36 was jointly established with partner hospitals. The dataset contains a total of 4369 computed tomography (CT) images obtained from lung scans of 36 patients, of which 18 had COVID-19 infection and 18 were normal patients. Clinical diagnosis demonstrates that COVID-19 has obvious imaging characteristics in lung CT images [35,36]. The dataset has a male-to-female ratio of 5:7, and the age distribution of the patients ranges from 6 to 66 years. Among the 18 patients, 11 had mild symptoms, 4 had moderate symptoms, and 3 had severe symptoms. The dataset was annotated by professional doctors and contained three diseases: ground-glass opacity (GGO), interstitial infiltration, and lung consolidation. In the experiment, 3496 CT images were used for training, and 873 images were used for testing, where 25% percent of the data were randomly selected as the test set. The split of the training and test sets was patient-independent.

(2) The second dataset is a gland segmentation dataset (GlaS) [37]. Glands are present in most organ systems, such as the prostate and breast, which are significant histological structures and are the primary mechanism for the production of carbohydrates and proteins. Adenocarcinomas are a common type of cancer. The morphology of the gland is frequently used by medical professionals to assess the aggressiveness of adenocarcinomas. Therefore, accurate gland segmentation is crucial for diagnosis. The GlaS dataset includes 165 images from H&E-stained slices, of which 74 are benign and 91 are malignant. In the experiment, 75 images were used for training, 10 images were used as validation samples, and the remaining 80 images were used for testing.

(3) The third dataset is the public dataset Synapse [38], which includes 30 abdominal CT scan sample sequences. It contains the annotations of 8 abdominal organs (aorta, gallbladder, spleen, kidney (L), kidney (R), liver, pancreas, stomach) and a total of 3779 axial CT images. In the experiment, 18 scans (2212 axial slices) were used as training samples and 12 (1567 axial slices) were used as test samples.

(4) The fourth dataset is the public dataset I2CVB [39]. I2CVB is a collaborative community of common datasets for computer vision that aims to provide common evaluation methods as a basis for data collection and sharing. The I2CVB dataset contains multiparametric magnetic resonance imaging (mpMRI) sequences and lesion labels of 17 prostate cancer (PCa) cases. Each patient had only one lesion, and each sequence contains 13–15 images. We randomly selected 20% of these as the test set and the remaining data as the training set.

3.7. Implementation and evaluation

For all the experiments, data enhancement operations, such as random rotation (0°, 90°, 180°, 270°), random scaling (0.8–1.2 times), and random flipping, were applied to increase the amount

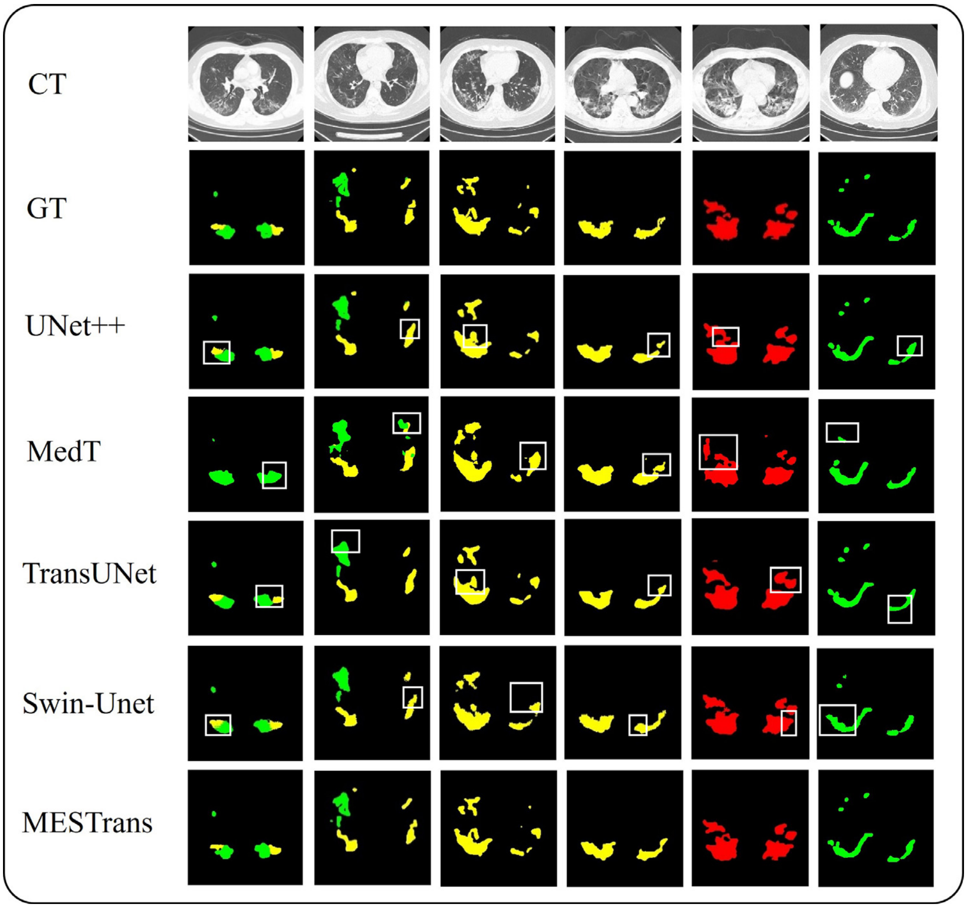

Fig. 5. Segmentation results on the COVID-DS36 dataset. The three color markers correspond to the three lesions, green for ground-glass opacity, yellow for interstitial infiltration, and red for lung consolidation.

Table 3
Binary segmentation comparison on the GlaS dataset. ↑ means the higher the better. (mean ± standard deviation of the Dice similarity coefficient and intersection over union).

Methods	DSC (%)↑	IoU (%)↑
U-Net [17]	86.34±0.65	76.81±0.79
UNet++ [19]	87.07±1.20	78.10±1.93
Attention U-Net [18]	86.98±1.05	77.53±1.59
MRUNet [42]	87.72±0.49	79.39±1.06
TransUNet [29]	87.63±0.44	79.10±0.93
MedT [32]	82.92±0.62	72.46±0.86
Swin-Unet [30]	88.25±0.74	79.86±0.90
PAE-Net [41]	89.63±0.71	82.08±0.79
MESTrans (Ours)	89.95±0.86	82.39±0.77

of data and improve the robustness of the model. The model was implemented using the PyTorch framework. The experimental settings and evaluation indicators of the four datasets are as follows:

(1) For the COVID-DS36 dataset, experiments were performed on a single Nvidia Titan RTX 24GB GPU. In the experiment, the batch size was set to 8, the input image size was 224 × 224, the stochastic gradient descent (SGD) optimizer was applied to train the network, and the initial learning rate was set to 0.001. The network used the DSC, sensitivity, and specificity as metrics. Sensitivity refers to the probability of not being missed when diagnosing a disease and specificity refers to the probability of not being misdiagnosed when diagnosing a disease. They are calculated as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{14}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{15}$$

$$Specificity = \frac{TN}{TN + FP} \tag{16}$$

where A and B indicate the predicted area and the ground truth, respectively. TP, TN, FP, and FN indicate true positives, true negatives, false positives, and false negatives, respectively.

(1) For the GlaS dataset, the network input image size was 224 × 224, the Adam optimizer was applied to train the network, the initial learning rate was set to 0.001, and the batch size was 4. Training was performed on an Nvidia GeForce GTX 1080Ti 11GB GPU. DSC and IoU were used as metrics, and the calculation of IoU can be expressed by formula (17):

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

(2) For the Synapse dataset, the network input image scale was 224 × 224, the batch size was 24, and the SGD optimizer was applied to train the model. The initial learning rate was set to 0.01, momentum was 0.9, and weight decay was 0.0001. The training was performed on three Nvidia GeForce

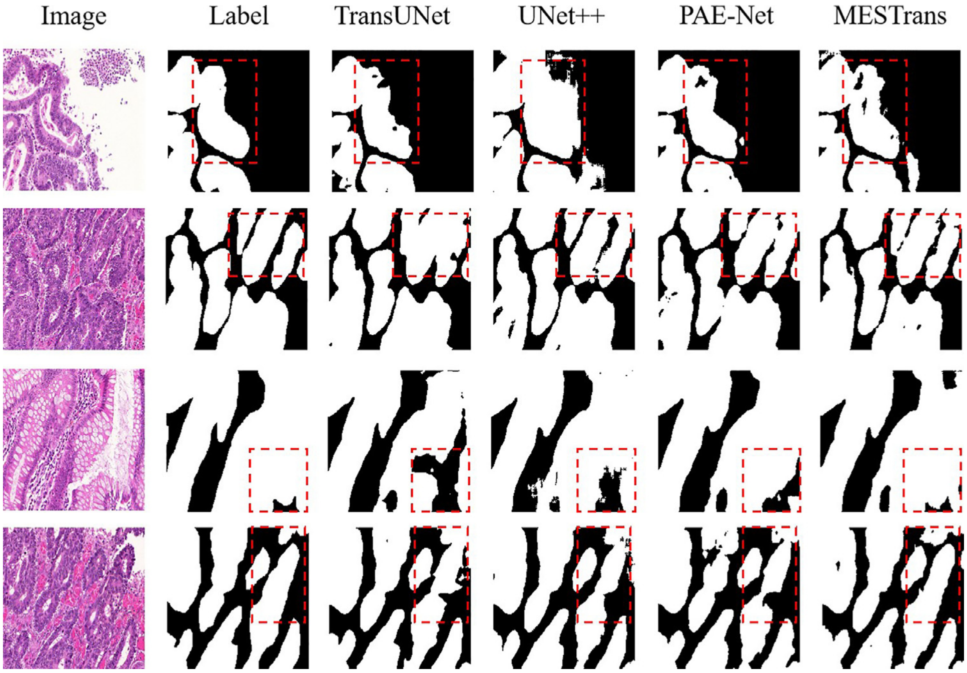

Fig. 6. Segmentation results on the GlaS dataset.

GTX 1080Ti 11GB graphics cards. DSC and HD were used as metrics, where HD defines the distance between two sets; therefore, the smaller the HD, the closer the prediction result is to the true label. For each point in set A, the minimum distance from set B was calculated and then the maximum value was taken as the distance from A to B. The distance from sets B to A was calculated in the same manner. Then the maximum value between the two was taken as HD, which can be expressed by formula (18):

$$H(A,B) = \max\left(\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right) \tag{18}$$

(3) For the I2CVB dataset, experiments were performed on a single Nvidia Titan RTX 24GB GPU. The batch size was set to 8, the input image size was 224 × 224, the SGD optimizer was applied to train the network, and the initial learning rate was set to 0.001. The network used DSC and IoU as metrics.

4. Results

4.1. Comparison with state-of-the-art methods

This section compares the proposed network MESTrans with other state-of-the-art methods on the four datasets. The experimental analysis verifies the excellent segmentation performance of the model.

(1) COVID-DS36 dataset. As presented in Table 2, Attention U-Net, UNet-CBAM [40], UNet++, MedT, TransUNet, Swin-Unet, and PAE-Net [39] are selected as the comparison networks. The DSC values obtained by segmenting the three lesion types using UNet++ are 0.6955, 0.6994, and 0.7880, respectively. In addition to traditional classical networks, we compare them with three state-of-the-art Transformer-based image segmentation networks. The DSC obtained by MedT in the experiment is (0.6521, 0.6045, and 0.7518), and the sensitivity is (0.6897, 0.6556, and 0.7496). TransUNet introduces a Transformer structure at the bottom of the U-Net encoder to extract deep semantic global information. The results for DSC and sensitivity are (0.7745, 0.7883, and 0.8013) and (0.7900, 0.8013, and 0.8038), respectively. The other network is Swin-Unet, which uses Transformer blocks to form a classic encoder-decoder structural model. The obtained DSC of the three lesions is (0.7814, 0.7678, and 0.8068) and the sensitivity is (0.8065, 0.7974, and 0.8161). The DSC and sensitivity obtained by the proposed network are (0.8123, 0.8627, and 0.8323) and (0.8104, 0.8689, and 0.8369). It can be observed that the proposed network MESTrans achieves excellent performance. Compared with TransUNet, the proposed method improves the segmentation accuracy of the three lesion types by an average of 4.77% on DSC and by an average of 4.04% on sensitivity; compared with Swin-Unet, the results have an average increase of 5.04% and 3.21% on DSC and sensitivity, respectively; compared with MedT, the average increase on DSC is 16.63% for three diseases and the average increase on sensitivity is 14.04%.

It can be observed that the proposed network exhibits advanced segmentation accuracy, particularly for the two disease types, GGO and interstitial infiltration. In the image, the texture features of GGO and interstitial infiltration are relatively close, and sometimes it is difficult for doctors to distinguish them, whereas MESTrans greatly improves the segmentation accuracy of the two diseases, reflecting the advanced performance of the model.

At the same time, the experimental results are also visualized on the COVID-DS36 dataset to verify the performance of the model. As shown in Fig. 5, several sample images are selected: the first line is the original image, the second line is the real label, and the next few lines are the prediction results of the UNet++, MedT, TransUNet, Swin-Unet, and MESTrans networks. The prediction results of the proposed MESTrans are closer to the real labels than those of the other networks and show stronger segmentation ability in more complex lesion areas. Green, yellow, and red in the figure represent the three types of lesions: GGO, interstitial infiltration, and lung consolidation.

(1) GlaS dataset. As presented in Table 3, the DSC and IoU metrics are used to evaluate the performance of the proposed network MESTrans on this dataset. The upward arrow indicates that the larger the value, the better the network performance. The proposed method is compared with two different types of networks. The first is the classic improved network based on U-Net, including UNet++, Attention U-Net, and MultiResUNet [42]. Another type of network based on the Transformer includes TransUNet, MedT, and Swin-Unet. In the classical network, compared with U-Net, the proposed network improves the DSC and IoU by 3.61% and 5.58%, respectively. Compared with UNet++, MESTrans improves the DSC and IoU by 2.88% and 4.29%. Compared with Attention U-Net, the proposed method improves the two indicators by 2.97% and 4.86%, whereas compared with MRUNet, the proposed network improves these two indicators by 2.23% and 3.00%. Compared with the classic U-Net based network, the proposed method exhibits a stronger segmentation performance.

In the network combining Transformer and convolution, the DSC and IoU of TransUNet, MedT, and Swin-Unet are (0.8763, 0.7910), (0.8292,0.7246) and (0.8825, 0.7986), respectively, and the indicators of the proposed network MESTrans are (0.8995, 0.8239). Compared with TransUNet, the DSC and IoU improves by 2.32 and 3.29%, respectively; compared with MedT, the two indicators increases by 7.03 and 9.93%; and compared with Swin-Unet, the two indicators increases by 1.70 and 2.53%, respectively. It can be observed that the proposed method still achieves better results than the advanced networks.

As shown in Fig. 6, the prediction results of each model are compared using the GlaS dataset. The first column is the original image, and the second column is the ground truth, followed by the prediction results of the proposed network MESTrans and the comparison networks TransUNet, UNet++, and PAENet. The red dashed boxes in the images reflect a more accurate segmentation performance of the MESTrans.

(1) Synapse public dataset. As presented in Table 4, 8 abdominal organs are segmented in this dataset to verify the multi-class segmentation capability of the proposed network. The values in the table are percentage data. The second and third columns are the average DSC and HD predicted by each model, respectively, and the next is the DSC of the network for each organ. The upward arrow indicates that the larger the DSC, the more accurate the segmentation; the downward arrow indicates that the smaller the HD, the higher the segmentation accuracy. R50-UNet and R50-Gated-UNet [43] replace the encoder parts of U-Net and Gated-UNet with ResNet50. ViT-CUP uses ViT as the encoder, and CUP uses continuous 2 × upsampling until the feature map restores the original image resolution. As presented in the table, the proposed MESTrans has a strong comprehensive segmentation ability. Compared with V-Net, the average DSC improves by 8.39%; compared with the classic network UNet++, the average DSC and HD improves by 7.44% and 43.63 mm; compared with R50-UNet, the two indicators improve by 2.52% and 13.88 mm, respectively. The DSC and HD obtained by ViT-CUP are 0.6786 and 36.11 mm, which are improved by 9.34% and 13.12 mm on the proposed network; compared with MedT, the average DSC is improved by 11.21% and the average HD is optimized by 19.52 mm; and compared with Swin-Unet, the two indicators are improved by 2.72% and 2.11 mm, respectively. At the same time, for the DSC of 8 organs, the proposed MESTrans achieves the best segmentation accuracy on the aorta and left kidney, and achieved the second highest accuracy on the right kidney, liver, pancreas, spleen, and stomach, demonstrating the advanced multi-class segmentation performance of the proposed method.

Table 4
Multi-organ segmentation comparison on the Synapse dataset. ↑ means the higher the better and ↓ represents the opposite. (mean ± standard deviation of the Dice similarity coefficient and Hausdorff distance).

Methods	Avg DSC↑	Avg HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [20]	68.81±-	–	75.34±-	51.87±-	77.10±-	80.75±-	87.84±-	40.05±-	80.56±-	56.98±-
UNet++ [19]	69.76±0.95	66.62±1.31	82.94±0.02	58.15±1.22	75.67±0.24	63.82±0.19	89.56±0.79	49.79±0.15	72.47±1.62	65.68±0.60
DARR [44]	69.77±-	–	74.74±-	53.77±-	72.31±-	73.24±-	94.08±-	54.18±-	89.90±-	45.96±-
R50-UNet [17]	74.68±-	36.87±-	84.18±-	62.84±-	79.19±-	71.29±-	93.35±-	48.23±-	84.41±-	73.92±-
R50-Gated-UNet [43]	75.57±-	36.97±-	55.92±-	63.91±-	79.20±-	72.71±-	93.56±-	49.37±-	87.19±-	74.95±-
ViT-CUP [13]	67.86±0.25	36.11±2.11	70.19±0.43	45.10±2.68	74.70±1.27	67.40±0.15	91.32±0.25	42.00±0.87	81.75±0.44	70.44±1.10
TransUNet [29]	77.48±-	31.69±-	87.23±-	63.13±-	81.87±-	77.02±-	94.08±-	55.86±-	85.08±-	75.62±-
MedT [32]	65.99±0.33	42.51±2.14	71.90±0.49	57.82±1.10	64.72±1.08	64.13±0.29	89.83±0.88	44.69±2.34	75.02±0.93	59.85±1.27
Swin-Unet [30]	74.48±0.41	25.10±1.99	80.12±0.06	62.22±1.12	81.31±1.55	72.64±0.08	92.58±0.29	52.54±1.31	84.85±0.41	69.58±0.50
PAE-Net [41]	76.24±0.12	26.68±1.06	83.32±0.51	62.61±1.61	81.64±1.10	76.92±0.11	93.52±0.16	53.80±1.90	85.57±1.16	72.52±0.51
MESTrans (Ours)	77.20±0.17	22.99±1.21	87.67±0.35	57.89±0.74	81.87±1.08	77.65±0.92	93.84±0.28	54.18±0.76	89.14±0.30	75.38±0.47

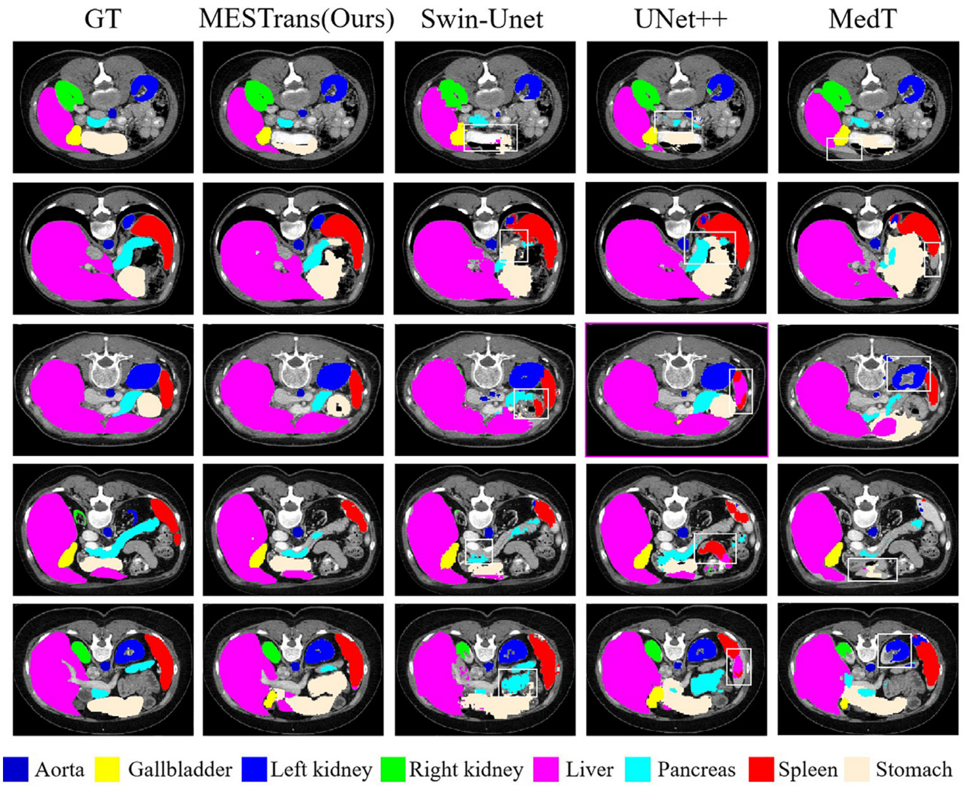

Fig. 7. Segmentation results on the Synapse dataset.

As shown in Fig. 7, the prediction results of each model are compared on the Synapse dataset. The first column is the real label, followed by MESTrans, the proposed network, and the comparison networks Swin-Unet, UNet++, and MedT. Different colors in the figure represent different abdominal organs. It can be observed that the prediction results of the proposed network are closer to the real labels than other networks. The white box in the figure demonstrates that MESTrans shows a more accurate prediction performance for the segmentation details. It can be observed that, first, MESTrans has better recognition ability for continuous regions, and more accurate segmentation performance for complex edges; second, MESTrans can segment target regions more accurately and make fewer false positive diagnoses. Third, MESTrans is better able to deal with the problem of imbalance among multiple classes and has greater average segmentation accuracy for each organ type.

(1) I2CVB public dataset. This dataset includes multi-parameter magnetic resonance (MR) images, of which we used the apparent diffusion coefficient (ADC) image, T2-weighted (T2W) image, and diffusion-weighted image (DWI). As presented in Table 5, our method achieved equally good results for the prostate organ segmentation task. Compared with the experimental results of Akardi et al. and Liu et al., our network improves by 2% on DSC and 1.5% on IoU, respectively.

Table 5
Organ segmentation comparison on the I2CVB dataset. ↑ means the higher the better. (mean ± standard deviation of the Dice similarity coefficient and intersection over union).

Methods	DSC↑	IoU↑
Liu et al. [45]	–	0.843±-
Wang et al. [46]	0.904±-	–
Alkadi et al. [47]	0.921±-	–
MESTrans (Ours)	0.923±0.013	0.858±0.002

Table 6
Ablation experiments on the GlaS dataset. ↑ means the higher the better. (mean ± standard deviation of the Dice similarity coefficient and intersection over union).

Methods	DSC(%)↑	IoU(%)↑
Backbone (U-Net)	86.34±0.65	76.81±0.79
Backbone + SATrans	88.99±1.11	81.35±0.80
Backbone + FFM	87.44±1.16	78.80±0.93
Backbone + SATrans + FFM	89.95±0.86	82.39±0.77

4.2. Ablation experiment

Through ablation experiments, the roles of each module in the network are verified. The addition of SATrans indicates that the MEB module is also added. As listed in Table 6, an ablation experiment was performed on the GlaS dataset. The first row is the benchmark network U-Net, which obtained DSC and IoU of 0.8634 and 0.7681, respectively; it can be observed in the second row that the addition of the SATrans module improved the two indicators by 2.65 and 4.54% respectively. SATrans enhances the attention of the target area by extracting multi-level and multi-scale spatial context information, which is critical for the overall segmentation performance of the network. Additionally, it demonstrates that the fusion of global and local features is crucial for resolving the complexity of object size variation in medical image segmentation. The third row adds the FFM to the network backbone. The FFM can enhance the feature expression ability of important areas. Compared to the first line, the DSC and IoU indicators increases by 1.10 and

Table 7
Ablation experiments on the Synapse dataset. ↑ means the higher the better and ↓ represents the opposite. (mean ± standard deviation of the Dice similarity coefficient and Hausdorff distance).

Methods	Avg DSC(%)↑	Avg HD(mm)↓
Backbone (U-Net)	71.77±0.43	53.04±3.66
Backbone + SATrans	75.57±0.16	36.45±1.29
Backbone + FFM	74.77±0.90	41.81±2.63
Backbone + SATrans + FFM	77.20±0.17	22.99±1.21

1.99% respectively. The fourth line combines the FFM and SATrans to form the proposed network MESTrans, which obtained DSC and IoU values of 0.8634 and 0.7681, respectively. Compared with the first row, the indicators are improved by 3.61% and 5.58%, which fully reflects the effectiveness of the proposed module for accurate segmentation.

As presented in Table 7, ablation experiments were performed using the Synapse dataset. The average DSC and HD of Backbone's multi-organ segmentation are 0.7177 and 53.04 mm, respectively; with the addition of the SATrans module, the average DSC increases by 3.80% and the HD decreases by 16.59 mm; with the addition of the FFM module, the average DSC and HD are 0.7477 and 41.81 mm, which are optimized by 3.00% and 11.23 mm compared to Backbone; finally, the SATrans and FFM modules are added to Backbone to form the proposed network MESTrans, and the average DSC is 0.7720 and the HD index is 22.99 mm. Compared with Backbone, the DSC is increased by 5.43%, and the HD is decreased by 30.05 mm, demonstrating the effectiveness of the proposed module in enhancing the network segmentation performance. DSC is sensitive to the internal filling of the segmentation results, whereas HD is sensitive to the boundaries of the segmentation results. SATrans improves significantly on both metrics, indicating that it constructs the spatial relationship between multilevel and multi-scale image patches and effectively captures the global context information.

4.3. T-SNE visualization results

To further demonstrate the effectiveness of the proposed network, we visualized the feature distribution trained by MESTrans and Backbone. The experiment was performed on several normal tissue regions and lesion regions randomly sampled from the test set of COVID-DS36. This was realized for two reasons: first, the segmentation task is equivalent to a pixel-by-pixel classification task, and the experiment cannot be performed on the entire test set; second, a single CT image generally contains only one to two lesions, and most images do not have lesions, so visualization on only one CT image cannot visually reflect the overall performance of the model.

As shown in Fig. 8, the features of false positives are marked as "blue" and the features of false negatives are marked as "orange". In clinical diagnosis, false negatives are more harmful to patients than false positives. Our model generally reduces the probability of misclassification, and among the misclassified features, the proportion of false negative features is substantially lower than that of false positive features. In addition to the misclassified features, we also focus on the model's ability to distinguish between all categories. It can be observed that the aggregation ability for each category is enhanced, and the distinction between lesion regions and normal tissues is more significant. In conclusion, the proposed module can correct errors and achieve improve this task.

5. Discussion

With technological progress and social development, deep learning has had a significant impact on various fields. Medical image analysis has become a very important part of computer vision. Medical image segmentation is a challenging and active research component that can be used in different tasks such as extraction of lesions or target tissue regions, image-guided interventions, and radiology-aided diagnosis.

In medical image segmentation, models based on codec architectures have made substantial progress over the last few years,

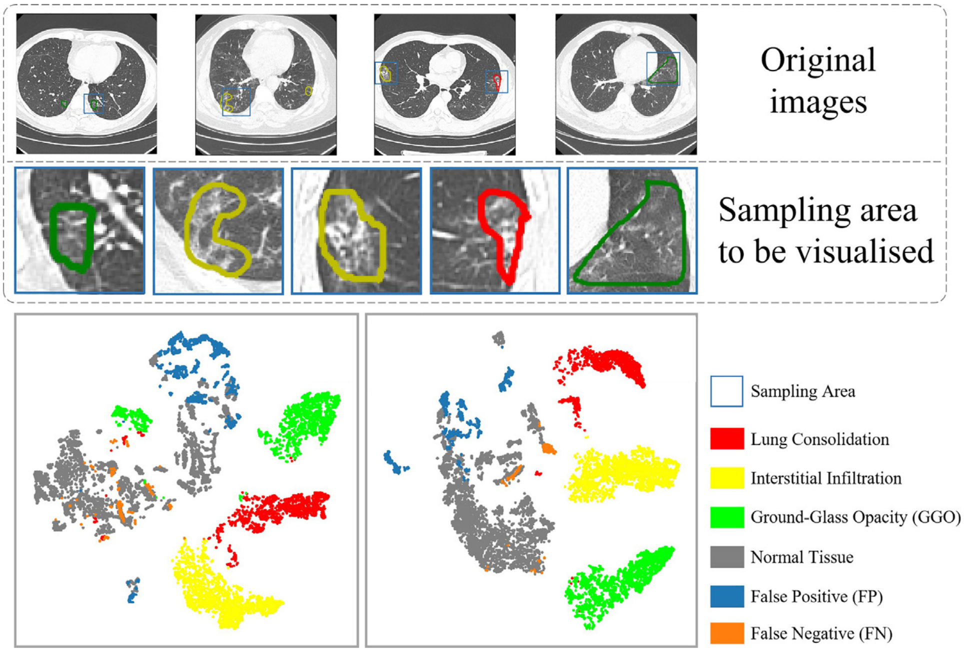

Fig. 8. The first row shows several CT images of the lungs from the COVID-DS36 test set. The second row is a random sampling of areas with lesions from the CT images used as T-SNE visualization. The markings in the image represent the true lesion area, green represents ground-glass opacity, yellow represents interstitial infiltration, and red represents lung consolidation. The bottom two diagrams show T-SNE embeddings of normal tissue areas and lesioned areas randomly. (a) Visualization of the potential space of Backbone. (b) Visualization of the latent space of MESTrans.

but there are inherent limitations to such models. The inductive bias of the convolution operation lacks the ability to extract explicit globally relevant features, whereas the Transformer architecture can capture globally relevant features due to the self-attention mechanism, and the two can complement each other effectively in feature extraction. It is worth noting that the Transformer architecture requires more overhead for training because the self-attention mechanism has $O(n^2)$ time and space complexity concerning the sequence length. How to adapt the self-attention mechanism to medical image segmentation tasks is also a current research topic. To address these issues, this study proposes a novel medical image segmentation network, MESTrans, which consists of a U-Net encoder, MEBs, a multi-layer SATrans, and a decoder with FFMs. Compared to the current methods, the network achieves good experimental results on four different types of datasets. A discussion of the experimental results is as follows.

(1) As can be observed from the experiments on the COVID-DS36 dataset in Table 2, the improvement in the segmentation accuracy of our model is greater for GGO and interstitial infiltration lesions. These two lesions show two main characteristics on CT images: first, the lesion area is discontinuous and has smaller areas; second, the texture features of both lesions are very similar. This result reflects the enhanced perception ability and segmentation accuracy of the model for smaller targets, as well as the improved discrimination ability for targets with similar texture features.
(2) Our model achieved a state-of-the-art performance on the GlaS dataset. On the Synapse dataset, our method achieved the second-best average DCS metric and still showed good segmentation advantages for some of the abdominal organs (e.g., the aorta and left kidney).
(3) I2CVB is a very small dataset with only 17 patients. As presented in Table 5, our model achieves good results on this small dataset. A better DSC indicates that the network can still perceive the target area and identify the texture features inside the target with a small amount of data training.

The segmentation advantages of this network make it a great clinical prospect; however, some shortcomings remain. In the future, further compression of the model should be considered to reduce the number of parameters to be calculated. A 3D medical image segmentation network can be considered to leverage the inter-layer correlation features of medical images to improve the network's ability to identify targets. The segmentation capability for the model on more types of medical images can be explored to make a greater contribution to intelligent healthcare.

6. Conclusion

In this study, a new medical image segmentation network called MESTrans is proposed. It introduces the Transformer structure based on the classic U-Net network, combines global and local information, and exhibits strong performance. In this study, we design and implements an MEB and multi-layer SATrans. The MEB divides the feature map of each layer of the encoder into image blocks of different sizes, generates embedding vectors with multi-scale information, and inputs them into a Transformer-based structure. In the SATrans, the spatial dependence relationship at multi-level and multi-scale is effectively modeled. Furthermore, an FFM is constructed, which combines deep and shallow features to select important features and gives them greater weights during training to improve performance. This study conducts experiments on four different types of medical image datasets, all of which achieve advanced segmentation levels, thereby reflecting the good generalization ability of the proposed network. Simultaneously, it is also compared with traditional segmentation networks, such as U-Net, UNet++, and Attention U-Net, as well as newly proposed Transformer-based networks, such as Swin-Unet, MedT, and TransUNet. The comparison results demonstrate the superior performance of the proposed network.

Declaration of Competing Interest

The authors declare no conflicts of interest with respect to the research, authorship, and publication of this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant No. 82170110), the Science and Technology Commission of Shanghai Municipality (Grant Nos. 20DZ22544000, 20DZ2261200, ZD2021CY001), and the Fujian Province Department of Science and Technology (Grant No. 2022D014)

References

[1] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Proceedings of the International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 424–432.
[2] C. Kou, W. Li, W. Liang, Z. Yu, J. Hao, Microaneurysms segmentation with a U-Net based on recurrent residual convolutional neural network, J. Med. Imaging 6 (2019) 025008.
[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, 86, 1998, pp. 2278–2324.
[4] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[5] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, C. Xu, Cmt: convolutional neural networks meet vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12175–12185.
[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[7] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[8] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122, (2015) doi: 10.48550/arXiv.1511.07122.
[9] K. Chowdhary, Natural language processing, Fundamentals of artificial intelligence, 2020, pp. 603–649.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, in: Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.
[11] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, arXiv preprint arXiv:1904.10509, (2019) doi: 10.48550/arXiv.1904.10509.
[12] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.C. Chen, Axial-deeplab: stand-alone axial-attention for panoptic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 108–126.
[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, (2020) doi: 10.48550/arXiv.2010.11929
[14] H. Wang, P. Cao, J. Wang, O.R. Zaiane, Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2022, pp. 2441–2449.
[15] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2016) 29.
[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[17] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing And Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[18] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, Attention u-net: learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, (2018) doi: 10.48550/arXiv.1804.03999.
[19] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, in: Unet++: A nested U-Net Architecture For Medical Image segmentation, Deep learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support, Springer, 2018, pp. 3–11.

[20] F. Milletari, N. Navab, S.A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: Proceedings of the Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[21] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062, (2014) doi: 10.48550/arXiv.1412.7062
[22] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the International Conference on Machine Learning (ICML), 2001.
[23] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 834–848.
[24] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[25] J. Shi, Y. Ye, D. Zhu, L. Su, Y. Huang, J. Huang, Automatic segmentation of cardiac magnetic resonance images based on multi-input fusion network, Comput. Methods Programs Biomed. 209 (2021) 106323.
[26] D.P. Fan, T. Zhou, G.P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-net: automatic COVID-19 lung infection segmentation from CT images, IEEE Trans. Med. Imaging 39 (2020) 2626–2637.
[27] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, Adv. Neural Inf. Process. Syst. 33 (2020) 12546–12558.
[28] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation, in: Proceedings of the International Conference On Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 171–180.
[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, (2021) doi: 10.48550/arXiv.2102.04306
[30] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: In Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, Cham: Springer Nature, Switzerland, 2023, pp. 205–218.
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[32] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 36–46.
[33] Y. Tang, D. Yang, W. Li, H.R. Roth, B. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20730–20740.
[34] H. Du, J. Wang, M. Liu, Y. Wang, E. Meijering, SwinPA-Net: swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12, doi:10.1109/TNNLS.2022.3204090.
[35] M.Y. Ng, E.Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M.M.S. Lui, C.S.Y. Lo, B. Leung, P.L. Khong, Imaging profile of the COVID-19 infection: radiologic findings and literature review, Radiol. Cardiothorac. Imaging 2 (1) (2020) e200034.
[36] Z.W. Zhao, Z. Zhong, X. Xie, Q. Yu, J. Liu, Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study, AJR Am. J. Roentgenol. 214 (2020) 1072–1077.
[37] K. Sirinukunwattana, J.P. Pluim, H. Chen, X. Qi, P.A. Heng, Y.B. Guo, L.Y. Wang, B.J. Matuszewski, E. Bruni, U. Sanchez, Gland segmentation in colon histology images: the glas challenge contest, Med. Image Anal. 35 (2017) 489–502.
[38] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, A. Klein, Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, 2015, p. 12.
[39] Z. Khan, N. Yahya, K. Alsaih, M.I. Al-Hiyali, F. Meriaudeau, Recent automatic segmentation algorithms of MRI prostate regions: a review, IEEE Access 9 (2021) 97878–97905.
[40] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference On Computer Vision (ECCV), 2018, pp. 3–19.
[41] F. Yu, Y. Zhu, X. Qin, Y. Xin, D. Yang, T. Xu, A multi-class COVID-19 segmentation network with pyramid attention and edge loss in CT images, IET Image Process. 15 (2021) 2604–2613.
[42] N. Ibtehaz, M.S. Rahman, MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[43] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.
[44] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, A. Yuille, Domain adaptive relational reasoning for 3d multi-organ segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 656–666.
[45] Z. Liu, W. Jiang, K.H. Lee, Y.L. Lo, Y.L. Ng, Q. Dou, V. Vardhanabhuti, K.W. Kwok, A two-stage approach for automated prostate lesion detection and classification with mask R-CNN and weakly supervised deep neural network, in: Proceedings of the Artificial Intelligence in Radiation Therapy: First International Workshop, AIRT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, Springer, 2019, pp. 43–51. October 17, 2019, Proceedings 1.
[46] H. Wang, Z. Zhang, B. Zhang, Y. Mi, J. Wu, H. Huang, Z. Ma, W. Wang, A feature regularization based meta-learning framework for generalizing prostate mri segmentation, in: Proceedings of the IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE, 2022, pp. 1–4.
[47] R. Alkadi, F. Taher, A. El-Baz, N. Werghi, A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images, J. Digit. Imaging 32 (2019) 793–807.