澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

EPT-Net: Edge Perception Transformer for 3D Medical Image Segmentation

YU Jiening
Yu Jiening@163.com

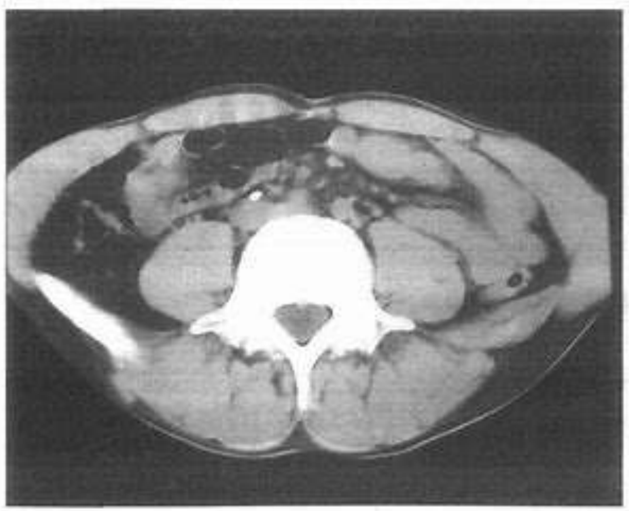
um 澳大

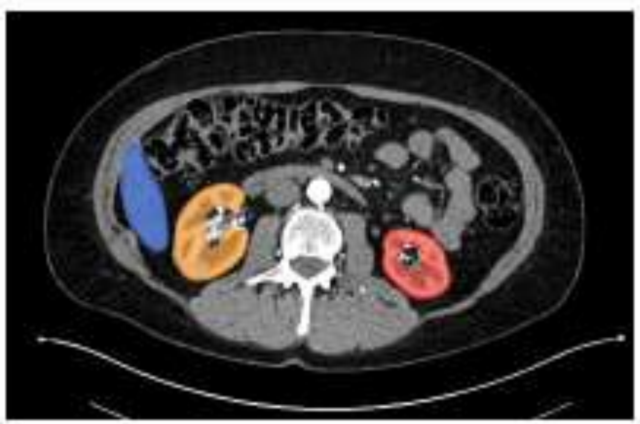
Catalogue

1. Introduction
2. Related work
3. Method
4. Experiments
5. Conclusion

1.Introduction

- **Medical image segmentation** is a process in which the original images are divided into regions of different categories, and then the region of interest is extracted.
- Organ segmentation can provide a visual representation for organ shape and size measurement, which is significant for detecting early manifestations of life-threatening diseases.
- Due to the influence of medical image acquisition devices, the **gray value** differences between different organs and tissues are close, resulting in relatively blurry boundaries between organs and surrounding tissues, as well as significant differences in position and morphology among different organs, making medical image segmentation a challenging task.





Solution

- Many segmentation methods have been proposed, including region-based segmentation, edge-based segmentation, level set segmentation, and fuzzy set segmentation. However, these methods rely on hand-craft features and have limited functional representation ability.
- Propose a novel hybrid framework of Edge Perception Transformer Network (EPT-Net) for accurate segmentation of 3D medical images. EPT-Net has an encoder-decoder structure.

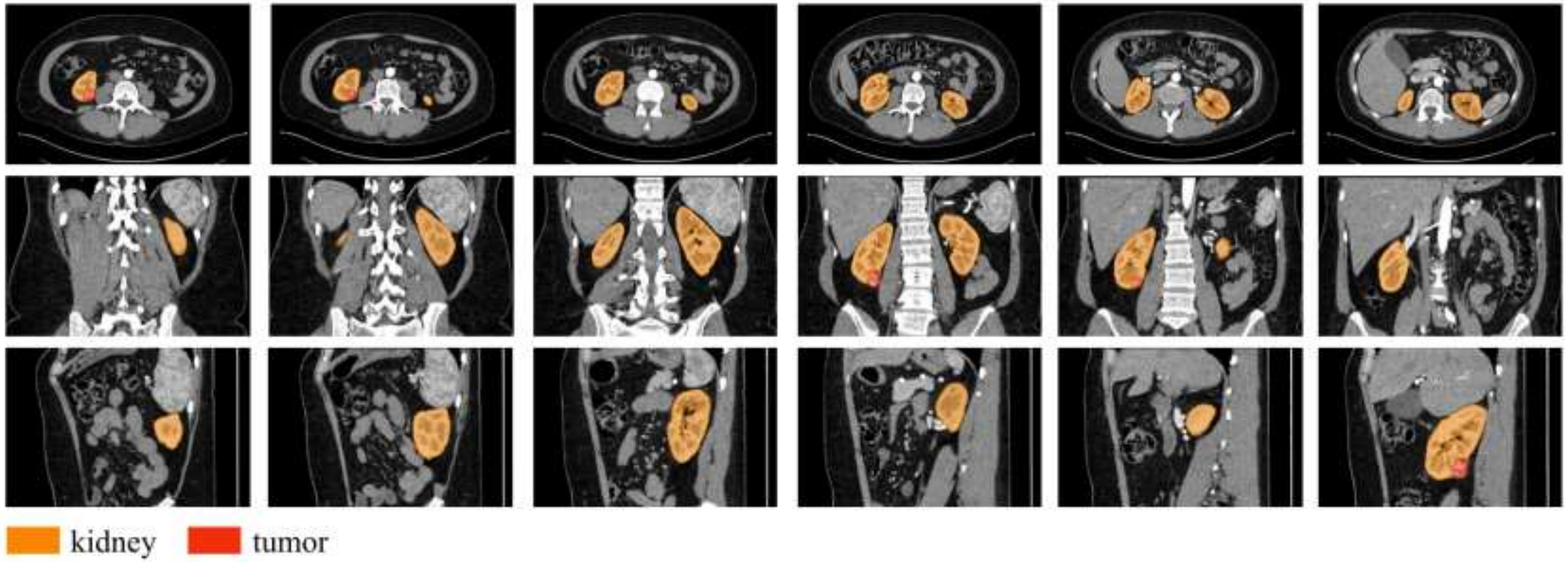


Fig. 1. KiTS19 dataset. The axial view (the first row), sagittal view (the second row) and coronal view (the third row) of CT images. The CT images of a patient from the KITS19 training set. Six angles of one case are selected for display.

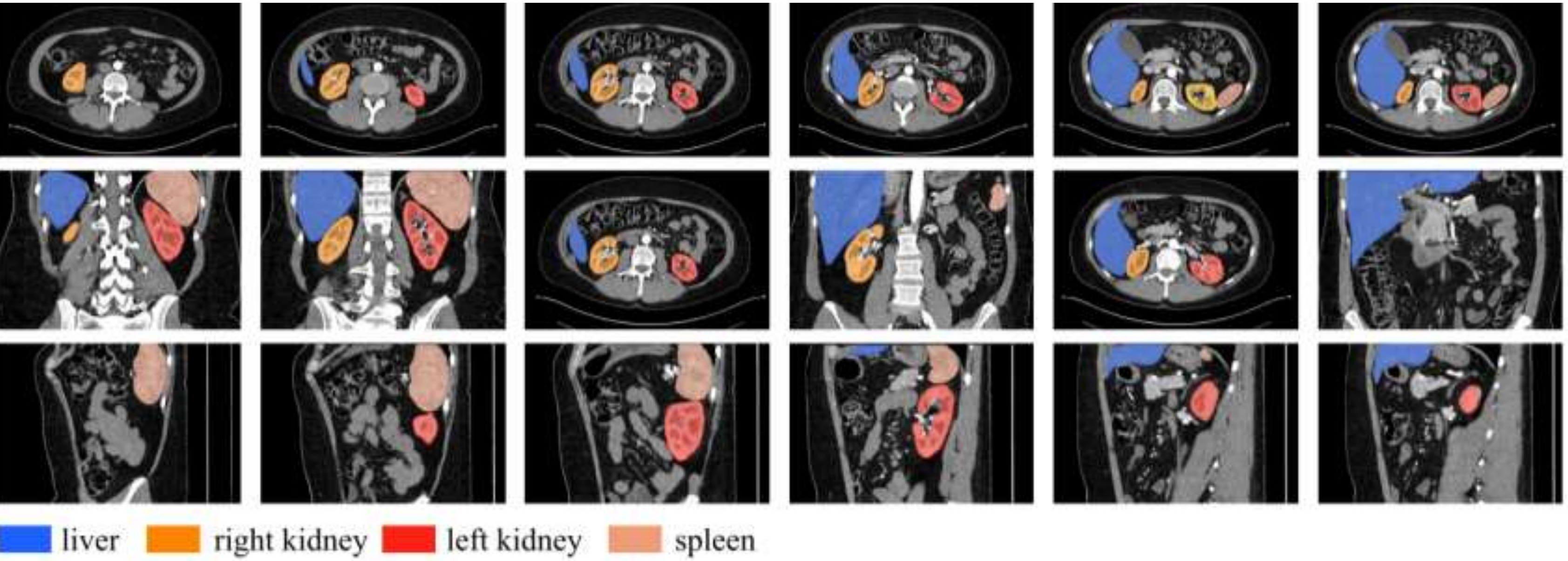


Fig. 2. KiTS19-M dataset. A slice example of the same case as Fig. 1 on KiTS19-M. In the original KiTS19 data, the liver, left kidney, right kidney, and spleen are labeled.

- Six angles of one case are selected for display.

Main contributions

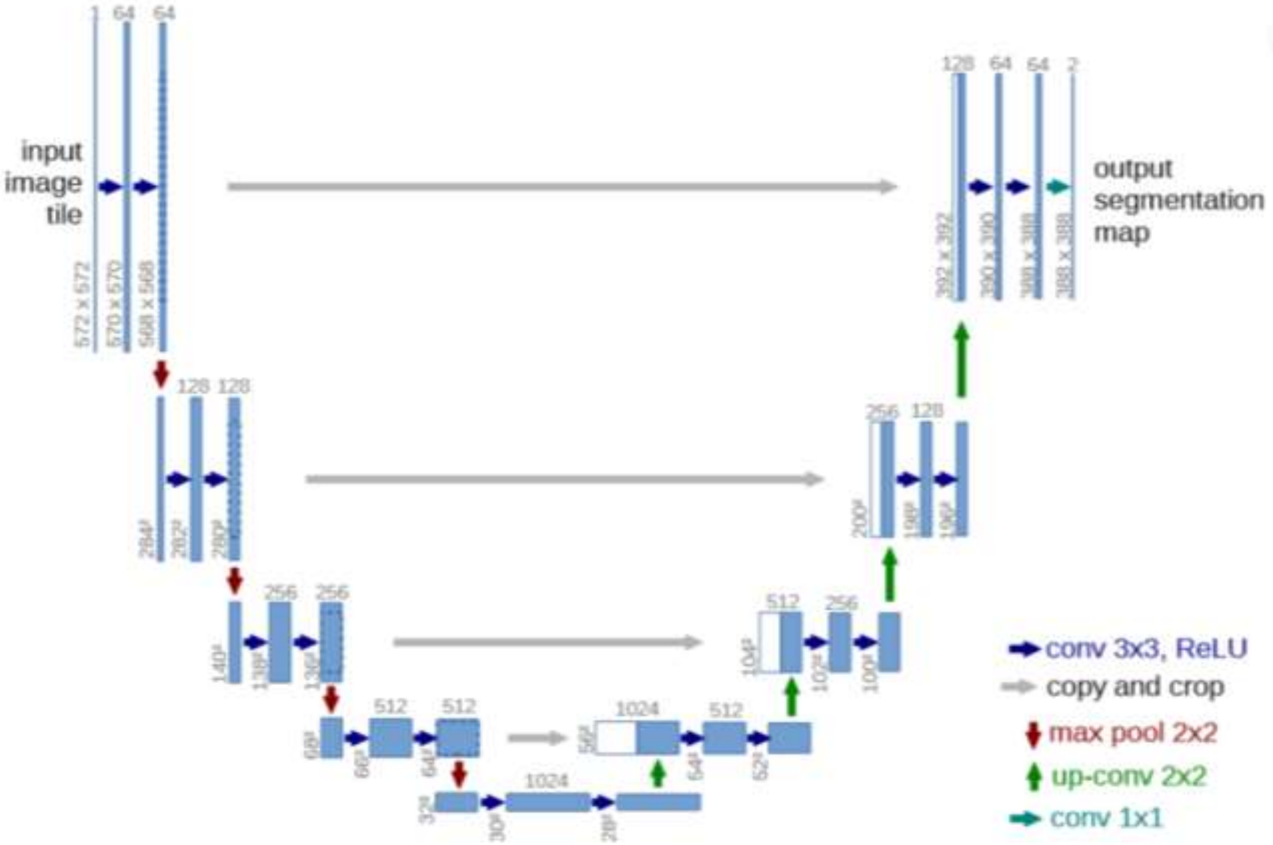
- This paper proposes a novel medical image segmentation network called EPT-Net, which hybrids improved Transformer and parameter-free attention mechanisms for modeling long-term dependencies of organ features.
- Propose a Dual Positional Embedding Transformer, including Learnable Positional Embedding and Voxel Spacial Positional Embedding. This method is used to optimize location coding, which can effectively capture the intrinsic correlation between different organ locations of medical images.
- Develop an Edge Weight Guidance Module to learn edge information in shallow features, which can capture tiny adhesions between neighboring organs. This design is to minimize the edge information function without increasing the network parameters.
- Validate the effectiveness and robustness of EPT-Net on three datasets, including the SegTHOR 2019, the Multi-Atlas Labeling Beyond the Cranial Vault, and the re-labeled KiTS19. Experiments demonstrate our method outperform state-of-the-art methods on these datasets.

2.Related work

- It is a gradual development process from manual segmentation to semi-automatic segmentation, and fully automatic segmentation.
- **Manual segmentation** means that experienced clinicians directly outline the boundary of the tissue on the original film image or use an image editor to use the mouse to outline the boundary of the tissue or the area of interest on the computer monitor.
- **Semi-automatic segmentation** technology is produced with the development of computer science. It organically combines the powerful data processing, algorithm analysis, and intelligent calculation capabilities of the computer and the automatic storage and memory functions with the knowledge and experience of medical experts. The image segmentation is done interactively.
- **Fully automatic segmentation** is entirely free of human intervention, and the entire process of image segmentation is completed automatically by a computer. However, most automatic segmentation algorithms are complicated and require extensive calculation.
- In many cases, the segmentation results are not yet ideal, and the segmentation speed and performance need to be improved.

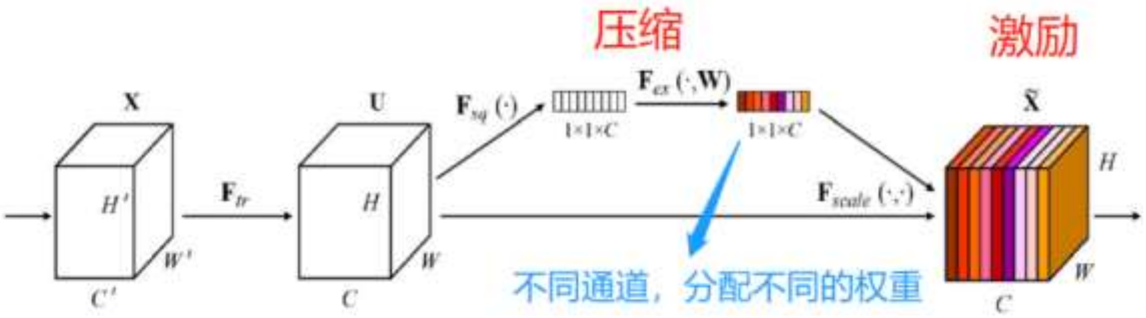
A.CNN-Based

- Since the advent of deep learning, **FCNNs** and in particular **encoder-decoder** architectures have achieved state-of-the-art results on various medical semantic segmentation task.
- 2D-based methods achieved promising performance on several 2D medical image segmentation tasks, **When applied to 3D medical images**, these 2D models perform the segmentation task in a **slice-by-slice** manner, and hence cannot capture inter-slice context information, leading to limited segmentation accuracy.



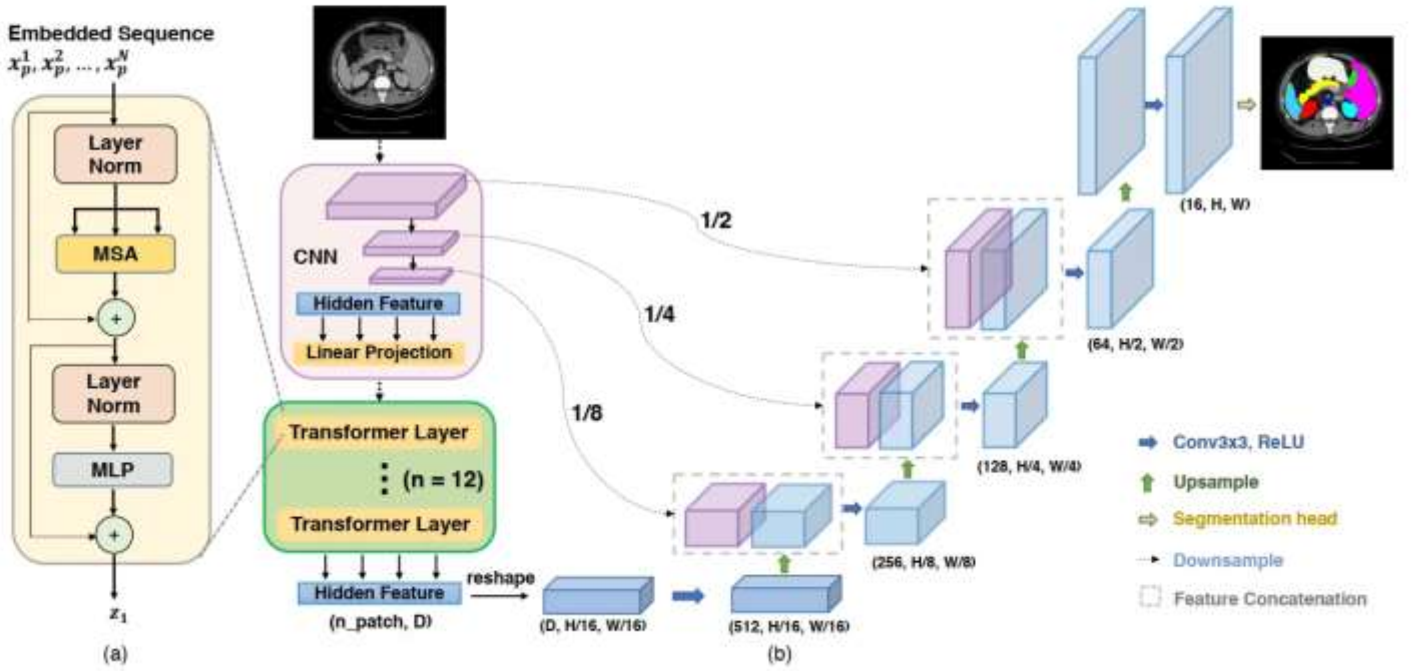
B. Attention mechanism

- There are different ways to achieve different types of attention, and there are three main modules: channel domains, spatial domains, and mixed domains.
- For channel domain attention mechanisms, it plays a vital role in image segmentation . In detail, it connects the relationship between different features, and the new dimensional information is injected into the network to boot the characteristics of the batch note. For the neural network, which conducts feature learning, adequate feature information will be better, and invalid feature information will be ignored and suppressed.



C. Transformer

- Medt applies the Transformer structure to the encoder for medical image segmentation and proposes a local-global (LoGo) training strategy . CoTr network based on CNN and Transformer for 3D medical image segmentation. This framework effectively bridges the convolutional neural network and Transformer and plays a very influential role in 3D medical image segmentation.
- TransBTS applies the Transformer in 3D CNN to the brain tumor segmentation field of MRI, which is a deep learning network composed of codecs .
- UTNet proposes a self-attention mechanism combined with positional encoding and applies it to convolutional neural networks to enhance medical image segmentation performance.
- U-NET Transformer combines the self-attention mechanism and cross-attention in the transformer network with the encoder and jump connection of the U-network, which solves the problem that the traditional network cannot model longrange context and spatial dependence.



D. Edge Information

- Most existing medical image segmentation methods focus on raw region extraction, ignoring edge information. Effective learning of edge information helps to obtain accurate object segmentation.
- Generally, **shallow layers** in the encoding path have richer detailed, and less semantic information. To effectively utilize edge information as a low-level feature,
- ET-Net designs an edge guidance module with an attention mechanism so that it utilizes edge information to monitor and guide the segmentation process . AEC-Net introduces an attention mechanism to learn edge and texture features in the encoding path simultaneously .
- MEA-Net generates new edge feature maps in the early stage of the network through-edge feature extraction and then uses multi-layer attention guidance to combine different individual feature maps with attention mechanisms to screen richer edge information.

3.Method: EPT-Net

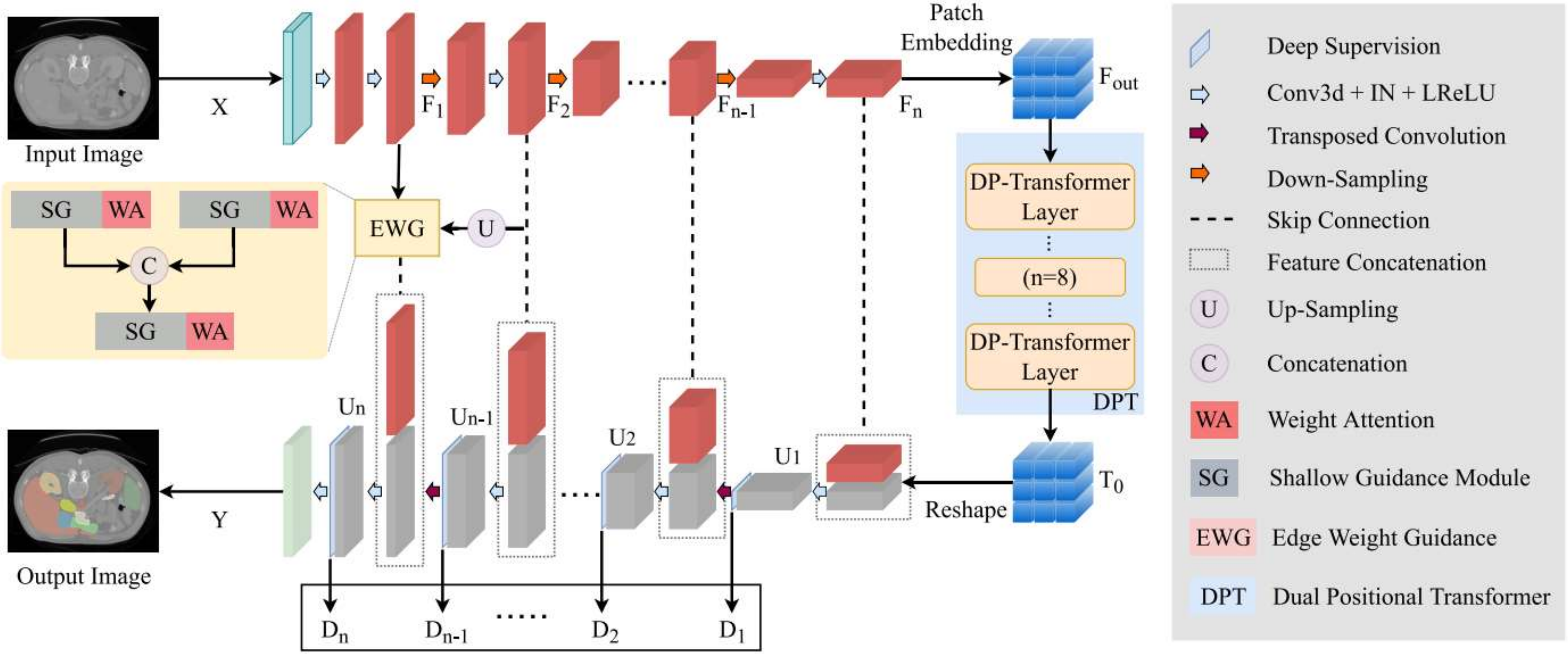
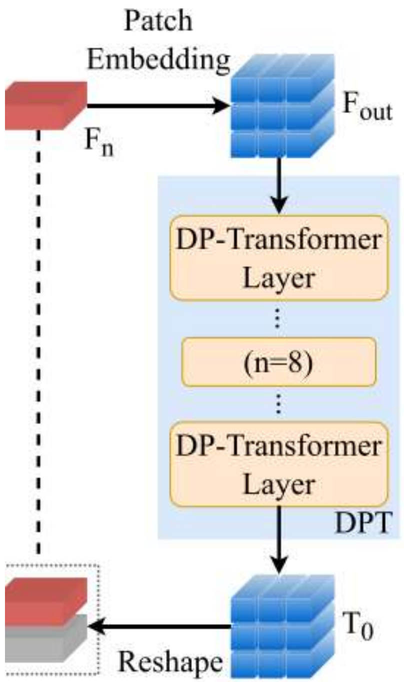


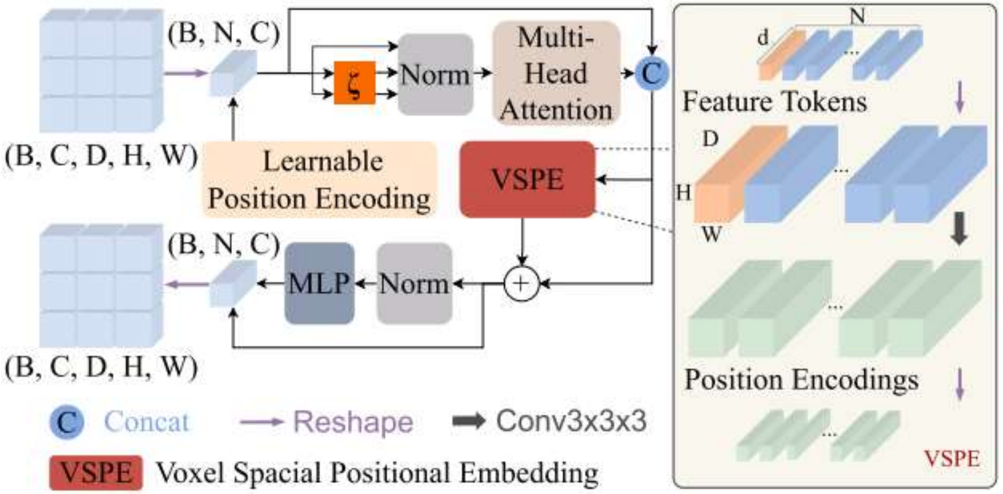
Fig. 3. The structure of EPT-Net. The network consists of the encoder-decoder structure, and the DPT module is in the last layer of the encoder to focus on the global information of features. The EWG module takes the features of the first two layers of the encoder as input and fuses the extracted fine-grained edge information with the decoder through a skip connection operation. D_i is the deep supervision part, which is used to accelerate the network convergence.

A. Dual positional transformer(DPT)

(1) Learnable Patch Embedding

- ViT cuts image into a fixed number of patch and obtains an operable patch sequence through linear encoding .
- All patch blocks in the sequence have no intersection. The patch serialization methods will lose some local information.
- The adjacent patch blocks are ensured to have a specific interactive part in the patch serialization process. Then, the number of patch blocks is not fixed and can be flexibly changed according to the size of the input data X.





A. Dual positional transformer(DPT)

(2) Voxel Spacial Positional Embedding(VSPE)

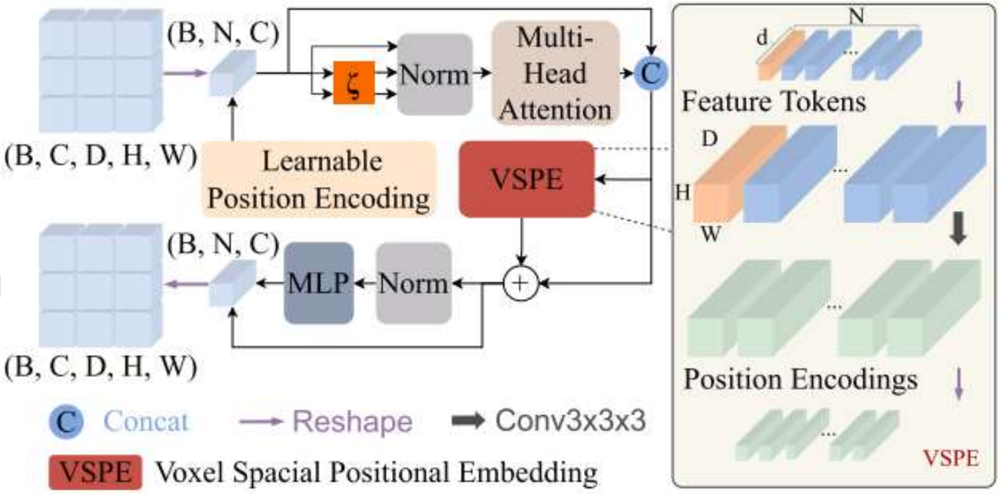
- Learnable position information is provided for the patch sequence before Self-Attention. Without position information, Self-Attention will be difficult to work.
- The core operation of Transformer is Self-Attention. The input vector X ′ ∈ R N×C (where N = D × H × W) of the encoder extracts new features through mutual attention. SelfAttention Layer is realized as follow:

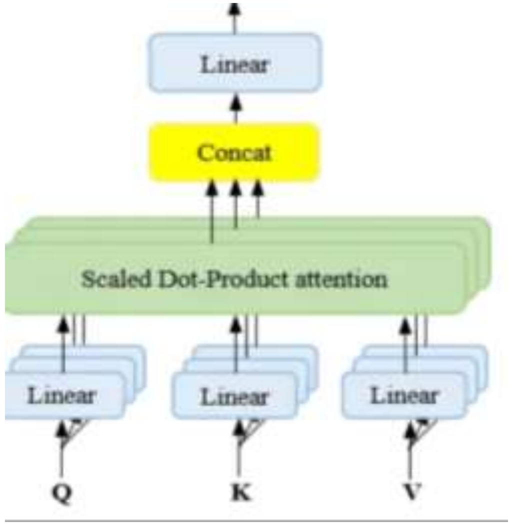
$$Att(Q,K,V) = Softmax\left(\frac{Q\cdot K^T}{\sqrt{n_{head}}}\right)\cdot V,$$

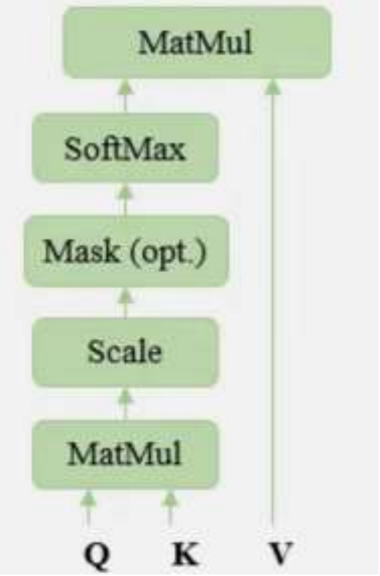
where $Q = W_Q \cdot X', K = W_K \cdot X', V = W_V \cdot X'$.

- The computational complexity of the method is O(N^2), For the sequence of patch blocks X ′ ∈ R N×C, we use the operation ζ to reshape the input sequence. Mak $W'_K \in \mathbb{R}^{(\sigma^3\cdot C)\times C}$ and $W'_V \in \mathbb{R}^{(\sigma^3\cdot C)\times C}$.

the computational complexity is reduced to O($\frac{N^2}{\sigma}$).



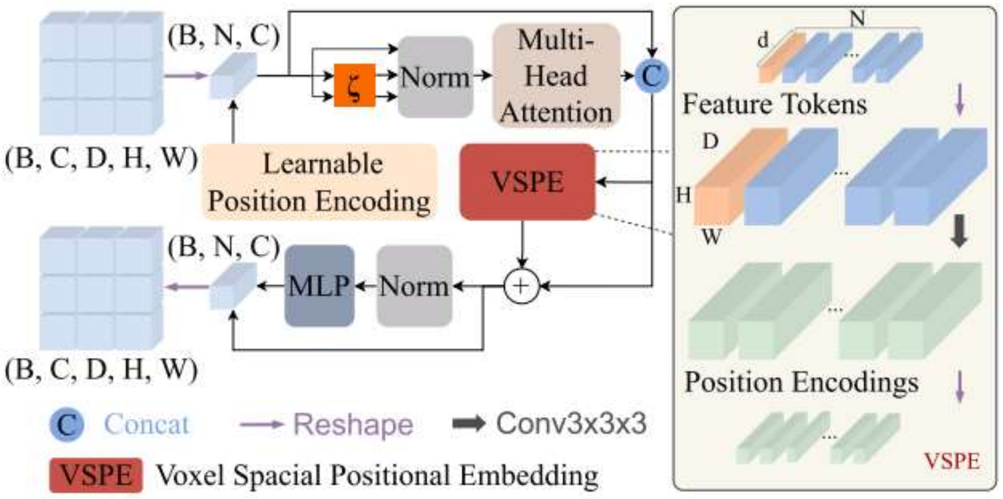




(2) Voxel Spacial Positional Embedding(VSPE)

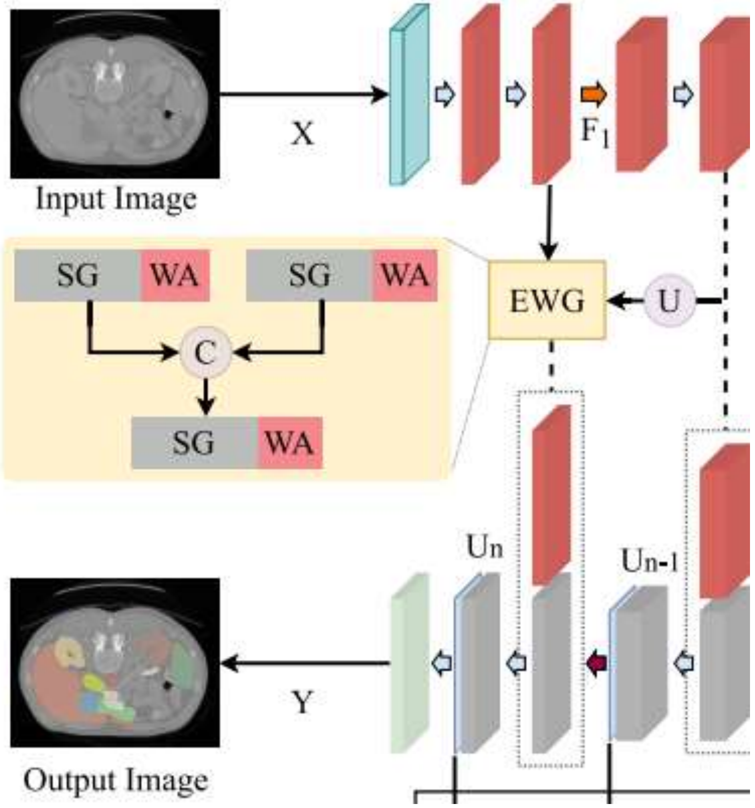
- SA is followed by the **Feed-Forward Network (FFN).** We consider providing additional axial position information in the FFN. The additional position information can help Transformer focus on the connections between different slices.
- The proposed VSPEcan generate the position codes containing multi-dimensional position information. Then the certain position information with the output of multi-head attention is put into the MLP together. The process is shown in Fig. 4. The VSPE can be realized by 3 × 3 × 3 DWConv with the kernel size k = 3, stride s = 1, and zero padc l is follov y

$$FFN(x) = MLP(LeakyReLU(VSPE(x) + x)).$$

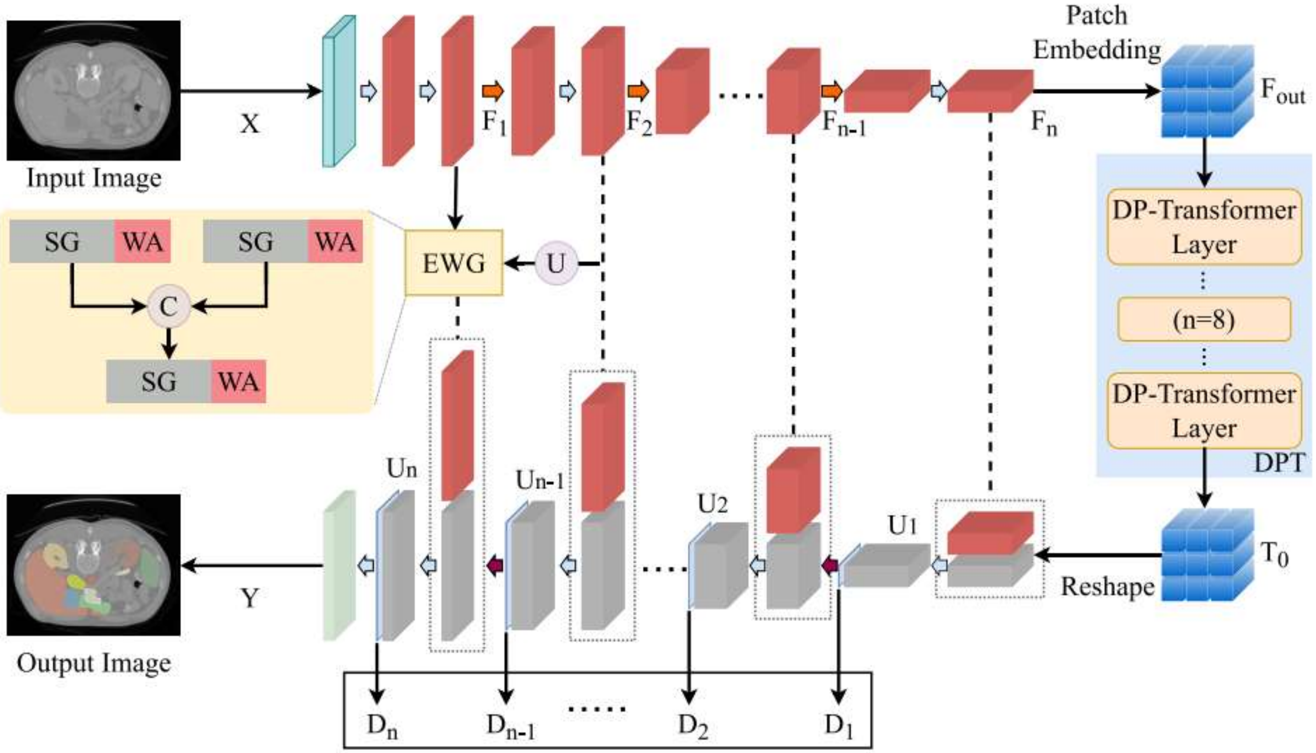


B. Edge Weight Guidance Module(EWG)

- **Semantic information** is usually divided into **shallow** and **deep** features. The shallow layers can obtain more detailed information, such as texture and boundary, while the deep layers have more abstract classification information.
- It is necessary to extract shallow edge information, and the concatenation of shallow and deep features can better guide network training.
- The Edge Weight Guidance module is proposed to guide the network to learn edge information. It consists of a **Shallow Guidance (SG)** module for extracting edge detail features and a Weighted Attention **(WA) module** for attention to edge weights.



B. Edge Weight Guidance Module(EWG)



(1) Shallow Guidance Module(SG)

In order to extract the edge information retained in the low-level features, only the early two layers of the encoder are operated, named D1 and D2 . The output of D2 is up-sampled to the exact resolution as the output of D1, then fed into a3 × 3 × 3 convolution layer and concatenated together. After that, the cascade features go through a 1 × 1 × 1 convolution layer to act as the shallow guidance features in the decoding path.

SG provides predicted edge detection results for early supervision.

(2) Weighted Attention Module

In WA, the **priority of each feature point** or the **importance of each neuron** needs to be evaluated. the priority function of each feature point is defined as:

$$\hat{p}_t(\omega_t,\ b_t) = (x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2 + (z_t - \hat{z}_t)^2 + \frac{1}{N-1}\sum_{i=1}^{N-1}[(x_0 - \hat{x}_i)^2 + (y_0 - \hat{y}_i)^2 + (z_0 - \hat{z}_i)^2],$$

the **weight** can be obtained by:

$$\omega_t = -\frac{2(x_t - m_i)}{(x_t - 2v_i^2 + m_i^2 + 2\xi)}. \qquad m_i = \tfrac{1}{M-1}\textstyle\sum_{i=1}^{N-1} x_i \qquad v_i = \tfrac{1}{N-1}\textstyle\sum_{i=1}^{N-1}(x_i - m_t)^2$$

According to the above derivation, the bias can be computed $b_t = -\frac{1}{2}(x_t + m_i)\omega_t.$

the minimum Eq. 3 can be calculated as follows: $\hat{p}_t = \frac{4(\hat{v}^2 + \xi)}{(x_t - \hat{m})^2 + 2\hat{v} + 2\xi},$

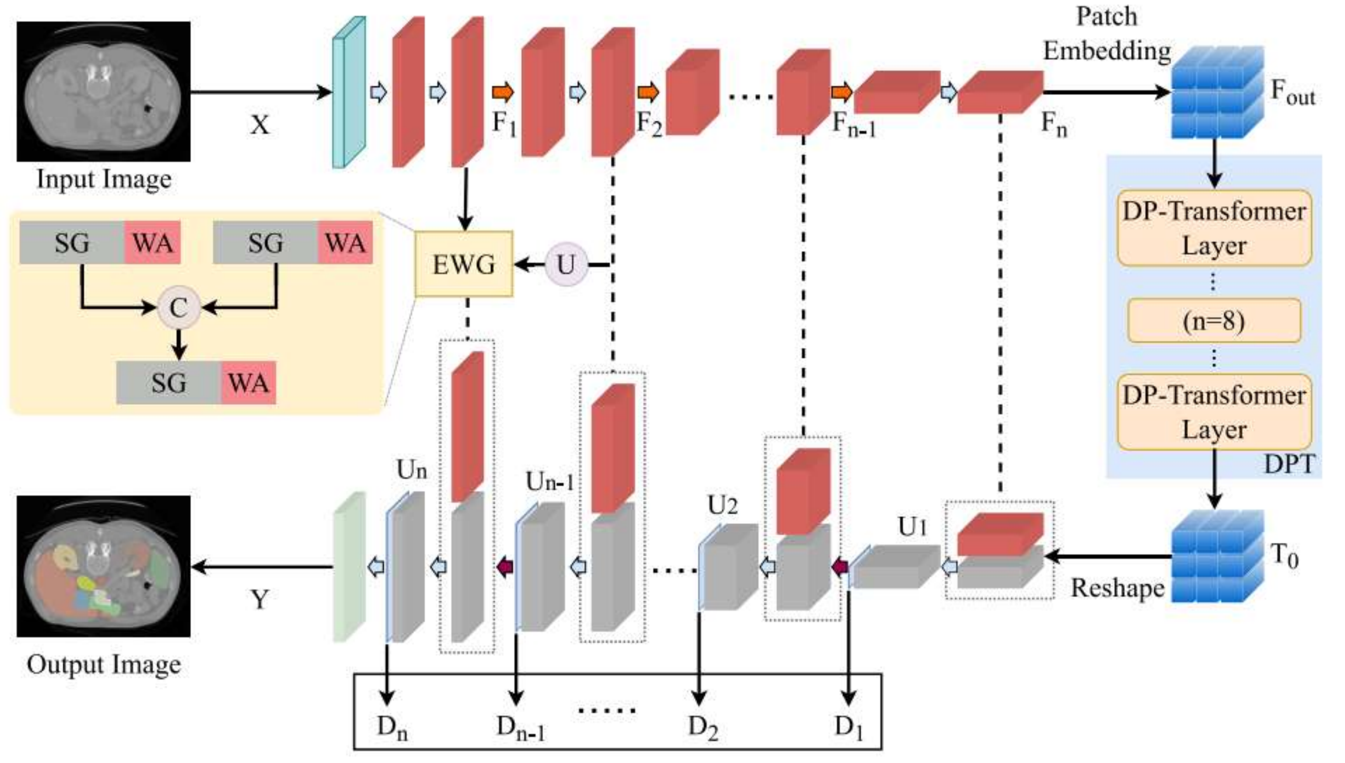
$$\hat{m} = \tfrac{1}{M-1}\textstyle\sum_{i=1}^{N-1} x_i \text{ and } \hat{v}^2 = \tfrac{1}{N-1}\textstyle\sum_{i=1}^{N-1}(x_i - \hat{m})^2$$

The final whole optimization is:

$$\tilde{X} = X \cdot sigmoid(P),$$

where X represents the set of all feature points. P combines all pˆt through channel and spatial dimensions, and sigmoid(·) is used to limit the value range of P.

- In this way, **WA and SG are combined to generate EA**. The WA can prioritize the whole feature map and can be easily combined with three-dimensional convolution. Inserting the WA-module into the SG module after each convolution operation can yield more representative edge features. Here, the extraction of edge information from shallow features is completed.



C. EPT-Net for Segmentation

- A 3D medical image segmentation learning architecture named EPT-Net is proposed. The nnUNet is taken as the backbone framework of semantic segmentation. Fig. 3 shows that EPT-Net is mainly composed of DP-Transformer encoder, EWG module, and CNNs decoder.
- **In the encoder stage**, the EPT-Net takes image X ∈ R D×H×W with a spatial resolution of D × H × W ×C and C the number of channels. The image X is fed into the encoder module to extract deep semantic features through the backbone network. In the SegTHOR 2019 dataset , the backbone **3**D-Unet **performs five convolution down-sampling** to extract depth feature information, and each layer is connected with an adaptive 3D convolution layer after down-sampling. At the same time, the output F1 and F2 of the first two layers in the encoding stage are selected to extract edge information. They are used as inputs to the SG. WA is added after each convolution of the SG. Finally, the extracted edge information is obtained for fusion with the upsampling feature information. $F_{out} = C_{en}(I),$ **eature of the absolute encoder is represented as**:

- The feature information at the bottom layer is taken as the input of DPT. Here, **Transformer is used to replace the original 3 × 3 × 3 convolution layer.** The size of the input feature entering the DPT is the same as the output size. The output of DPT is defined as:

$$T_0 = DPT(F_{out}).$$

- **In the training stage**, the deep supervision strategy is used to optimize the problems such as the disappearance of training gradient and slow convergence speed of the deep neural network and plays a specific role in regularization. It is defined as Di . The total loss function is a combination of Dice loss and cross-entropy loss, and it can be computed in a voxel-wise manner according to:

$$\mathcal{L}_{total}(Y, G) = 1 - \frac{\vartheta}{J}\sum_{j=1}^{J}\frac{2\sum_{i=1}^{I} Y_{i,j}G_{i,j}}{\sum_{i=1}^{I} Y_{i,j}^2 + \sum_{i=1}^{I} G_{i,j}^2} - \frac{\hat{\vartheta}}{J}\sum_{j=1}^{J}\sum_{i=1}^{I} G_{i,j}\log Y_{i,j},$$

Experiments

A. Datasets

- Three data sets are selected for experimental verification, including two of the public datasets **SegTHOR 2019** , **Multi-Atlas Labeling Beyond the Cranial Vault (BCV)** and a relabeled KiTS19 dataset, called **KiTS19-M**.
- The **SegTHOR 2019** dataset specifically used to segment the dangerous organs in the thoracic cavity around the tumor during radiotherapy, including the heart(hea), tracheak(tra), aorta(aor), and esophagus(eso) . The dataset includes 60 CT scans, divided into 40 training sets and 20 test sets. Experienced radiation therapists have manually drawn the heart, trachea, aorta and esophagus.
- **MultiAtlas Labeling Beyond the Cranial Vault** contains 13 labels of abdominal organs, including the spleen(spl), left and right kidneys(L-K and R-K), gallbladder(gal), esophagus(eso), liver(liv), and stomach(sto), aorta(aor), inferior vena cava (I-V-C), portal vein and splenic vein(P-V/S-V), pancreas(pan), right and left adrenal gland(R-A-G and L-A-G), hand-labeled by two experienced college students and radiologists. This dataset contains 30 labeled CT images and 20 test images.

A. Datasets

The **KiTS19-M** dataset comes from the KiTS19 dataset kidney tumor segmentation competition. It contains 300 samples, including 210 training samples and 90 test samples. In order to promote our research on multi-organ segmentation, we re-label multiple organs in the dataset. After relabeling by experienced experts, we have finely labeled four abdominal organs including liver, spleen, right and left kidney.

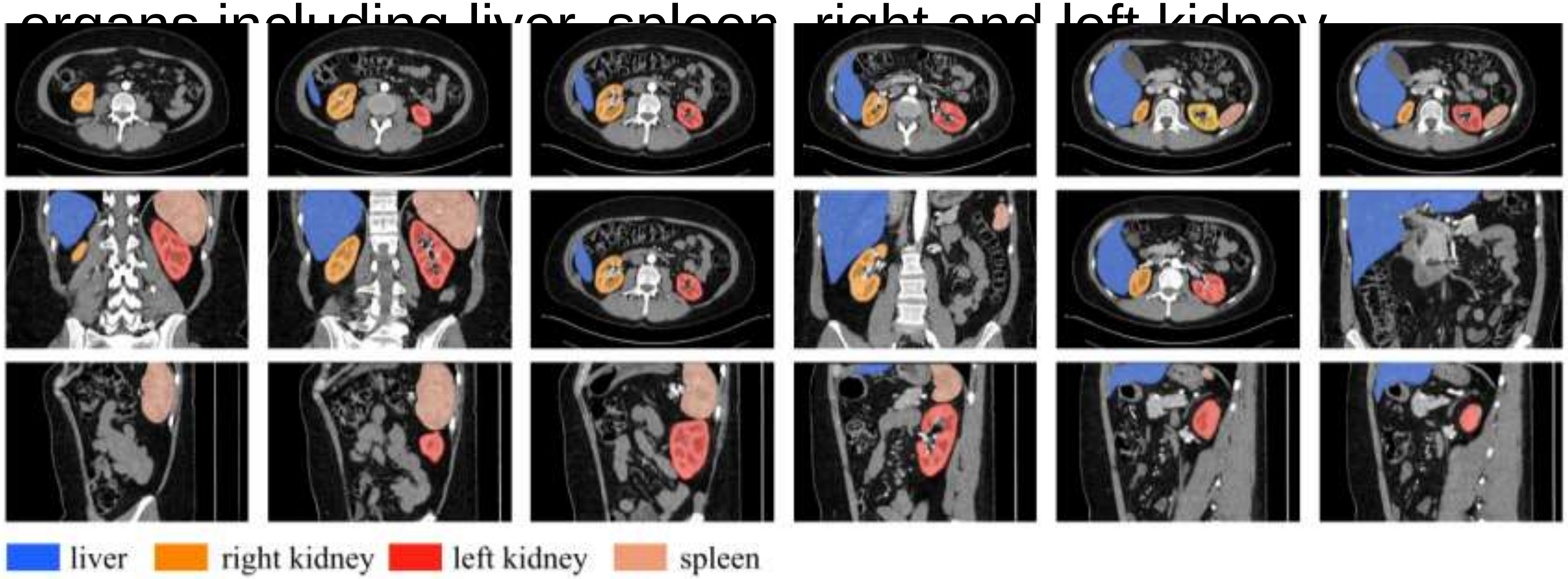


Fig. 2. KiTS19-M dataset. A slice example of the same case as Fig. 1 on KiTS19-M. In the original KiTS19 data, the liver, left kidney, right kidney, and spleen are labeled.

B. Implementation Details

- In order to solve the problem that the actual space size of a single voxel of data is inconsistent, a resampling operation is performed in the preprocessing stage.
- At the same time, to make reasonable use of the memory space, the image is cropped to the main area containing the label.
- Cropping the surrounding area that does not have a label can effectively improve the calculation efficiency.
- Preprocessing the image for normalization is also an important step. First, collect the foreground information of the entire training set. Then, based on the mean and standard deviation of the entire training set, z-scoreing normalization is performed to ensure that the gray values of each image can have a similar distribution.
- To alleviate overfitting during training, data augmentation is performed through a series of operations, including rotation, zooming, Gaussian noise, Gaussian blur, brightness processing, and contrast processing

B. Implementation Details

- After preprocessing, the data are randomly cropped as input. Due to the limit of memory, the batch size b = 2 is set, and the method of instance normalization is adopted. This small-batch training method is more suitable for medical data training.
- The total number of training epochs is 1000, and each epoch has 250 iterations. The loss function is a combination of Dice loss and cross-entropy loss. A stochastic gradient descent algorithm is used as the optimization algorithm (where the momentum is 0.99, and the initial learning rate is 0.01). The decline of the learning rate follows the poly principle followed as: l = linit ∗ (1 − epochindex /epochmax) 0.9 . In this paper, our model is trained using 2 RTX 3090 with 24GB memory each. The system version is Ubuntu 16.04 LTS, and the platform version is PyTorch 1.7.0.

C. Hyperparameter Settings

1) Hyperparameter ξˆ Setting:

The hyperparameter ξˆ is not a robust variable in the energy function and significantly impacts the segmentation performance.

The Weighted Attention module performs edge segmentation with the help of fine-grained features guided by the Shallow Guidance module.

Table I demonstrates the validity of Weighted Attention under different ξˆ in the KITS19-M dataset, with optimal accuracy achieved when ξˆ = 0.9.

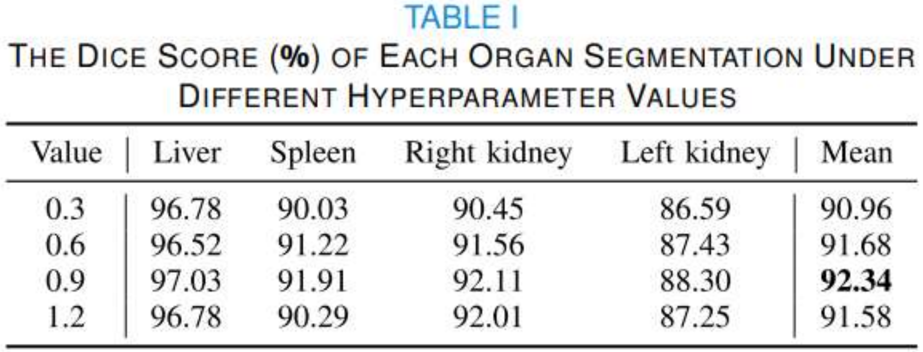


TABLE I

THE DICE SCORE (%) OF EACH ORGAN SEGMENTATION UNDER DIFFERENT HYPERPARAMETER VALUES

Value	Liver	Spleen	Right kidney	Left kidney	Mean
0.3	96.78	90.03	90.45	86.59	90.96
0.6	96.52	91.22	91.56	87.43	91.68
0.9	97.03	91.91	92.11	88.30	**92.34**
1.2	96.78	90.29	92.01	87.25	91.58

C. Hyperparameter Settings

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}$$

2) Input Setting of DPT: In EPT-Net, the DPT extracts depth feature information to improve edge subdivision.

Table II compares the network segmentation accuracy, memory occupation, and running time, respectively.

The results shown that the mean Dice score for the segmentation is 92.12% when the input size is 4 × 6 × 5. Increasing the input size to 8 × 12 × 10 and 16 × 24 × 40 improved the Dice scores by 0.12 and 0.36 points, respectively. Howerver, memory is increase by 34.1% and 69.1%, and time improve by 19.7% and 79%.

DPT consumes less memory than the CNN when the input feature size is around 4 × 6 × 5 and accelerates the convergence of the network

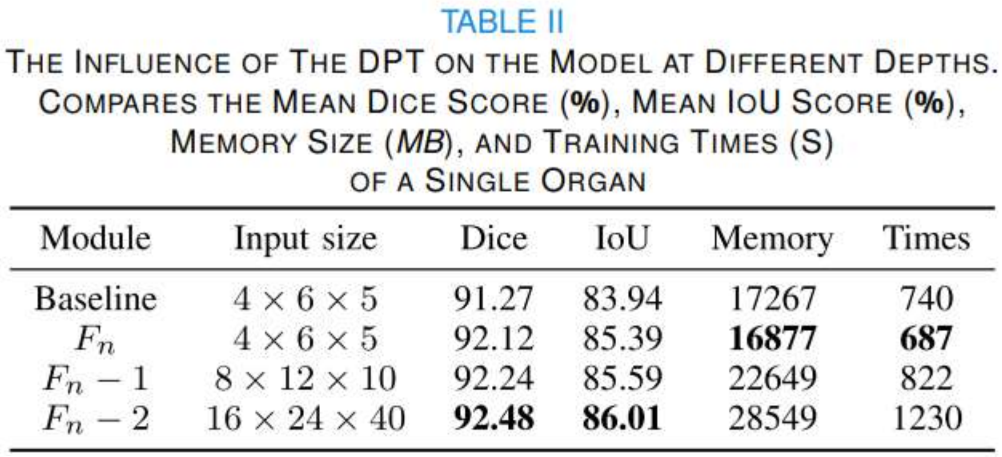


TABLE II

THE INFLUENCE OF THE DPT ON THE MODEL AT DIFFERENT DEPTHS. COMPARES THE MEAN DICE SCORE (%), MEAN IOU SCORE (%), MEMORY SIZE (*MB*), AND TRAINING TIMES (S) OF A SINGLE ORGAN

Module	Input size	Dice	IoU	Memory	Times
Baseline	$4 \times 6 \times 5$	91.27	83.94	17267	740
F_n	$4 \times 6 \times 5$	92.12	85.39	**16877**	**687**
$F_n - 1$	$8 \times 12 \times 10$	92.24	85.59	22649	822
$F_n - 2$	$16 \times 24 \times 40$	**92.48**	**86.01**	28549	1230

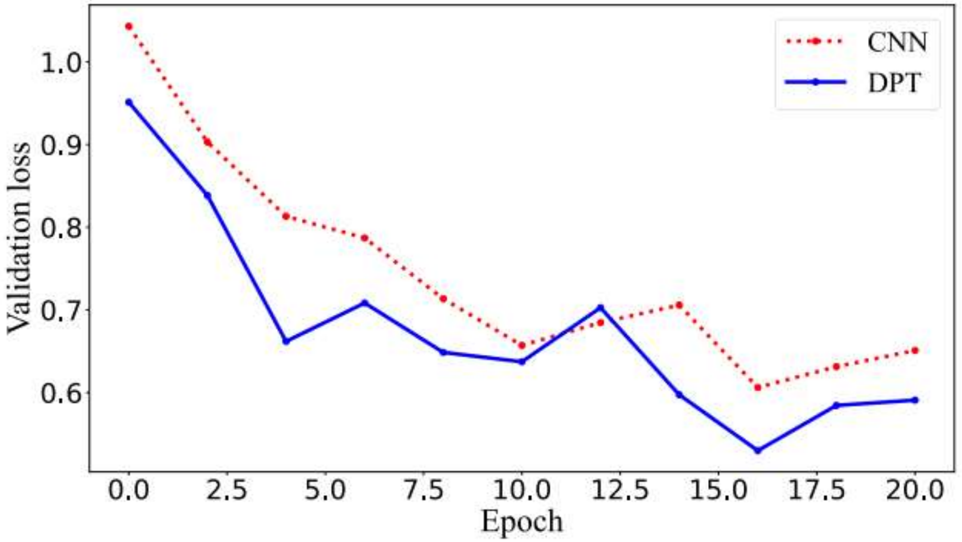
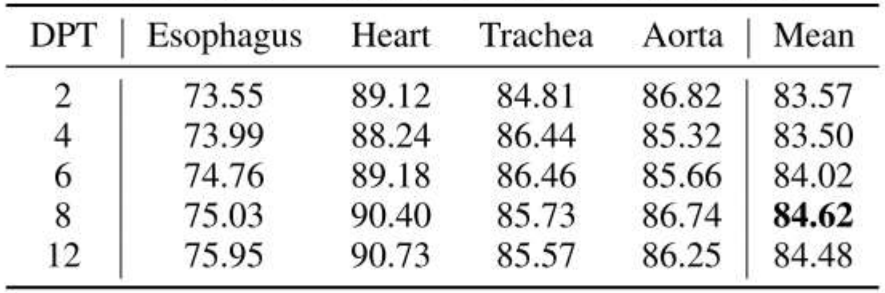


Fig. 5. The curve shows the effectiveness of DPT. The blue curve represents the decreasing trend of the validation loss of the network under the DPT module. The red curve represents the influence of CNN in the same setting.

D. Ablation Study

- A series of ablation studies on the dataset of SegTHOR 2019 and KiTS19-M have been conducted. During ablation experiments, we set the batch size to 1 and the total number of training epochs to 100.
- 1) DPT Evaluation: In order to compare the impact of different layers of Transformer on the model performance, the number of layers of 2, 4, 6, 8, and 12 is seted, respectively.
- As shown in Table III, stacking Transformer can improve the results of accuracy. The experimental results show that the best performance can be achieved when the number of DPT layers is 8.
- Fig. 5 compares the decreasing loss trend on the verification set between DPT and CNN during training. Compared with CNN, DPT is more conducive to network convergence

TABLE III
THE PERFORMANCE OF THE NETWORK UNDER DIFFERENT DPT LAYER NUMBERS ON THE SEGTHOR 2019 DATASET



DPT	Esophagus	Heart	Trachea	Aorta	Mean
2	73.55	89.12	84.81	86.82	83.57
4	73.99	88.24	86.44	85.32	83.50
6	74.76	89.18	86.46	85.66	84.02
8	75.03	90.40	85.73	86.74	**84.62**
12	75.95	90.73	85.57	86.25	84.48

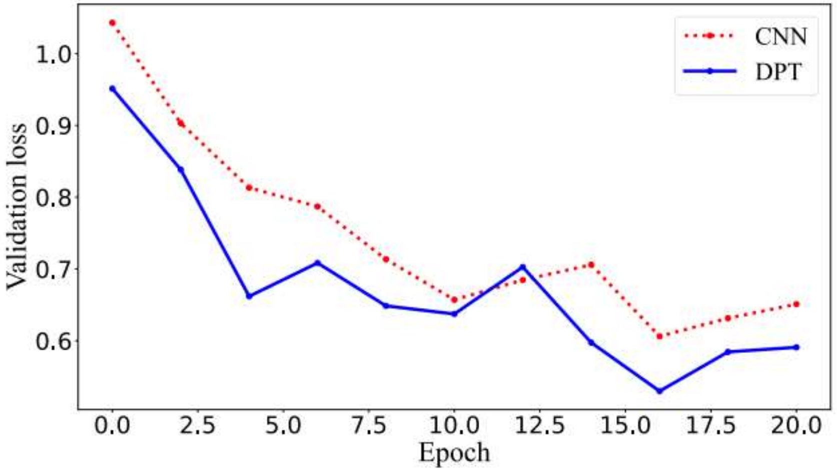


Fig. 5. The curve shows the effectiveness of DPT. The blue curve represents the decreasing trend of the validation loss of the network under the DPT module. The red curve represents the influence of CNN in the same setting.

1) DPT Evaluation

- In order to verify the effectiveness of the VSPE module, the experiments set up different position schemes of the VSPE module on the DPT. Table IV shows that the segmentation accuracy obtained with VSPE is higher than without it.
- In order to verify the influence of the convolution kernel size in the VSPE module on network feature extraction, we set the convolution kernels 5 × 5 × 5, 7 × 7 × 7, 3 × 3 × 3. Table IV shows that the size of the convolution kernels used for feature extraction in the VSPE module does not greatly affect the segmentation results. Moreover, using 3 × 3 × 3 convolution consumes the least network parameters.

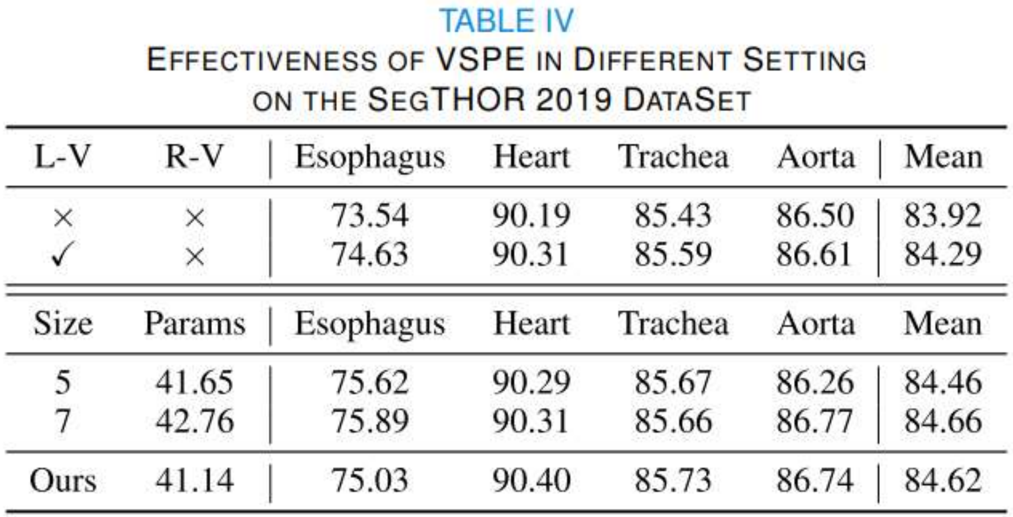


TABLE IV
EFFECTIVENESS OF VSPE IN DIFFERENT SETTING
ON THE SEGTHOR 2019 DATASET

L-V	R-V	Esophagus	Heart	Trachea	Aorta	Mean
×	×	73.54	90.19	85.43	86.50	83.92
✓	×	74.63	90.31	85.59	86.61	84.29
Size	Params	Esophagus	Heart	Trachea	Aorta	Mean
5	41.65	75.62	90.29	85.67	86.26	84.46
7	42.76	75.89	90.31	85.66	86.77	84.66
Ours	41.14	75.03	90.40	85.73	86.74	84.62

D. Ablation Study

2) Module Evaluation

- The DPT can effectively model long-range information through dual location information. EPT-Net make better use of the multi-dimensional spatial information of 3D medical data. The EWG can provide more refined edge features, which helps pay attention to the boundaries of multiple organs and get smoother boundary segmentation results.
- Table V shows the experimental results of each module. The average Dice and IoU score of single and all organs in 100 epochs under each module is calculated. DPT helps to integrate low-level and high-level features better to get segmentation results.

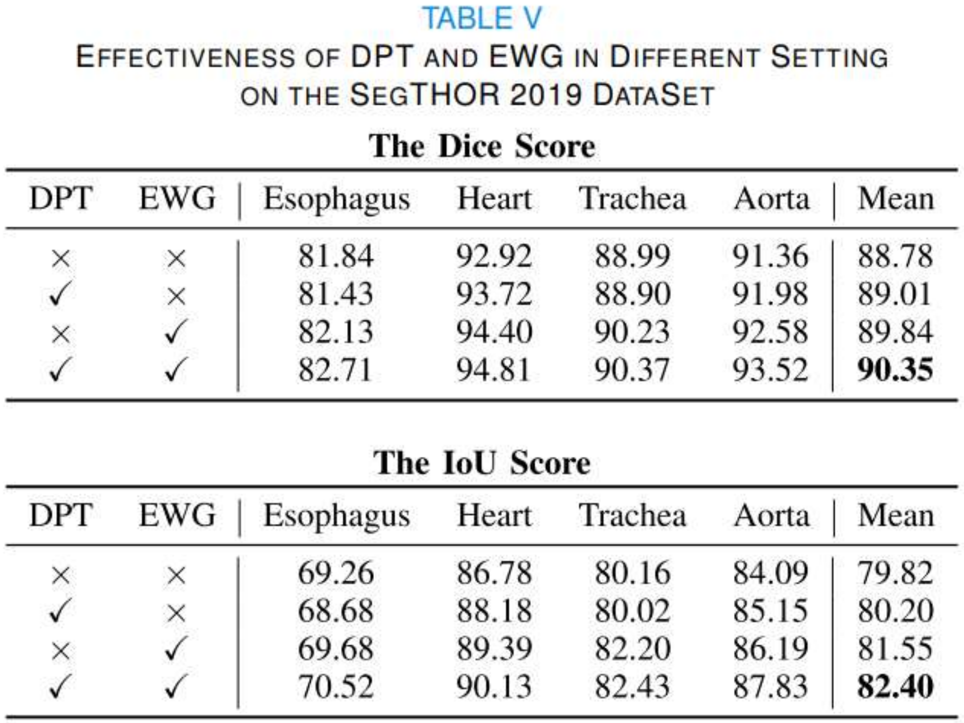


TABLE V
EFFECTIVENESS OF DPT AND EWG IN DIFFERENT SETTING ON THE SEGTHOR 2019 DATASET

The Dice Score

DPT	EWG	Esophagus	Heart	Trachea	Aorta	Mean
×	×	81.84	92.92	88.99	91.36	88.78
✓	×	81.43	93.72	88.90	91.98	89.01
×	✓	82.13	94.40	90.23	92.58	89.84
✓	✓	82.71	94.81	90.37	93.52	**90.35**

The IoU Score

DPT	EWG	Esophagus	Heart	Trachea	Aorta	Mean
×	×	69.26	86.78	80.16	84.09	79.82
✓	×	68.68	88.18	80.02	85.15	80.20
×	✓	69.68	89.39	82.20	86.19	81.55
✓	✓	70.52	90.13	82.43	87.83	**82.40**

2) Module Evaluation

- Compared with the baseline, the predicted Dice and IoU score in the experiment has limited improvement in accuracy, and the visualizationeffect of edge segmentation is not obvious.
- As shown in Fig. 6, the brightness of the image represents the priority of feature points, and the brighter the pixel point, the higher the priority here. It can be seen that the most prominent bright spots are often on the boundary line of the organ and the boundary line between the organ and the background. Therefore, with the help of DPT and EWG, the network can focus on organ and edge information more efficiently.
- As shown in Fig. 7, the curves of different colors represent the segmentation results of different methods Obviously, compared with the baseline, the network after mixing the DPT and EWG modules is the closest to the edge segmentation results of the true value,

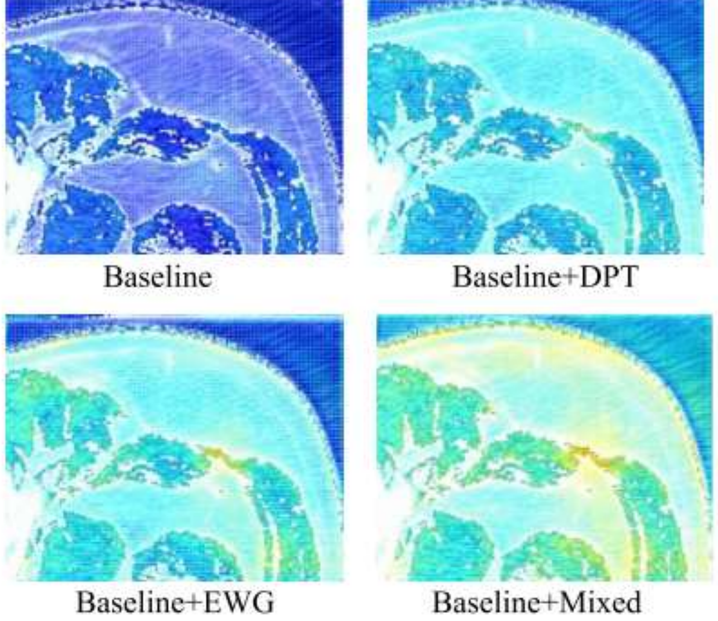


Fig. 6. An illustration of the thermodynamic map from the same patch perspective under different networks. The dark blue in the figure represents the feature point with a smaller weight, and the bright yellow is the area with enormous weight. Areas with more brilliant colours also represent areas of greater interest to the network.It can be seen from the figure that after the integration of DPT and EWG, the network has significantly improved its attention to the boundary of organs.

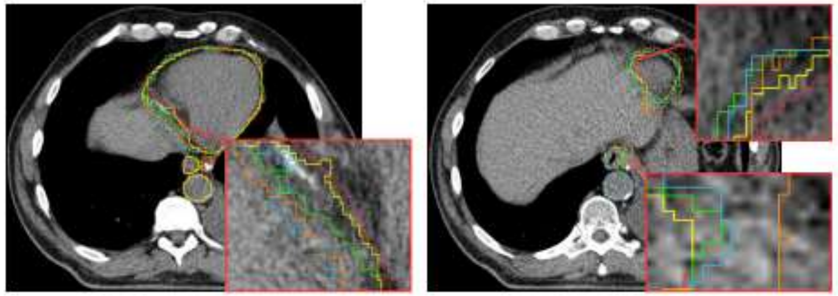


Fig. 7. The edge segmentation results of DPT and EWG modules are visualized in the figure. In the figure, the red curve is the true value, and the brown curve is the baseline, the blue curve is the network after adding DPT, the green curve is the network after adding EWG, and the yellow curve is the network after mixing DPT and EWG. The left image shows the segmentation results of different methods in large organs, and the right image visualizes the different segmentation results in small organs.

E. Comparison With State-of-the-Arts Methods

1) Results on SegTHOR 2019

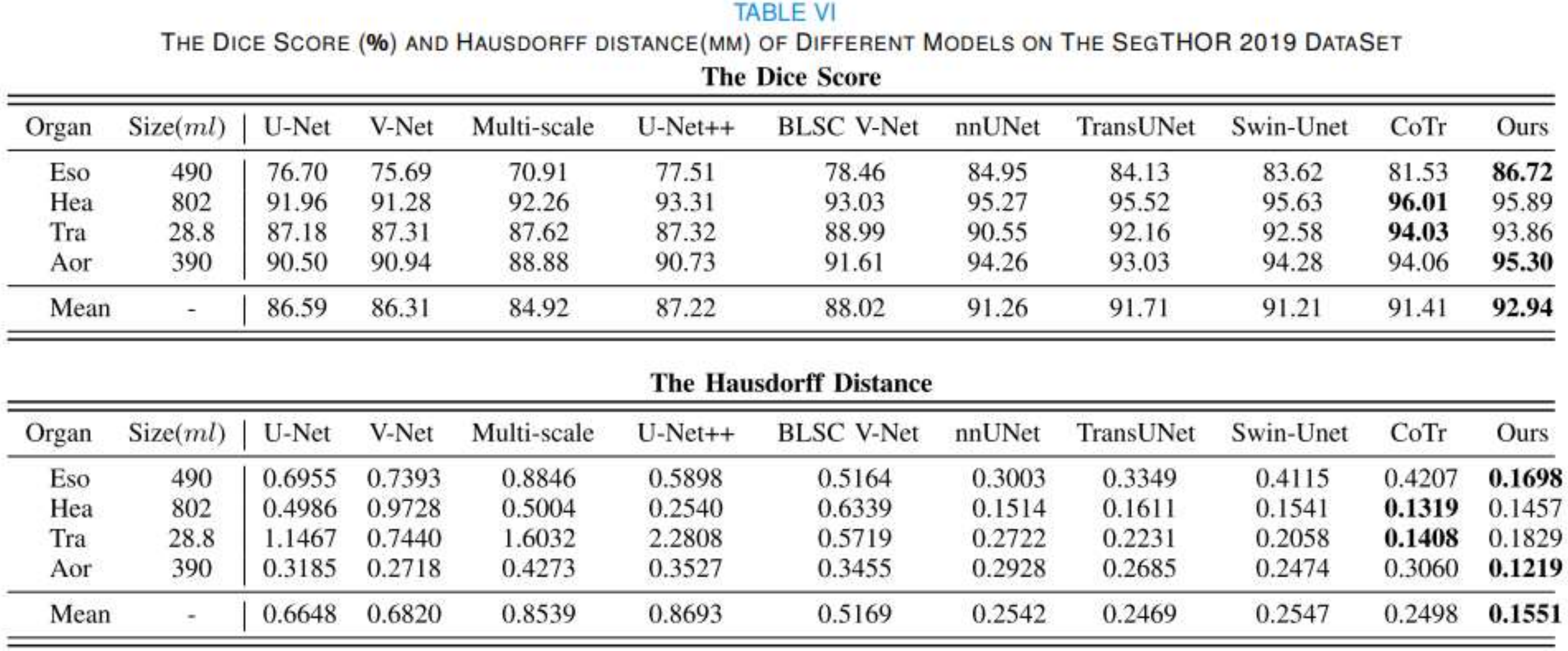


TABLE VI

THE DICE SCORE (%) AND HAUSDORFF DISTANCE(MM) OF DIFFERENT MODELS ON THE SEGTHOR 2019 DATASET

The Dice Score

Organ	Size(ml)	U-Net	V-Net	Multi-scale	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Eso	490	76.70	75.69	70.91	77.51	78.46	84.95	84.13	83.62	81.53	**86.72**
Hea	802	91.96	91.28	92.26	93.31	93.03	95.27	95.52	95.63	**96.01**	95.89
Tra	28.8	87.18	87.31	87.62	87.32	88.99	90.55	92.16	92.58	**94.03**	93.86
Aor	390	90.50	90.94	88.88	90.73	91.61	94.26	93.03	94.28	94.06	**95.30**
Mean	-	86.59	86.31	84.92	87.22	88.02	91.26	91.71	91.21	91.41	**92.94**

The Hausdorff Distance

Organ	Size(ml)	U-Net	V-Net	Multi-scale	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Eso	490	0.6955	0.7393	0.8846	0.5898	0.5164	0.3003	0.3349	0.4115	0.4207	**0.1698**
Hea	802	0.4986	0.9728	0.5004	0.2540	0.6339	0.1514	0.1611	0.1541	**0.1319**	0.1457
Tra	28.8	1.1467	0.7440	1.6032	2.2808	0.5719	0.2722	0.2231	0.2058	**0.1408**	0.1829
Aor	390	0.3185	0.2718	0.4273	0.3527	0.3455	0.2928	0.2685	0.2474	0.3060	**0.1219**
Mean	-	0.6648	0.6820	0.8539	0.8693	0.5169	0.2542	0.2469	0.2547	0.2498	**0.1551**

$$h(A,B) = \max_{a \in A}\{\min_{b \in B} ||a-b||\}$$

Transformer-based models have higher average Dice, in which long-range modeling can more effectively improve the attention to edge information.

2) Results on Multi-Atlas Labeling Beyond the Cranial Vault

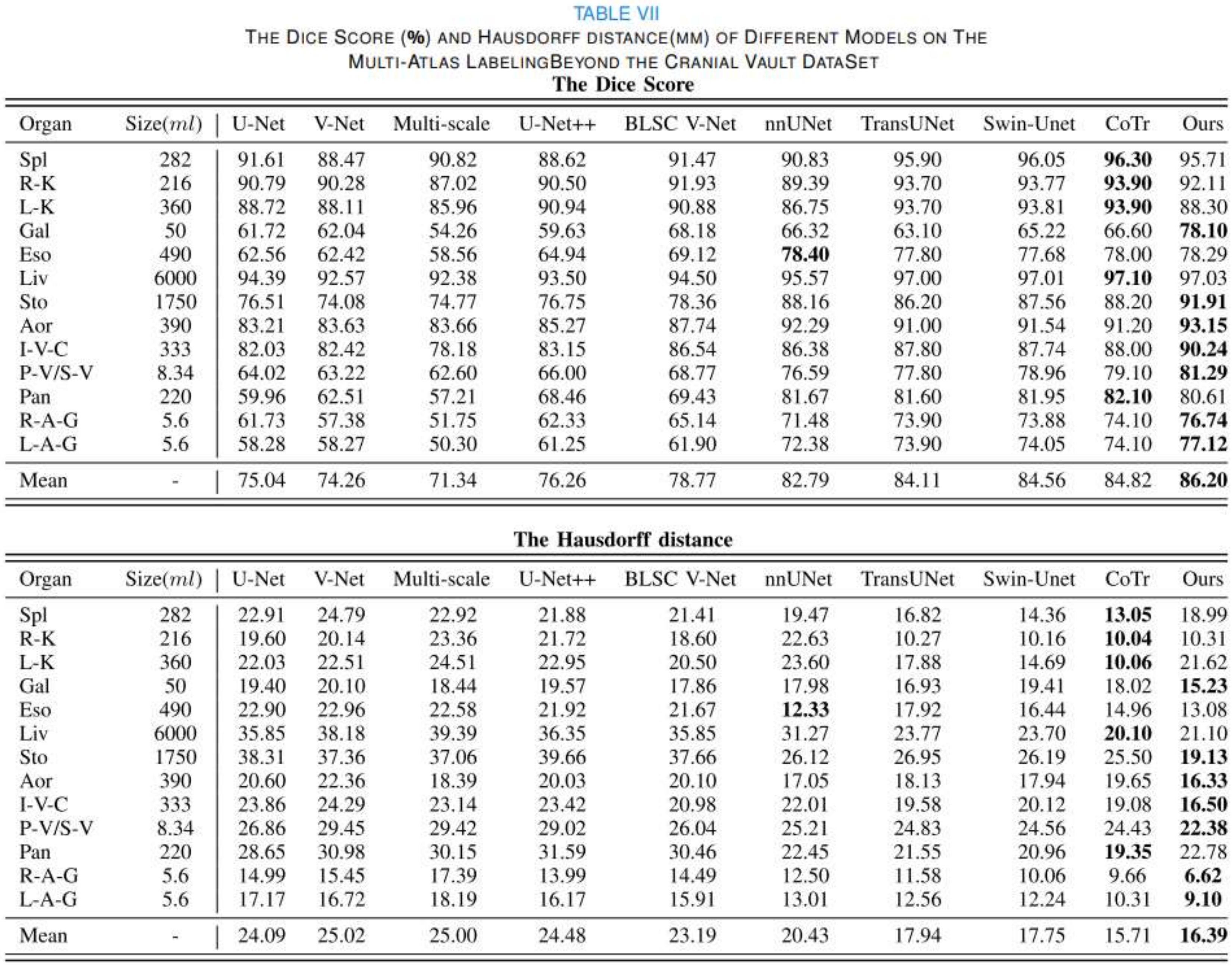


TABLE VII

THE DICE SCORE (%) AND HAUSDORFF DISTANCE(MM) OF DIFFERENT MODELS ON THE MULTI-ATLAS LABELINGBEYOND THE CRANIAL VAULT DATASET

The Dice Score

Organ	Size(ml)	U-Net	V-Net	Multi-scale	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Spl	282	91.61	88.47	90.82	88.62	91.47	90.83	95.90	96.05	**96.30**	95.71
R-K	216	90.79	90.28	87.02	90.50	91.93	89.39	93.70	93.77	**93.90**	92.11
L-K	360	88.72	88.11	85.96	90.94	90.88	86.75	93.70	93.81	**93.90**	88.30
Gal	50	61.72	62.04	54.26	59.63	68.18	66.32	63.10	65.22	66.60	**78.10**
Eso	490	62.56	62.42	58.56	64.94	69.12	**78.40**	77.80	77.68	78.00	78.29
Liv	6000	94.39	92.57	92.38	93.50	94.50	95.57	97.00	97.01	**97.10**	97.03
Sto	1750	76.51	74.08	74.77	76.75	78.36	88.16	86.20	87.56	88.20	**91.91**
Aor	390	83.21	83.63	83.66	85.27	87.74	92.29	91.00	91.54	91.20	**93.15**
I-V-C	333	82.03	82.42	78.18	83.15	86.54	86.38	87.80	87.74	88.00	**90.24**
P-V/S-V	8.34	64.02	63.22	62.60	66.00	68.77	76.59	77.80	78.96	79.10	**81.29**
Pan	220	59.96	62.51	57.21	68.46	69.43	81.67	81.60	81.95	**82.10**	80.61
R-A-G	5.6	61.73	57.38	51.75	62.33	65.14	71.48	73.90	73.88	74.10	**76.74**
L-A-G	5.6	58.28	58.27	50.30	61.25	61.90	72.38	73.90	74.05	74.10	**77.12**
Mean	-	75.04	74.26	71.34	76.26	78.77	82.79	84.11	84.56	84.82	**86.20**

The Hausdorff distance

Organ	Size(ml)	U-Net	V-Net	Multi-scale	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Spl	282	22.91	24.79	22.92	21.88	21.41	19.47	16.82	14.36	**13.05**	18.99
R-K	216	19.60	20.14	23.36	21.72	18.60	22.63	10.27	10.16	**10.04**	10.31
L-K	360	22.03	22.51	24.51	22.95	20.50	23.60	17.88	14.69	**10.06**	21.62
Gal	50	19.40	20.10	18.44	19.57	17.86	17.98	16.93	19.41	18.02	**15.23**
Eso	490	22.90	22.96	22.58	21.92	21.67	**12.33**	17.92	16.44	14.96	13.08
Liv	6000	35.85	38.18	39.39	36.35	35.85	31.27	23.77	23.70	**20.10**	21.10
Sto	1750	38.31	37.36	37.06	39.66	37.66	26.12	26.95	26.19	25.50	**19.13**
Aor	390	20.60	22.36	18.39	20.03	20.10	17.05	18.13	17.94	19.65	**16.33**
I-V-C	333	23.86	24.29	23.14	23.42	20.98	22.01	19.58	20.12	19.08	**16.50**
P-V/S-V	8.34	26.86	29.45	29.42	29.02	26.04	25.21	24.83	24.56	24.43	**22.38**
Pan	220	28.65	30.98	30.15	31.59	30.46	22.45	21.55	20.96	**19.35**	22.78
R-A-G	5.6	14.99	15.45	17.39	13.99	14.49	12.50	11.58	10.06	9.66	**6.62**
L-A-G	5.6	17.17	16.72	18.19	16.17	15.91	13.01	12.56	12.24	10.31	**9.10**
Mean	-	24.09	25.02	25.00	24.48	23.19	20.43	17.94	17.75	15.71	**16.39**

- For dense organ segmentation tasks, CNN is easy to overfit during the training process, making it difficult to improve the segmentation accuracy.
- EPT-Net achieves the highest score of 86.2%, which is 1.38 and 3.41 points higher than CoTr and nnUNet, respectively.
- The shortest Hausdorff distance of EPT-Net is 16.39, which is 0.09 and 4.04 lower than CoTr and nnUNet.
- the segmentation accuracy of EPT-Net for small size organs is greatly improved, but the improvement of segmentation accuracy for large size organs is not obvious.

3) Results on KiTS19-M

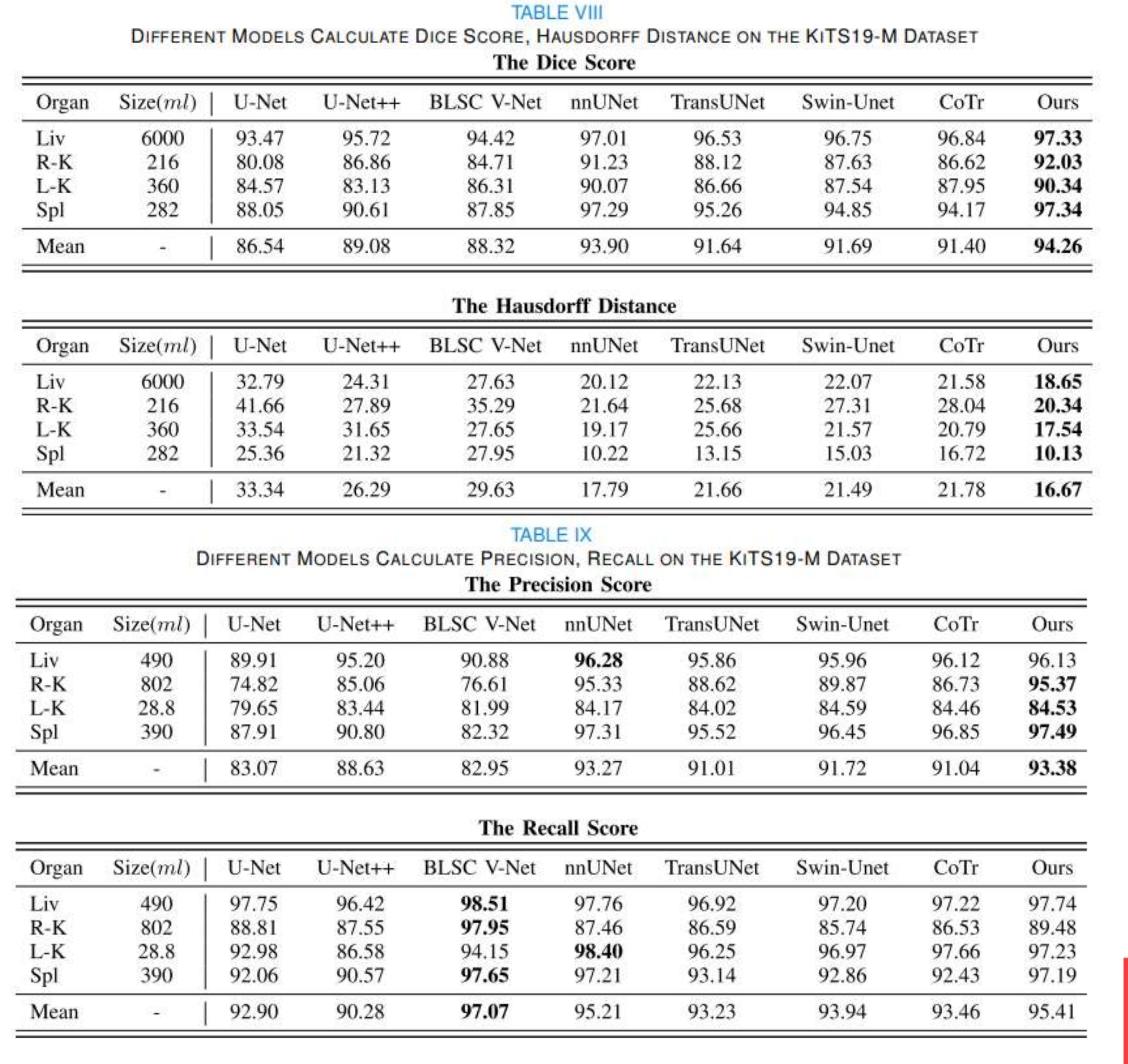


TABLE VIII

DIFFERENT MODELS CALCULATE DICE SCORE, HAUSDORFF DISTANCE ON THE KITS19-M DATASET

The Dice Score

Organ	Size(ml)	U-Net	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Liv	6000	93.47	95.72	94.42	97.01	96.53	96.75	96.84	**97.33**
R-K	216	80.08	86.86	84.71	91.23	88.12	87.63	86.62	**92.03**
L-K	360	84.57	83.13	86.31	90.07	86.66	87.54	87.95	**90.34**
Spl	282	88.05	90.61	87.85	97.29	95.26	94.85	94.17	**97.34**
Mean	-	86.54	89.08	88.32	93.90	91.64	91.69	91.40	**94.26**

The Hausdorff Distance

Organ	Size(ml)	U-Net	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Liv	6000	32.79	24.31	27.63	20.12	22.13	22.07	21.58	**18.65**
R-K	216	41.66	27.89	35.29	21.64	25.68	27.31	28.04	**20.34**
L-K	360	33.54	31.65	27.65	19.17	25.66	21.57	20.79	**17.54**
Spl	282	25.36	21.32	27.95	10.22	13.15	15.03	16.72	**10.13**
Mean	-	33.34	26.29	29.63	17.79	21.66	21.49	21.78	**16.67**

TABLE IX

DIFFERENT MODELS CALCULATE PRECISION, RECALL ON THE KITS19-M DATASET

The Precision Score

Organ	Size(ml)	U-Net	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Liv	490	89.91	95.20	90.88	**96.28**	95.86	95.96	96.12	96.13
R-K	802	74.82	85.06	76.61	95.33	88.62	89.87	86.73	**95.37**
L-K	28.8	79.65	83.44	81.99	84.17	84.02	84.59	84.46	**84.53**
Spl	390	87.91	90.80	82.32	97.31	95.52	96.45	96.85	**97.49**
Mean	-	83.07	88.63	82.95	93.27	91.01	91.72	91.04	**93.38**

The Recall Score

Organ	Size(ml)	U-Net	U-Net++	BLSC V-Net	nnUNet	TransUNet	Swin-Unet	CoTr	Ours
Liv	490	97.75	96.42	**98.51**	97.76	96.92	97.20	97.22	97.74
R-K	802	88.81	87.55	**97.95**	87.46	86.59	85.74	86.53	89.48
L-K	28.8	92.98	86.58	94.15	**98.40**	96.25	96.97	97.66	97.23
Spl	390	92.06	90.57	**97.65**	97.21	93.14	92.86	92.43	97.19
Mean	-	92.90	90.28	**97.07**	95.21	93.23	93.94	93.46	95.41

- Table VIII shows that the Hausdorff distance of EPT-Net is lower than other networks.
- It can be concluded from the above results that our prominent model has strong versatility, can deal with various medical image segmentation problems, and help effectively reconstruct the three-dimensional model.
- EPT-Net has the most noticeable improvement in the segmentation accuracy of the spleen and right kidney, and the sizes of these two organs are relatively small, which also verifies that our method has a better segmentation effect for small-sized organs.
- We believe that the enhanced 3D spatial localization capability of DPT helps the network to capture and segment small organs.

4) Results on Edge Information Based

As shown in Table X, our method achieves the best segmentation results, with 7.72 points improvement in the dice score and 16.67 points improvement in the Hausdorff distance score. This may be because the edge guidance module proposed by ET-Net and the attention mechanism introduced by AEC-Net at the encoder stage do not sufficiently learn the boundary information between multiple organs, while EPT-Net can learn more edge representations. Such results demonstrate that our method has good predictive performance and generalization ability in edge-based segmentation methods.

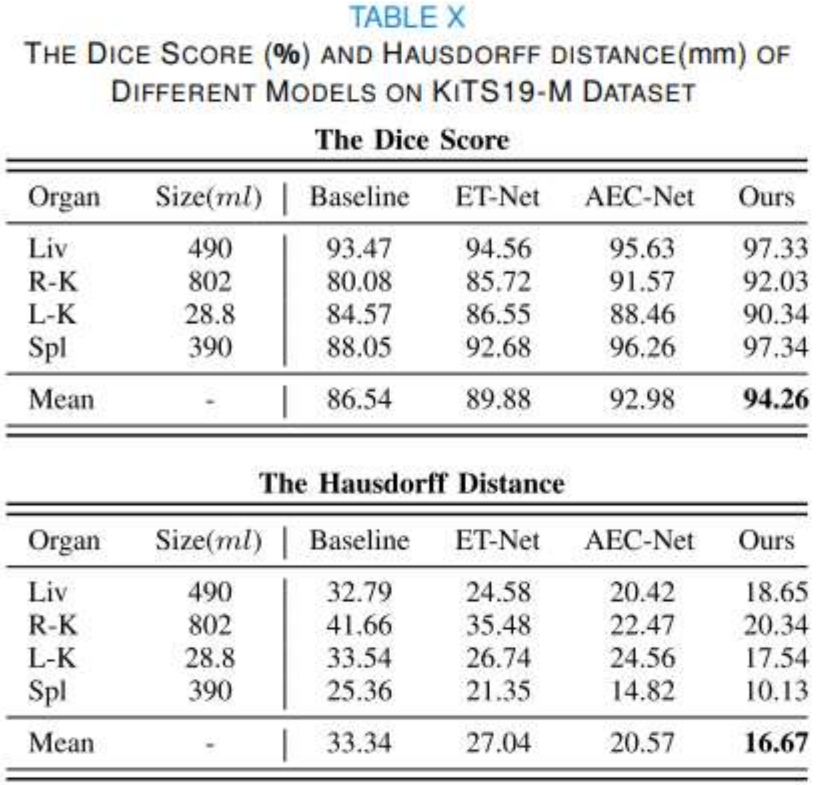


TABLE X
THE DICE SCORE (%) AND HAUSDORFF DISTANCE(mm) OF DIFFERENT MODELS ON KITS19-M DATASET

The Dice Score

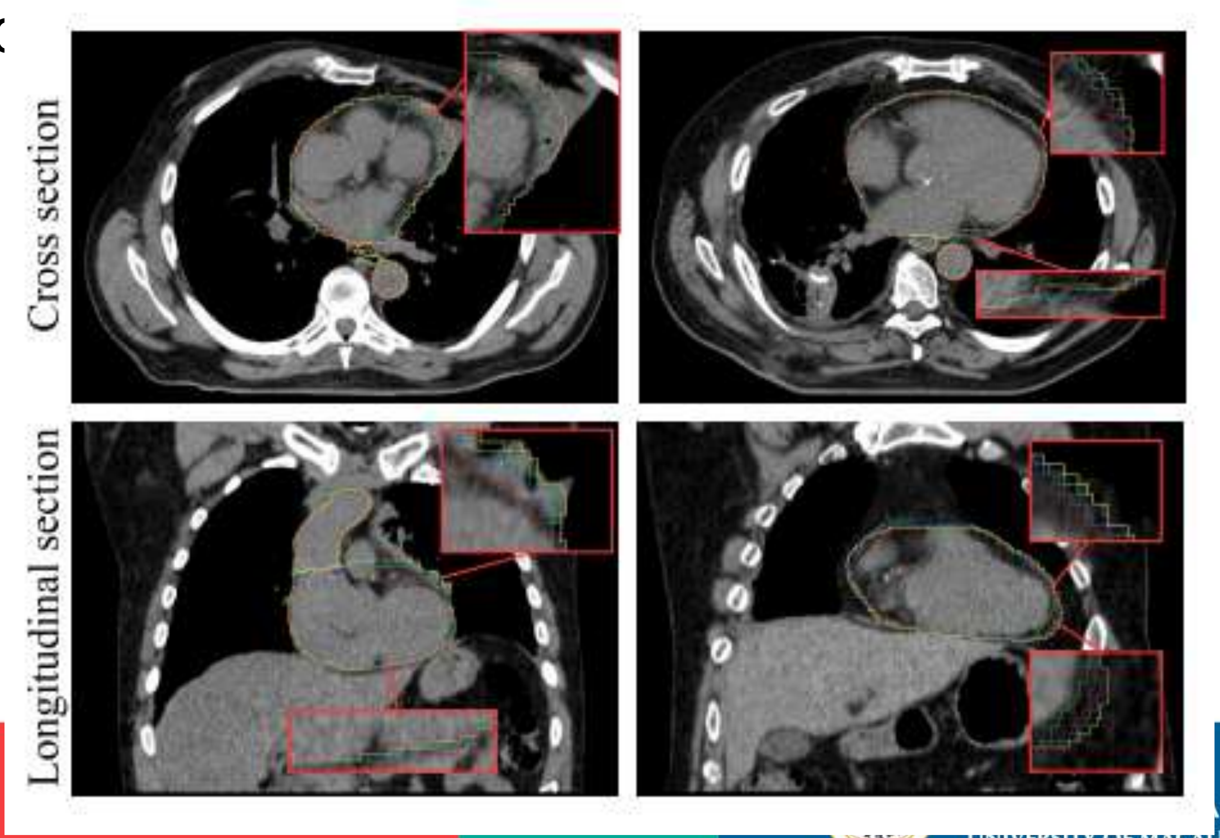
Organ	Size(ml)	Baseline	ET-Net	AEC-Net	Ours
Liv	490	93.47	94.56	95.63	97.33
R-K	802	80.08	85.72	91.57	92.03
L-K	28.8	84.57	86.55	88.46	90.34
Spl	390	88.05	92.68	96.26	97.34
Mean	-	86.54	89.88	92.98	**94.26**

The Hausdorff Distance

Organ	Size(ml)	Baseline	ET-Net	AEC-Net	Ours
Liv	490	32.79	24.58	20.42	18.65
R-K	802	41.66	35.48	22.47	20.34
L-K	28.8	33.54	26.74	24.56	17.54
Spl	390	25.36	21.35	14.82	10.13
Mean	-	33.34	27.04	20.57	**16.67**

F. Qualitative Results

- In Fig. 8 is a visualization of the segmentation results of several edge-based methods.
- The first row is the cross-section of different situations. The second row is the longitudinal section of different situations. The red line represents the ground truth. The blue line represents our method. The green line represents AEC-Net, and the yellow line represents the baseline. It can be seen from the two different views that the segmentation boundaries predicted by



- In Fig. 9, the inference results on the SegTHOR 2019 test set are shown, and we visualize the difference between the inference results of some models and Ground Truth. The red boxes show areas with significant differences from the Ground Truth. The green circles represent the difference between our method and the Ground Truth in the same region.

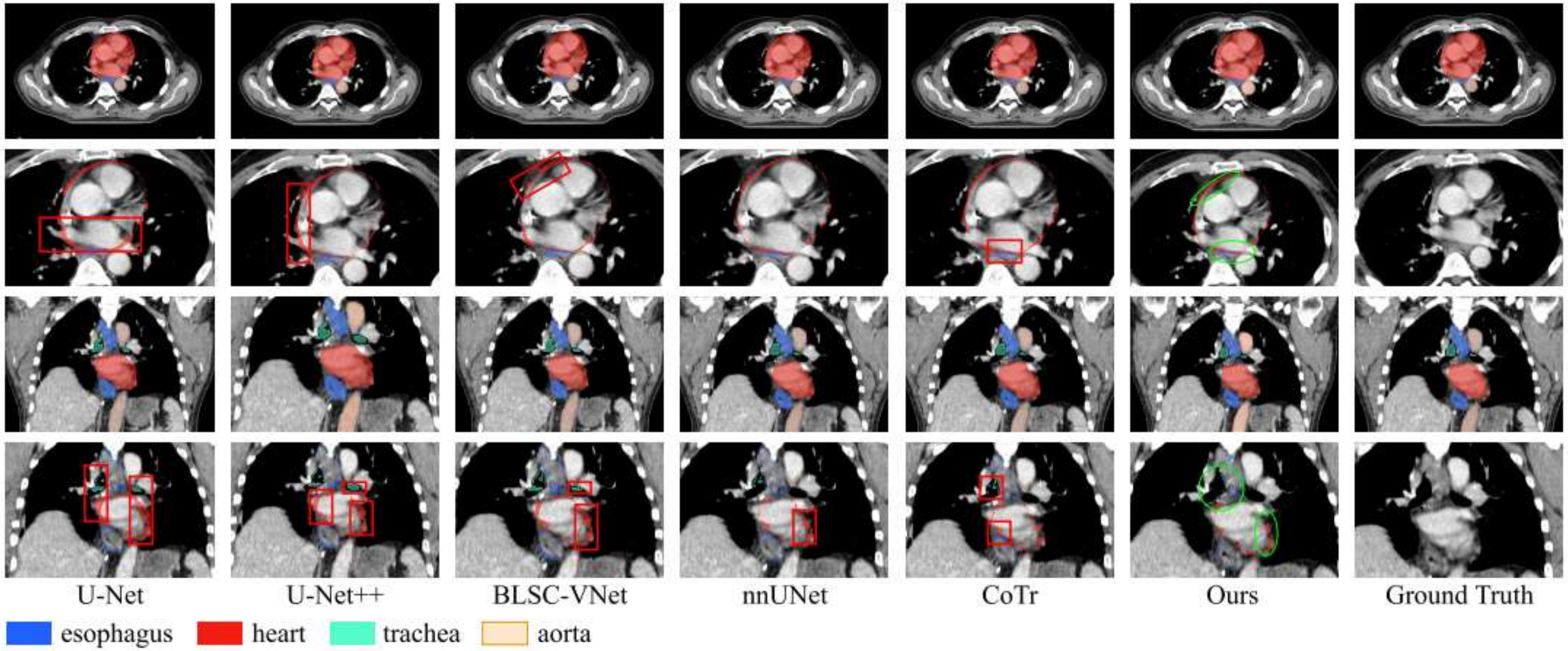


Fig. 9. The segmentation results of different models on SegTHOR 2019 dataset. The first two columns are axial views, and the last two columns are sagittal views. The second and fourth columns respectively show the difference between the prediction results and the actual background value. Different colors represent the location of the wrong segmentation of different organs.

- Fig. 10 is the inference results of different models on the KITS19-M validation set. The regions with significant segmentation errors in the inference results are shown in the red boxes. The segmentation effects of UNet, U-Net++, BLSCVNet and CoTr on small organs show some defects. EPTNet can effectively process spatial position information and performs better in both contour segmentation and edge segmentation of organs. Especially in the marginal segmentation of small organs, our method can reduce the occurrence of false positives. The above results show that our model has strong robustness.

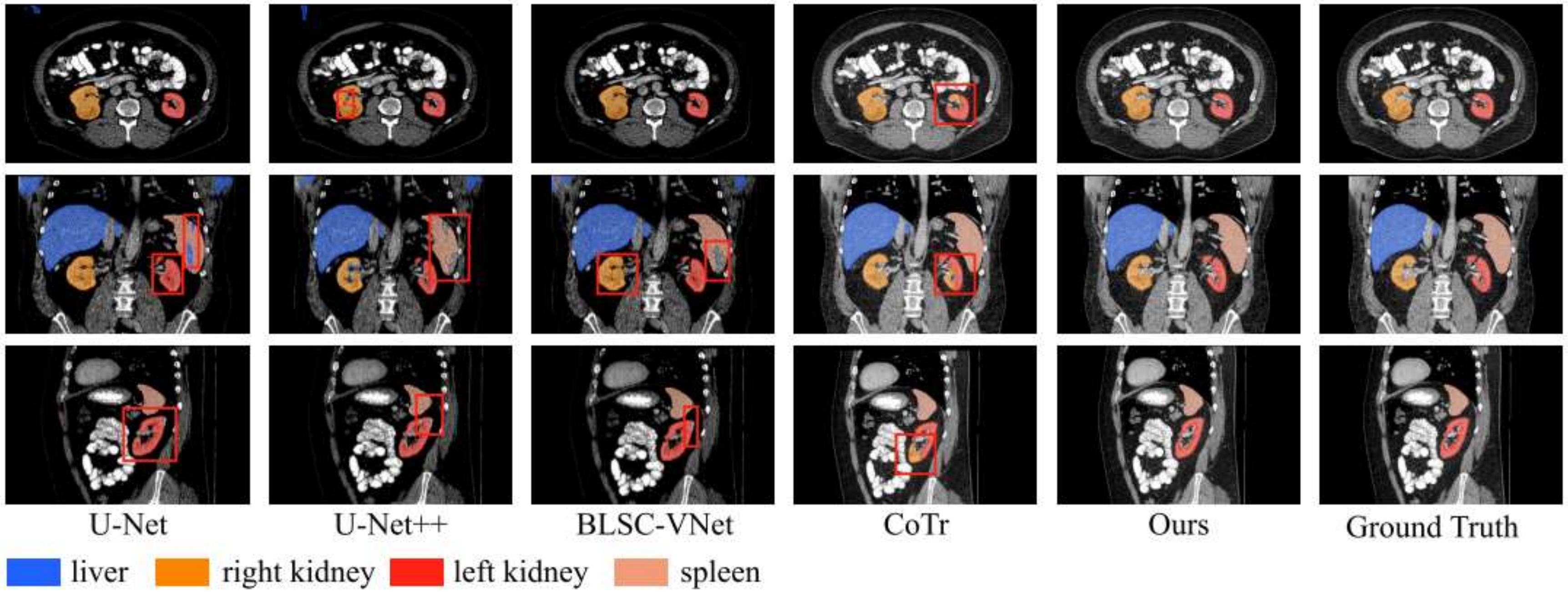


Fig. 10. The segmentation results on KiTS19-M dataset. Visual inferencing results: Each row is a medical image in a different direction. Each column is a segmentation result of a network model.

- Fig. 11 shows the different segmentation results in KiTS19-M and SegTHOR 2019 for several Transformerbased methods. In KiTS19-M, the segmentation results ofTransUNet, Swin-Unet, CoTr, and our method for large organs are similar. However, our method and CoTr have fewer incorrect segmentation regions and higher correct segmentation performance for small organs. It can be seen that our method has better segmentation results on target edges as well as target morphology compared to other methods.

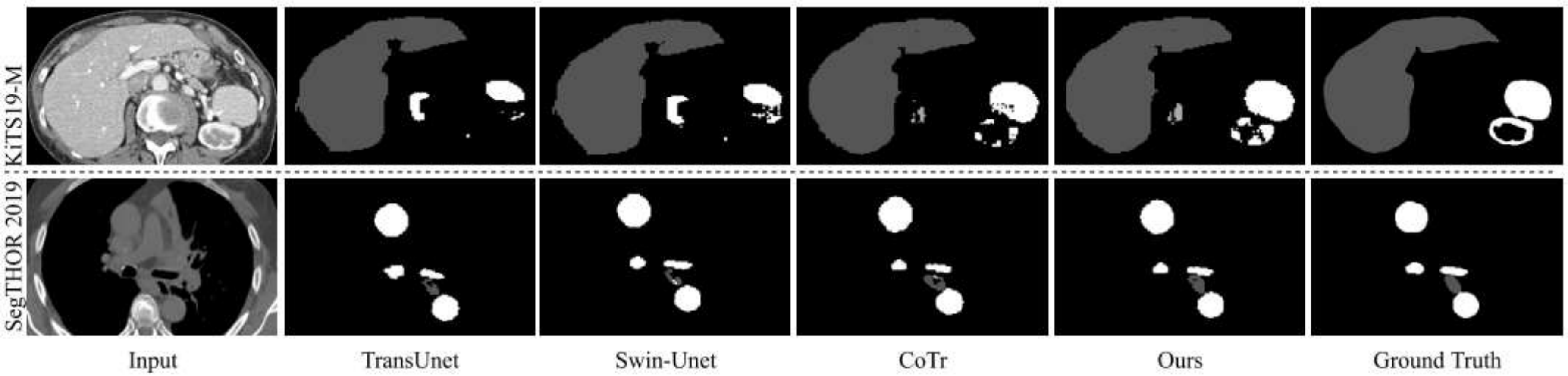


Fig. 11. Different results based on the Transformer method are visualized. Where black represents the background and gray and white represent the different organs.

Conclusion

- This paper introduces a new Transformer-based network structure, called EPT-Net, for semantic segmentation of medical images.
- We use a CNN encoder to learn edge information in high-resolution features of shallow networks and efficiently capture contextual representations at multiple scales by an EWG module without increasing network parameters.
- We learn global abstract features in deep networks through a DPT module that enhances localization capabilities. Our network is suitable for multi-organ and single-organ tasks and is also practical for small organs that are difficult to segment.
- However, the lack of a large amount of labelled data limits the performance and generalization ability of supervised segmentation tasks. How to combine this research with self-supervised learning requires further research.

Thank You!