Applied Intelligence (2023) 53:28542–28554
https://doi.org/10.1007/s10489-023-05032-2

Hybrid transformer-CNN with boundary-awareness network for 3D medical image segmentation

Jianfei He[1] · **Canhui Xu**[1]

Accepted: 18 September 2023 / Published online: 6 October 2023

Abstract
3D volumetric medical image segmentation is a crucial task in computer-aided diagnosis applications, but it remains challenging due to low contrast and boundary ambiguity between organs and surrounding tissues. Considering that accurate boundary voxels are of importance for organ segmentation, which relies on rich detailed features information. The most recent convolutional neural networks and transformer networks have attempted to enhance 2D boundaries during feature extraction. Few approaches focus on boundary voxels preservation for 3D scenarios. To address these issues, we propose the Hybrid Transformer-CNN with Boundary-awareness(HTCB-Net) network, which follows an encoder-decoder segmentation paradigm with learnable boundary modules. The 3D swin-transformer encoder is embedded with auxiliary object-related boundary map by designing a learnable boundary extracting module(BEM), which assists model in obtaining rich and discriminative feature representations. Our boundary map obtained from BEM supervises explicitly feature extraction process. Subsequently, boundary preserving module(BPM) adopts a novel fusion strategy, which integrates extracted boundary map and the corresponding encoder features. Through this module, the boundary position awareness is combined for feature enhancement with spatial complement and channel attention. We evaluate the performance of the proposed method with quantitative experiments on three public available datasets in both CT and MRI modalities: OAI-ZIB, Spleen, Pancreas. The comparative experimental results demonstrate that our HTCB-Net preserves more precise 3D boundaries and obtains significant improvements, particularly in terms of Average Symmetric Surface Distance(ASSD).

Keywords 3D medical image segmentation · Transformer · Boundary preserving

1 Introduction

Medical image segmentation plays a crucial role in clinical applications, such as disease diagnosis, surgical planning, and prognosis evaluation. This technology allows doctors to accurately segment organs and lesions from CT or MRI images and create three-dimensional models which can detail their location, size, and shape. However, the obscure boundaries caused by low contrast between organs and their surrounding tissues make accurate segmentation challenging. As shown in Fig. 1, the intensities of voxels in the targeted organs are very similar to those of the neighboring tissues, and the low intensity variations lead to indistinguishable boundaries, even for human vision [1, 2]. As one of the most fundamental research topic in computer-aided diagnosis, 3D volumetric medical image segmentation has attracted increasing attention over the past decades. Existing methods can be roughly divided into two branches, namely tradition based and deep-learning based methods. Most traditional methods utilize contrast variation, orientation, active contour and shape attributes [3–5]. However, the high dependence on handcrafted features and manual intervention limits the segmentation performance of these methods.

With the development of deep learning, numerous convolutional neural networks(CNNs) have achieved remarkable progress due to the excellent capability of feature extraction, and encoder-decoder architecture has became the most popular solution in 3D medical segmentation. These "U-shape" models apply cascaded 3D convolution and deconvolution operations on multiple scales feature layers to generate

✉ Canhui Xu
ccxu09@yeah.net

Jianfei He
2594297576@qq.com

[1] School of Information Science and Technology, Qingdao University of Science and Technology, 266000, ShanDong, Qingdao, China

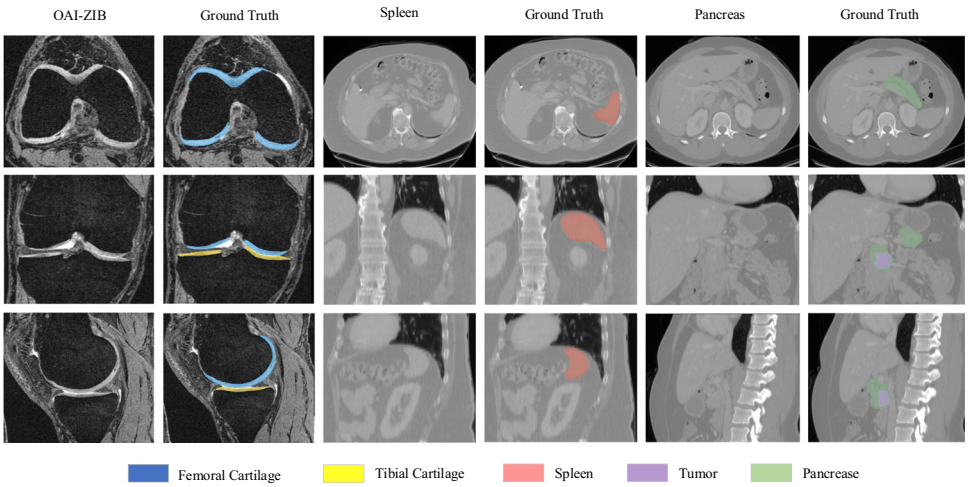

Fig. 1 Illustration of three representative examples from three separate datasets. It shows the axial view (first row), sagittal view (second row), and coronal view (third row) of CT images. Images are grouped in pairs within each row, with each pair consisting of an original image and its corresponding ground truth. It is observed that the boundaries between organs and their neighboring tissues are relatively ambiguous and indistinguishable

voxel-wise segmentation. Compared to 2D CNNs, 3D CNNs offer a notable advantage in capturing the inter-slice relationship, which plays a significant role in improving segmentation performance, especially in tasks involving volumetric data. In the initial stages, the implementation of the 3D CNN model is carried out straightforwardly by substituting 2D convolution and pooling operations with their 3D counterparts. Later on, to extract discriminative features, both residual connections and skip connections are incorporated into the feature extraction process. For example, Jo et al. [6] propagate the features from the encoder to decoder using skip connections with attention gate, which can automatically learns target structures of varying shapes and size. Recent works [7–11] have followed encoder-decoder architecture and made improvements in multi-scale feature integrating strategy.

However, limited by intrinsic localized receptive fields, CNNs-based networks tend to neglect global spatial context and long-range pixel relationships [12–15]. To overcome this limitation, recent studies have attempted to extend transformer to vision tasks [16–18]. Compared with CNNs-based models, transformer-based models first reshape image into sequences of flattened tokens with position embedding and effectively learn features using multi-head self-attention(MSA). This mechanism allows capturing long-range dependencies among a sequence of tokens, and facilitates information global communication. In these models, some employ cascaded transformer layer as encoder to learn the global dependencies of tokenized image patches [19–21]. Furthermore, it has been demonstrated that both local and global features are crucial for dense prediction. Therefore, some models utilize self-attention to fuse multi-scale representations that contain diverse information [22, 23]. It is claimed that transformer-based encoder can learn richer feature representations in 3D volumetric medical image due to the relational nature of voxels along different axes.

Despite the improved performance, previous methods often suffer from edge degradation and limitations in high-dimensional feature extraction, thereby restricting their prediction accuracy. Undoubtedly, boundary prior is an effective essential complement information which contains abundant detail, and it is significant for accurate image segmentation. There are successful boundary-involved segmentation methods in 2D medical image applications [24–27]. For instance, Lee et al. [28] combine boundary point maps with corresponding features utilizing boundary preserving blocks. Notably, the boundary point maps are directly generated from hierarchical feature maps while being guided by boundary point labels. To summarize, leveraging boundary information is beneficial as it enables networks to extract discriminative contextual features, thereby improving segmentation accuracy. However, it is worth mentioning that most of the aforementioned boundary methods are primarily focused on CNN-based approaches for 2D medical images, and limited work has been done to enhance boundaries in the context of 3D medical images.

In this paper, we propose a hierarchical hybrid transformer-CNN with boundary-awareness network(HTCB-Net) in encoder-decoder architecture which explicitly embeds boundary semantic for 3D volumetric medical image segmentation. Firstly, with cascaded swin-transformer as encoder, HTCB-Net hierarchically extracts multi-level features. Then, we design a simple yet effective boundary extracting module(BEM) to explore boundary semantics under boundary label supervision. With the learned boundary map, the boundary preserving module(BPM) integrates its downsampling boundary maps into its corresponding transformer features. Finally, fused features with boundary enhancement connect with the convolutional decoder via skip connection at various resolutions to generate the final prediction. To evaluate the performance, we conduct experiments on three datasets: OAI-ZIB, Spleen, Pancreas. More detailed

experiment results will be elaborated in result comparison section. These experiments prove that our proposed BEM and BPM component can effectively preserve local boundary information and improve segmentation accuracy. The main contributions of this work are as follows:

- Proposed a 3D volumetric medical image segmentation architecture, namely hierarchical hybrid transformer-CNN with boundary-awareness network(HTCB-Net). It adopts swin-transformer as encoder and embeds auxiliary object-related boundary map by designing a learnable boundary extracting module(BEM), which assists model in obtaining rich and discriminative feature representations.
- Introduced a novel fusion strategy in boundary preserving module(BPM) to integrate boundary map and corresponding hierarchical features. With the learned boundary map, this strategy complements and enriches boundary detail information into encoded features, significantly increasing the accuracy of boundary segmentation.
- Proposed a boundary generation algorithm in 3D space for generating boundary labels, which are used as explicit supervision for auxiliary boundary task in BEM. By exploring the boundary label semantics, the learning process of feature enhancement could be guided explicitly.
- To the best of our knowledge, HTCB-Net is one of the very first attempts to introduce boundary awareness into 3D convolutional and transformer-based model. The comparative experimental results demonstrate that our HTCB-Net preserves more precise 3D boundaries and achieves an improvement.

2 Related work

2.1 3D volumetric medical image segmentation

3D volumetric medical image segmentation, serving as a crucial task in computer-aided diagnosis, has been extensively studied and a number of models have been proposed over the past decades. Among these models, encoder-decoder based convolutional neural networks(CNNs) are proposed and became the de-facto standard. For example, Ran et al. [29] introduce CANet, which integrates multiple attentions into the standard U-Net model. The goal of this integration is to emphasize the importance of spatial positions, channels, and scales simultaneously. The nnU-Net proposed by Isensee et al. [30] involves data attributes and network design using empirical knowledge. It can automatically handle preprocessing, training, and reasoning. To further explore the relationship between multi-scale features, Zongwei et al. [9] redesign skip concatenation by adding dense skip connection between feature maps. Information comes from multi-scale features can be aggregated effectively and flexibly through this novel fusion scheme. Besides that, some work make attempts on enlarging limited CNN receptive field. For example, Zaiwang et al. [31] embed a context extractor between encoder and decoder to capture more global context and preserve detailed spatial information for capturing long-range dependencies. Feng et al. [32] propose CPFNet which exploits and fuses rich context information by designing a scale-aware pyramid fusion module. This module dynamically fuses multi-scale context information in high-level features.

Recently, along with the merits of long range dependencies and global context, transformer-based methods achieve remarkable success and unprecedented progress. There are several types of configuration for transformer-based medical image segmentation models. The mostly used architecture applies cascaded transformer blocks as encoder and convolutional blocks as decoder [17, 33, 34]. These methods directly feed token patches from image into transformer-based encoder to model global context features. For example, UNETR [17] employs a transformer-based encoder to learn sequence representations. After that, in order to improve the effectiveness of transformer-based models, some works envision the use of local self-attention in shifted windows to model hierarchical features [19–21]. On basis of Swin Transformer [19], Hatamizadeh et al. propose Swin UNETR [21], a 3D segmentation model combines Swin Transformer layer and UNETR's architecture. It computes self-attention in shifted windows and connects to an convolutional-based decoder via skip connection. Though pure transformer architecture is more intuitive and consistent in design, the problem of high complexity and the lack of inductive bias are challenging to be solved for 3D dense prediction. Hence, hybrid design combining transformer and CNN tends to be a profitable encoder-decoder architecture.

2.2 Boundary information in segmentation

In medical images, boundaries are usually ambiguous due to low intensity contrast between targeted organs and neighboring tissues. Intuitively, boundary prior information is essential for accurate image segmentation. Therefore, many boundary-involved segmentation methods have been recently proposed in 2D medical segmentation. For example, Kervadec et al. [35] propose a boundary loss which assigns different weights to pixels by measurement of boundary pixel distance. After that, Zhu et al. [24] introduce a novel boundary-weighted segmentation loss function which enhances the sensitivity of the trained model to blurred boundaries. Similar strategy is adopted in the work proposed by Karimi et al. [36]. In their method, the introduced distance

loss is used to reduce the hausdorff distance for boundary enhancement. This practice guarantees the connectedness of topology in network-like structures, such as vessels or neurons. The above methods improve the segmentation accuracy by introducing boundary-involved loss function, and other works directly predict boundary as an auxiliary task that shares features with the segmentation task. For example, Ma et al. [37] propose a novel boundary-aware reward mechanism in the form of relative cross-entropy and relative weight for interactive image segmentation. To effectively leverage boundary information, Wang et al. [25] employ a pyramid edge extraction module. This module is capable of extracting multi-granularity boundary information and offers strong cues. Sun et al. [26] point out that embedding additional extracted edge clues is able to enhance boundary semantic description. Yang et al. [27] conduct an edge weight guidance module which leverages low-level features to extract rich and fine-grained edge information for effective edge detection. All above mentioned methods are mostly CNN-based for 2D medical image. To the best of our knowledge, few work focus on embedding extra 3D boundary information to 3D models.

3 Method

The overall architecture of our proposed hybrid transformer-CNN with boundary-awareness network(HTCB-Net) is illustrated in Fig. 2. It comprises of four components: 3D swin transformer encoder, boundary extracting module(BEM), boundary preserving module(BPM) and convolutional decoder. Specifically, 3D swin transformer encoder is responsible for extracting hierarchical feature representations $F_i (i = 1, 2, 3, 4)$ at different resolutions using self-attention mechanism. As an auxiliary task, the BEM extracts object-related boundary map from low-level features and high-level features. This boundary map is then combined with the corresponding transformer features through multiple BPMs in order to complement spatial boundary detail information. Finally, a convolutional decoder is employed to aggregate multi-level fused features with boundary enhancement in a top-down manner to obtain predicted segmentation result.

3.1 3D swin transformer block

Inspired by previous works, our 3D swin transformer encoder extracts multi-layer features with different resolution. The encoder contains four stages. Each stage consists of a swin transformer block for calculating self-attention in shifted windows and patch merging block for reducing resolution. Given a sub-volume image sized $H \times W \times D \times C$ and a volumetric patch $P_h \times P_w \times P_d$, patch partition layer is utilized to project image into 3D tokens F_{l-1} with dimension of $\frac{H}{P_h} \times \frac{W}{P_w} \times \frac{D}{P_d} \times C'$, where C' is the embedded feature dimension. Then, the F_{l-1} is fed into swin transformer blocks. Every block calculates multi-head self-attention in regular and shifted windows respectively, followed by multi-layer perceptron(MLP) and layer normalization(LN). In Eq.1, swin transformer blocks output hierarchically feature

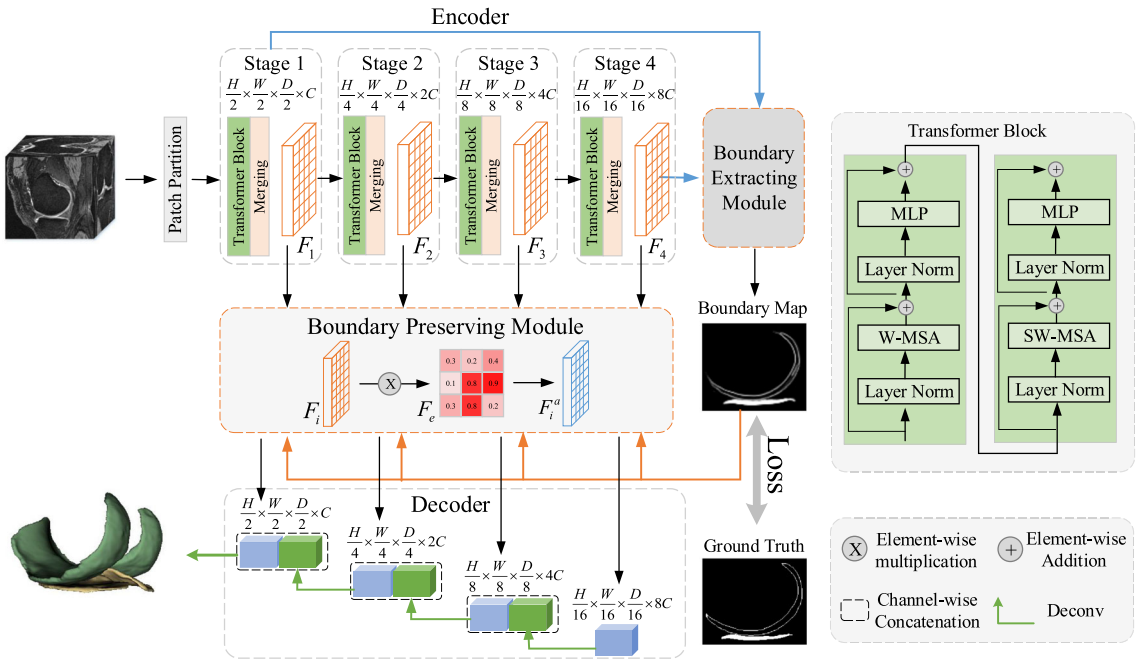

Fig. 2 The overall architecture of our Hybrid Transformer-CNN with Boundary-awareness network(HTCB-Net) which consists of four key parts: 3D swin transformer encoder, boundary extracting module(BEM), boundary preserving module(BPM) and convolution decoder

representations $F_i, \{i = 1, 2, 3, 4\}$. It is worth noting that swin transformer blocks greatly reduce computational cost by introducing attention calculation within local window, rather than calculating in all patches.

$$\begin{aligned}
\hat{F}_l &= \boldsymbol{W-MSA}(LN(F_{l-1})) + F_{l-1} \\
F_l &= \boldsymbol{MLP}(LN(\hat{F}_l)) + \hat{F}_l \\
\hat{F}_{l+1} &= \mathbf{SW{-}MSA}(LN(F_l)) + F_l \\
F_{l+1} &= \boldsymbol{MLP}(LN(\hat{F}_{l+1})) + \hat{F}_{l+1}
\end{aligned} \tag{1}$$

where $W\text{-}MSA$ and $SW\text{-}MSA$ are regular and window partitioning multi-head self-attention modules, $\hat{F}_l$ and $\hat{F}_{l+1}$ denote the output of $W\text{-}MSA$ and $SW\text{-}MSA$ respectively, LN refers to layer normalization, and MLP is multi-layer perceptron.

3.2 Boundary extracting module

Edge prior information, in the form of precise boundaries, can significantly improve the performance of semantic segmentation. It provides guidance to capture finer details and complement spatial information. Inspired by [26], we propose boundary extracting module(BEM) to predict the 3D object-related boundary. Due to the fact that low-level feature contains redundant detail information, while the high level feature contains more semantic information for locating target, we incorporate low-level feature F_1 and high-level feature F_4 to obtain more accurate boundary in this module, as shown in Fig. 3. Specifically, 3D convolutional layers are applied separately at the feature maps with different resolution to ensure uniform channel size. In addition, 3D convolution with bilinear upsampling operation is utilized on F_4 to maintain same resolution with F_1. Next, concatenation operation is employed to integrate the low level feature $F_1^{'}$ and the upsampled feature $F_4^{'}$ along the channel dimension. Finally, we use 3D convolution layers F_{conv} with sigmoid function σ, followed by bilinear upsampling operation $F_{upsampling}$ to obtain the final predicted boundary map F_e. This process can be formulated as,

$$\begin{aligned}
F_1^{'} &= F_{conv}(F_1) \\
F_4^{'} &= F_{upsampling}(F_{conv}(F_4)) \\
F_e &= F_{upsampling}(\sigma(F_{conv}(F_{concat}(F_1^{'}, F_4^{'}))))
\end{aligned} \tag{2}$$

where F_{conv} denotes convolution layer, and $F_{upsampling}$ denotes bilinear upsampling operation. F_{concat} represents concatenation operation along the channel dimension, and σ is non-linear activate function, $sigmoid$. Note that boundary labels are utilized in BEM training as an auxiliary task, which provides explicit supervision and enhances the feature representation capability. It is expected that BEM can enrich the model to learn discriminative features by focusing on complementary boundary information.

3.3 Boundary preserving module

The Boundary Preserving Module (BPM) is designed to incorporate object-related boundary cues into transformer features in order to enhance feature representations. It is achieved by using spatial complement and channel attention, as illustrated in Fig. 4. Specifically, given feature $F\hat{i} \in R^{h^i \times w^i \times d^i \times c^i}$ and boundary map F_e, where h^i, w^i, d^i, and c^i denote the height, width, depth, and channels of the vision feature from the i-th swin transformer block, the BPM performs an element-wise multiplication between $F\hat{i}$ and F_e, followed by an additional skip connection and 3D convolution to obtain the initial fused features F_i^e. Mathematically, this can be expressed as:

$$F_i^e = F_{conv}((F_i \otimes D(F_e)) \oplus F_i) \tag{3}$$

where D represents a downsampling operation, and F_{conv} denotes a convolution layer with a $3 \times 3 \times 3$ kernel size and padding of 1. The symbols $\oplus$ and $\otimes$ represent element-wise addition and multiplication, respectively. It should be noted that $D(F_e)$ will automatically be broadcasted to the same size as F_i in channel dimension. Since different channels often contain different semantic information [38], we further

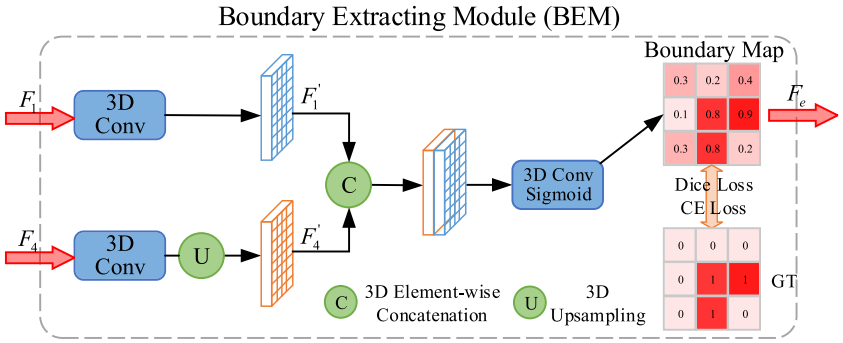

Fig. 3 Illustration of boundary extracting module(BEM). This module aims at extracting object-related boundary map by incorporating low-level spatial details and high-level semantic features

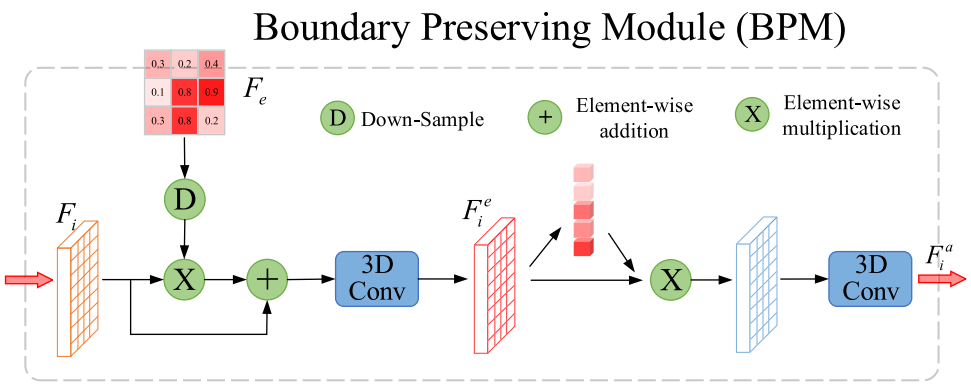

Fig. 4 Illustration of boundary preserving module which is designed to complement spatial information for enhancing the feature representation

introduce channel attention to highlight critical channels and suppress redundant channels and noise information. Specifically, we aggregate the convolutional feature F_i^e by using channel-wise global average pooling. The corresponding channel attention weight is obtained through one dimensional fully connected layer, followed by non-linear sigmoid activation function. We then multiply the channel attention weight with the input feature F_i^e and reduce the number of channels using a 3D convolutional layer to obtain the final output F_i^a, which is fed into the convolutional decoder. It can be denoted as follows:

$$F_i^a = F_{conv}(\sigma(F_{1D}(GAP(F_i^e))) \otimes F_i^e) \tag{4}$$

where F_{conv} denotes 3D convolution, F_{1D} denotes one dimensional fully-connection operation and σ is the sigmoid function.

3.4 Boundary point generation algorithm

The boundary extracting module is trained by computing the loss between the predicted boundary map and the ground truth boundary labels. However, obtaining accurate boundary labels can be challenging. To address this issue, we have developed an algorithm that adopts morphological erosion operations to extract boundaries from the original segmentation labels. The 3D erosion operation can be mathematically expressed using the formula shown in Eq. (5), where (i, j, k) represents a pixel position in the output image C, $A_{i,j,k}$ indicates a structural element in the original image A with (i, j, k) as its center, and $B \subseteq A_{i,j,k}$. It represents the fact that the structural element B is completely contained within A.

$$C_{i,j,k} = \begin{cases} 1, if\ B \subseteq A_{i,j,k} \\ 0, otherwise \end{cases} \tag{5}$$

Specifically, as shown in Fig. 5, given 3D segmentation S_{label}, blue cubes represent background and orange cubes are foreground. We first apply morphological erosion to remove boundary and generate intermediate I_{label}. The light green cubes in intermediate I_{label} are boundaries that have been removed using morphological operations. Finally, we subtract intermediate label from the original label to derive the final boundary B_{label}. The entire process is illustrated in Fig. 5.

3.5 Training segmentation network

The segmentation network is designed to produce: boundary map predicted by BEM and final overall segmentation result. To optimize the model, we need to calculate loss for both outputs. The main objective during training phase is to minimize discrepancy between predicted segmentation output and ground truth.

To achieve this objective, we use a combination of two different loss functions: DICE loss and Cross Entropy (CE) loss. The DICE loss function is effective in addressing class imbalance issue by emphasizing on correctly segmenting the minority class. Its effectiveness has been demonstrated in various medical imaging applications such as tumor segmentation and vascular segmentation. On the other hand, CE loss measures difference between predicted probability distribution and true probability distribution. It is commonly used in multiclass classification tasks, where each class is considered separately.

For DICE loss, given predicted segmentation and ground truth, we first compute the intersection between two segmentation masks. Then it divides sum of pixels in both masks. The resulting DICE coefficient is used for DICE loss. And this process can be formulated as:

$$DL(y, \hat{p}) = 1 - \frac{2y\hat{p} + \epsilon}{y + \hat{p} + \epsilon} \tag{6}$$

where y represents label, $\hat{p}$ denotes prediction, and ϵ is a small value which is normally assigned 1.

CE loss can be formulated as follows:

$$CE(y, \hat{p}) = -ylog(\hat{p}) - (1 - y)log(1 - (\hat{p})) \tag{7}$$

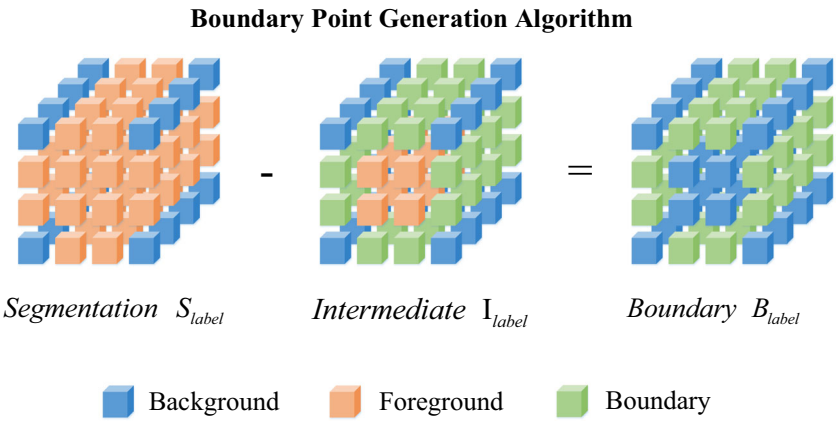

Fig. 5 Illustration of 3D boundary point generation algorithm. The blue cubes represent background, while the orange cubes represent foreground. By applying morphological operations, the boundary B_{label}, marked as light green cubes, can be calculated for training process

where y represents label, and $\hat{p}$ denotes prediction. Finally, the whole loss can be depicted as:

$$L_{Total} = DL_{Seg} + CE_{Seg} + DL_{Boundary} + CE_{Boundary} \quad (8)$$

where L_{Total} is total loss, DL_{Seg} and CE_{Seg} denote DICE loss and CE loss between segmentation result and segmentation label. $DL_{Boundary}$ and $CE_{Boundary}$ present DICE loss and CE loss between boundary map and boundary label, respectively.

4 Experimental results and discussion

4.1 Datasets and evaluation metrics

Three publicly available datasets: OAI-ZIB, Spleen, and Pancreas, are used to evaluate our method. OAI-ZIB [39] is a knee joint cartilage dataset from the Osteoarthritis Initiative(OAI), which contains 507 3D MRI scans. Each MRI image has a size of 384 × 384 × 160 and a voxel spacing of 0.37 × 0.37 × 0.7. It is annotated femoral cartilage and tibial cartilage manually by experts. The dataset is split into 253 MRI scans for training and 254 for testing.

The Spleen dataset [40] is part of the Medical Segmentation Decathlon (MSD) and consists of 61 three-dimensional spleen CT scans, with 41 images for training and 20 for testing. The average image size is 187 × 512 × 512 with a spacing of 1.6 × 0.79 × 0.79.

The Pancreas dataset [40] is also part of the MSD and aims to automatically segment pancreas and tumors from CT scans. The dataset includes 420 portal-venous phase CT scans of patients undergoing resection of pancreatic masses, with 281 for training and 129 for testing. The pancreatic parenchyma and pancreatic mass (cyst or tumor) are manually segmented in each slice by an expert abdominal radiologist using the Scout application [41]. The average image size is 96 × 512 × 512 with a spacing of 2.5 × 0.8 × 0.8.

To compare the performance of different models, dice similarity coefficient (DSC) and average symmetric surface distance (ASSD) are adopted between predicted results and ground truth.

4.2 Implementation details

Our HTCB-Net network is implemented using PyTorch 1.8.0 and Monai 1.7.0. We train it on a single RTX 3090Ti GPU using the AdamW optimizer with momentum of 0.9. The initial learning rate was set to 0.0002 and decreased to 0.8 times the previous value every 100 epochs. The training was conducted for 600 epochs with batch size of 1. The Monai's DiceCELoss is adapted as loss function in the training.

Preprocessing is a critical step that significantly affects the final segmentation accuracy [30, 42]. Therefore, we adopt two different preprocessing strategies based on the image modality of CT and MRI images. For CT images, we first identify the maximum value, minimum value, mean value, and variance of all foreground voxels in training set, as intensity of HU value can reflect the physical properties of different tissues. Subsequently, clip the [0.05, 99.5] value of all foreground voxels to effectively utilize the extra information of HU value. Finally, all data are normalized using z-scoring, where mean value is subtracted and the result is divided by the standard deviation. For MRI images, the z-scoring is directly adopted. Additionally, all images are resampled to the same spacing to achieve a uniform voxel distribution: [0.7, 0.36, 0.36] for OAI-ZIB, [1.6, 0.79, 0.79] for Spleen, and [2.5, 0.8, 0.8] for Pancreas.

During training phase, considering the memory limitations, the randomly cropped patch size of 96 × 96 × 96 is fed into network. In inference phase, the overlapping sliding windows are used to crop sub-volumes, and the averages of probability maps of these sub-volumes are used to generate whole volume prediction. The sub-volume size is also set to 96 × 96 × 96, and overlap rate of each window is set to 0.5.

Table 1 Comparative results on OAI-ZIB testing set. DSC and ASSD are calculated for femoral cartilage and tibial cartilage, respectively

Model	Femoral Cartilage		Tibial Cartilage	
	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)
V-Net	87.71	1.1520	84.11	1.7328
3D U-Net	89.03	0.7766	84.77	1.0323
nnU-Net	89.05	0.6825	**85.53**	0.8340
SegFormer	85.55	1.8193	83.79	3.2474
UNETR	88.34	0.6962	85.27	0.9987
Swin-UNETR	89.27	0.6274	85.30	0.8436
ours	**89.33**	**0.6073**	85.49	**0.7201**

The bold entries indicates the representation of the most notable experimental outcomes. These entries, thereby, signify the results that have achieved the highest level of performance or yielded the most significant findings within the respective tables. The utilization of bold significance emphasizes distinctive characteristics of different algorithms, enabling readers to easily identify the most remarkable or note worthy outcomes within the presented data with visual distinction

4.3 Results

In the experiment, several classical 3D medical image segmentation networks with different architecture are evaluated, such as V-Net, 3D U-Net, nnU-Net, SegFormer, UNETR and Swin-UNETR. The former three are all CNN-based, among which nnU-Net has achieved excellent results on several famous medical image segmentation challenging. The latter three are all Transformer-based, in which the decoder of UNETR and Swin-UNETR are both CNN decoder, while SegFormer's decoder adopts several MLP layers to reduce the parameters.

4.3.1 OAI-ZIB segmentaion

The evaluation results on OAI-ZIB are presented in Table 1. Our method achieved a DSC of 89.33% and 85.49% for Femoral Cartilage (FC) and Tibial Cartilage (TC), respectively, surpassing all competing approaches on FC and falling 0.04% lower than nnU-Net on TC. The average $ASSD$ values of our method are 0.6073 in FC and 0.7201 in TC, compared to nnU-Net with 0.6825 in FC and 0.8340 in TC. Our model also achieves an improvement over the Transformer-based model Swin-UNETR, reducing the ASSD in FC and TC by 0.02 and 0.12, respectively. Overall, our method achieves significant performance compared to other models, especially in $ASSD$, which can be attributed to the boundary preserve capability of BPM and BEM. The boundary preserving module enables the network to capture more representative features by complementing boundary information. Furthermore, nnU-Net and Swin-Transformer show competitive performance in these experiments, with nnU-Net achieving the best segmentation result in terms of DSC for TC.

Figures 6 and 7 present visual comparison of several methods implemented under the same experimental configurations. Each row in Fig. 6 represents a visualization

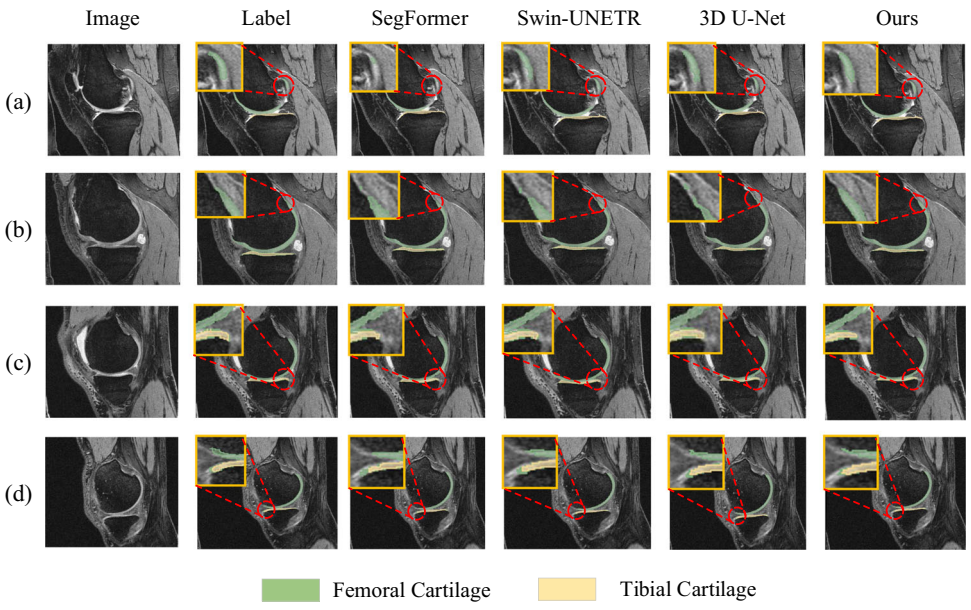

Fig. 6 Comparative result examples of image segmentation between our HTCB-Net and other baseline methods on OAI-ZIB dataset. The region with visual improvements are highlighted with yellow dashed lines and amplified, where green represents the femoral cartilage and yellow represents the tibial cartilage

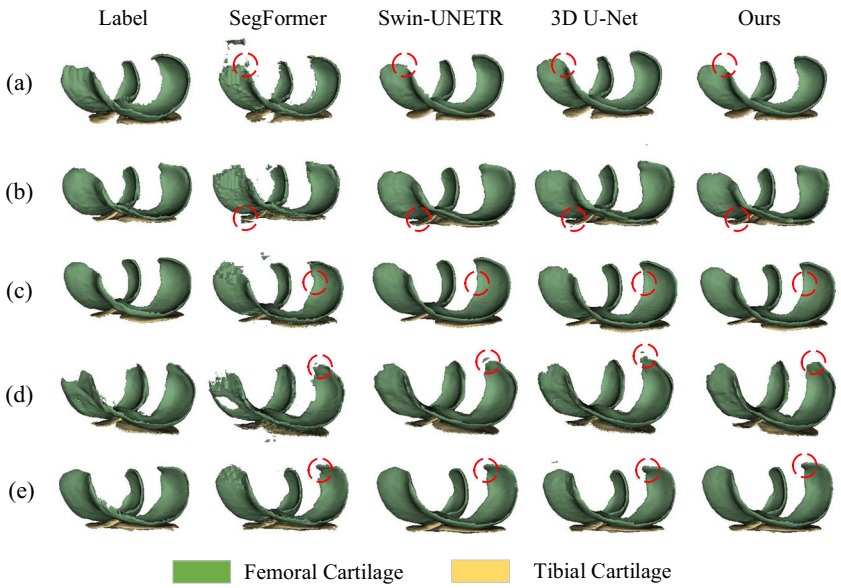

Fig. 7 Comparison of 3D visualization with other methods on the OAI-ZIB testing set reveals the performance of the segmentation results, which can be intuitively observed from the parts circled in red dashed lines

comparison of a segmentation example using different methods, where green represents the femoral cartilage and yellow represents the tibial cartilage. The segmentation results of SegFormer, Swin-UNETR, 3D U-Net, and our method are displayed from left to right. To better observe the segmentation details, we have circled the parts that require closer attention with red dashed lines and amplified them in the upper left corner of each image. Observing the row (a) and (b), our method exhibits smoother and more continuous boundaries compared to other methods, and the overall shape is closer to the ground truth label. This is mainly attributed to the fact that the BEM and BPM focus more attention on the boundaries, allowing for better perception of shape and boundary information. However, although the continuity of the intermediate boundaries has been improved, the segmentation results in the corner parts (row c and d) exhibit under-segmentation or over-segmentation. It is highly probable that the less refined edges generated by the boundary generation algorithm in the corner regions are responsible for inaccurate predictions by the network. Figure 7 shows 3D visualization of the segmentation results, where each row represents the case segmentation result using different models. This allows for intuitive observation of the segmentation quality and is also a commonly used form of presentation in clinical applications. The row (d) indicates that the segmentation results of HTCB-Net are smoother and more continuous overall, with no additional isolated falsely predicted parts compared to other methods.

4.3.2 Spleen segmentaion

As shown in Table 2, the dice score and average symmetric surface distance of HTCB-Net are 95.26% and 0.4201mm, respectively. Compared with other methods, such as SegFormer, UNETR, Swin-UNETR, and 3D U-Net, HTCB-Net exhibits significant improvement. Furthermore, we can observe that the performance of the transformer-based method Swin-UNETR surpasses that of the convolutional-based 3D U-Net by 0.41% and 0.3386 on dice score and average symmetric surface distance, respectively, indicating that the transformer-based model can extract more representative features via self-attention mechanism.

Figure 8 shows visual comparison between HTCB-Net and other baseline models on Spleen testing set. Notably, in row (a), Swin-UNETR produces additional false positives in the top left corner, whereas our model performs accurately. On the other hand, compared to other methods, our model obtains smoother boundaries in row (b) and (c), without additional noisy discontinuity. This is attributed to the network emphasizing the information surrounding the boundaries of tissues. However, in row (d), we notice that our model

Table 2 Comparative results on spleen testing set

Model	Spleen	
	DSC(%)	ASSD(mm)
SegFormer	94.18	0.8701
UNETR	93.95	0.9084
3D U-Net	94.16	0.8929
Swin-UNETR	94.57	0.5543
ours	**95.26**	**0.4201**

The bold entries indicates the representation of the most notable experimental outcomes. These entries, thereby, signify the results that have achieved the highest level of performance or yielded the most significant findings within the respective tables. The utilization of bold significance emphasizes distinctive characteristics of different algorithms, enabling readers to easily identify the most remarkable or note worthy outcomes within the presented data with visual distinction

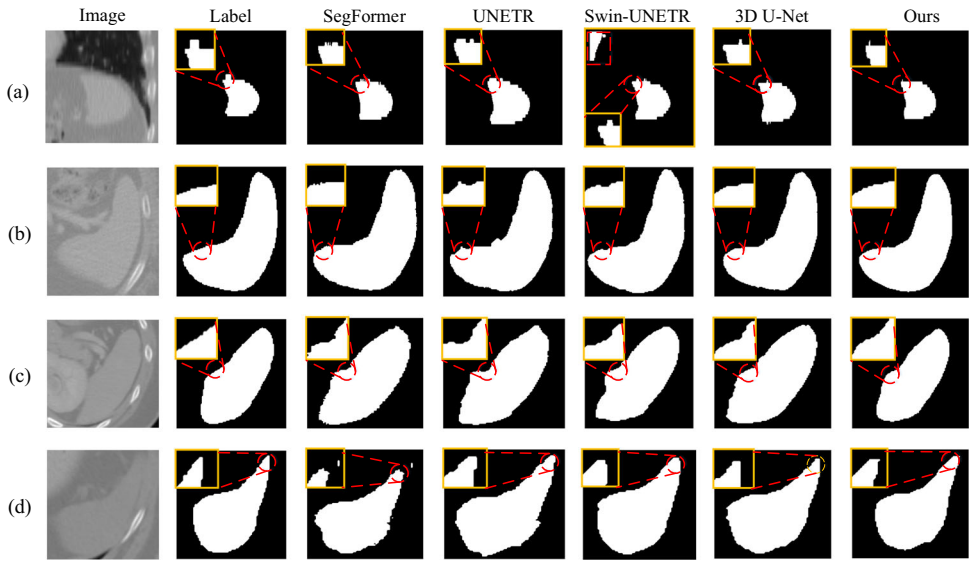

Fig. 8 Visual comparison between HTCB-Net and other baseline models on Spleen testing set. Each row represents a case, and the segmentation improvements are highlighted with yellow dashed lines and amplified, particularly near the boundary of the spleen

does not perform well in corner regions, whereas 3D U-Net achieves better results. The probable reason behind may lie in the inherent issues associated with the erosion operation in boundary generation algorithm.

4.3.3 Pancreas segmentaion

Table 3 presents the results of several models on the pancreas dataset. This task is particularly challenging, and all models struggle to achieve excellent results, especially in tumor segmentation. The dice scores of HTCB-Net on the pancreas and tumor are 81.47% and 51.83%, respectively. The average symmetric surface distances are 1.7725mm and 17.13mm in the pancreas and tumor, respectively. Compared to 3D U-Net, our model's performance increases the mean DSC score and ASSD score from 79.03 to 81.47 and from 51.77 to 51.83, respectively. Additionally, our model outperforms V-Net and Swin-UNETR by 0.76% and 0.22% on the pancreas. However, for small targets such as tumors, nnU-Net achieves the best DSC performance at 52.86%, surpassing our model. One possible analysis is that tumors in this dataset are considered small targets, and extracting their boundaries poses a challenge for our boundary extraction algorithm, which may result in the model not achieving optimal results. Additionally, it is worth noting that all the distance-based ASSD metrics significantly outperform other methods because the boundaries are more accurately preserved.

4.4 Ablation study

We conduct several experiments on Spleen dataset to validate the effectiveness of each module. The baseline model consists of a swin transformer encoder and a convolutional decoder. Then boundary extracting module(BEM) and boundary preserving module(BPM) are separately added into baseline model, named model A and model B. Notably, as the boundary map predicted by BEM is used as input of BPM, the element-wise multiplication operation is removed, retaining only the channel attention in model B. Finally, we add both the BEM and BPM to model simultaneously, resulting in our final HTCB-Net architecture.

Table 3 Comparative results on pancreas dataset

Model	Pancreas		Tumor	
	DSC(%)	ASSD(mm)	DSC(%)	ASSD(mm)
V-Net	80.71	2.1520	50.11	18.7328
3D U-Net	79.03	2.7766	51.77	18.2523
nnU-Net	81.18	2.6825	**52.86**	18.0740
Swin-UNETR	81.25	2.4401	52.65	18.59
ours	**81.47**	**1.7725**	51.83	**17.13**

The bold entries indicates the representation of the most notable experimental outcomes. These entries, thereby, signify the results that have achieved the highest level of performance or yielded the most significant findings within the respective tables. The utilization of bold significance emphasizes distinctive characteristics of different algorithms, enabling readers to easily identify the most remarkable or note worthy outcomes within the presented data with visual distinction

Table 4 Ablation experiments on Spleen dataset

Model	DSC(%)	ASSD(mm)
baseline	94.17	0.8929
baseline + BEM	94.32	0.8519
baseline + BPM	94.89	0.8367
baseline + BEM + BPM	**95.26**	**0.4201**

The bold entries indicates the representation of the most notable experimental outcomes. These entries, thereby, signify the results that have achieved the highest level of performance or yielded the most significant findings within the respective tables. The utilization of bold significance emphasizes distinctive characteristics of different algorithms, enabling readers to easily identify the most remarkable or note worthy outcomes within the presented data with visual distinction

1) With boundary extracting module(BEM): Table 4 demonstrates that the segmentation results of baseline model equipped with boundary extracting module (BEM) outperform those of baseline model alone by 0.15% (94.17% $\rightarrow$ 94.32%) in terms of DSC. This improvement might be attributed to the fact that auxiliary boundary extraction module enables the shared feature encoder to learn more discriminative representations.

2) With boundary preserving module(BPM): As is widely recognized, channels are interrelated both in 2D and 3D domains. Therefore, we attempt to introduce channel attention mechanism to explore the dependencies between channels in boundary preserving module (BPM). Compared to baseline model, the segmentation results of baseline model with BPM increased DSC by 0.72% (94.17% $\rightarrow$ 94.89%) and decreased $ASSD$ by 0.0562 (0.8929 $\rightarrow$ 0.8367), achieving better performance. It might be attributed to the fact that the channel attention mechanism highlights critical channels and suppresses redundant channels, thereby improving model's ability to capture important features.

3) With boundary extracting module(BEM) and boundary preserving module(BPM): Table 4 shows that our HTCB-Net achieves a DSC of 95.26% and an $ASSD$ of 0.4201, outperforming other models significantly. The superior performance of our HTCB-Net suggests that the boundary preserving module (BPM) can effectively integrate objective-related boundary information extracted by the boundary extracting module (BEM) with feature maps, thereby enhancing the model's ability to capture spatial detail information.

4.5 Limitation and discussion

Medical image segmentation plays a critical role in accurately identifying and locating lesions and organs from CT or MRI images. By utilizing our proposed method, doctors can establish three-dimensional models of organs, allowing for a detailed understanding of their location, size, and shape, ultimately improving the efficiency of diagnosis and treatment. Our method exhibits smoother and more continuous boundaries, reducing the disconnected regions and making the overall external contour more similar in shape to the ground truth label and without additional noisy discontinuity. However, the high computational cost associated with this process currently limits its practical application. Additionally, extracting boundaries of small targets, such as blood vessel cells, is a challenging task in the field of medical image processing. Moving forward, it will be essential to conduct further research to explore more efficient and accurate algorithms for boundary extraction of small targets. Moreover, investigating how to better leverage the boundary information extracted from these targets could also yield important insights.

5 Conclusion

In this paper, we propose a hybrid transformer-CNN architecture with a boundary-awareness network (HTCB-Net) for 3D volumetric medical image segmentation. Firstly, we use a 3D swin transformer encoder to extract multi-scale feature representations. To incorporate 3D boundary information, the boundary extracting module (BEM) utilizes features to obtain an object-related boundary map. Furthermore, the boundary preserving module (BPM) integrates the downsampled boundary map into corresponding hierarchical features extracted by swin transformer blocks. Extensive experiments on three public datasets: OAI-ZIB, Spleen, Pancreas, show that our HTCB-Net preserves more precise 3D boundaries and achieves significant improvements in segmentation performance. Our model has the potential to alleviate the effects of outliers, over-segmentation, and mislabeling.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant No. 61806107 and 61702135, Shandong Key Laboratory of Wisdom Mine Information Technology, and the Opening Project of State Key Laboratory of Digital Publishing Technology.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhou S, Nie D, Adeli E, Yin J, Lian J, Shen D (2019) High-resolution encoder-decoder networks for low-contrast medical image segmentation. IEEE Trans Image Process. 29:461–475
2. Hu H, Shen L, Guan Q, Li X, Zhou Q, Ruan S (2022) Deep co-supervision and attention fusion strategy for automatic covid-19 lung infection segmentation on ct images. Pattern Recog. 124:108452
3. Hsu W-Y, Lu C-C, Hsu Y-Y (2020) Improving segmentation accuracy of ct kidney cancer images using adaptive active contour model. Medicine 99
4. Sedghi Gamechi Z, Bons LR, Giordano M, Bos D, Budde RP, Kofoed KF, Pedersen JH, Roos-Hesselink JW, Bruijne M (2019) Automated 3d segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced ct. Eur Radiol. 29:4613–4623
5. Myller KAH, Honkanen JTJ, Jurvelin JS, Saarakkala S, Töyräs J, Väänänen SP (2018) Method for segmentation of knee articular cartilages based on contrast-enhanced ct images. Ann Biomed Eng 46:1756–1767
6. Schlemper J, Oktay O, Schaap M, Heinrich MP, Kainz B, Glocker B, Rueckert D (2018) Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 53:197–207
7. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A (2018) H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Trans Med Imaging. 37:2663–2674
8. Zhou C, Ding C, Wang X, Lu Z, Tao D (2019) One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. IEEE Trans Image Process. 29:4516–4529
9. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2019) Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging. 39:1856–1867
10. Xia H, Ma M, Li H, Song S (2021) Mc-net: multi-scale context-attention network for medical ct image segmentation. Appl Intell 52:1508–1519
11. Hu Q, Wei Y, Li XL, Wang C, Wang H, Wang S (2022) Svf-net: spatial and visual feature enhancement network for brain structure segmentation. Appl Intell 53:4180–4200
12. Hu H, Zhang Z, Xie Z, Lin S (2019) Local relation networks for image recognition. IEEE/CVF International Conference on Computer Vision (ICCV) 2019:3463–3472
13. Xue C, Zhu L, Fu H, Hu X, Li X, Zhang H, Heng P-A (2021) Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal 70:101989
14. Lin X, Yu L, Cheng K-T, Yan Z (2022) The lighter the better: Rethinking transformers in medical image segmentation through adaptive pruning. IEEE transactions on medical imaging, PP
15. Shi C, Xu C, He J, Chen Y, Cheng Y, Yang Q, Qi H (2022) Graph-based convolution feature aggregation for retinal vessel segmentation. Simul Model Pract Theory 121:102653
16. Wu Y, Liao K-Y, Chen J, Wang J, Chen DZ, Gao H, Wu J (2022) D-former: a u-shaped dilated transformer for 3d medical image segmentation. Neural Comput Appl 35:1931–1944
17. Hatamizadeh A, Yang D, Roth HR, Xu D (2021) Unetr: Transformers for 3d medical image segmentation. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022:1748–1758
18. Lin A-J, Chen B, Xu J, Zhang Z, Lu G, Zhang D (2021) Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE Trans Instrum Meas 71:1–15
19. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF International Conference on Computer Vision (ICCV) 2021:9992–10002
20. Yuan F, Zhang Z, Fang Z (2022) An effective cnn and transformer complementary network for medical image segmentation. Pattern Recognit 136:109228
21. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D (2022) Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: BrainLes@MICCAI
22. Heidari M, Kazerouni A, Kadarvish MS, Azad R, Aghdam EK, Cohen-Adad J, Merhof D (2022) Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023:6191–6201
23. Lin J, Lin J, Lu C, Chen H, Lin H, Zhao B, Shi Z, Qiu B, Pan X, Xu Z, Huang B, Liang C, Han G, Liu Z, Han C (2022) Ckd-transbts: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. IEEE transactions on medical imaging, PP
24. Zhu Q, Du B, Yan P (2019) Boundary-weighted domain adaptive neural network for prostate mr image segmentation. IEEE Trans Med Imaging 39:753–763
25. Wang R, Chen S, Ji C, Fan J, Li Y (2022) Boundary-aware context neural network for medical image segmentation. Medical image analysis. 78:102395
26. Sun Y, Wang S, Chen C, Xiang T (2022) Boundary-guided camouflaged object detection. In: IJCAI
27. Yang J, Jiao L, Shang R, Liu X, Li R, Xu L (2023) Ept-net: Edge perception transformer for 3d medical image segmentation. IEEE transactions on medical imaging, PP
28. Lee HJ, Kim JU, Lee S, Kim HG, Ro YM (2020) Structure boundary preserving segmentation for medical image with ambiguous boundary. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020:4816–4825
29. Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest JA, Ourselin S, Vercauteren T, Zhang S (2020) Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans Med Imaging 40:699–711
30. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18(2):203–211
31. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J (2019) Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans Med Imaging 38:2281–2292
32. Feng S, Zhao H, Shi F, Cheng X, Wang M, Ma Y, Xiang D, Zhu W, Chen X (2020) Cpfnet: Context pyramid fusion network for medical image segmentation. IEEE Trans Med Imaging 39(10):3008–3018
33. He Q, Sun X, Diao W, Yan Z, Yin D, Fu K (2022) Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing. ISPRS J Photogramm Remote Sens
34. Tang Y, Yang D, Li W, Roth HR, Landman BA, Xu D, Nath V, Hatamizadeh A (2022) Self-supervised pre-training of swin transformers for 3d medical image analysis. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022:20698–20708
35. Kervadec H, Bouchtiba J, Desrosiers C, Granger É, Dolz J, Ayed IB (2019) Boundary loss for highly unbalanced segmentation. Med Image Anal 67:101851
36. Karimi D, Salcudean SE (2020) Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans Med Imaging 39:499–513
37. Ma C, Xu Q, Wang X, Jin B, Zhang X, Wang Y, Zhang Y (2020) Boundary-aware supervoxel-level iteratively refined interactive 3d image segmentation with multi-agent reinforcement learning. IEEE Trans Med Imaging 40:2563–2574
38. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42:2011–2023

39. Ambellan F, Tack A, Ehlke M, Zachow S (2019) Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. Med Image Anal 52:109–118
40. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider Annette, Landman BA, Litjens GJS, Menze BH, Ronneberger O, Summers RM, Ginneken B, Bilello M, Bilic P, Christ PF, Do RKG, Gollub MJ, Heckers S, Huisman HJ, Jarnagin WR, McHugo M, Napel S, Pernicka JSG, Rhode KS, Tobon-Gomez C, Vorontsov E, Meakin JA, Ourselin S, Wiesenfarth M, Arbeláez P, Bae B, Chen S, Daza LA, Feng J-J, He B, Isensee F, Ji Y, Jia F, Kim N, Kim I, Merhof D, Pai A, Park B, Perslev M, Rezaiifar R, Rippel O, Sarasua I, Shen W, Son J, Wachinger C, Wang L, Wang Y, Xia Y, Xu D, Xu Z, Zheng Y, Simpson AL, Maier-Hein L, Cardoso MJ (2022) The medical segmentation decathlon. Nat Commun 13: 4128
41. Dawant BM, Li R, Lennon B, Li S (2007) Semi-automatic segmentation of the liver and its evaluation on the miccai 2007 grand challenge data set
42. Huang Q, Sun J, Ding H, Wang X, Wang G (2018) Robust liver vessel extraction using 3d u-net with variant dice loss function. Comput Biol Med 101:153–162