Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5689301, 16 pages
https://doi.org/10.1155/2020/5689301

Research Article

BE-FNet: 3D Bounding Box Estimation Feature Pyramid Network for Accurate and Efficient Maxillary Sinus Segmentation

Zhuofu Deng[1], **Binbin Wang**[1,2] **and Zhiliang Zhu**[1]

[1]*College of Software, Northeastern University, Shenyang 110169, China*
[2]*Deep NEU Technology Co., Ltd., Shenyang 110169, China*

Correspondence should be addressed to Zhuofu Deng; dengzf@swc.neu.edu.cn

Received 3 September 2019; Revised 22 December 2019; Accepted 8 January 2020; Published 28 January 2020

Academic Editor: Daniel Zaldivar

Maxillary sinus segmentation plays an important role in the choice of therapeutic strategies for nasal disease and treatment monitoring. Difficulties in traditional approaches deal with extremely heterogeneous intensity caused by lesions, abnormal anatomy structures, and blurring boundaries of cavity. 2D and 3D deep convolutional neural networks have grown popular in medical image segmentation due to utilization of large labeled datasets to learn discriminative features. However, for 3D segmentation in medical images, 2D networks are not competent in extracting more significant spacial features, and 3D ones suffer from unbearable burden of computation, which results in great challenges to maxillary sinus segmentation. In this paper, we propose a deep neural network with an end-to-end manner to generalize a fully automatic 3D segmentation. At first, our proposed model serves a symmetrical encoder-decoder architecture for multitask of bounding box estimation and in-region 3D segmentation, which cannot reduce excessive computation requirements but eliminate false positives remarkably, promoting 3D segmentation applied in 3D convolutional neural networks. In addition, an overestimation strategy is presented to avoid overfitting phenomena in conventional multitask networks. Meanwhile, we introduce residual dense blocks to increase the depth of the proposed network and attention excitation mechanism to improve the performance of bounding box estimation, both of which bring little influence to computation cost. Especially, the structure of multilevel feature fusion in the pyramid network strengthens the ability of identification to global and local discriminative features in foreground and background achieving more advanced segmentation results. At last, to address problems of blurring boundary and class imbalance in medical images, a hybrid loss function is designed for multiple tasks. To illustrate the strength of our proposed model, we evaluated it against the state-of-the-art methods. Our model performed better significantly with an average Dice 0.947 ± 0.031, VOE 10.23 ± 5.29, and ASD 2.86 ± 2.11, respectively, which denotes a promising technique with strong robust in practice.

1. Introduction

Maxillary sinus is an important part of the body which has multiple functions including olfaction, filtering, heating, and humidifying the inhaled air. People who suffer from nasal function impairment may have a reduced quality of life [1]. In the last few years, functional endoscopic sinus surgery (FESS) has been established as the state-of-the-art technique for the treatment of endonasal pathologies. Recently, robot-assisted FESS replaces the traditional one that grows inconvenient for the surgeon. To exactly define the workspace, the knowledge about the anatomical structure of maxillary sinus is required. Manual segmentation costs about 900 minutes for one patient's CT scans which become infeasible for daily practice [2]. Consequently, automatic segmentation approaches with high accuracy should be imperative. However, there are some difficulties in practice. At first, the high rate of structure variations exists in maxillary sinus like location, size, and shape. In addition, plenty of lesions frequently appear in the cavity with different intensities, scales, and positions, which lead to extremely heterogeneous textures in volume of interests (VoI). Figure 1 illustrates some distinct cases indicating a general segmentation method encountering more challenges.

Medical image analysis has played a crucial role in clinical practice for a long period, and related techniques

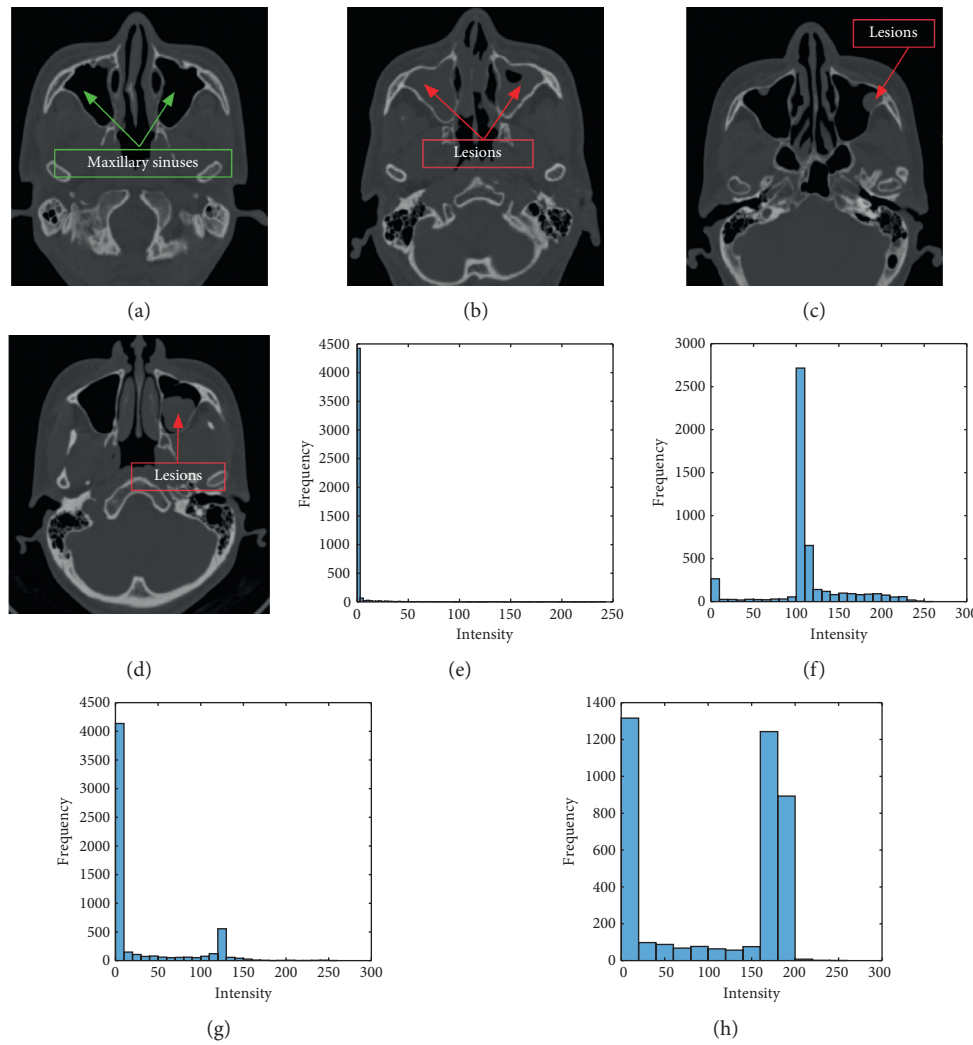


Figure 1: Inhomogeneous context of maxillary sinus reflects great challenges in research of segmentation. (a) Pair of healthy sinuses, (b)–(d) filled with lesions with disparate scales, locations, and shapes, and (e)–(h) average of the normalized intensity histograms of the corresponding CT scans above.

including partial differential equation (PDE), machine learning, and deep learning models have achieved rapid and efficient development to promote computer-aided application in clinics. For example, Wei et al. [3] proposed an adaptive variational PDE model for image reconstruction. Compared with the state-of-the-art models, their work could obtain more accurate reconstruction results and increase the probability of the doctor's diagnosis being correct avoiding future follow-on examinations. Since large scale and complexity of magnetic resonance imaging (MRI) data require effective preprocessing approaches, Ke et al. [4] presented an adaptive independent subspace analysis (AISA) method to discover meaningful electroencephalogram activity in MRI scans for supporting diagnostics. In experiments, their proposed model achieved 94.7% of accuracy and 0.9356 of f-score from the real autism spectrum disorder dataset. Połap and Woźniak [5] proposed a model of bacteria recognition based on a composition of region covariance with convolutional neural networks. The process of recognition is divided into two stages. In the first stage, an input microscopy image is segmented by the use of the region covariance model. Then, these segments are forwarded to CNN for recognition of visible bacteria strains. Experimental results show high potential of the proposed methodology. Połap Woźniak [6] gave a segmentation technique based on medical image processing methods and swarm algorithm for lung segmentation on X-ray images for the subsequent diagnosis. The swarm methodology was used for extraction of interested portions with the convolutional neural network as a detector. Khan et al. [7] proposed a novel deep learning framework for the detection and classification of breast cancer in breast cytology images using the concept of transfer learning. This technique facilitates detection and classification of breast cancer in the early stages of its development that may allow patients to have proper treatment. With benefits from related progress of computer-aided

skills, people enjoy more and more efficient health services and protections to improve the quality of lives as far as possible.

Specifically, medical image segmentation has become an essential component that gives more contributions to region of interest (ROI) [8], lesion measurement [9, 10], 3D visualization reconstruction [11, 12], medical data compression, and transmission [13, 14]. Many fully automatic segmentation methods have been proposed over the last decades, including intensity thresholding, region growing, and deformable models. These methods, however, rely on hand-crafted features with the limited feature learning capability. For example, level set models [15–17] are sensitive to noises like lesions in sinus leading to unexpected results. Recently, fully convolutional neural networks (FCNs) have achieved great success on a broad array of segmentation problems in medical images [18–30]. The literature can be classified into two categories broadly. One is based on 2D FCNs depending on multiscale feature map fusion [18, 19, 23]. The other category involves 3D FCNs, where 2D convolutions are replaced by 3D kernels with the volumetric data input [21, 22, 27].

3D convolutional kernel reflects competitiveness at extracting discriminative features along X, Y, and Z directions for 3D classification, detection, and segmentation. However, 3D FCNs, in general, produce an explosion of investment into training parameters and related computational resources. Previous works discuss technical limitations when employing a 3D CNN on medical imaging data [31–33]. In order to incorporate 3D contextual information, multiple works optimize FCN baseline with 2D [18], 2.5D [20, 33], or small 3D patches [34, 35]. Although alleviating the pressure by 3D kernels, these methods pay more attention on irrelevant backgrounds, waste lots of computation resources, and cause a large number of undersegmentations or oversegmentations. To resolve above problems, region of interest (RoI) localization modules are individually designed as a discrete part of the workflow, such as image registration based on atlas with prior knowledge [36, 37]. However, in cases of maxillary sinus segmentation, because of inhomogeneous intensity distribution, their registration performances are poor with a slow speed workflow. Then, RoI detection based on deep learning shows great potentials [25, 38–42]. Some works [38, 39, 42] extract region proposals using external modules like selective search strategy or multiscale combinatorial grouping (MCG) [43], drawbacks of which include time-consuming process of searching the best candidates with limited features considered. Later works introduce an additional well-trained segmentation model in low dimension for RoI localization slice by slice [25, 44]. Although the approaches reduce computation costs dramatically, they have limitations on extracting features between adjacent slices and tend to provide more false positives in localization of the 3D objective.

Multitask models combining bounding box (bbox) localization and segmentation emerge as promising development, such as Mask R-CNN [34] and Multitask Cascaded Convolutional Networks (MTCNN) [45]. Related methods can analyze significant features within bbox and save computation resources to achieve better speed and accuracy. For accurate localization, the region proposal network (RPN) and RoIAlign are introduced to detect and refine bbox localization. Then, multitask networks can manage classification or segmentation work within the bbox to obtain advanced outcomes. In the state-of-the-art, feature pyramid network (FPN) [46] holds competitiveness in multilevel object detection by fusing different levels of feature maps to preserve better details. In this work, authors extend FPN on the Mask R-CNN model. By extracting ROI on each level feature map with RPN, the evaluation of segmentation with intra-RoI FCN acquires more accurate results. However, extending Mask R-CNN with FPN to the 3D mode directly encounters some difficulties. At first, 3D FPN that has symmetric encoder-decoder construction serves excessive GPU memory especially for the high-resolution dataset. Meanwhile, the 3D RPN network produces many 3D anchors of different sizes that also cause a great cost of computation resources. In addition, the distributions of ground truth and background are always imbalanced in medical images. Plenty of different 3D anchors overfit the object so that the 3D bbox cannot have stable localization estimation. Moreover, in Mask R-CNN, the RoIAlign module for localization of RoIs runs bilinear interpolation to resample feature tensors in the anchors to fixed dimensions. Such mechanism results in losing features of details, giving challenges to medical images with the low-level resolution.

To address these issues and inspired by the Mask R-CNN with ResNet-FPN [34] and residual attention network (RAN) [47], we propose a novel multitask framework for segmentation with 3D bounding box estimation, named as 3D bounding box estimation feature pyramid network (BE-FNet), which is designed to effectively extract 3D volumetric maxillary sinus from CT scans in an end-to-end manner. Compared with traditional 3D segmentation models, our proposed model serves more advanced accuracy and computation efficiency as a result of crucial components in deep convolutional neural network architecture. Sufficient ablation studies on collected 50 CT scans demonstrated the superiority of our proposed model with the following main contributions:

(1) We propose a deep neural network with multitask of 3D bounding box estimation and in-region segmentation branches. BE-FNet holds symmetric encoder-decoder architecture with shared parameters. Image encoder is responsible for bounding box estimation and decoder for in-region 3D segmentation similar with U-net. As a result of exploring the target in more significant shrunk space, our proposed model can reduce the computation cost remarkably compared with traditional 3D semantic segmentation neural networks.

(2) To avoid overfitting problems in the state-of-the-art multitask models, we design an overestimation strategy to generate a reasonable 3D bounding box that cannot cause any redundant memory cost. In addition, to increase the depth of the network, we

introduce residual dense blocks as the backbone to enhance the flow of residuals, substantially increasing the depth of the neural network. Moreover, we design a mechanism of attention excitation to improve salient detection applied in bounding box estimation process, which does not give any computation burden for 3D deep neural networks. Especially, the structure of multilevel feature fusion in the pyramid network strengthens the ability of identification to global and local discriminative features in foreground and background, achieving more advanced segmentation results in space.

(3) To resolve the problems of class imbalance and blurring boundary of sinus cavity for segmentation, we define a hybrid loss function of Dice and contour-aware loss. Besides, a multiresolution model ensemble strategy has been introduced to boost segmentation robustness, generating more reliable results and suppressing false positives.

(4) Our model does not depend on any pretrained model or commonly used postprocessing techniques such as 3D conditional random field (CRF). The generalization of the proposed approach is demonstrated through testing on extensive experiments. Not only does our model extract accurate maxillary sinus volume but also achieve competitive performances in related research areas compared with the state-of-the-art methods, which can be generalized in other applications and proved a great promising technique in future.

To the best of our knowledge, this is the first use of attention excitation mechanism to locate and estimate the 3D bounding box for maxillary sinus segmentation with a remarkable performance using the multitask neural network, resulting in a generalized segmentation solution than methods available to date. The entire model is built up based on the backbone of FPN with residual dense blocks depending on different hierarchical feature fusions. These innovations make sure that our model pays attentions on more significant VoI reducing massive computing resource costs and providing more advanced segmentation results.

Our paper is organized as follows. In Section 2, we describe our model in detail and report the experimental results compared with the state-of-the-art methods in Section 3. Section 4 further discusses some insight as well as issues of the proposed method. The conclusions are drawn in Section 5.

2. Materials and Methods

2.1. Overview of Our Proposed Architecture. Our multitask neural network architecture for segmentation is depicted in Figure 2. The proposed architecture consists of three main stages which are responsible for preprocessing, bounding box estimation, and in-region segmentation, respectively. At first, in order to reduce computational cost on redundant information of CT image backgrounds, we adopt the Otsu segmentation algorithm to extract RoIs coarsely based on connectivity analysis [48] in each slice. Besides, the backbone of the network is divided into image encoder and decoder branches. Image encoder focuses on salient detection to estimate the 3D bounding box with a fixed size. Through cropping feature fusions between the encoder and the decoder, the image decoder branch employs in-region segmentation for maxillary segmentation in an end-to-end manner.

2.2. Data Preprocessing. For a medical image, Hounsfield units (HU) are a measurement of relative densities determined by CT. Normally, the HU values range from −1000 to 1000. Any smoothing method is not adopted in our work. Since inhomogeneous texture in the maxillary sinus cavity with lesions provides the significant character in 3D segmentation, noise management is not necessary and destroys potential rules for effective segmentation. Especially, deep neural networks have the capability of learning discriminative features in background or foreground of original medical images, serving great adaptiveness in complicated condition. Therefore, for generalization of our proposed model, we kept the original range of intensity without any preprocessing methods to avoid possible artifacts from image resampling, preserving original details for segmentation. To CT scans, some contents of the image that belong to background waste too much computation resources and are possible to ignore. Therefore, we use Otsu segmentation to quickly extract foreground and its coarse bounding box with connectivity analysis. Then, these cropped images are fed into the BE-FNet and performed data augmentation like scaling, flipping, intensity jittering, and translation. A few examples of the comparison between original and cropped images are illustrated in Figure 3.

2.3. BE-FNet Architecture-Building Deeper Network. To implement fully automatic segmentation for maxillary sinus in CT scans, we design a hybrid neural network unifying 3D bounding box estimation and in-region segmentation with shared features and weights over different tasks. The baseline of BE-FNet borrows spirits from FPN [46], where different level feature maps are fused to promote discriminative feature extraction. Image encoder branch is responsible for exploring the attention of objective to estimate the 3D bounding box. With the decoder in significant VoI, in-region segmentation for maxillary sinus grows benefited from shared features from different levels in the pyramid network. Deeper networks have greater discriminative power due to the additional nonlinearities and better quality of local optima [49]. However, convolutions with 3D kernels are computationally expensive in comparison to the 2D variants, which hamper the addition of more layers. Moreover, 3D architectures have a large number of more trainable parameters, with each layer adding $C_l C_{l-1} \prod_{i=\{x,y,z\}} k_l^{(i)}$ weights to the model. C_l is the number of feature maps in layer l, and $k_l^{\{x,y,z\}}$ is the size of its kernel in the respective spatial dimension. Overall this makes the network increasingly prone to overfitting, which increases GPU memory dramatically.

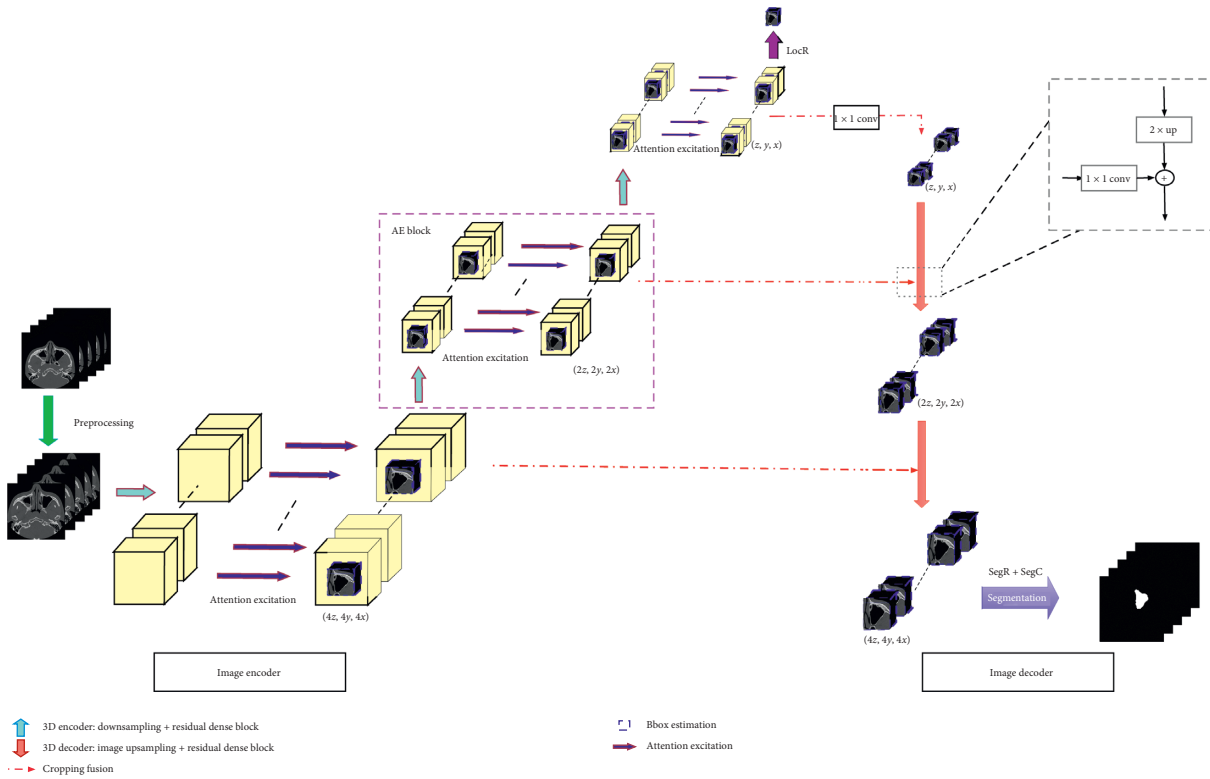
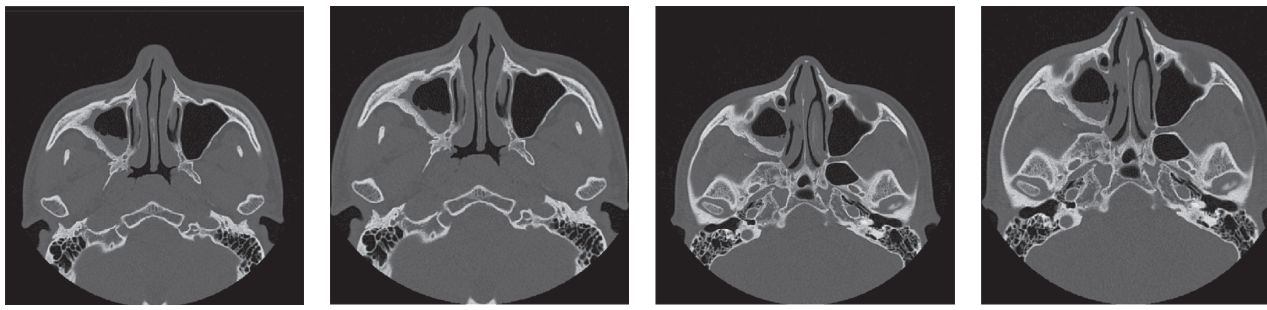


FIGURE 2: Overview of the proposed multitask network for maxillary sinus segmentation. The network consists of three parts: preprocessing, image encoder for bounding box estimation, and decoder for in-region segmentation. Preprocessing extracts coarse RoIs that are passed to the image encoder to estimate the 3D bounding box attributes. Then, cropped bounding boxes are fused to the image decoder following FPN architecture for sinus segmentation with softmax function at the end.



(a) (b) (c) (d)

FIGURE 3: (a and c) Original images. (b and d) Cropped images. Cropped images reduce the GPU memory cost.

In order to set up a deeper 3D architecture, we adopt the sole use of small 3^3 kernels that are faster to convolve with and contain less parameters. This design approach was previously found beneficial for classification of natural images but its effect is even more drastic on 3D networks [22]. When compared to other size choices such as 5^3, the 3^3 kernels reduce the element-wise operations and regarding trainable parameters in a large scale. Therefore, 3D deeper network that use smaller kernels are more efficient to deal with 3D medical image segmentation maintaining acceptable accuracy. However, deeper networks are more difficult to train, where the forward or backwards propagated signal may explode or vanish if care is not given to retain its variance [50]. This phenomenon especially appears in the 3D neural network. Consequently, we introduce a variant of the residual dense block [51] to build the backbone of BE-FNet for 3D image encoder-decoder extracting significant feature maps. The residual dense block holds more depth concatenations and fewer parameters, which serves great discriminative capability and more efficient end-to-end training process. Besides, to avoid the problem of "internal covariate shift" [52], more seriously in the 3D network, we adopt Batch Normalization (BN) technique to all hidden layers [52], which allows normalization of the FM

activations at every optimization step in order to better preserve the training signal. The structure of the residual dense block is described in Figure 4.

2.4. 3D Bounding Box Estimation. This strategy includes salient object detection and the size estimation of the 3D bounding box within the image encoder branch. In semantic segmentation, Share-net is used to set up an attention probability map assisting evaluation to foreground, achieving an excellent performance [53]. More recently, in medical image analysis, many works adopt this idea to localize the objective effectively [54, 55]. Residual attention network (RAN) integrates the residual blocks into architecture of attention branch to enhance the saliency in backgrounds [47]. However, converting RAN to 3D version directly costs massive computation resource in soft mask branch (SMB). Inspired by SE-Net [56], instead of element-wise operations between SMB and trunk branch (TB), we apply the block of attention excitation (AE) for strengthening significant 3D feature expression that forms the attention probability map for 3D bounding box localization.

Figure 5 describes the structure of the AE block in details using attention excitation strategy. At the beginning, the AE block receives the output $z \times h \times w \times c$ of the residual dense block. In the upper branch, the 3D feature maps are squeezed to $z \times c$ numbers with global pooling. For example, feature maps with $z \times h \times w \times$ produce $z \times$ values to concatenate a vector feeding one fully connection network, where hidden layers first squeeze the input size to $(1/8) \times z$ and the output layer restore it to z. The last sigmoid function makes sure the final trainable weights to fall in $[0, 1]$. At the end of both branches, feature fusion happens relying on element-wise multiplication. With the AE block, the trainable attention weights enhance the salient features of the foreground, which facilitates the effective localization of the 3D bounding box eliminating false positives remarkably on maxillary sinus segmentation.

Traditional methods on size estimation, such as RPN, generally set up a trainable regression network to predict the geometric attributes of the bounding box. However, this approach leads to overwhelming computing resource cost for automatic anchor generations and overfitting, which brings unacceptable issues in the deeper 3D neural network. In this paper, we extract the largest salient area with connection analysis. Based on prior knowledge to maxillary sinus volume, we estimate a fixed and overdesigned size $d \times h \times w$ such as $100 \times 150 \times 150$ in the original cube. In other layers of the image encoder, the corresponding sizes are calculated easily according to the up or down sampling ratios. The image encoder branch takes over the entire training process for salient object detection, and to overcome the class imbalance problem in medical images, we introduce Dice loss and weighted crossentropy loss over pixels for RoI attention. Its advantages include free hyperparameter and weak saliency detection. The Dice loss is defined as follows:

$$\mathrm{Loss}(P, G) = 1 - 2 \times \frac{\sum_{i=1}^{N} p_i g_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \varepsilon}, \tag{1}$$

where p and g represent predicted mask and ground truth, respectively. The sums are computed over the N voxels of the predicted volume. ϵ is a smoothness term which prevents from devision by 0. In the optimization stage, the Dice loss is minimized by gradient descend using the following derivate equation (2). Equation (3) gives the hybrid loss for 3D salient detection, where L_{wce} denotes the weighted crossentropy loss and L_d the Dice loss. $\|W\|_2^2$ is the sum of squares of two norms in subnetwork and λ_1, λ_2, and λ_3 weight three terms in L_{roi}:

$$\frac{\partial \mathrm{Loss}(P, G))}{\partial p_k} = -2 \times \frac{\sum_{i=1}^{N} p_i g_i - g_k \sum_{i=1}^{N} (p_i + g_i)}{\left| \sum_{i=1}^{N} (p_i + g_i) \right|^2}, \tag{2}$$

$$L_{\mathrm{roi}} = \lambda_1 L_{\mathrm{wce}}(P_r, G_r) + \lambda_2 L_d(P_c, G_c) + \lambda_3 \|W\|_2^2. \tag{3}$$

2.5. Cropping Fusion Layer. Feature fusion with the skip connection in different levels has promoted the convolutional neural network to acquire more advanced learning ability [18, 46, 57]. For better 3D maxillary sinus segmentation, we set up the network designing multilevel feature fusion layers such as FPN, where the estimated bounding box is cropped directly without any resampling and fused into the decoder branch. In every fusion node, the higher level features are convolved up two times by the residual dense block to concatenate the cropped one of the low level. Related details are illustrated in Figure 2. This mechanism reduces the size of fed data improving efficacy of training and inference in BE-FNet, getting rid of limitations of computing resources to the 3D deeper neural network.

2.6. In-Region Segmentation with Hybrid Loss. Relying on significant 3D bounding box estimation and shared features fusion, the decoder branch is constructed for in-region segmentation with two trainable tasks. One is responsible for evaluation on the performance of entire segmentation. The other pays more attention on identifying the blurring boundary. Consequently, at the end of BE-FNet, a hybrid loss function is designed increasing more constraints to the network, followed by

$$L_{\mathrm{seg}} = \lambda_1 L_d(P_r, G_r) + \lambda_2 L_c(P_c, G_c) + \lambda_3 \|W\|_2^2, \tag{4}$$

where L_d and L_c denote Dice loss and contour-aware loss. λ_1, λ_2, and λ_3 are weighted coefficients and $\|W\|_2^2$ belongs to the regularization term of equation (4). The Dice loss discussed before ensures the performance even though the foreground accounts for a relatively smaller portion in the background. In addition, the object boundary plays a critical role in segmentation task. Especially, in maxillary sinus cavity, as a result of interference caused by lesions, some parts of the sinus boundary are ambiguous and lacks necessary information for feature extraction. Some examples are illustrated in Figure 6. To address this problem, we adopt the recently

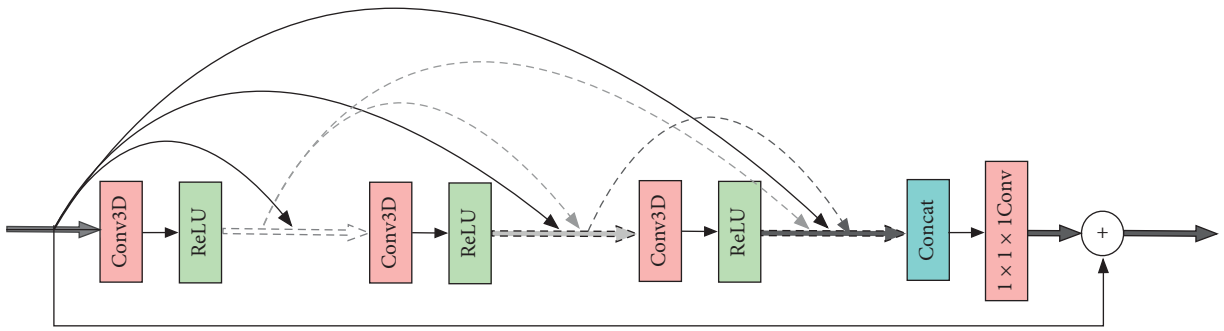


Figure 4: Example of the residual dense block as the building module for our proposed network, which holds the obvious characteristics of residual and dense blocks.

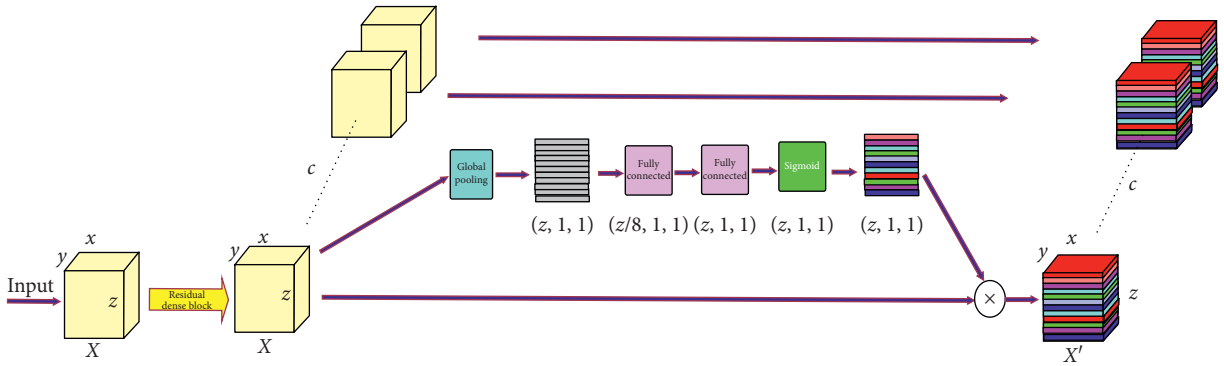


Figure 5: Example of attention excitation block that consists of parameter estimation and feature fusion processes. Compared with traditional methods, it facilitates 3D objective localization in the deeper network.

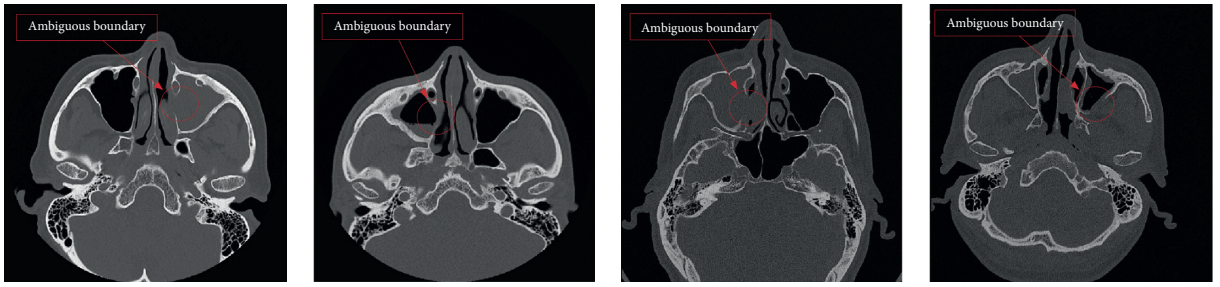


Figure 6: (a–d) Examples of maxillary sinuses images with ambiguous boundaries, which are caused by cavity lesions resulting in troubles for segmentation.

proposed strategy contour-aware loss of the deep contour-aware network (DCAN) to learn and predict the blurring boundary in medical images [26]. Both of the tasks are trained in a parallel and end-to-end way together.

2.7. Multiscale Pathway. The final version of the proposed network BE-FNet is built by extending the primary model with multiple resolution pathways that are identical with the architecture completely, which includes high, normal, and low resolution branches, named H-BE-FNet, N-BE-FNet, and L-BE-FNet, respectively. At the end of each network, we resample the 3D images to original resolution rate and vote the final segmentation based on three predicted probability maps. The spacing between pixels along z, y, and x axes of acquired CT scans fall from $0.5 \times 0.35 \times 0.35$ mm to $0.625 \times 0.39 \times 0.39$ mm in our dataset. Then, we resample the input images spacing ranging from the original to $1.0 \times 1.0 \times 1.0$ mm, $1.5 \times 1.5 \times 1.5$ mm, and $2.0 \times 2.0 \times 2.0$ mm for H-BE-FNet, N-BE-FNet, and L-BE-FNet, respectively.

2.8. Implementation Details. The BE-FNet architecture was implemented using Pytorch [58] and Tensorflow [59] libraries. All the models were trained from scratches. The

parameters of the network were initialized with random values and trained with backpropagation based on Adam [60], using an initial learning rate (LR) of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate would be reduced by 0.1 if the network went to plateau after 20 epochs. Five-fold cross-validation was conducted on 50 scans. The detailed architecture of the BE-FNet network is shown in Table 1. In general, based on prior experience to maxillary sinus anatomical structure, we preferred $3 \times 3 \times 3$ and $1/2 \times 1/2 \times 1/2$ for 3D convolution kernel and max pooling. Preprocessing is responsible for extracting coarse VoI and dropping redundant backgrounds. With the prediction of well-trained LocR, we localized the objective and estimated an over-designed 3D bounding box in the RDBlock4 layer. To the bounding box sizes in RDBlock3 and RDBlock2, we up sampled them along x, y, and z axes by two or four times with trilinear interpolation in order to match the dimensions of different level pyramids for feature fusions. SegR and SegC represent two different tasks with hybrid loss. The AE block denotes a mechanism to excite the object's attention in maximum that does not change the scales of the feature maps. In this multitask network, we first train the LocR branch for 3D bounding box estimation, and then SegR and SegC branches for maxillary sinus segmentation in a parallel way.

3. Results

3.1. Dataset. In our study, approved by an institutional review board for restricted domain in our project, we used 50 CT volume scans (12.13 GB) by SOMATOM Definition AS + SIEMENS containing maxillary sinus to evaluate the proposed multitask network BE-FNet. All of them have the same 512×512 in-plane resolution but with different number of axial slices. The spacing between pixels along z, y, and x axes of the acquired dataset falls within from $0.5 \times 0.35 \times 0.35$ mm to $0.625 \times 0.39 \times 0.39$ mm. The corresponding ground truth is provided by two experienced radiologists manually. The training and inference of our proposed model are run with two NVIDIA GTX1080 Ti 11 GB GPUs, 32G RAM and Intel i7-7700K CPU with 8 cores 4.20 GHz. Especially, we did not adopt any preprocessing of noise management for our dataset in order to preserve original details and avoid possible artifacts for training and inference process.

3.2. Evaluation Metrics. Our segmentation method was evaluated using four quantitative metrics, including Dice Similarity Coefficient (Dice) [61], Volumetric Overlap Error (VOE), Average Symmetric Surface Distance (ASD), and Inference time cost on GPU. In such case, we assume the maxillary sinus as the foreground and the others as the background. The ground truth and predicted region of a maxillary sinus is denoted as A and B, respectively. The Dice is used for precise evaluation of the segmentation results, with a higher number indicating a better result, which is an important indicator for the evaluation of segmentation. Dice $\in [0, 1]$. As for perfect segmentation, Dice = 1:

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{5}$$

The VOE is the nonoverlapping ratio of the segmentation result and ground truth data. It is also used to evaluate the precision of the results, with a lower number indicating a better result, as shown in the following equation:

$$\mathrm{VOE}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{6}$$

ASD (in millimeters) evaluates the distance of the border voxels of segmentation and ground truth. For ASD, a lower value denotes a better segmentation. Related metric is defined as follows:

$$\mathrm{ASD}(A, B) = \frac{1}{|S(A)| + |S(B)|} \left(\sum_{s_A \in S(A)} d(s_A, S(B)) + \sum_{s_B \in S(B)} d(s_B, S(A)) \right). \tag{7}$$

ASD is calculated based on the surface voxels. $S(A)$ denotes the set of surface voxels. The shortest distance from a voxel v to the set $S(A)$ is defined as $d(v, S(A)) = \min_{s_{A \in S(A)}} \|v - s_A\|$, where $\|\cdot\|$ is the Euclidean norm. As for perfect segmentation, this quantity is zero. Inference time is used to evaluate the cost of computational resource and complexity of the BE-FNet model in the inference process.

3.3. Ablation Studies. The challenges regarding the maxillary sinus segmentation fully automatic process include (1) abnormal and ambiguous anatomy structure of maxillary sinus in CT scans, (2) ubiquitous lesions resulting in interferences to accurate segmentation, and (3) the relationship of overwhelming scale of data with the 3D deeper neural network. In experiments, we have evaluated our proposed model for comparison with other state-of-the-art methods on the performances of maxillary sinus segmentation. Examples of predicted vs. ground truth of comparison methods are shown in Figure 7. Eight volume predictions are illustrated in Figure 8.

Our proposed BE-FNet multitask network is divided into two subnetworks bounding box estimation and in-region segmentation. By bounding box estimation, the backbone of the entire network enjoys the acceleration based on the effective VOI eliminating false positives notably. To stick out our proposed model for improved segmentation on speed and quality, we compared our proposed model with the state-of-the-art methods 3D U-Net [21], V-Net [29], HL-FCN [62], 3D CNN + CRF [22], 2D FCN + RNN [63], 3D CNN + Level Set [64], and 3D Deep Nested Level Set [65] to demonstrate the predominance of BE-FNet with the efficient bounding box estimation and in-region segmentation strategy. For the sake of fairness, all models were evaluated on the same receptive field of $92 \times 92 \times 92$. Specifically, in the design of 3D U-Net or 3D FCN, the baseline of the image encoder and decoder was built according to BE-FNet's

Table 1: Architecture of the proposed BE-FNet, consisting of preprocessing, 3D image encoder and decoder, bounding box estimation, cropping fusion, and logits parts. The symbol denotes no information about this item. The first and second columns indicate the descriptions of modules and their sublayers, respectively. The forth and fifth columns tell the size of kernels and their output channels, respectively.

Module name	Layer name	Input layer (s)	Kernel	Out channel (s)	Receptive field
Preprocessing	Coarse RoI extraction	Image	—	—	—
	RDBlock1	Coarse RoI extraction	$3 \times 3 \times 3$	32	$7 \times 7 \times 7$
Image encoder	RDBlock2	RDBlock1	$3 \times 3 \times 3$	32	$13 \times 13 \times 13$
	AEBlock1	RDBlock2	—	32	—
	MaxPooling1	AEBlock1	$1/2 \times 1/2 \times 1/2$	32	$14 \times 14 \times 14$
	RDBlock3	MaxPooling1	$3 \times 3 \times 3$	64	$26 \times 26 \times 26$
	AEBlock2	RDBlock3	—	64	—
	MaxPooling2	AEBlock2	$1/2 \times 1/2 \times 1/2$	64	$28 \times 28 \times 28$
	RDBlock4	MaxPooling2	$3 \times 3 \times 3$	64	$52 \times 52 \times 52$
	AEBlock3	MaxPooling3	—	64	—
	MaxPooling3	RDBlock4	$1/2 \times 1/2 \times 1/2$	64	$56 \times 56 \times 56$
	RDBlock5	MaxPooling3	$1 \times 1 \times 1$	96	$56 \times 56 \times 56$
Cropping fusion	CroppingFusion1	AEBlock1, EB1	—	32	$13 \times 13 \times 13$
	CroppingFusion2	AEBlock2, EB2	—	64	$26 \times 26 \times 26$
	CroppingFusion3	AEBlock3, EB3	—	64	$52 \times 52 \times 52$
Image decoder	UpConv1	CroppingFusion3 UpConv1	$2 \times 2 \times 2$	64	—
	Concat1	CroppingFusion2 Concat1	—	64	—
	RDBlock6	RDBlock6	$3 \times 3 \times 3$	64	$80 \times 80 \times 80$
	UpConv2	UpConv2	$2 \times 2 \times 2$	32	—
	Concat2	CroppingFusion1	—	32	—
	RDBlock7	Concat2	$3 \times 3 \times 3$	32	$92 \times 92 \times 92$
Logits	LocR (Softmax-Task1)	RDBlock5	$1 \times 1 \times 1$	2	$56 \times 56 \times 56$
	SegR (Softmax-Task2)	Concat2	$1 \times 1 \times 1$	2	$92 \times 92 \times 92$
	SegC (Softmax-Task2)	Concat2	$1 \times 1 \times 1$	2	$92 \times 92 \times 92$

Figure 2 and all hyperparameters kept consistent. Meanwhile, in this section we set BE-FNet with Dice loss only instead of hybrid loss, which emphasizes the strength of the bounding box estimation in the neural network including performance and efficient computing speed. Statistical results are listed in Table 2.

3D U-Net [21] is the 3D version of 2D U-Net with multilevel feature maps concatenation. In experiments, we evaluated the 3D U-Net model with different resolutions on Dice loss over pixels. Without the significant VoI, the encoder-decoder branches suffered from more false positives in the same receptive field and acquired relatively lower Dice 0.816 ± 0.084 of three resolutions ensemble strategy. V-net [29] optimizes 3D U-Net using Dice loss and a novel training set augmentation strategy with random nonlinear transformations and histogram matching. Consequently, it gave an obvious improvement Dice 0.883 ± 0.053, VOE 13.87 ± 8.06, and ASD 3.95 ± 3.73. HL-FCN [62] presents the hybrid loss function that is designed under a multitask learning framework to tackle the class imbalance issue and improve the discrimination capability, providing a remarkable Dice of 0.905 ± 0.059. 3D CNN + CRF [22] pays more attention on how to facilitate 3D segmentation model efficiently on CT scans with a small 3D convolutional kernel. Meanwhile, the CRF method is selected to join the end of the network for optimization for outputs. 2D FCN + RNN [63] is derived from the 2D model and adopts RNN to extract features between slices. Both of the methods lack insufficient context utilization, resulting in the excessive oversegmentation phenomenon with Dice 0.828 ± 0.087 and 0.835 ± 0.073, respectively. The level set model relies on curve evolution that runs competent in complicated shape of object segmentation but sensitive to noise independently. To address the problem, 3D CNN + Level Set [64] tries to predict subgrid areas on the probabilities of foreground or background with the deep learning network and allocates weights to the energy, which prevent the level set functional from being trapped into local minima. 3D Deep Nested Level Set [65] lies on 3D CNN to generate proper initial contours to guarantee evolutions happening in target regions. However, the lesions in maxillary sinus cavity appear stochastic to locations and the outside is filled with organs with different densities. Both of models cannot fit these complicated conditions showing lower Dice 0.719 ± 0.140 and 0.783 ± 0.106. In contrast, our proposed model is qualified the ability of advanced bounding box estimation and more accurate in-region segmentation with remarkable results Dice 0.929 ± 0.035, VOE 10.89 ± 5.67, and ASD 3.04 ± 2.48 over five-fold crossvalidations on average. Specifically, with the same baseline of encoder-decoder based on FCN or U-Net, BE-FNet reflects obvious predominance that could be generalized in related research fields.

In addition, we also evaluated BE-FNet in comparisons of the aforementioned methods on inference costs to discuss the time complexity of our proposed model. 3D CNN + Level Set and 3D Deep Nested Level Set were not included since they belong to semiautomatic algorithms interacted by users and related achievements were not satisfied enough. Based on the same configuration of training, V-Net and HL-FCN are fundamentally derived

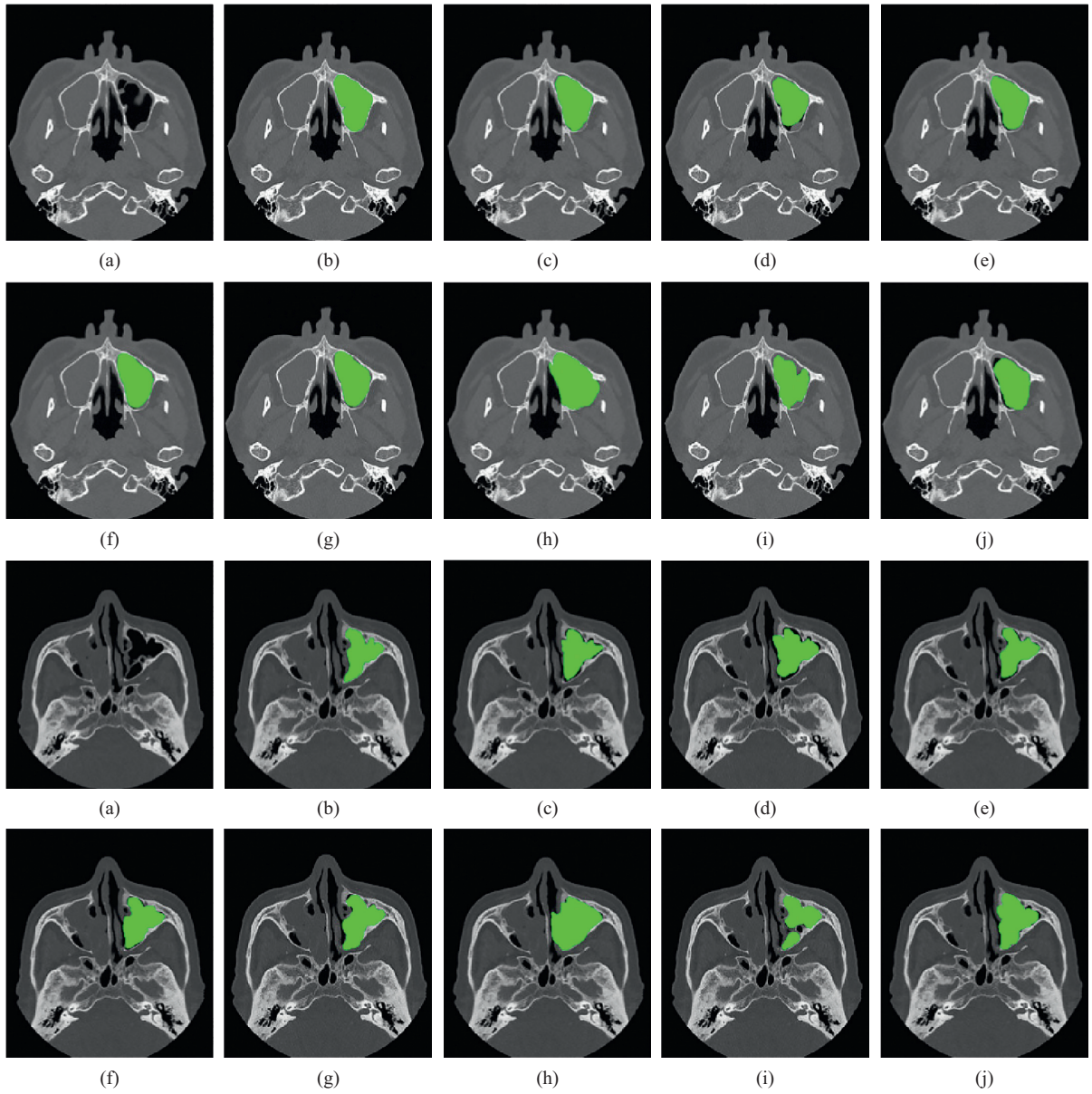


Figure 7: Comparisons of BE-FNet with the state-of-the-art methods. (a) Original CT image, (b) expert delineation, (c) proposed model BE-FNet, (d) 3D U-Net, (e) V-Net, (f) HL-FCN), (g) RA-UNet, (h) 3D Mask R–CNN, (i) 3D CNN + Level Set, and (j) 2D FCN+3D FCN.

from U-Net, and their performances of time complexity were similar, 11.127 s, 11.173 s, and 11.125 s for inference, respectively. As a result of superior bounding box estimation by encoder branch, proper VoI can be extracted significantly, which tremendously scales down the input size fed into the segmentor and helps our proposed neural network to restrain false positives achieving faster inference process 0.511 s and more accurate results.

3.4. Comparison to Other State-of-the-Art Multitask Networks. Furthermore, we also compared BE-FNet to a discrete VoI localization-based method of the multitask network. In detail, 3D Mask R-CNN [34], RA-UNet [66], 2D FCN + 3D FCN [67], and 3D + 2D FCN [68] were considered and results are listed in Table 3. We provide these results for reference and emphasize benefits of our optimized 3D bounding box estimation strategy to maxillary sinus segmentation that supplies generalization in similar tasks. Among the approaches, 3D Mask R-CNN [34] utilizes the RPN network to produce plenty of anchors for fitting foregrounds. In practice, the objects in medical image backgrounds have characters of low contrast and abnormal anatomy structure that cause RPN to generate overestimated or underestimated 3D bounding box leading to failures of bounding box detection and regression, 0.765 ± 0.121, 30.08 ± 12.39, and 9.39 ± 10.72 for Dice, VOE, and ASD, respectively. 2D FCN+3D FCN [67] employs 2D FCN to localize the possible objective on each 2D slice with the predicted probability map, whereas 2D FCN serving as an

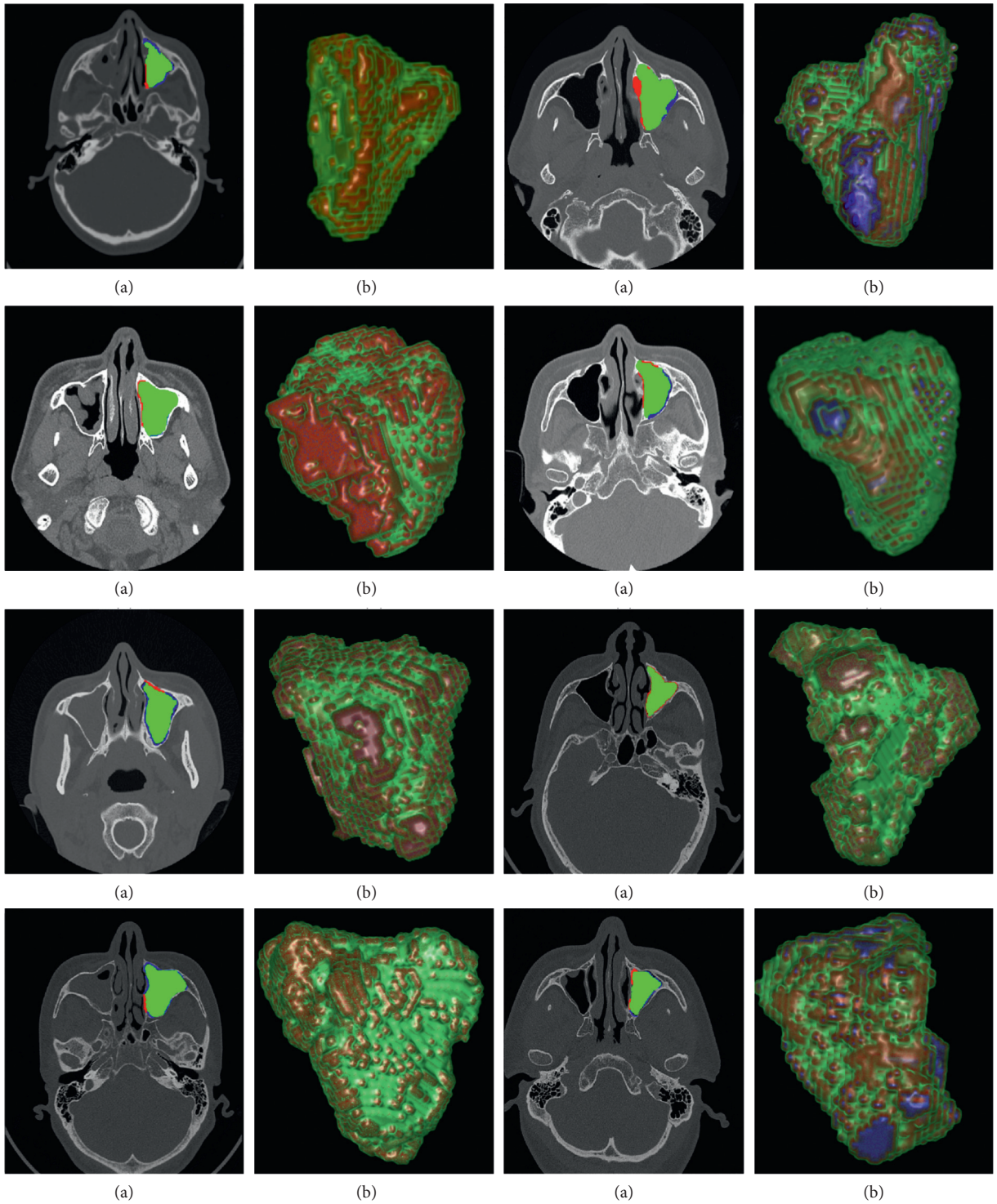


FIGURE 8: (a) Selected 2D slices. (b) 3D segmentation results displayed with volume rendering. Green indicates true positives; red indicates false positives; blue indicates false negatives.

RoI locator produces more false positive candidates due to weak ability of extracting features along z-axis. 3D + 2D FCN [68] estimates the 3D bounding box depending on 3D FCN. For facilitating GPU memory cost, it uses 2D FCN to form the final result that equally cannot deal with volume data at last. RA-UNet [66] designs a novel multitask network for attention localization and in-region segmentation. The attention mechanism lives on the strategy of Residual Attention Network (RAN) [47] to improve VoI accuracy. However, their proposed RAN architecture only focuses on 2D slices with lower relative accuracy, and especially the probability map predicted by RAN is selected

Table 2: Comparisons of the state-of-the-art methods with different metrics. Results are represented as mean ± standard deviation.

Models	Dice	VOE (%)	ASD (mm)	GPU inference time (s)
BE-FNet + ensemble	0.929 ± 0.035	10.89 ± 5.67	3.04 ± 2.48	0.511
BE-FNet (HighRes)	0.889 ± 0.048	14.11 ± 7.92	4.27 ± 3.87	0.302
BE-FNet (MidRes)	0.919 ± 0.044	12.91 ± 6.68	3.32 ± 2.96	0.138
BE-FNet (LowRes)	0.910 ± 0.048	12.96 ± 6.95	3.88 ± 3.25	0.071
3D U-Net + ensemble	0.816 ± 0.084	29.55 ± 10.11	6.91 ± 4.67	11.125
3D U-Net (HighRes)	0.772 ± 0.119	33.80 ± 14.25	9.34 ± 10.26	6.428
3D U-Net (MidRes)	0.791 ± 0.095	30.31 ± 11.67	8.18 ± 8.90	3.176
3D U-Net (LowRes)	0.803 ± 0.088	30.71 ± 10.33	7.13 ± 8.46	1.521
V-Net + Ensemble	0.883 ± 0.053	13.87 ± 8.06	3.95 ± 3.73	11.127
V-net (HighRes)	0.825 ± 0.080	24.04 ± 11.58	6.67 ± 5.16	6.429
V-net (MidRes)	0.856 ± 0.062	15.61 ± 8.83	5.38 ± 3.89	3.176
V-net (LowRes)	0.871 ± 0.069	13.79 ± 8.11	4.59 ± 4.25	1.522
HL-FCN + Ensemble	0.905 ± 0.059	12.62 ± 7.98	3.32 ± 3.49	11.173
HL-FCN (HighRes)	0.859 ± 0.071	16.36 ± 9.14	5.56 ± 3.90	6.466
HL-FCN (MidRes)	0.882 ± 0.055	14.13 ± 8.48	4.14 ± 3.87	3.182
HL-FCN (LowRes)	0.876 ± 0.060	14.82 ± 10.22	4.46 ± 4.01	1.525
3D CNN + CRF (LowRes)	0.828 ± 0.087	22.39 ± 11.99	6.74 ± 4.38	—
2D FCN + RNN (LowRes)	0.835 ± 0.073	20.27 ± 10.74	6.26 ± 3.92	—
3D CNN + Level set (LowRes)	0.719 ± 0.140	41.42 ± 20.16	10.48 ± 11.24	—
3D deep nested level set (LowRes)	0.783 ± 0.106	34.96 ± 13.65	8.72 ± 9.85	—

Table 3: Comparisons of the state-of-the-art multitask networks for VoI localization and segmentation. Results are represented as mean ± standard deviation.

Models	Dice	VOE (%)	ASD (mm)	GPU inference time (s)
3D mask R–CNN (HighRes)	0.751 ± 0.139	31.45 ± 13.34	9.84 ± 11.28	0.614
3D mask R–CNN (MidRes)	0.737 ± 0.136	33.29 ± 14.83	10.61 ± 12.11	0.277
3D mask R–CNN (LowRes)	0.726 ± 0.150	34.73 ± 16.46	12.47 ± 14.63	0.129
3D mask R–CNN + Ensemble	0.765 ± 0.121	30.08 ± 12.39	9.39 ± 10.72	1.020
RA-UNet (HighRes)	0.844 ± 0.067	19.06 ± 9.91	5.83 ± 4.41	0.298
RA-UNet (MidRes)	0.869 ± 0.058	17.24 ± 8.66	4.74 ± 3.96	0.135
RA-UNet (LowRes)	0.857 ± 0.062	17.52 ± 8.97	5.01 ± 4.15	0.063
RA-UNet + Ensemble	0.877 ± 0.059	15.83 ± 8.18	4.65 ± 3.83	0.496
2D FCN+3D FCN (LowRes)	0.829 ± 0.074	24.61 ± 11.74	6.41 ± 4.02	0.068
3D FCN+2D FCN (LowRes)	0.832 ± 0.076	21.23 ± 10.95	6.48 ± 3.77	0.067
BE-FNet (HighRes)	0.889 ± 0.048	14.11 ± 7.92	4.27 ± 3.87	0.302
BE-FNet (MidRes)	0.919 ± 0.044	12.91 ± 6.68	3.32 ± 2.96	0.138
BE-FNet (LowRes)	0.910 ± 0.048	12.96 ± 6.95	3.88 ± 3.25	0.071
BE-FNet + Ensemble	0.929 ± 0.035	10.89 ± 5.67	3.04 ± 2.48	0.511
BE-FNet + HL (HighRes)	0.898 ± 0.049	14.15 ± 7.96	4.17 ± 3.92	0.317
BE-FNet + HL (MidRes)	0.925 ± 0.041	12.87 ± 6.62	3.31 ± 2.89	0.149
BE-FNet + HL (LowRes)	0.916 ± 0.045	12.79 ± 6.83	3.82 ± 3.14	0.073
BE-FNet + HL + Ensemble	0.947 ± 0.031	10.23 ± 5.29	2.86 ± 2.11	0.539

as a coarse segmentation giving the problem of underestimated sizes for the bounding box. In this section, BE-FNet + HL refers to BE-FNet with hybrid loss to stress advantages of contour-aware loss. BE-FNet + HL adopts the AE block to explore 3D target attention effectively and estimate an overestimated but proper size of the 3D bounding box for preventing from more false positives. In addition, similar with skip connection of FPN architecture, shared features of different levels in pyramid hierarchy with cropping fusion connections enable our proposed model to achieve the best performance Dice 0.947 ± 0.031, VOE 10.23 ± 5.29, and ASD 2.86 ± 2.11, respectively, emphasizing the advantages of tasks joint training and cross-module feature sharing. Besides, the outperformance of BE-FNet + HL than BE-FNet indicates benefits of hybrid loss attractively.

Besides, we compared our proposed BE-FNet with the state-of-the-art algorithms for evaluations on time complexity in order to verify our achievement of multitask. Because of generating multiple anchors on each pixels in feature maps, RPN had to infer proper bounding box resulting in great cost of computation with 1.020s for 3D Mask R-CNN + Ensemble. Although RA-UNet + Ensemble depended on salient detection to explore candidate bounding boxes saving plenty of time complexity, this method lacked an effective strategy of estimation that causes excessive false positives in prediction. 2D FCN + 3D FCN and 3D FCN + 2D FCN tried 2D convolution operations to

learn discriminative features in context that is not competent in space and leads to unsatisfactory results, even though they served faster speed 0.068 s and 0.067 s, respectively. Since BE-FNet + Ensemble adopts efficient and effective mechanism of bounding box estimation, they provided more advanced result of Dice 0.929 ± 0.035 with time cost 0.511 s. BE-FNet + HL + Ensemble with hybrid loss addressed issues of class imbalance and blurring boundary and ran the best performance of Dice 0.947 ± 0.031, which gave more time complexity 0.539 s in an inference process. Statistics demonstrate that our proposed BE-FNet can not only outperform the state-of-the-art models in segmentation accuracy but also reflect a lower time complexity that could be facilitated in computer-aid diagnosis.

4. Discussion

For accurate and robust maxillary sinus segmentation in clinical diagnosis, we propose a novel multitask neural network to implement an end-to-end training and inference process. There are some difficulties for traditional methods including inhomogeneous intensity, plenty of lesions, abnormal anatomical structure, blurring boundary of sinus cavity, and excessive computing costs in the deeper 3D neural network. For a fully automatic segmentation skill in generalization, we provide a novel model BE-FNet adaptive to maxillary sinus in low contrast CT scans. The main advantages of the proposed approach are demonstrated:

(1) To facilitate 3D segmentation of large data in CNN, we design an efficient and effective deeper neural network with multitask of estimating 3D bounding box and in-region segmentation. The 3D bounding box estimation helps to reduce great computing cost and eliminate false positives remarkably enhancing capability of generalization in our proposed network.

(2) To prevent overfitting problems happened in lots of research studies, an overestimation strategy is devised to generate a proper 3D bounding box that is able to extract the most significant space for in-region segmentation. Besides, for increasing depth of the deep neural network, we design residual dense blocks as the backbone of the model to improve the capability of learning.

(3) We supply a mechanism of attention excitation to improve salient detection applied in bounding box estimation process, which does not give any computation burden for 3D deep neural networks. Especially, the structure of multilevel feature fusion in the pyramid network strengthens the ability of identification to global and local discriminative features in foreground and background achieving more advanced segmentation results in space.

(4) To resolve the problem of blurring boundary in sinus cavity, we design a hybrid loss function with Dice and contour-aware loss. Moreover, a multiresolution model ensemble strategy has been introduced to boost segmentation robustness, generating more reliable results and constraining false positives tremendously.

In addition, the whole baseline of our proposed model is fully automatic. At the beginning, we need to train the image encoder branch for effective bounding box estimation. Then, with the prediction, the entire network completes an end-to-end process for in-region segmentation. This novel automatic framework combining hybrid tasks and loss functions provides more accurate maxillary sinus segmentation especially in low contrast and noisy CT scans. To show the generalization capability of our method in the clinical practice, we tested our trained model on dataset with five crossfold evaluations. First of all, we compared BE-FNet with the state-of-the-art frameworks to stress the importance of 3D bounding box estimation. Figure 7 illustrates that our proposed model can deal with cases in low contrast, heterogeneous, noisy backgrounds, and outperforming commonly used frameworks based on deep learning. As proven, Table 2 and Figure 7 demonstrate that BE-FNet has more accuracy and robustness regardless of possible lesions, holding an average Dice 0.929 ± 0.035, VOE 10.89 ± 5.67, and ASD 3.04 ± 2.48 with obvious superiority among approaches. Meanwhile, we also evaluated different models on the inference time cost of GPU. As a result of significant VoI extraction, the magnitude of trained data is reduced by an exciting extent, which facilitates our research in deeper 3D network. Furthermore, to emphasize the efficacy of our proposed strategy on bounding box estimation, 3D Mask R-CNN [34], RA-UNet [66], 2D FCN + 3D FCN [67], and 3D + 2D FCN [68] based on multitask networks with localization joined in comparison and BE-FNet achieved the state-of-the-art results on maxillary sinus segmentation. These findings indicate three key points. At first, AE block mechanism benefits the accurate salient object localization. Besides, the setting of experienced overdesigned size for the bounding box eliminates false positives as far as possible. At last, hybrid loss functions explore the optimized balance of extracting blurring boundary and small object segmentation in noisy texture. Consequently, our proposed network that combines efficient 3D bounding box estimation and in-region segmentation tasks overcomes the aforementioned issues of popular methods, serving a significant advanced result.

The presented work has some limitations. At first, for salient detection with multiple objectives, if they are overlapped or close, our proposed mechanism tends to make mistakes of identification, which influences the performances of 3D bounding box estimations and in-region segmentation. Consequently, we should further discuss how to effectively estimate multiple bounding boxes and segment multiple objects in practice. In addition, in cases, we found that areas of salient detection account for low percentages of ground truth, which causes the estimated center of the 3D bounding box to deviate from excepted position away and more false positives in segmentation. A possible solution could incorporate dilated convolutions to enlarge local receptive fields for exploring the complete attention picture.

5. Conclusion

In this paper, we present a multitask neural network for 3D maxillary sinus segmentation from CT scans, which consists of 3D bounding box estimation and in-region 3D segmentation. With AE block mechanism, the proposed model is able to detect the maxillary sinus effectively. Then, based on geometrics, an overdesigned size of the 3D bounding box is estimated. Compared with the state-of-the-art methods, this strategy hinders from inappropriate VoIs resulting in oversegmentation or undersegmentation. Moreover, through cropping fusion layers the shared features in different hierarchy of the pyramid network improve the in-region segmentation results remarkably. At last, to address the issue of the blurring cavity boundary, the hybrid loss function guarantees advanced extraction of candidate boundaries and small objective segmentation in noisy backgrounds of the medical image. Compared with the state-of-the-art methods, our BE-FNet is benefited from bounding box localization which saves computing resources and improves the performance of in-region segmentation. To further evaluate the proposed estimation mechanism, we tested it to compete with popular models such as 3D Mask R-CNN. After extensive experiments, the competitive results were found, respectively. Some limitations are presented for future work to be optimized.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by the Liaoning Provincial Natural Science Foundation of China (Grant no. 2019-MS-112).

References

[1] J. S. Rhee, D. T. Book, M. Burzynski, and T. L. Smith, "Quality of life assessment in nasal airway obstruction," *The Laryngoscope*, vol. 113, no. 7, pp. 1118–1122, 2003.

[2] K. Tingelhoff, A. I. Moral, M. Elizete Kunkel et al., "Comparison between manual and semi-automatic segmentation of nasal cavity and paranasal sinuses from ct images," in *Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5505–5508, IEEE, Lyon, France, August 2007.

[3] W. Wei, B. Zhou, D. Połap, and M. Woźniak, "A regional adaptive variational pde model for computed tomography image reconstruction," *Pattern Recognition*, vol. 92, pp. 64–81, 2019.

[4] Q. Ke, J. Zhang, W. Wei, R. Damasevicius, and M. Wozniak, "Adaptive independent subspace analysis of brain magnetic resonance imaging data," *IEEE Access*, vol. 7, pp. 12252–12261, 2019.

[5] D. Połap and M. Woźniak, "Bacteria shape classification by the use of region covariance and convolutional neural network," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Budapest, Hungary, July 2019.

[6] D. Połap and M. Woźniak, "Lung segmentation on x-ray images with neural validation," in *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Honolulu, HI, USA, December 2017.

[7] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. J. P. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019.

[8] Y. Yu, P. Decazes, J. Lapuyade-Lahorgue, I. Gardin, P. Vera, and R. Su, "Semi-automatic lymphoma detection and segmentation using fully conditional random fields," *Computerized Medical Imaging and Graphics*, vol. 70, no. 1–7, 2018.

[9] M. van Eijnatten, R. van Dijk, D. Johannes, G. Streekstra, J. Koivisto, and J. Wolff, "Ct image segmentation methods for bone used in medical additive manufacturing," *Medical Engineering & Physics*, vol. 51, no. 6–16, 2018.

[10] E. Abdulhay, M. A. Mohammed, N. A. Dheyaa Ahmed Ibrahim, and V. Venkatraman, "Computer aided solution for automatic segmenting and measurements of blood leucocytes using static microscope images," *Journal of Medical Systems*, vol. 42, no. 4, p. 58, 2018.

[11] E. Soodmand, D. Kluess, P. A. Varady et al., "Interlaboratory comparison of femur surface reconstruction from ct data compared to reference optical 3d scan," *Biomedical Engineering Online*, vol. 17, no. 1, p. 29, 2018.

[12] I. Mehmood, M. Sajjad, K. Muhammad et al., "An efficient computerized decision support system for the analysis and 3d visualization of brain tumor," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 12723–12748, 2019.

[13] X. Baik, Q. Huang, S. Chang, J. He, and H. Wang, "Lossless medical image compression using geometry-adaptive partitioning and least square-based prediction," *Medical & Biological Engineering & Computing*, vol. 56, no. 6, pp. 957–966, 2018.

[14] Z. Fan, L. Sun, X. Ding, Y. Huang, C. Cai, and J. Paisley, "A segmentation-aware deep fusion network for compressed sensing mri," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 55–70, Munich, Germany, September 2018.

[15] C. Chunming Li, C. Chenyang Xu, C. Changfeng Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[16] C. Li, R. Huang, Z. Ding, J. C. Gatenby, D. N. Metaxas, and J. C. Gore, "A level set method for image segmentation in the presence of intensity inhomogeneities with application to mri," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2007–2016, 2011.

[17] S. Lankton and A. Tannenbaum, "Localizing region-based active contours," *IEEE Transactions on Image Processing*, vol. 17, no. 11, pp. 2029–2039, 2008.

[18] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing, Berlin, Germany, 2015.

[19] M. F. Stollenga, W. Byeon, L. Marcus, and J. Schmidhuber, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2998–3006, Montreal, Canada, June 2015.

[20] H. R. Roth, L. Lu, A. Farag et al., "Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 556–564, Springer, Munich, Germany, October 2015.

[21] Ö. Çiçek, A. Ahmed, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Springer, Athens, Greece, October 2016.

[22] K. Kamnitsas, C. Ledig, V. F. J. Newcombe et al., "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[23] M. Havaei, A. Davy, D. Warde-Farley et al., "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[24] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[25] X. Li, H. Chen, X. Qi, D. Qi, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[26] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: deep contour-aware networks for object instance segmentation from histology images," *Medical Image Analysis*, vol. 36, pp. 135–146, 2017.

[27] H. Chen, D. Qi, X. Wang, J. Qin, J. C. Y. Cheng, and P.-A. Heng, "3d fully convolutional networks for intervertebral disc localization and segmentation," in *Proceedings of the International Conference on Medical Imaging and Virtual Reality*, pp. 375–382, Springer, Bern, Switzerland, August 2016.

[28] Q. Dou, L. Yu, H. Chen et al., "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[29] F. Milletari, N. Navab, and A. Seyed-Ahmad, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, Stanford, CA, USA, October 2016.

[30] L. Yu, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images," *AAAI*, vol. 66–72, 2017.

[31] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 246–253, Springer, Nagoya, Japan, September 2013.

[32] R. Li, W. Zhang, H.-I. Suk et al., "Deep learning based imaging data completion for improved brain disease diagnosis," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 305–312, Springer, Boston, MA, USA, September 2014.

[33] H. R. Roth, L. Lu, A. Seff et al., "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 520–527, Springer, Boston, MA, USA, September 2014.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, IEEE, Venice, Italy, October 2017.

[35] N. Dong, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-d fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1123–1136, 2018.

[36] F. van der Lijn, T. den Heijer, M. M. B. Breteler, and W. J. Niessen, "Hippocampus segmentation in mr images using atlas registration, voxel classification, and graph cuts," *Neuroimage*, vol. 43, no. 4, pp. 708–720, 2008.

[37] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy," *Neuroimage*, vol. 46, no. 3, pp. 726–738, 2009.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[39] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[40] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3992–4000, Boston, MA, USA, June 2015.

[41] F. Liao, Xi Chen, X. Hu, and S. Song, "Estimation of the volume of the left ventricle from mri images using deep neural networks," 2017, https://arxiv.org/pdf/1702.03833.pdf.

[42] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 447–456, Boston, MA, USA, June 2015.

[43] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 328–335, Columbus, OH, USA, June 2014.

[44] M. Tang, Z. Zhang, C. Dana, J. Martin, and J. L. Jaremko, "Segmentation-by-detection: a cascade network for volumetric medical image segmentation," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1356–1359, IEEE, Washington, DC, USA, April 2018.

[45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[46] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CVPR*, vol. 1, p. 4, 2017.

[47] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," 2017, https://arxiv.org/abs/1704.06904.

[48] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[49] C. Anna, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proceedings of the Artificial Intelligence and Statistics*, pp. 192–204, San Diego, CA, USA, 2015.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, Sardinia, Italy, May 2010.

[51] Y. Zhang, Y. Tian, Yu Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, Salt Lake City, UT, USA, June 2018.

[52] Sergey Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, https://arxiv.org/abs/1502.03167.

[53] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, Las Vegas, NV, USA, June 2016.

[54] O. Oktay, J. Schlemper, L. Le Folgoc et al., "Attention u-net: learning where to look for the pancreas," 2018, https://arxiv.org/abs/1804.03999.

[55] S. Jo, O. Oktay, L. Chen et al., "Attention-gated networks for improving ultrasound scan plane detection," 2018, https://arxiv.org/abs/1804.05338.

[56] J. Hu, Li Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[57] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[58] P. Adam, S. Gross, S. Chintala, and G. Chanan, "Pytorch: tensors and dynamic neural networks in python with strong GPU acceleration," 2017.

[59] M. Abadi, B. Paul, J. Chen et al., "Tensorflow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283, Savannah, GA, USA, 2016.

[60] D. P. Kingma and B. Jimmy, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[61] C. Li, X. Wang, S. Eberl, M. Fulham, Y. Yin, and D. Dagan Feng, "Supervised variational model with statistical inference and its application in medical image segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 196–207, 2015.

[62] Y.-J. Huang, D. Qi, Z.-X. Wang et al., "Hl-fcn: hybrid loss guided fcn for colorectal cancer segmentation," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 195–198, IEEE, Washington, DC, USA, April 2018.

[63] G. Liang, H. Hong, W. Xie, and L. Zheng, "Combining convolutional neural network with recursive neural network for blood cell image classification," *IEEE Access*, vol. 6, pp. 36188–36197, 2018.

[64] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical Physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[65] J. Duan, S. Jo, W. Bai et al., "Deep nested level sets: fully automated segmentation of cardiac mr images in patients with pulmonary hypertension," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 595–603, Springer, Granada, Spain, September 2018.

[66] Q. Jin, Z. Meng, C. Sun, L. Wei, and R. Su, "Ra-unet: a hybrid deep attention-aware network to extract liver and tumor in ct scans," 2018, https://arxiv.org/abs/1811.01328.

[67] M. Rezaei, H. Yang, and C. Meinel, "Instance tumor segmentation using multitask convolutional neural network," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, Rio de Janeiro, Brazil, July 2018.

[68] S. Rafiei, E. Nasr-Esfahani, K. Najarian, N. Karimi, S. Samavi, and S. M. Reza Soroushmehr, "Liver segmentation in ct images using three dimensional to two dimensional fully convolutional network," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2067–2071, IEEE, Athens, Greece, October 2018.