Efficient Multiple Organ Localization in CT Image Using 3D Region Proposal Network

Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai

***Abstract*— Organ localization is an essential preprocessing step for many medical image analysis tasks, such as image registration, organ segmentation, and lesion detection. In this paper, we propose an efficient method for multiple organ localization in CT image using a 3D region proposal network. Compared with other convolutional neural network-based methods that successively detect the target organs in all slices to assemble the final 3D bounding box, our method is fully implemented in a 3D manner, and thus, it can take full advantages of the spatial context information in CT image to perform efficient organ localization with only one prediction. We also propose a novel backbone network architecture that generates high-resolution feature maps to further improve the localization performance on small organs. We evaluate our method on two clinical datasets, where 11 body organs and 12 head organs (or anatomical structures) are included. As our results shown, the proposed method achieves higher detection precision and localization accuracy than the current state-of-the-art methods with approximate 4 to 18 times faster processing speed. Additionally, we have established a public dataset dedicated for organ localization on http://dx.doi.org/10.21227/df8g-pq27. The full implementation of the proposed method has also been made publicly available on https://github.com/superxuang/caffe_3d_faster_rcnn.**

***Index Terms*— Organ localization, CT image, convolutional neural network, region proposal network.**

I. Introduction

ORGAN localization plays an important role in clinical practice. It could be used as a start point for many automatic medical image analysis tasks, such as image registration [1]–[3], organ segmentation [1]–[9] and lesion detection [6]. Appropriate initial estimation of organ position and extent can largely improve the performance of the subsequent processing procedures. For instance, by discriminating the presence of target organs (*detecting*) and predicting their bounding boxes (B-box) (*locating*), organ segmentation methods can discard most of non-relevant information and focus on regions that are more likely to contain the target organs. It not only improves the computation and memory efficiency but also reduces the risk of false positive segmentations [8]. Moreover, organ localization is also important for efficient data retrieval and visual navigation of CT scans [2], [3], [9], [10]. With auto-generated regions of interest, there is less need for retrieving the entire CT scan from the picture archiving and communication systems (PACS) database and search the target organ slice by slice manually. Note that, for the brevity, we use the word "localization" to present the meaning of detecting as well as locating in this paper.

However, organ localization in CT image is a challenging task. The target organ usually has a large variation of appearance across different patients. Neighboring low-contrast soft-tissues and noise also cloud the definition of organ boundaries. The high dimensionality of CT image makes the localization procedure require large amounts of time and memory but it is just a preprocessing step of the subsequent algorithms. In addition, due to the variation of scanning range, the target organ may be truncated in the CT image, which is ambiguous to the classifier.

In the early times, many classical machine learning-based methods are proposed to tackle this challenging problem. These methods estimate the B-box of the organs using different classifiers (e.g. AdaBoost [5], probabilistic boosting-tree [4], [11], [12], and random forests [2], [3]) trained on various handcraft features (e.g. intensity and gradient features [2], [3], Haar-like features [4], [5], [11], [12]). Zheng *et al.* [4], [11], [12] propose a marginal space learning algorithm and exploit it to perform efficient 3D anatomical structure localization in CT image. In this method, the geometry of 3D anatomical structures are described as nine pose parameters (three for position, three for orientation, and three for anisotropic scaling) which are successively predicted in three consecutive stages. Zhou *et al.* [5] introduce an ensemble learning method for automatic organ localization. They train a set of Adaboost classifiers using Haar-like features extracted from 2D image slices, and combine the output of these

Manuscript received November 15, 2018; revised January 15, 2019; accepted January 17, 2019. Date of publication January 24, 2019; date of current version July 31, 2019. This work was supported by the National Natural Science Foundation of China under Grant No. 61601012 and the National Key R&D Program of China under Grant No. 2017YFC0113100. *(Corresponding authors: Bo Liu; Xiangzhi Bai.)*

X. Xu is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: xuang199085@163.com).

F. Zhou, B. Liu, and X. Bai are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191 China (e-mail: zhfugen@buaa.edu.cn; bo.liu@buaa.edu.cn; jackybxz@buaa.edu.cn).

D. Fu is with the Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China (e-mail: dfu@rayertech.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2894854

classifiers through a collaborative majority voting procedure to produce the final B-box of the target organ. Criminisi *et al.* [2] propose a method which utilizes random regression forests to predict multiple organ B-boxes based on a set of handcraft intensity and gradient features. Gauriau *et al.* [3] further develop this method by cascading the random forests classifiers to locate organs in a coarse-to-fine manner.

In recent years, deep convolutional neural networks (ConvNets) with hierarchical feature learning capability has become a popular methodology for medical image analysis. Many ConvNet-based methods for organ localization are proposed [6]–[9], [13], outperforming the classical machine learning-based methods which rely on the handcraft features. de Vos *et al.* [7] propose a ConvNet-based method for automatic multi-organ localization in CT image. In this work, three independent ConvNets are trained to predict target organs' presence in 2D slices extracted along three orthogonal directions (i.e. axial, sagittal, and coronal). By combining the output of these three ConvNets, the centroid and extent of the organs can be determined. As a further research [9], they merge the three independent ConvNets into one and introduce spatial pyramid pooling (SPP) layers [14] into their network to process slices of variant sizes. Humpire-Mamani *et al.* [13] propose an aggregated orthogonal decision ConvNet performing voxel-wise predictions to locate kidneys from CT image. Three orthogonal slices are fed to a ConvNet to determine whether the voxel where the three input slices intersect is inside or outside the B-box of kidney. Humpire-Mamani *et al.* [6], [8] employ three ConvNets to predict the extent of the target organs along three orthogonal directions. Compared with the method by de Vos *et al.* [7], their ConvNets take several adjacent slices as input to predict the presence of target organs in the central slice and achieve the current state-of-the-art performance in this field. However, almost all these methods employ 2D ConvNets to perform slice-wise predictions on the volumetric CT images. There could be two limitations in this strategy. Firstly, the ConvNet needs to be run separately for each slice in three directions. It is time consuming and may be redundant since most of the adjacent slices present almost identical contents. Secondly, the ConvNets implemented in 2D manners cannot make full use of the 3D spatial contextual information in CT image, and may be incapable to perform accurate organ localization, especially for the small organs. Thus, we argue that handling this problem using 3D ConvNets could be more efficient and robust.

At the early stage of our research, we attempt to exploit a 3D version of faster R-CNN [15] to tackle the problem of organ localization in CT image. Faster R-CNN [15] is a ConvNet-based method for general object detection. Currently top leading methods on natural object detection challenges (e.g. PASCAL VOC [16], [17] and MS COCO [18]) almost all utilize the faster R-CNN framework or its variants [19], [20]. The major workflow of faster R-CNN is illustrated in the left of Fig.1. Given an input image, the faster R-CNN firstly employs a ConvNet, called backbone network, to extract its feature map. This feature map is shared by a subsequent region proposal network (RPN) [15] and a fast R-CNN classifier [21] as input. The RPN generates a number of class-agnostic candidate B-boxes that are more likely to contain foreground objects, and the fast R-CNN performs multi-classification and bounding refinement on these candidate B-boxes. Finally, non-maximum suppression (NMS) is adopted to eliminate redundancy and the top-scored candidate B-box is kept as the final output. We implement the faster R-CNN fully in 3D manner and successfully exploit it to locate organs in CT image. According to our experiment results, this 3D faster R-CNN achieves comparable localization accuracy and much faster processing speed comparing with the current state-of-the-art methods [7]–[9].

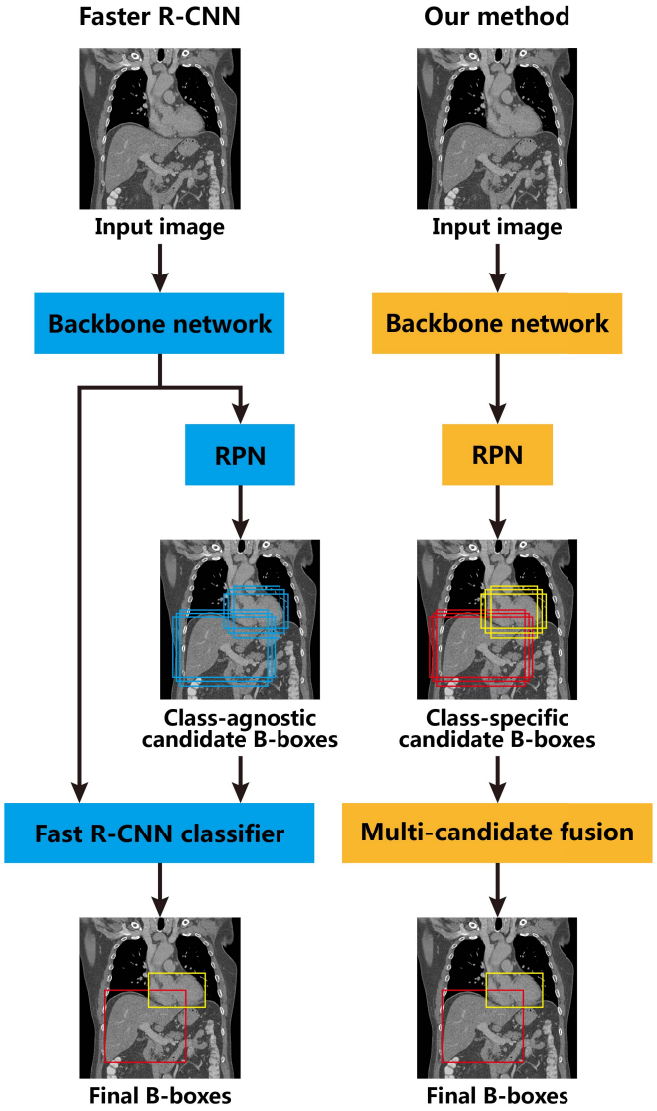

Fig. 1. Comparison between the workflow of faster R-CNN (left) and our proposed method (right).

However, the faster R-CNN framework is specially designed for general object detection in natural image, thus it may not fully consider the peculiarity of organ localization in CT image. Compared with the general object detection task that needs to detect an arbitrary number of instances for each class, there is at most one instance for each organ in CT image. By utilizing this characteristics, we simplify

the 3D faster R-CNN and propose a simple but effective method for multi-organ localization in CT image. As the workflow shown in the right of Fig.1, we remove the fast R-CNN classifier from the faster R-CNN framework and directly exploit the RPN to generate class-specific candidate B-boxes. Since there is at most one instance for each organ, we directly fuse the candidate B-boxes that are assigned with same labels to produce the final B-box. As our experiments results shown, this multi-candidate fusion strategy not only eliminates redundant candidate B-boxes, but also plays a role of bounding refinement which is originally performed by the fast R-CNN classifier in the faster R-CNN. In order to further improve the localization accuracy of our method, we also propose a novel backbone network to generate high-resolution feature maps. It is intuitive that higher-resolution feature maps can lead to richer spatial information and larger number of candidate B-boxes, both of which can contribute to the improvement of localization accuracy, especially for the small organs. We successively evaluate our method on an abdominal clinical CT dataset and a head clinical CT dataset, where the localization of 11 body organs and 12 head organs (or anatomical structures) are performed. Compared with the current state-of-the-art method, our method achieves higher detection precision and localization accuracy with approximate 4∼18 times faster processing speed.

Another contribution of this work is that we build and share a set of annotations (i.e. the B-boxes of 11 body organs) based on a public abdominal clinical CT dataset. Currently, there is no dedicated public dataset for organ localization. Most of the previous works [2], [3], [7], [9], [13] perform evaluations on their private datasets. This makes it difficult to conduct a direct comparison between different methods in this field. To this end, we build this annotation set based on a public clinical CT dataset and open it to other researchers.

In summary, the major contributions of this work are in three folds:

1) We propose an efficient method for organ localization in CT image based on the 3D RPN. Unlike the 2D ConvNet-based methods that successively detect the target organs in all slices to assemble the final 3D B-box, our method is fully implemented in 3D manner, thus can take full advantages of the spatial context information in CT image to perform fast organ localization with only one prediction. As our results shown, the proposed method achieves approximate 4∼18 times faster processing speed than the current state-of-the-art methods.
2) We design a novel backbone network architecture to generate high-resolution feature maps for accurate small organ localization. By combining the hierarchical feature maps in different resolutions through deconvolution operation and skip connection, the proposed backbone network can produce high-resolution feature maps that contain strong semantic information as well as fine spatial details. Benefiting from this design, our method achieves higher organ detection and localization accuracy than the current state-of-the-art methods, especially for the small organs.
3) We have built a dataset for body organ localization based on a public clinical abdominal CT image dataset and made this dataset publicly available to other researchers. The B-box annotations of 11 body organs are done by us and included in this dataset. To the best of our knowledge, this is the first dedicated public dataset for organ localization.

It is well-known that the R-CNN based methods are the state of the arts for 2D object detection. However, their performance for the problem of 3D organ localization is still unknown. Although it is not the first time that the faster R-CNN framework is extended to 3D versions [22]–[24], to the best of our knowledge, this work is the first attempt to locate multiple organs in CT image using R-CNN based method. And the experimental results show that it achieves compelling performance in comparison to previous methods.

The remaining sections are organized as following: Section II introduces the datasets and data augmentation strategies used in this work. Section III presents our proposed method, its training methodology and implementation details. In Section IV, we conduct extensive experiments to evaluate the proposed method and justify the choice made in our design. Some issues are specially discussed in Section V. Finally, we conclude this work in Section VI.

II. Materials

A. Datasets

In this work, we conduct experiments on an abdominal clinical CT dataset and a head clinical CT dataset. The information about these two datasets are summarized in Table I and introduced in detail next.

The abdominal clinical CT dataset is built on the MICCAI Liver Tumor Segmentation (LiTS) challenge[1] dataset, an ongoing benchmark for evaluating algorithms on liver and liver lesion segmentation. It consists of 201 contrast-enhanced abdominal CT scans collected from various clinical sites around the world, 131 for training and 70 for testing. 13 CT scans are randomly chosen from the training set for validation. Based on these CT scans, we annotate B-boxes for 11 body organs: heart (50/3/28), left lung (49/3/21), right lung (49/3/21), liver (118/13/70), spleen (118/13/70), pancreas (118/13/70), left kidney (117/12/70), right kidney (118/13/69), bladder (98/11/67), left femoral head (98/11/66) and right femoral head (94/11/66). The number in the parentheses indicates the number of the organ annotated in training, validation and testing sets, respectively. The B-box of liver in the training set are directly computed according to their segmentation masks which are provided by the MICCAI LiTS challenge. Other B-boxes are annotated by two radiologists and verified/modified by a third radiologist who is more experienced. Note that, all truncated organs, which are not fully contained in the CT image, are considered as background because in clinical practice the organ of interest would be completely included in the scanning range. We use these annotations as ground-truth for training and evaluation. As mentioned

[1] https://competitions.codalab.org/competitions/17094

TABLE I
PARAMETERS OF THE DATASETS USED IN THIS WORK

Dataset	Subset	Image number	Slice size	Slice number	In-plane resolution[mm]	Slice thickness[mm]
Abdomen	Training	118	512×512	74-987	0.56-1.00	0.70-5.00
	Validation	13	512×512	122-846	0.68-1.00	0.70-3.00
	Testing	70	512×512	42-1026	0.60-0.98	0.45-5.00
Head	Training	80	512×512	109-202	0.62-1.27	3.00
	Validation	9	512×512	130-151	0.88-1.56	3.00
	Testing	30	256×256 512×512	118-208	0.90-1.95	3.00

before, these annotations are made publicly available[2] to other researchers.

The head clinical CT dataset is composed of 119 head clinical CT scans which are provided by one hospital. It is randomly split in 80 for training, 9 for validation and 30 for testing. Each CT scan has a segmentation mask of 12 head organs (or anatomical structures): left eye, right eye, brainstem, oral cavity, left optic nerve, right optic nerve, left inner ear, right inner ear, left joint, right joint, left parotid gland and right parotid gland. All the segmentation masks are manually delineated by two radiologists and verified/modified by a third radiologist who is more experienced. We directly compute the organs' B-boxes according to their segmentation masks.

B. Data Preprocessing and Augmentation

In the data preprocessing stage, the input CT images are resampled to a uniform spatial resolution of $2.0\times2.0\times2.0\,mm^3$ using bilinear interpolation. Axial slices are center cropped with a maximum physical size of $300\times300\,mm^2$. Voxel intensities are rescaled from $[-1000, 1600]$ Hounsfield Unit (HU) to [0, 1]. Intensities outside this HU range are clipped.

In order to improve the model's generalization ability and mitigate over-fitting, data augmentation strategies are applied on the input CT images. We randomly resample each input CT image by one of the following options:

a. Keep the whole CT image as input;
b. Sample a sub-scan of the input CT image.

The minimum slice number of the sub CT scan is set to 50 to keep enough input context information. Ground-truth B-boxes truncated in the sub-sampled CT image are considered as background. Furthermore, we randomly translate the input CT volume by a maximum of $10\,mm$ along the normal direction of the coronal and sagittal planes.

III. METHOD

In Fig.2 we provide a schematic representation of our method. For an input CT image, we firstly exploit a ConvNet to extract its feature map and assign a set of reference B-boxes with different sizes and shapes to each feature map cell. This feature map is then fed into a subsequent RPN to predict multi-class scores and B-box adjustment parameters for each reference B-box. By applying the adjustment parameters on the reference B-boxes, a substantial number of class-specific candidate B-boxes are generated. Finally, we apply a multi-candidate fusion strategy on the candidate B-boxes to eliminate redundancy and produce the final B-box of the target organ. The entire method is fully implemented in 3D manner. Note that, there is no fully connected layer involved in this deep network, thus it is free to the size of the input CT image.

A. Backbone Network for Feature Extraction

As the first step of our method, we employ a ConvNet, called backbone network, to extract feature map from the input CT image. In the original faster R-CNN framework, a number of standard networks (e.g. AlexNet [25], VGG-16 [27], and ResNet-34 [28]) show excellent performance when used as the backbone network. It is straightforward to use a 3D version of these standard networks (truncated before any fully connected layers) in our method. However, the low-resolution feature map generated by these backbone networks is too coarse for accurate organ localization, especially for the organs with small size. To address this problem, we propose a novel backbone network to generate high-resolution feature map. As the architecture illustrated in Fig.3, we build our backbone network based on the AlexNet architecture. While the input CT image proceeds through the early convolutional and pooling layers, hierarchical feature maps with different resolutions are generated. The low-level feature maps contain rich spatial information while the high-level feature maps capture strong semantic knowledge. Inspired by the design of skip connections used in semantic segmentation ConvNets [29]–[32], we aggregate the feature maps in different levels to hallucinate high-resolution feature map. Specifically, we employ a deconvolutional layer (deconv_1) to upsample the high-level feature map and combine it with the low-level feature map to produce a high-resolution feature map that retains strong semantic information as well as fine spatial details. Theoretically, we can aggregate all the feature maps in different levels to generate a feature map in the same size with the input CT image. However in practice, due to the limited memory size and the trade-off between accuracy and efficiency, we only combine the top feature map to the second-to-top one. Batch normalizations [33] are applied after each convolutional layer to accelerate the learning convergence of the network.

[2]http://dx.doi.org/10.21227/df8g-pq27

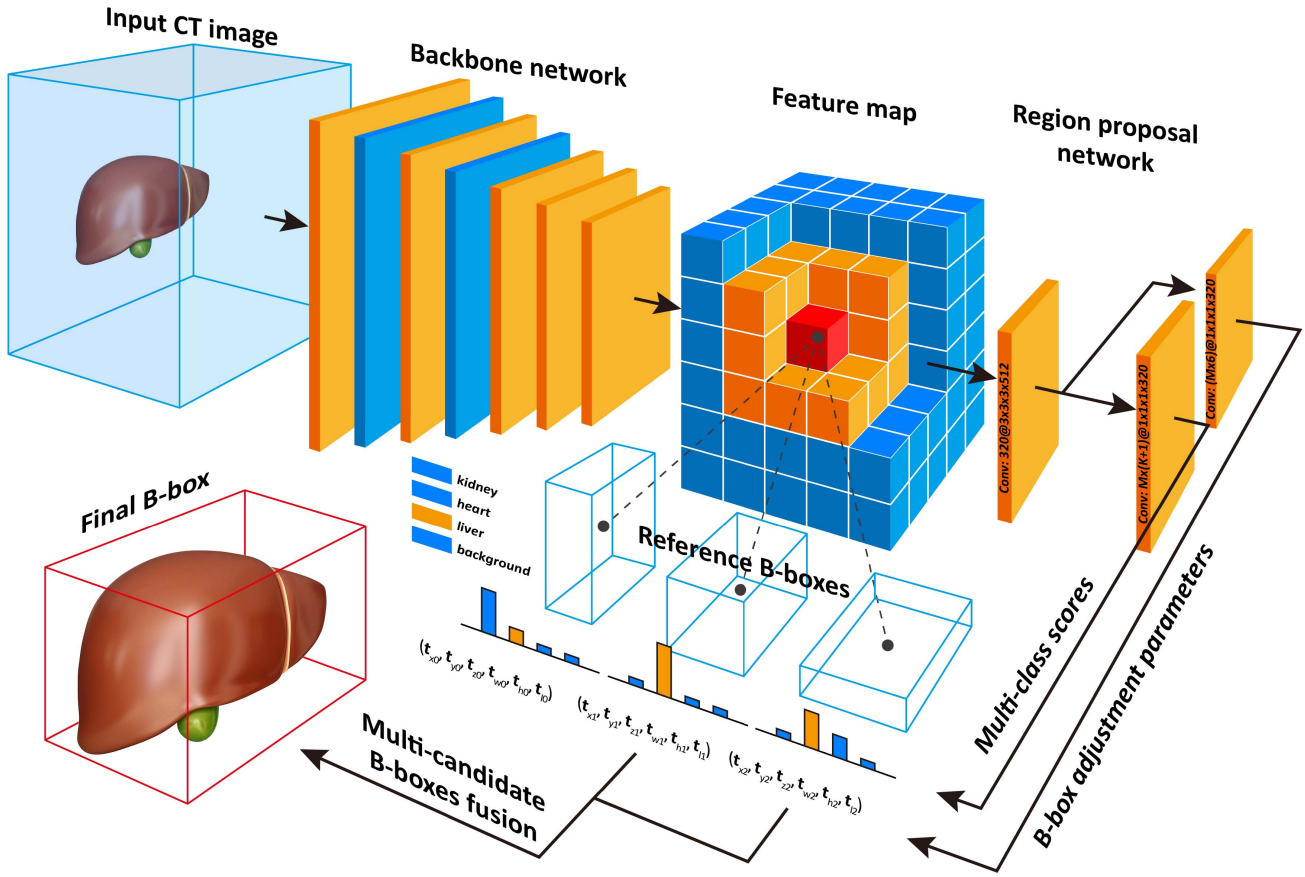

Fig. 2. Schematic representation of our method.

B. Region Proposal Network for B-Box Prediction

When we get the feature map of the input CT image, we feed it into a subsequent 3D RPN to generate candidate B-boxes. As illustrated in Fig.2, the 3D RPN consists of a $3 \times 3 \times 3$ convolutional layer followed by two sibling $1 \times 1 \times 1$ convolutional layers. Next, we provide a brief overview of the RPN's working principle.

For a feature map composed of $W \times H \times L$ cells, each cell corresponds to a spatial point in the coordinate system of the input CT image. We assign M reference B-boxes with different shapes and scales to each feature map cell, thus there is a total of $W \times H \times L \times M$ reference B-boxes and the input CT image can be completely covered by these reference B-boxes. Each reference B-box is associated with $K+1$ class scores, which indicate the presence of K target organs and background objects in it, and 6 adjustment parameters $(t_x, t_y, t_z, t_w, t_h, t_l)$ adjusting the reference B-box to better fit the potential target organ. The ultimate goal of RPN is to predict the class scores and the adjustment parameters for each reference B-box, and produce $W \times H \times L \times M$ candidate B-boxes that can be expressed as:

$$\begin{cases} x = x_r + t_x w_r \\ y = y_r + t_y h_r \\ z = z_r + t_z l_r \end{cases} \quad \begin{cases} w = w_r e^{t_w} \\ h = h_r e^{t_h} \\ l = l_r e^{t_l} \end{cases} \tag{1}$$

where the tuple (x, y, z, w, h, l) denotes a candidate B-box centered at (x, y, z) with a size of (w, h, l). The subscript r indicate the reference B-box. We perform a $3 \times 3 \times 3$ sliding window search at each feature map cell and map each sliding window to a 320-d feature vector. According to this feature vector, $M \times (K+1)$ class scores and $M \times 6$ adjustment parameters can be predicted. This fashion is naturally implemented with a $3 \times 3 \times 3$ convolutional layer followed by two sibling $1 \times 1 \times 1$ convolutional layers (One predicts the multi-class scores and the other one predicts the adjustment parameters). Rectified linear units [26] are used for activation after the $3 \times 3 \times 3$ convolutional layer.

It is intuitive that the localization accuracy strongly depends on the default size of the reference B-boxes. In this work, we use 4 base sizes, i.e. $\{30\,mm, 60\,mm, 120\,mm, 240\,mm\}$ for body organs and $\{10\,mm, 20\,mm, 40\,mm, 80\,mm\}$ for head organs, in each spatial dimension to define 64 reference B-boxes at each feature map cell. These base sizes are selected empirically according to the target organs. In practice, one can also design other distributions of the reference B-box to better fit the targets.

C. Multiple Prediction Strategy

Because an organ's B-box can be simultaneously expressed by several neighboring reference B-boxes with different adjustment parameters, there would be multiple candidate B-boxes

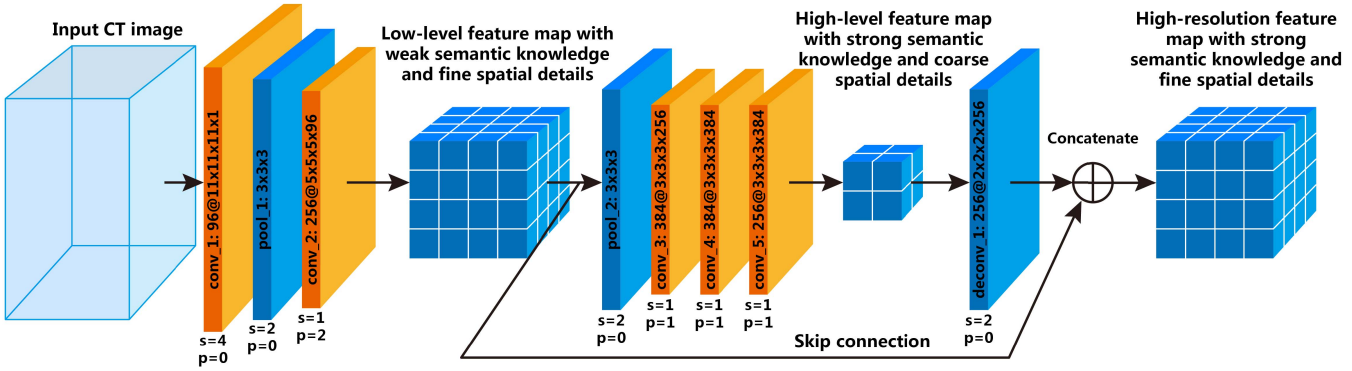

Fig. 3. Structure of our proposed backbone network. This network is build based on the AlexNet [25] architecture. Rectified linear units [26] are used for activation after each convolutional and deconvolutional layer. The convolutional kernel size is expressed in the form of "256@$5 \times 5 \times 5 \times 96$," indicating 256 different kernels of size $5 \times 5 \times 5 \times 96$. "s" and "p" below each layer denote the corresponding stride and padding size, respectively.

generated for the same organ through the RPN. To eliminate this redundancy and further improve the localization accuracy, we apply following multi-candidate fusion strategy on the candidate B-boxes at testing stage.

For each organ, we keep the associated candidate B-boxes $\boldsymbol{B}_i$ that meets a combination of the following two conditions:

a. the class score p_i is higher than an absolute threshold T_1;
b. the class score p_i ranked in the top T_2 percent of all the associated candidate B-boxes.

The hyper-parameters T_1 and T_2 are set to 0.9 and 10%, respectively, which are determined by grid searching on the validation set. Since there is at most one instance for each organ in the image, we directly use a weighted mean of these candidate B-boxes to produce the final B-box $\boldsymbol{B}$:

$$\boldsymbol{B} = \frac{\sum_i p_i \boldsymbol{B}_i}{\sum_i p_i} \tag{2}$$

D. Training Loss Functions

In the training stage of our method, we firstly assign class labels to the reference B-boxes. Unlike the class-agnostic (foreground/background) label assignment used in the original RPN, we assign the reference B-boxes with class-specific labels. Specifically, for each reference B-box, we separately compute its Intersection-over-Union (IoU) overlap with every ground-truth B-box in the input CT image. If the highest IoU value is higher than a foreground threshold T_f, we assign the corresponding ground-truth label ($u > 0$) to the reference B-box; if the highest IoU value is lower than a background threshold T_b, we assign a background label ($u = 0$) to the reference B-box. The reference B-boxes associated with neither any ground-truth B-box nor the background are assigned with an ignored label ($u = -1$), which means that they would not contribute to the training objective. The value of T_f and T_b are empirically selected to make sure there are enough reference B-boxes associated with target organs. In our experiment configurations, we use $T_f = 0.35$ and $T_b = 0.25$ for body organs, and $T_f = 0.15$ and $T_b = 0.05$ for head organs. To make sure there is at least one reference B-box associated with each ground-truth B-box, we directly associate each ground-truth B-box with the reference B-box that achieves the highest IoU overlap with it. For the reference B-box assigned with positive labels, we also compute its target adjustment parameters $\boldsymbol{t}^*$ relative to the associated ground-truth B-box according to Equation 1.

With above label assignment, we can get the optimal model by minimizing following multi-task objective function:

$$Loss = L_{cls} + L_{reg} \tag{3}$$

where L_{cls} and L_{reg} denote the classification loss function and the bounding regression loss function, respectively.

For the original RPN, the classification loss L_{cls} is defined as a cross entropy loss function which assigns same weight to each class. However, in our application there is an extreme inter-class imbalance. Most of the reference B-boxes are labeled as background while only few of them are associated with the organs. Especially for the small organ that occupy a very small region of the CT scan, there could be only a trickle of reference B-boxes associated with it. This inter-class imbalance leads to inefficient training as most reference B-boxes are easy background or ignored that contribute no useful learning signal to the final objective function. Inspired by the focal loss function [34], we propose to use the following classification loss function to rebalance the loss that comes from different classes and focus training on hard-classified reference B-boxes:

$$L_{cls} = -\sum_i [u_i \geq 0] \frac{1}{N_{u_i}} (1 - p_i)^{\gamma} \log p_i \tag{4}$$

Here, i is the index of a reference B-box in a training batch and u_i denotes its ground-truth label. [.] is the Iverson bracket. N_{u_i} represents the total number of the reference B-boxes that are labeled with u_i in the training batch. p_i is the predicted probability of reference B-box i belonging to class u_i. The modulating factor $(1 - p_i)^{\gamma}$ controls the weights of the hard-classified examples. The hyper parameter γ is determined by cross validation and analyzed in Section IV. In our experiments, we set $\gamma = 2$ for body organ localization and $\gamma = 0$ for head organ localization.

The bounding regression loss function L_{reg} is directly extended from the L_1 loss function that is used in the faster

R-CNN [15]. It is defined as

$$L_{reg} = \frac{1}{N_{total}} \sum_i [u_i > 0] \; smooth_{L_1}(\boldsymbol{t}_i - \boldsymbol{t}_i^*) \tag{5}$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise}, \end{cases} \tag{6}$$

where $\boldsymbol{t}$ denotes the predicted adjustment parameters. N_{total} is the number of the foreground reference B-boxes in the training batch.

E. Training Configurations and Implement Details

Our method is implemented in Caffe framework [35] with Insight Toolkits 4.10 on Windows 7 (64 Bit) platform. All experiments are conducted on a workstation equipped with an Intel® Core™ i7-5960X CPU working at 3.00 GHz and a NVIDIA GTX 1080 Ti graphic card with 11 GB of memory. In the training stage of our method, the deep network is trained end-to-end by back-propagation [36] and stochastic gradient descent algorithm. We use a learning rate of 10^{-4} for 1000 training epochs (which has been proven to be sufficient for the convergence of the modules in the experiments) and save the model parameters every 20 epochs. Among the 50 saved models, the one that achieves the highest global IoU metric on the validation set is selected as the final model to be evaluated on the testing set. In each training batch, we feed one CT image to the network. The weight decay and momentum parameter are set to 5×10^{-4} and 0.99, respectively. Layers in the RPN are initialized by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01, and other layers are initialized using the Xavier algorithm [37]. For better reproducibility, the full implementation of our method are made publicly available.[3]

IV. Experiments

A. Metrics

In this work, the performance of methods is comprehensively evaluated from three aspects: the detection precision, the localization accuracy and the processing efficiency.

For the detection precision, we use the average precision (AP) (i.e. the area under precision-recall curve) as the major metric. To compute this metric, the IoU threshold of 0.35 is used to tell whether the predicted B-box is true or false positive.

To evaluate localization accuracy, we calculate the IoU overlap ratio between the predicted B-box and the ground-truth B-box. The IoU overlap ratio is defined as:

$$IoU = \frac{V_p \cap V_{gt}}{V_p \cup V_{gt}} \tag{7}$$

where V_p and V_{gt} denote the volume of the predicted B-box and the ground-truth B-box, respectively. To conduct direct comparison with previous methods, we calculate the widely reported metrics, i.e. the mean and standard deviation of the absolute wall distance and centroid distance between the predicted B-box and the ground-truth B-box. For the brevity, all the wall/centroid distance error are expressed using the format of "mean (standard deviation)" in the following contents. For the head CT dataset which has the segmentation mask for each

[3]https://github.com/superxuang/caffe_3d_faster_rcnn

TABLE II
DETECTION AND LOCALIZATION RESULTS OF 11 BODY ORGANS ON THE ABDOMINAL CLINICAL CT DATASET

Organs	IoU [%]		Wall dist. [mm]	Centroid dist. [mm]
	Mean	Worst		
Left lung	84.84	72.73	5.09(3.83)	7.55(2.64)
Right lung	86.88	72.31	4.87(4.93)	7.91(4.80)
Heart	80.52	66.46	4.07(4.63)	6.53(4.05)
Liver	77.83	58.62	8.46(9.36)	12.26(6.69)
Spleen	70.01	42.17	6.28(6.65)	9.95(5.84)
Pancreas	58.56	16.89	9.23(7.95)	13.25(6.48)
Left kidney	75.29	48.04	4.31(4.18)	6.19(3.75)
Right kidney	76.46	48.35	3.89(3.47)	5.59(2.86)
Bladder	58.23	6.24	7.32(6.53)	10.23(4.58)
Left femoral head	77.26	43.86	2.10(1.89)	3.25(1.79)
Right femoral head	79.77	61.72	1.85(1.62)	3.01(1.53)
Global	73.01	6.24	5.38(6.25)	7.94(5.73)
Precision[%]	Recall[%]		AP[%]	Time[s]
97.91	98.71		98.24	0.29

TABLE III
DETECTION AND LOCALIZATION RESULTS OF 12 HEAD ORGANS (OR ANATOMICAL STRUCTURES) ON THE HEAD CLINICAL CT DATASET

Organs	IoU [%]		Overlap [%]	Wall dist. [mm]	Centroid dist. [mm]
	Mean	Worst			
Brainstem	70.26	52.10	98.77	2.70(2.34)	3.56(1.48)
Left eye	69.18	57.11	98.84	1.78(1.27)	2.78(1.02)
Right eye	68.09	56.09	98.98	1.87(1.28)	2.86(1.08)
Left inner ear	49.23	33.92	97.03	2.14(1.56)	2.99(1.41)
Right inner ear	49.90	29.11	96.62	2.03(1.64)	2.86(1.54)
Left joint	44.99	26.66	95.59	2.42(1.76)	2.98(1.49)
Right joint	49.15	30.32	93.67	2.19(1.56)	3.07(1.46)
Left optic nerve	42.05	13.57	92.42	2.35(1.88)	3.49(1.57)
Right optic nerve	40.00	23.74	93.55	2.40(1.84)	3.61(1.60)
Oral cavity	79.00	58.80	98.72	2.55(2.27)	3.62(1.80)
Left parotid	63.25	39.16	97.74	4.46(4.00)	7.04(2.90)
Right parotid	62.46	39.08	96.39	4.91(5.35)	7.88(4.37)
Global	57.30	13.57	96.53	2.65(2.68)	3.89(2.57)
Precision[%]	Recall[%]			AP[%]	Time[s]
91.11	91.11			84.78	0.25

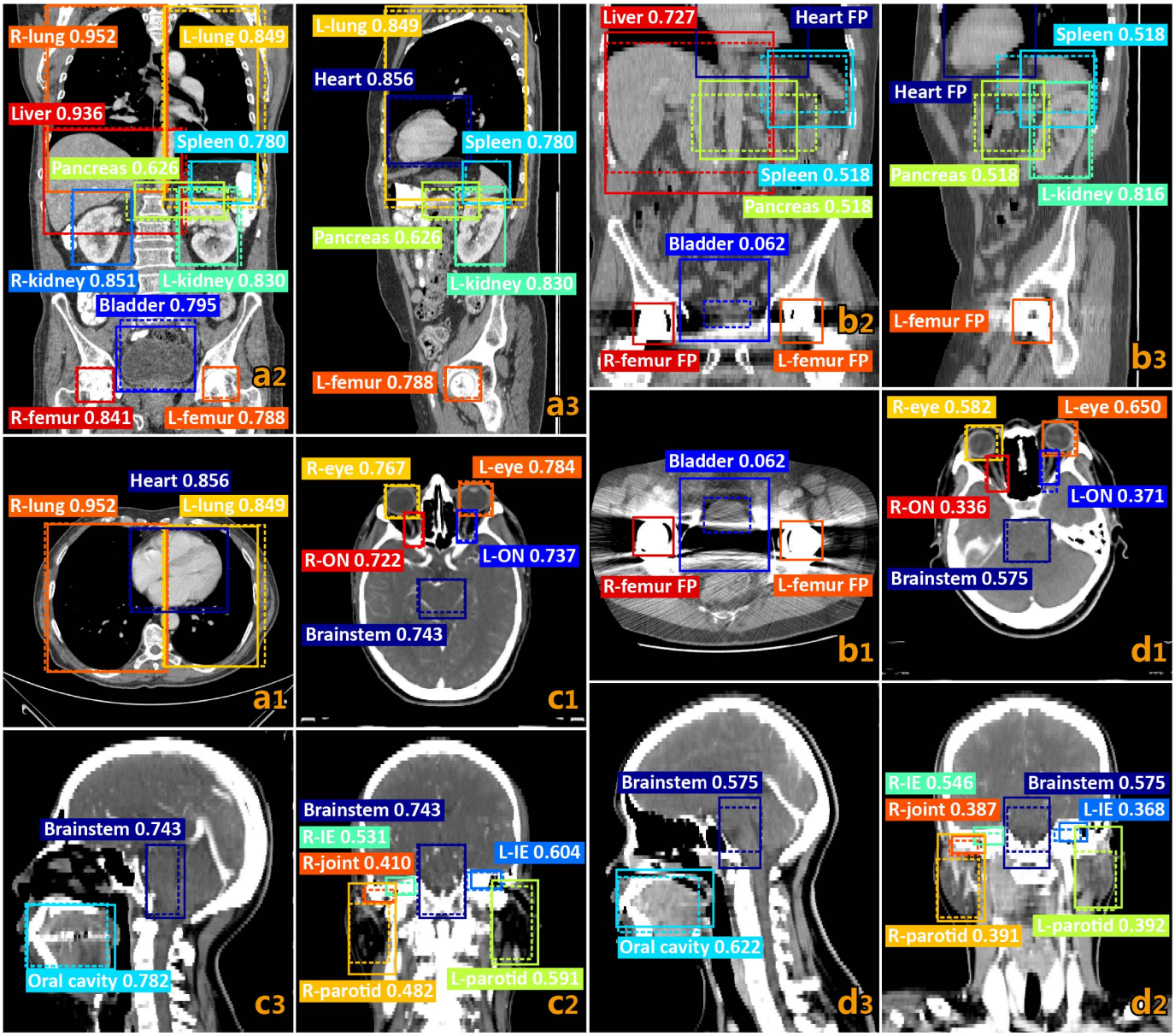

Fig. 4. The best (a and c) and worst (b and d) cases in our results of body (a and b) and head (c and d) organ localization viewed in axial (1), coronal (2) and sagittal (3) planes. Different organs are represented by different colors. Solid line presents our predicted B-box. Dash line indicates the ground-truth B-box. The IoU overlaps between the predicted B-boxes and the ground-truth B-boxes are specified following the organ labels. False positives (FP) caused by truncated heart and artificial femoral heads appear in the worst case (b) of body organ localization. Best viewed in color.

target organ, we also calculate the overlapping ratio between the organ and the predicted B-box.

Processing efficiency is evaluated by the average processing time for detecting and locating all the target organs in one CT image. We also count the model size for further comparison.

B. Localization of Body Organs

We firstly evaluate the performance of our method for body organ localization on the abdominal clinical CT dataset. The results are summarized in Table II and partly visualized in Fig.4. For the detection performance, the precision, recall and AP are 97.91%, 98.71% and 98.24%, respectively. Zero false negative and 13 false positives appear in the results. Among the 13 false positives, 8 B-boxes achieve the IoU overlap lower than the threshold of 0.35. The other 5 false positives are caused by the truncated organs and the artificial structures (e.g. the artificial femoral heads) which are considered as background in the ground-truth. For the localization accuracy, the global average IoU is 73.01%, and almost all the organs achieve high IoU except for the bladder (58.23%) and the pancreas (58.56%). We attribute the relatively poor results on these two organs to the fact that the bladder and the pancreas suffer large variations of appearance and low boundary contrast. The wall/centroid distance error of relatively small organs (e.g. femoral head and kidney) are generally lower than that of the large organs (e.g. liver and lung). The average

TABLE IV
WALL DISTANCE AND AVERAGE PROCESSING TIME OF DIFFERENT METHODS FOR BODY ORGAN LOCALIZATION. TOP ROWS: THE REPORTED RESULTS OBTAINED FROM DIFFERENT DATASETS. BOTTOM ROWS: THE RESULTS OBTAINED FROM THE ABDOMINAL CT DATASET

Methods	Wall dist. [mm]											Time [s]
	Left lung	Right lung	Heart	Liver	Spleen	Pancreas	Left kidney	Right kidney	Bladder	Left femoral head	Right femoral head	
Criminisi et al. [2] (2013)	12.9(12.0)	10.1(10.0)	13.4(10.5)	15.7(14.5)	15.5(14.7)	-	13.6(12.5)	16.1(15.5)	-	10.6(14.4)	11.0(15.7)	4.0
Gauriau et al. [3] (2015)	-	-	-	10.7(4.0)	7.9(4.0)	-	5.5(4.0)	5.6(3.0)	-	-	-	3.2
Samarakoon et al. [1] (2017)	-	-	-	15.8(-)	14.8(-)	-	11.5(-)	11.0(-)	-	7.7(-)	7.4(-)	2.2
Hussain et al. [13] (2017)	-	-	-	-	-	-	6.2(6.0)	5.9(6.4)	-	-	-	-
Humpire et al. [6] (2017)	2.9(-)	2.6(-)	-	8.2(-)	7.2(-)	-	5.7(-)	5.8(-)	8.7(-)	1.8(-)	1.9(-)	-
De Vos et al. [7], [9] (2017)	-	-	**3.2(4.0)**	8.9(15.0)	-	-	-	-	-	-	-	6.4
Humpire et al. [8] (2018)	**2.3(3.1)**	**2.0(2.6)**	-	**5.8(12.7)**	**3.4(8.4)**	-	**2.7(7.2)**	**3.0(9.3)**	**4.7(7.9)**	**1.0(2.3)**	**1.0(2.5)**	4.0
De Vos et al. [7], [9] (2017)	6.8(13.2)	6.3(14.2)	9.2(14.1)	11.7(15.8)	11.8(13.5)	13.1(12.0)	10.1(12.4)	8.7(10.5)	11.1(9.7)	4.9(4.8)	4.9(4.8)	1.2
Humpire et al. [8] (2018)	5.5(5.4)	**4.6(5.1)**	6.1(6.3)	10.7(13.6)	10.1(12.3)	11.9(10.4)	8.5(9.9)	8.2(9.2)	10.4(8.7)	5.8(4.2)	6.0(4.6)	5.4
3D faster R-CNN	7.3(5.7)	8.2(6.7)	6.6(6.0)	9.9(10.0)	7.8(7.2)	10.2(8.5)	5.2(5.5)	5.4(4.7)	8.1(6.9)	3.3(3.0)	3.7(4.0)	0.3
Ours	**5.1(3.8)**	4.9(4.9)	**4.1(4.6)**	**8.5(9.4)**	**6.3(6.7)**	**9.2(7.9)**	**4.3(4.2)**	**3.9(3.5)**	**7.3(6.5)**	**2.1(1.9)**	**1.9(1.6)**	**0.3**

TABLE V
WALL DISTANCE AND AVERAGE PROCESSING TIME OF DIFFERENT METHODS OBTAINED FROM THE HEAD CT DATASET

Methods	Wall dist. [mm]											Time [s]
	Brainstem	Left eye	Right eye	Left inner ear	Right inner ear	Left joint	Right joint	Left optic nerve	Right optic nerve	Oral cavity	Left parotid	Right parotid	
De Vos et al. [7], [9] (2017)	4.0(3.3)	3.0(3.2)	3.9(8.1)	2.7(2.5)	3.6(8.4)	3.5(3.2)	4.7(10.4)	3.6(2.9)	3.7(2.8)	3.7(4.6)	6.1(6.0)	7.9(11.2)	1.5
Humpire et al. [8] (2018)	4.7(3.6)	3.9(2.7)	4.1(2.9)	4.1(2.7)	4.2(2.7)	4.7(3.0)	4.4(2.7)	4.7(3.2)	4.7(2.7)	4.7(3.2)	6.2(5.1)	6.8(5.7)	3.7
3D faster R-CNN	3.7(3.0)	2.5(2.1)	2.9(2.8)	4.5(4.0)	4.2(4.0)	4.3(3.8)	3.6(2.8)	4.2(3.9)	4.1(3.6)	4.9(4.0)	5.9(4.7)	6.1(6.0)	0.4
Ours	**2.7(2.3)**	**1.8(1.3)**	**1.9(1.3)**	**2.1(1.6)**	**2.0(1.6)**	**2.4(1.8)**	**2.2(1.6)**	**2.4(1.9)**	**2.4(1.8)**	**2.6(2.3)**	**4.5(4.0)**	**4.9(5.3)**	**0.3**

processing time for detecting and locating all the 11 organs in one CT image is 0.29s.

C. Localization of Head Organs

The localization results of head organs (or anatomical structures) on the head clinical CT dataset are summarized in Table III. Due to the small size of the targets, it is more challenging to detect and locate the head organs than the body organs. Thus, when using the identical IoU threshold for detection, the average precision of head organs (84.78%) is obviously lower than that of the body organs (98.24%). Zero false negative and 32 false positives appear in the results. All the false positives are caused by the IoU lower than the detection threshold of 0.35. For the localization accuracy, the global average IoU of head organs (57.30%) is also lower than that of the body organs (73.01%). However, it is still acceptable considering the strictness of the 3D IoU metric. The global overlapping ratio between the organ and the predicted B-box is 96.53%, indicating that most regions of the target organs have been located inside the predicted B-boxes. Similar to the case of the body organs, there is a roughly positive correlation

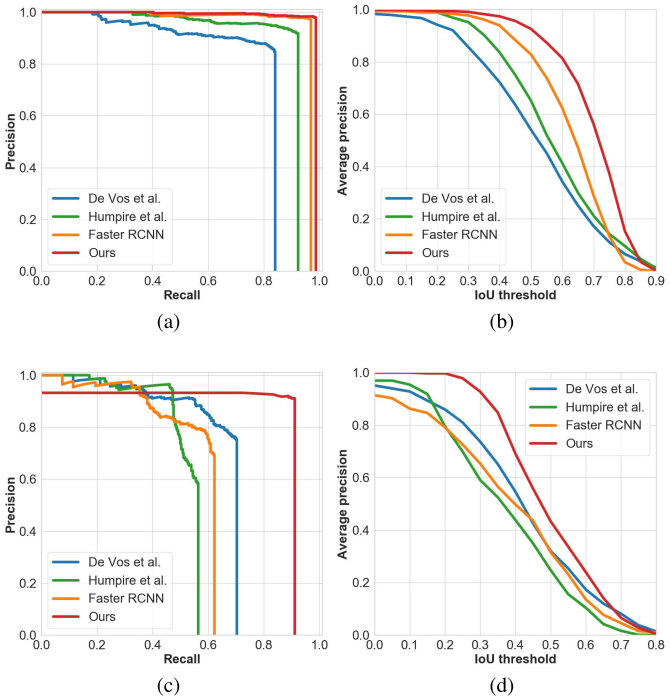

Fig. 6. Precision-recall curves (a and c) and AP-IoU threshold curves (b and d) of different methods on the abdominal CT dataset (a and b) and head CT dataset (c and d).

between the organ size and the absolute wall/centroid distance error. The testing speed for the localization of 12 head organs is 0.25s per CT image.

D. Comparison With Other Methods

In this section, we successively conduct experiments to compare the performance of our proposed method with that of other methods. In the top of Table IV, we firstly list the wall distance error and the average processing time of body organ localization reported in some previous works [1]–[3], [6]–[9], [13]. It can be seen that the recent deep learning-based methods (e.g. [7]–[9]) substantially improve the organ localization accuracy in comparison to the classical machine learning-based methods [1]–[3]. When compared with these previous methods, our method achieves comparable results. However, these results are obtained from different datasets thus can only be used for qualitative analysis. To make the comparison more convincing, we directly compare two current state-of-the-art methods (i.e. the method by de Vos *et al.* [7], [9] and the method by Humpire-Mamani *et al.* [8]) with the proposed method on our datasets. The models of these two methods are fully implemented and trained along the lines of [7]–[9]. The hyper-parameters and training configurations are specially fine-tuned on our datasets to optimize their performance. In addition, as the baseline of our method, the 3D faster R-CNN is also evaluated as a competitor in this comparison.

In Fig.6a and Fig.6c, we give the precision-recall curves (using the IoU threshold of 0.35 for organ detection) of these methods on the abdominal CT dataset and the head CT dataset, respectively. We also plot the AP measure as a function of the IoU threshold in Fig.6b and Fig.6d. It can be seen that, for the body organs, both the 3D faster R-CNN and our method outperform the other two 2D ConvNet-based methods. This result confirms our hypothesis that handling the problem of organ localization fully in 3D manner could be more robust than the slice-wise manner. Moreover, for the head organs, our method also achieves higher AP in a wide range of the IoU threshold than other competitors, indicating the efficiency of our backbone network designed for small organ localization.

For the localization accuracy, we list the wall distance error of these methods on each organ in the bottom of Table IV and Table V. The wall distance error of our method is lower than that of other three methods on most of the 23 target organs, especially for the small organs. We also give the box-plots of the IoU overlap ratio in Fig.5. It can be seen that our method achieves higher average IoU than other methods. These results demonstrate the generalizability of our method and its advantage for small organ localization. Note that, the performance of Humpire-Mamani *et al.* [8] for body organ localization is lower than its original results in [8]. We attribute this to the different characteristics of the two datasets used in [8] and this work. Compared with the CT images used in [8] which was collected from one medical center, our image data (i.e. the LiTS challenge dataset) comes from several clinical sites around the world using different scanners and protocols. The CT slice thickness in [8] varies from $1.00\,mm$ to $2.00\,mm$, while the CT slice thickness varies from $0.45\,mm$ to $5.00\,mm$ in our image data.

For the computation efficiency, both of the 3D faster R-CNN and our method achieve much faster processing speed comparing with the other two methods on the abdominal CT dataset and the head CT dataset, indicating the efficiency of processing CT image in 3D manner. Moreover, the model size of our method (68 MB) is considerably smaller than that of the 3D faster R-CNN (986 MB), demonstrating the effectiveness of our simplifications made on the 3D faster R-CNN framework.

E. Ablation Experiments

In this section, we conduct a set of ablation experiments on our method using the abdominal clinical CT dataset and the head clinical CT dataset to justify the choices made in our design. To avoid the impact of other factors, the ablation modules are kept identical with our final method except for the specified changes in each ablation experiment. The training strategy is also the same as our final method introduced in Section III. The experiment results are summarized in Table VI and discussed in detail next.

In the first experiment, we successively evaluate the performance of our method when combined with different backbone networks, including the AlexNet [25], the VGG-16 [27] and the ResNet-34 [28]. Note that, because of the large memory footprints of the 3D versions of VGG-16 and ResNet-34, the input CT images are down-sampled to a uniform spatial resolution of $4.0 \times 4.0 \times 4.0\,mm^3$ in this experiment. We only perform this experiment on the abdominal CT dataset since

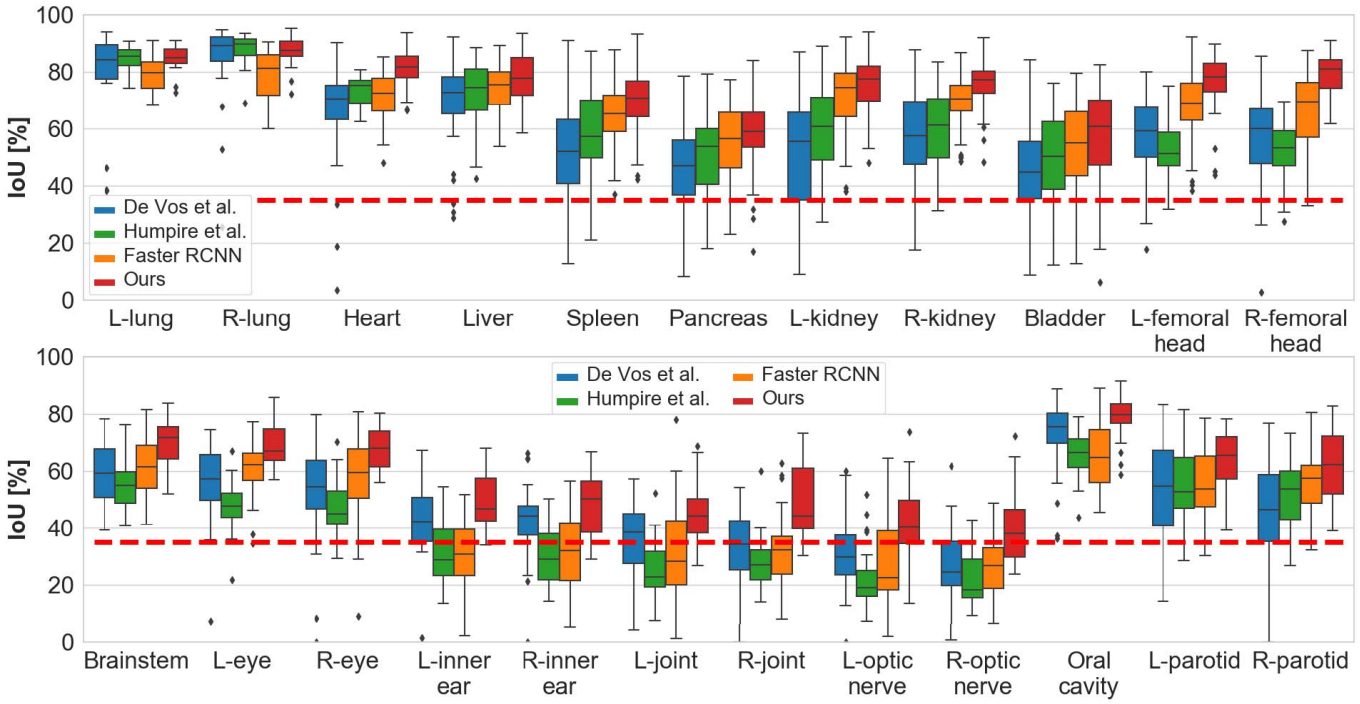

Fig. 5. Box-plots of the IoU overlap metric of different methods for body organ (top) and head organ (bottom) localization. The red dash line indicates the IoU threshold of 0.35 used for organ detection.

TABLE VI
RESULTS OF ABLATION EXPERIMENTS ON OUR METHOD. ASTERISK INDICATES THAT ONLY A FEW OF TARGETS ARE SUCCESSFULLY DETECTED AND USED TO CALCULATE THE RESULTS

Conditions	Abdominal clinical CT dataset					Head clinical CT dataset				
	AP [%]	IoU [%]	Wall dist. [mm]	Centroid dist. [mm]	Time [s]	AP [%]	IoU [%]	Wall dist. [mm]	Centroid dist. [mm]	Time [s]
AlexNet backbone ($4\,mm$)	67.87	55.61	10.85(12.59)	18.68(16.02)	0.01	-	-	-	-	-
VGG-16 backbone ($4\,mm$)	85.49	66.48	6.97(6.93)	10.79(6.92)	0.20	-	-	-	-	-
ResNet-34 backbone ($4\,mm$)	81.83	63.70	7.75(7.90)	12.02(7.65)	0.14	-	-	-	-	-
AlexNet backbone	96.36	70.01	5.87(6.37)	8.83(5.74)	0.05	62.17	49.19	3.59(3.46)	5.70(3.41)	0.05
Our backbone without skip connection	97.57	71.72	5.56(6.21)	8.37(5.80)	0.21	77.56	55.40	2.78(2.70)	4.21(2.55)	0.19
Softmax loss	22.80	73.91*	7.55(8.32)*	11.52(6.90)*	0.29	33.33	68.71*	3.67(3.83)*	5.51(3.48)*	0.26
Our classification loss ($\gamma = 0$)	98.15	72.68	5.48(6.28)	8.07(5.76)	0.29	82.69	57.30	2.65(2.68)	3.89(2.57)	0.25
Our classification loss ($\gamma = 1$)	98.24	72.88	5.39(6.23)	7.95(5.63)	0.29	75.93	55.46	2.80(2.77)	4.09(2.63)	0.25
Our classification loss ($\gamma = 2$)	98.24	73.01	5.38(6.25)	7.94(5.73)	0.29	70.06	55.05	2.82(2.76)	4.05(2.60)	0.25
Our classification loss ($\gamma = 3$)	97.80	72.98	5.38(6.15)	7.80(5.46)	0.29	72.49	55.27	2.80(2.69)	4.08(2.56)	0.25
Without multi-candidate fusion	98.06	72.41	5.54(6.30)	8.13(5.94)	0.29	67.67	55.52	2.89(3.07)	4.50(2.81)	0.25
Using partial reference B-boxes	88.50	71.35	5.87(6.95)	8.73(6.82)	0.16	34.54	38.65	6.82(8.52)	11.12(8.69)	0.14
Our method	**98.24**	**73.01**	**5.38(6.25)**	**7.94(5.73)**	**0.29**	**82.69**	**57.30**	**2.65(2.68)**	**3.89(2.57)**	**0.25**

this spatial resolution is too coarse for the head organ localization. As the results shown in the first row of Table VI, deeper and more complex network architectures (Table VI, VGG-16 and ResNet-34 backbone ($4\,mm$)) bring higher detection precision and localization accuracy but longer computation time.

We then proceed to investigate the behavior of the proposed backbone network in our method. For this purpose, we degenerate our backbone network to its base architecture, i.e. the AlexNet, and compare the resulting model with our final method. As the results shown (Table VI, AlexNet backbone), replacing our backbone network with the AlexNet in our

method leads to less processing time but lower detection precision and localization accuracy on both two datasets. Especially for the head organs, the AP and IoU decrease by a large margin of 20.52% and 8.11%, respectively. This result demonstrates that our proposed backbone network can improve the performance on small organs by generating high-resolution feature map that retains strong semantic information as well as fine spatial details. When we remove the skip connection from our backbone network (Table VI, Our backbone without skip connection), the detection precision and the localization accuracy of our method decreases, indicating that not only the high resolution of the feature map but also the feature combination from different levels contribute to the final improvement. The core idea of our proposed backbone network is to fuse feature maps in different resolutions through deconvolutional operation and skip connection. This design can also be applied to other base networks such as the VGG-16 [27] and the ResNet-34 [28]. However, due to the large memory overhead of volumetric CT image, we just implement it on the AlexNet in this work.

As mentioned before, we redesign the classification loss function in the training objective to rebalance different classes and focus training on hard-classified instances. To justify this design, we replace the classification loss function with the cross entropy loss function that is used for training the original RPN in the faster R-CNN. The resulting model (Table VI, Softmax loss) shows poor performance, indicating the importance of weight rebalance for our training procedure. We then investigate the impact of the hyper parameter γ in our classification loss function. As the results shown (Table VI, Our classification loss ($\gamma = 0, 1, 2, 3$)), the model achieves the best performance on the abdominal CT dataset and the head CT dataset when trained with $\gamma = 2$ and $\gamma = 0$, respectively. It indicates that focusing training on the hard-classified examples allows our method achieve better performance on body organ localization but plays a limited role in locating small head organs.

Next, we analyze the influence of the multi-candidate fusion strategy in our method. When we turn off the multi-candidate fusion at the testing stage and directly use the top-scored candidate B-box as the final output (Table VI, Without multi-candidate fusion), the detection precision and the localization accuracy of our method degrades on both two datasets. This result demonstrates that the multi-candidate fusion strategy can not only eliminate the redundant candidate B-boxes but also improve the final accuracy of our method.

As described in Section III-B, we use 4 base sizes in each spatial dimension to combine 64 reference B-boxes at each feature map cell. When we only use two of these base sizes ($\{60\,mm, 240\,mm\}$ for body organs and $\{40\,mm, 80\,mm\}$ for head organs, Table VI, Using partial reference B-boxes), resulting in 8 reference B-boxes for each feature map cell, the performance of our method degrades on both the abdominal dataset and the head dataset. This result shows that better fitted distribution of the reference B-box can lead to higher detection precision and localization accuracy for our method.

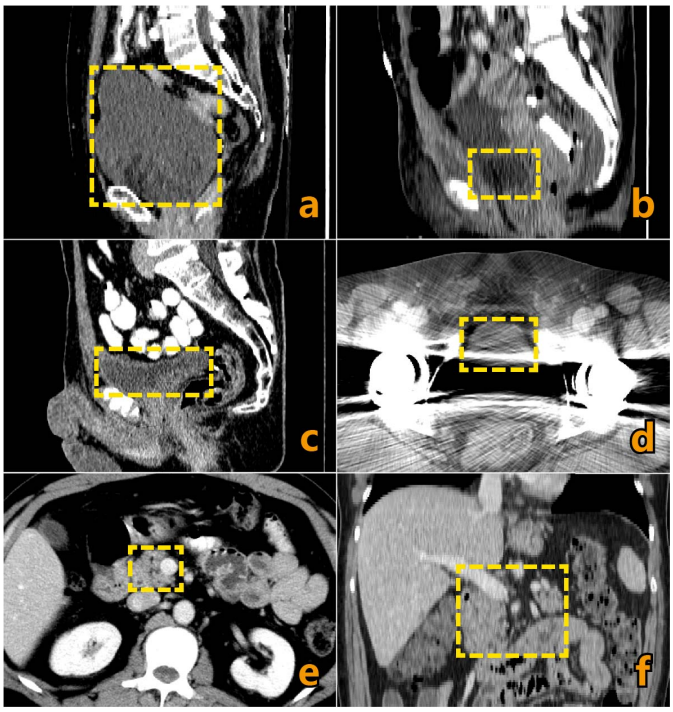

Fig. 7. Examples of the various appearance and low contrast boundaries of the bladder (a, b, c and d) and the pancreas (e and f).

V. Discussions

As our experiment results shown, the bladder and the pancreas achieves the lowest two IoU overlap ratios (58.23% and 58.56%, respectively) among the 11 body organs in the abdominal CT dataset. We attribute these relatively poor results to the various appearance and low contrast boundaries of these two organs. As mentioned in Section II, the abdominal CT images in the LiTS dataset are collected from various clinical sites around the world. The different states of bladder (Fig.7 (a), (b) and (c)), the metal artifact (Fig.7 (d)) and the neighboring low-contrast structures (Fig.7 (e) and (f)) make it difficult to achieve the same localization accuracy on these two organs as other organs with sharp boundaries (e.g. liver, lung, heart and kidney). Since the proposed method is based on the ConvNets driven by data, its performance could be further improved by enlarging the dataset, especially for the organ suffering from large variation of appearance. However, it is difficult to obtain a large scale of annotated training data in CT images due to the limited data source. How to improve the localization performance on this type of organ with limited training data will be a future research topic for our study.

Segmenting the organ of interests is quite necessary for clinical applications. Since the organ localization is usually used as a preprocessing step of the segmentation, the proposed method would be more generalizable if it could perform organ localization and segmentation simultaneously. Fortunately, the proposed method has a similar architecture to the faster R-CNN framework. Thus it is not difficult to integrate the ability of segmentation into the deep network, just like the style of the mask R-CNN [19], by adding an extra branch for segmentation mask prediction at the end of the RPN.

However, technical adjustments and optimizations may be required to make the extended model achieve comparable performance to the methods dedicated to organ segmentation. This will be one goal of our research in the next stage.

As mentioned in Section II, the abdominal CT dataset is built on the MICCAI LiTS dataset. The CT images in this dataset are collected from various patients with liver lesions. Thus the results of liver localization in our experiments could be used to evaluate the performance of our method when applied to patients with lesions in organs. Additionally, the CT images in the head CT dataset are collected from patients who are undergoing radiotherapy with tumors in the head and neck bodypart. There are also a number of abnormalities appearing in the CT images. As the final results shown in our experiments, the performance of our method is still acceptable when dealing with these abnormalities, demonstrating its robustness and generalizability.

Due to the higher data dimensionality and larger number of weight parameters, training 3D R-CNN based model is more time-consuming than the 2D ConvNets. However, significant advantages, such as higher localization accuracy and much faster testing speed, still encourage us to handle this problem using 3D R-CNN. To speed up the training procedure of the proposed method, we apply batch normalization [33] after each convolutional layer in the backbone network to improve the model convergence, and conduct most calculations on GPU in parallel. Through these measures, the training time for the proposed method on the abdominal CT dataset is shortened to around 14 hours.

VI. Conclusion

An efficient method for multiple organ localization in CT image based on 3D RPN is proposed. In this method, the input CT image is fully processed in 3D manners and all the target organs could be detected and located within one prediction. Benefiting from the design of multi-level feature map combination, the proposed method can also be used for accurate small organ localization. As the experiment results shown, the proposed method achieves higher detection precision and localization accuracy with approximate 10 times faster processing speed than the current state-of-the-art method for both body and head organ localization.

References

[1] P. N. Samarakoon, E. Promayon, and C. Fouard, "Light random regression forests for automatic multi-organ localization in CT images," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 371–374.

[2] A. Criminisi *et al.*, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[3] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Med. Image Anal.*, vol. 23, no. 1, pp. 70–83, 2015.

[4] Y. Zheng, B. Georgescu, and D. Comaniciu, "Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Williamsburg, VA, USA: Springer, 2009, pp. 411–422.

[5] X. Zhou *et al.*, "Automatic localization of solid organs on 3D CT images by a collaborative majority voting decision based on ensemble learning," *Comput. Med. Imag. Graph.*, vol. 36, no. 4, pp. 304–313, Jun. 2012.

[6] G. E. Humpire-Mamani, A. A. A. Setio, B. van Ginneken, and C. Jacobs, "Organ detection in thorax abdomen ct using multi-label convolutional neural networks," *Proc. SPIE*, vol. 10134, 2017, Art. no. 1013416.

[7] B. D. de Vos, J. M. Wolterink, P. A. de Jong, M. A. Viergever, and I. Išgum, "2D image classification for 3D anatomy localization: Employing deep convolutional neural networks," *Proc. SPIE*, vol. 9784, Mar. 2016, Art. no. 97841Y.

[8] G. E. Humpire-Mamani, A. A. A. Setio, B. van Ginneken, and C. Jacobs, "Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen CT scans," *Phys. Med. Biol.*, vol. 63, no. 8, 2018, Art. no. 085003.

[9] B. D. de Vos, J. M. Wolterink, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Isgum, "ConvNet-based localization of anatomical structures in 3-D medical images," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1470–1481, Jul. 2017.

[10] A. Criminisi, J. Shotton, and S. Bucciarelli, "Decision forests with long-range spatial context for organ localization in CT volumes," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.* London, U.K.: Springer, 2009, pp. 69–80.

[11] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Fast automatic heart chamber segmentation from 3D CT data using marginal space learning and steerable features," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[12] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features," *IEEE Trans. Med. Imag.*, vol. 27, no. 11, pp. 1668–1681, Nov. 2008.

[13] M. A. Hussain, A. Amir-Khalili, G. Hamarneh, and R. Abugharbieh, "Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision CNNs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Quebec City, QC, Canada: Springer, 2017, pp. 612–620.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[17] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[18] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[20] T.-Y. Lin and P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, vol. 1, no. 2, pp. 936–944.

[21] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[22] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 808–816.

[23] X. Chen *et al.*, "3D object proposals for accurate object class detection," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[24] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-level 3D object reconstruction via render-and-compare," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3559–3568.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[27] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[29] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[31] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Athens, Greece: Springer, 2016, pp. 424–432.

[32] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.

[35] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, 2014, pp. 675–678.

[36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.