TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/18 pp103–113
DOI: 10.26599/TST.2021.9010028
Volume 27, Number 1, February 2022

Two-Stage Lesion Detection Approach Based on Dimension-Decomposition and 3D Context

Jiacheng Jiao, Haiwei Pan*, Chunling Chen, Tao Jin, Yang Dong, and Jingyi Chen

Abstract: Lesion detection in Computed Tomography (CT) images is a challenging task in the field of computer-aided diagnosis. An important issue is to locate the area of lesion accurately. As a branch of Convolutional Neural Networks (CNNs), 3D Context-Enhanced (3DCE) frameworks are designed to detect lesions on CT scans. The False Positives (FPs) detected in 3DCE frameworks are usually caused by inaccurate region proposals, which slow down the inference time. To solve the above problems, a new method is proposed, a dimension-decomposition region proposal network is integrated into 3DCE framework to improve the location accuracy in lesion detection. Without the restriction of "anchors" on ratios and scales, anchors are decomposed to independent "anchor strings". Anchor segments are dynamically combined in accordance with probability, and anchor strings with different lengths dynamically compose bounding boxes. Experiments show that the accurate region proposals generated by our model promote the sensitivity of FPs and spend less inference time compared with the current methods.

Key words: lesion detection; Computed Tomography (CT); dimension-decomposition; 3D context; computer-aided diagnosis

1 Introduction

Cancers have become one of the major public health problems that seriously threaten the health of people. According to the latest statistics, the report was made by the National Health Commission of People's Republic of China (PRC) in 2019, malignant tumor accounts for 23.91% of all deaths among residents in PRC. Nowadays, the incidence of malignant tumors has maintained an increase of about 3.9% and the mortality rate increases by 2.5% each year. Detecting lesions with higher accuracy is helpful for radiologists.

In current clinical methods, Computed Tomography (CT) scans are normally used to model internal organs. Computer-aided diagnosis plays an important role in improving the efficiency of cancer detection. Since AlexNet has made exciting progress in the ILSVRC 2012 challenge[1], deep learning has become a popular issue in computer vision. The family of Convolutional Neural Networks (CNNs) are mainly used in detection, classification, and segmentation tasks. In addition to research on natural images, CNNs have become a powerful method to detect diseases in different kinds of medical images, e.g., lesion detection in color retinal images[2] and disease detection in X-ray[3].

Lesion detection in CT images is a challenging task. Low-resolution CT images usually contain few interclass variances because of the image-forming principle. Lesion and nonlesion areas often have similar appearances. Learning representative features in CT images is a central issue. On the basis of the image-forming principle of CT images, 3D CNNs are modified on medical image detection and segmentation[4–6]. The 3D Region Proposal Network (3D-RPN) is used to process volumetric CT data[7]. 3D CNNs encode rich spatial and context information

• Jiacheng Jiao, Haiwei Pan, Chunling Chen, Yang Dong, and Jingyi Chen are with the College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China. E-mail: geogrechen@hrbeu.edu.cn; panhaiwei2006@hotmail.com; ccl@hrbeu.edu.cn; dongyang@hrbeu.edu.cn; chenjingyi@hrbeu.edu.cn.

• Tao Jin is with the School of Software, Tsinghua University, Beijng 100084, China. E-mail: jintao05@gmail.com.

* To whom correspondence should be addressed.

Manuscript received: 2020-06-12; revised: 2020-08-13; accepted: 2020-08-31

from discriminative features to improve detection. However, 3D CNNs usually need more inference time and computational resources than 2D CNNs for similar tasks. In addition, annotating 3D bounding boxes is not as easy as 2D ones. To overcome the shortcomings of 3D CNNs, methods based on 2D CNNs have been proposed to aggregate multislice features and solve the problem of learning representative features[7, 8]. Lesion detection results also benefit from spatial and contextual attention mechanisms[9]. With feature-enhanced methods, sensitivity at different False Positives (FPs) is improved.

However, current studies normally adopt Region Proposal Networks (RPNs) to generate Regions of Interest (RoI)[8, 9]. RPNs predict RoI through anchors, which are a set of bounding boxes with fixed ratios and edge lengths. The imbalance problem in sample levels is reflected on the distribution of the Intersection of Union (IoU). The shape of lesions is usually irregular. Thus, current studies have two shortcomings.

- First, RPNs are widely used in object detection; however, it is designed to generate region proposals on natural images without involving the characteristic of medical images.
- Second, predicting bounding boxes with anchors restricts the quality of region proposals. Therefore, the inflexible method limits the location accuracy and inference speed in detection. The size of lesions can be extremely small in natural image object detection; small lesions have the risk of being neglected because models are designed for natural images. The two shortcomings mainly restrict precision and inference time.

In this study, we propose a new method of detecting lesions. In our model, anchors are decomposed to two segments (width and height). A length dictionary is set in accordance with the distribution of the length of the ground truth's edges. Region proposals are generated in accordance with the lengths of segments that are dynamic and flexible. The mechanism produces more accurate region proposals and enclosed lesions more tightly than RPN. Experiments show that our model's inference speed and sensitivity of FPs are increased by approximately 3% compared with the baseline 3D Context-Enhanced (3DCE) frameworks[8].

2 Related Work

Object detection is normally divided into two aspects: one-stage and two-stage methods. One-stage approaches need less computing resource and time compared with two-stage ones, because it does not use resampling operations. Nevertheless, two-stage algorithms generally have higher detection accuracy than one-stage ones. Two-stage detection networks mainly involve two components: region proposal generation and detection. Uijlings et al.[10] proposed selective search to produce RoIs. Girshick et al.[11] proposed Regions with CNN features (RCNN) that apply selective search to generate candidates in object detection. Ren et al.[12] introduced an RPN that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. The RPN is trained end-to-end to generate high-quality region proposals, which are used by faster RCNN for detection[12]. Dai et al.[13] proposed Region-based Fully Convolutional Networks (R-FCN) and Position-Sensitive ROI Pooling (PSROI-pooling) to address a dilemma between translation invariance in image classification and translation variance in object detection. Pang et al.[14] proposed the Libra-RCNN, which aims at balanced learning for object detection. Several two-stage methods have achieved remarkable progress on accuracy[15, 16]. One-stage detection methods benefit the speed of inference without resampling operations[17–19].

Region proposal methods have a substantial effect on the final detection results. Some methods apply grouping pixels[10], whereas others use sliding windows to generate region proposals[20]. These methods are trained independently. RPNs outperform previous proposal methods, such as selective search[11] and EdgeBoxes[20]. Li[21] presented Gaussian proposal networks, which propose bounding ellipses as 2D Gaussian distributions on the image plane. Some detection methods that use keypoints for detection were proposed[16, 22, 23]. In addition, several methods were proposed to optimize the algorithms in bounding-box regression or some metrics when assigning labels[24–26].

In lesion detection on CT images, some state-of-the-art researches have focused on the spatial feature of CT images. Dou et al.[4] proposed a method that uses a 3D CNN for FP reduction in automated pulmonary nodule detection from volumetric CT scans. Liao et al.[7] proposed 3D-RPN to generate 3D bounding boxes. 3D CNNs aggregate the spatial information, ameliorate the representative feature[7], and consume more inference time and computing resources than 2D networks[4–6]. 3D CNNs normally lack pretrained networks and need to be trained from scratch. Aggregating multiple CT scan features into the same feature map overcomes the above-mentioned shortcomings with the help of a 2D detection network[9, 21]. A group of slices is fed into a 2D detection network to generate feature maps separately; these maps are then aggregated in the channel for the final detection procedure.

3 Method

The pipeline of our network is shown in Fig. 1. Although most mainstream detection networks only support three-channel images, we follow the processing guideline of raw images in DeepLesion[29]. Every slice is converted to an image, as shown in Fig. 2. With the increasing number of neighboring slices, the feature extracted from neighboring slices is aggregated into the same feature map by concatenating operation.

3.1 Data preprocessing and feature extraction

The raw CT images of DeepLesion[27] are one-channeled, thus dissatisfying the input format of our backbone, i.e., a pretrained VGG-16[28] on ImageNet[30]. In Fig. 1, CT images $I_{\{1,2,\ldots,M\}}$ with red, yellow, and green bounds are the one-channel-processed images, and M is the number of CT images. Every three slices is regarded as the three channels' data of RGB format images to compose a three-channel image. In our model, M is set to multiples of 3. Among all slices, the $(M+1/2)$-th one is the only key slice that is annotated with ground truth. In the pretrained backbone, layers from conv1 (that means the first convolutional layer, so do the following names) to conv5 are used to extract the base features of images. Among the layers in the backbone, pool4 (the fourth pooling layer) and pool5 (the fifth pooling layer) are moved to keep the richness of information. In the backbone, kernel size and padding are set to 3 and 0, respectively. The weights from conv1 to conv6 are shared for different images in each sample. The number of channels is $\{3, 64, 128, 256, 512, 512\}$ successively. A normal CT image in DeepLesion[27] is shown in Fig. 2.

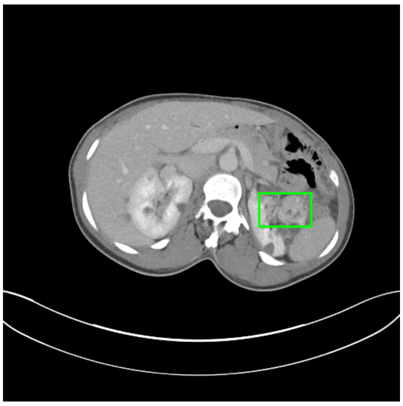


Fig. 2 Key slice with ground truth, the green box is the ground truth which restricts the bound of lesion.

In Fig. 2, the green box represents a typical ground truth in a key slice.

Every convolutional layer is followed by a Rectified Linear Unit (ReLU) and pooling layer, except for conv4 and conv5. To keep the resolution of the feature map, we remove the pooling layer of conv4 and conv5. In Fig. 1, nine slices are inputted into the backbone, thus generating three feature maps. The feature map extracted from the key slice supports the region proposal procedure. The reason for grouping the one-channel slice is not only to adopt the restriction of dimension, but also to fuse information from other slices. The mechanism of fusion improves the detection results. Futher analysis is provided in Section 4.3.

3.2 Region proposals algorithm

The main structure of DeRPN[29] follows that of the RPN. After acquiring the feature map generated by the key slice, DeRPN operates a 3×3 convolution operation on it. The vector produced by the above-mentioned layers is fed into two sibling fully connected layers: regression and classification layers. The classification layer predicts $2 \times 2N$ scores to estimate the probability of the matched strings. The regression layer predicts $2 \times 2N$ elements,

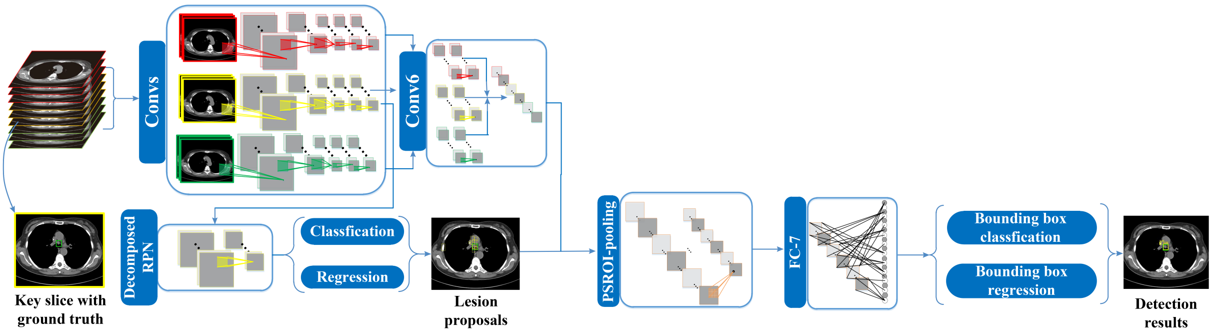


Fig. 1 Overview of the network. Our model uses a pretrained VGG-16[27] to extract base features and generates candidate boxes through DeRPN[28]. FC-7 means the seventh fully connected layer.

which present the locations of width and height segments. N is the size of the length dictionary. We break the bounding boxes into a set of independent segments, called anchor strings. Anchor strings are inputted into the two fully connected layers to predict the RoIs dynamically. Therefore, we set the length dictionary a_n as $\{4, 8, 16, 32, 64, 128, 256, 512\}$, which represents the length of anchor strings. The progression fits most lesion shapes in the DeepLesion dataset.

Current methods extract the features $\boldsymbol{x}$ from the image and input features to a classification layer and a regression layer, separately. Compared with the original regression references in RPN, which mainly include classification and regression of anchors, our method decomposes bounding boxes with respect to dimension as the regression references. The original vectors, denote the top-left coordinate, width, and height of the bounding boxes, are replaced with a pair of anchor strings $S_w(x_a, w_a)$ and $S_h(y_a, h_a)$, where $x_a, y_a, w_a,$ and h_a represent the top-left, width, and height of the anchor string. The original regression and classification methods are updated to adapt the new reference in the following:

$$\boldsymbol{t}^{\omega} = \boldsymbol{W}_r^{\omega}\boldsymbol{x} + \boldsymbol{b}_r^{\omega};\ \hat{G}_w(x_a, w_a) = \psi(\boldsymbol{t}^{\omega}, S_w(x_a, w_a)) \tag{1}$$

$$\boldsymbol{t}^{h} = \boldsymbol{W}_r^{h}\boldsymbol{x} + \boldsymbol{b}_r^{h};\ \hat{G}_h(y_a, h_a) = \psi(\boldsymbol{t}^{h}, S_h(y_a, h_a)) \tag{2}$$

$$P_w = \tau(\boldsymbol{W}_c^{\omega}\boldsymbol{x} + \boldsymbol{b}_c^{\omega});\ P_h = \tau(\boldsymbol{W}_c^{h}\boldsymbol{x} + \boldsymbol{b}_c^{h}) \tag{3}$$

where $\boldsymbol{W}_r^{\dagger}(\dagger = h \text{ or } \omega)$ is the weight of the regression layer, $\boldsymbol{b}_r^{\dagger}$ is the bias the of regression layers, $\boldsymbol{W}_c^{\dagger}$ and $\boldsymbol{b}_c^{\dagger}$ are the weight and bias of the classification layer, respectively. The parameterized coordinate sets for two anchor segments are $\boldsymbol{t}^{\omega}$ and $\boldsymbol{t}^{h}$. $S_w(x_a, w_a)$ and $S_h(y_a, h_a)$ represent regression items for the object's width and height, respectively. $\hat{G}_w()$ and $\hat{G}_h()$ serve as the prediction of ground truth of width and height, respectively. τ and ψ are the classification and decoded functions, respectively[11]. Anchor string contains a property of confidence P_w (width segment) or P_h (height segment) for every anchor string.

In RPN, we filter anchors with the NonMaximum Suppression (NMS) mechanism. Anchors are allocated to positive labels if their IoUs are greater than 0.7 or the anchors hold the largest IoU in a certain category. The anchor strings replace the anchors. In replace of IoUs, anchor strings are matched with the references of length and IoUs simultaneously,

$$\begin{aligned} M_j = & \{i \mid \arg\min_i \left|\log e_j - \log a_i\right|\} \cup \\ & \{(i, i+1) \mid \left|\frac{e_j}{a_i}\right| - \sqrt{q} \leqslant \beta\} \end{aligned} \tag{4}$$

where i is the index of a_n, and M_j denotes the index set of the selected anchor strings for the j-th object. In our model, j is set as 2. e_j is the lesion edge (width or height), and q represents the common ratio. In our experiment, q is set to 2. In the second term of Eq. (4), we choose anchor strings in a range of $(a_i(\sqrt{q} - \beta), a_i(\sqrt{q} + \beta))$. β is 0.1 in our experiments which is used to adjust the balance of range of hyperparameters. The procedure exploits the scope of a_n, thus decreasing the risk of overfitting.

On the basis of heuristic thoughts, positive labels are allocated to anchor strings, which are located on the center of the lesion. In addition to the above-mentioned anchor strings, we use a mechanism called observe-to-distribute[28]. First, we observe the corresponding regression results. Second, the regressed anchor strings are combined with boxes. If the IoUs of the boxes are greater than 0.7, positive labels are distributed. Anchor strings are allocated to negative labels in other circumstances. The loss function is designed to count the confidence of predicted anchors. In the detection procedure, tiny lesions are usually neglected for occlusion with other lesions. We apply the scale-sensitive loss function shown in the following:

$$\begin{aligned} L(\{p_i\}, \{t_i\}) = & \sum_{j=1}^{N}\sum_{i=1}^{B}\frac{1}{|R_j|}L_{\text{cls}}(p_i, p_i^*) \times 1\left\{i \in R_j\right\} + \\ & \lambda\sum_{j=1}^{N}\sum_{i=1}^{B}\frac{1}{|G_j|}L_{\text{reg}}(\boldsymbol{t}_i, \boldsymbol{t}_i^*) \times 1\left\{i \in G_j\right\} \end{aligned} \tag{5}$$

$$R_j = \left\{k | s_k = a_j, k = 1, 2, \ldots, B\right\} \tag{6}$$

$$G_j = \left\{k | s_k = a_j, s_k \in \boldsymbol{A}, \text{ and } p_i^* = 1, k = 1, 2, \ldots, B\right\} \tag{7}$$

where B is the size of the batch, s_k is the k-th anchor string in a training batch, and p_i represents the predicted probability of the i-th anchor string in a batch. The ground truth label p_i^* is set to 1 if the anchor string is positive; Otherwise, p_i^* is 0, the subscript $*$ means the true value of relative variable. $\boldsymbol{t}_i$ is a predicted vector representing the parameterized coordinates. $\boldsymbol{A}$ is the set of aligned anchor strings. R_j denotes an index set containing anchor strings of the same scale, and j is used to indicate the scale corresponding to term a_j in a_n, that is $\{8, 16, 32, 64, 128, 256, 512\}$. Similarly, G_j is an index set containing positively aligned anchor strings of the same scale. The classification loss L_{cls} is a cross-entropy loss, and the regression loss L_{reg} is designed as a smoothed $L1$ loss. λ is a balancing parameter between the regression and classification losses and is set to 10 in our experiment.

Every anchor string is a certain edge of box. We combine the 2 edges to a box. In all anchor strings, we select the top-N items. Every width anchor string selects height anchor string successively. We employ the NMS to select the bounding boxes composed by anchor strings. The probability P^B of bounding boxes is calculated in the following:

$$P^B = \frac{1}{\dfrac{1}{P^w} + \dfrac{1}{P^h}} \tag{8}$$

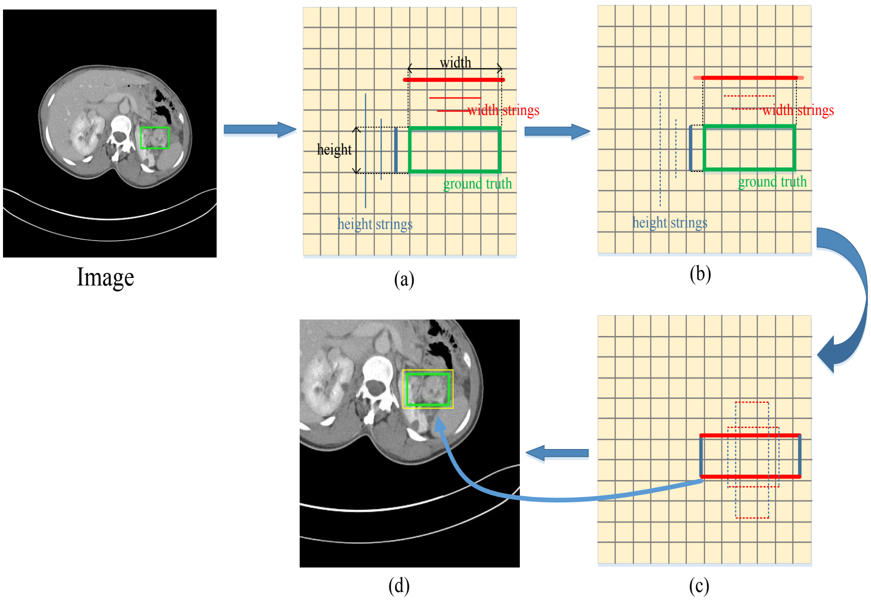
The mechanism is more dynamic and flexible than methods generating boxes directly. The procedure of DeRPN is shown in Fig. 3.

3.3 Feature fusion and detection

The backbone generates three feature maps, which are annotated with red, yellow, and green solid lines. The three feature maps undergo conv6, whose kernel size is 3 × 3. Then, the outputs are concatenated to generate the S^2DM channels' feature map (S is the size of the pooled feature map for each proposal and D is set to 10 in this paper). The S^2DM channels' feature map aggregates the 3D information; it contains all features extracted by our framework. In our model, information aggregation is used in Section 3.1. Neighbouring slices are grouped as three-channel images to generate fusing feature maps. The feature map derived from the key slice is sent to DeRPN to generate region proposals. M feature maps are concatenated to generate conv6. In Fig. 1, M is 3. The concatenated feature map is sent to PSROI-pooling, together with the region proposals. Every region proposal is mapped on the corresponding position of concatenation feature map to evaluate PSROI-pooling operation. PSROI-pooling summarizes these scores on lesion proposals. The object classification and bounding-box regression results are finally obtained by PSROI-pooling. After PSROI-pooling, we add three fully-connected layers, a 2048-dimension fully connected layer and two fully connected layers, for classification and bounding-box regression. Through the three new fully-connected layers, the results of PSROI-pooling are optimized again to improve detection results. Our model has three loss terms: scale-sensitive loss, latter regression loss, and classification loss. They are optimized jointly in our framework.

4 Experiment

We describe the implementation result of our approach for detecting lesion on the DeepLesion dataset. In



Image

(a)

(b)

(d)

(c)

Fig. 3 Mechanism of DeRPN[28]. The entire image is cropped to highlight the region proposal. (a) CT images are transferred to feature maps. The area of lesion is painted in green grids. (b) The algorithm searches the best-matched anchor strings independently, the bond ones represent the well-matched anchor strings. (c) Classification and regression are executed in accordance with Eqs. (1)–(3); the dash lines indicate low probability. (d) The anchor strings are grouped to compose anchors. The restructuring anchors are filtered on the basis of NMS. Anchors with yellow solid edges are region proposals.

Section 4.1, some descriptions about the DeepLesion dataset are introduced. Experiment results are shown in Section 4.2. Next, we analyze experiment results in Section 4.3.

4.1 DeepLesion dataset

DeepLesion is a large-scale CT image dataset released by NIH[27]. It contains 32 735 lesions in 32 120 CT slice images. Each pixel of CT images corresponds to 0.8 mm. In DeepLesion, 32 slices are 768 pixel × 768 pixel, and eight are 1024 pixel × 1024 pixel. Except for the above-mentioned 40 slices, other slices in DeepLesion are 512 pixel × 512 pixel. Every lesion is annotated with a four-dimension vector (x_1, y_1, x_2, y_2) denoting the top-left and bottom-right coordinates of ground truth. The length distribution of the ground truths' edges is shown in Fig. 4, where a_n is set on the basis of distribution above.

4.2 Implementation details

We adopt the offical split of DeepLesion: training (70%), validation (15%), and test (15%). We convert every CT slice to int32 format, and then subtract 32 768 Houndsfield Unit (HU) values for each pixel, thereby acquiring HU values. With the help of windowing parameters, we generate the image shown in Fig. 1. The network starts with a pretrained VGG-16[27] on ImageNet[30]. In the training procedure, the batch size is set to 2, meaning a batch contains 2 samples, indicating that a batch contains 2 samples. Each sample contains M pre-processing images. Stochastic gradient descent is applied in our model, with a momentum of 0.9 and decay of 5×10^{-5}. We train all models with eight epochs. In the first four epochs, the learning rate is frozen at 0.001. From the fourth to the eighth epoch, the learning rate decreases to 0.0001. We train our end-to-end model on five Telsa K80 GPUs.

4.3 Network performance

The widely used sensitivity is a statistical measure of the performance of algorithms. A new metric Region Proposal Proportion (RPP) is defined to measure the quality of region proposals,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{RPP} = \frac{R_i}{R_{\text{sum}}} \tag{10}$$

The number of true positives is denoted as TP. FN represents the number of false negatives. i is a term of S: $\{0-0.1, 0.1-0.2, \ldots, 0.9-1.0\}$. R_i represents the number of region proposals whose IoU is in the range of i. The complete presupposed IoU group is shown in the horizontal axis of Fig. 5. R_{sum} is the sum of the region proposals. We first evaluate our model on the official dataset. The framework of our model is shown

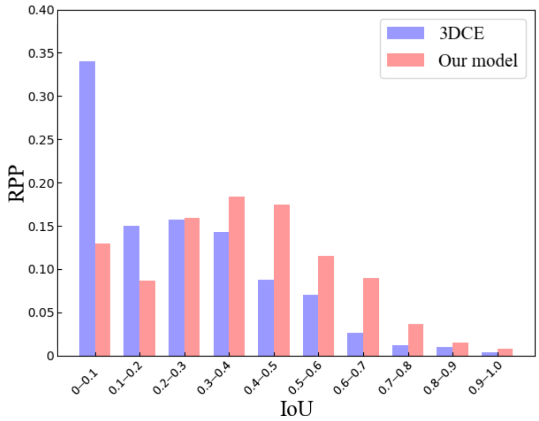
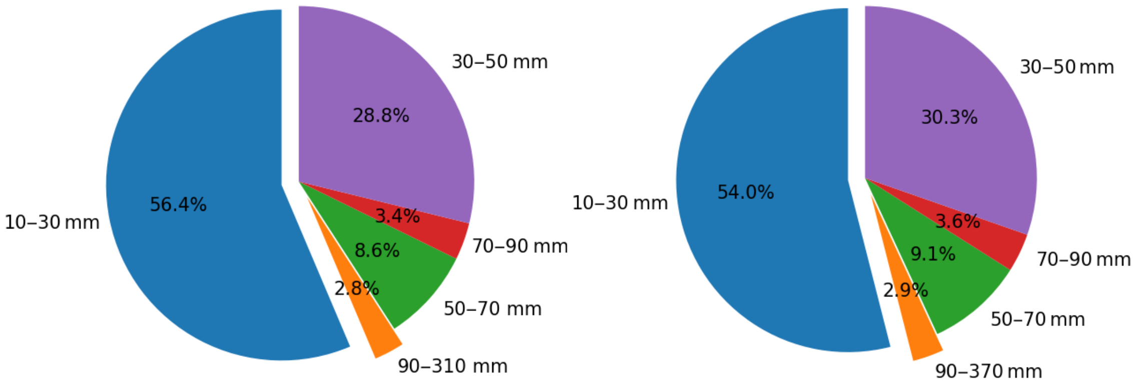


Fig. 5 RPP distribution of region proposals in training. The horizontal axis represents the IoU interval of region proposals.



(a) (b)

Fig. 4 Length distribution of the ground truth's edge, (a) width distribution and (b) height distribution.

in Fig. 1. Our model improves the sensitivity of FPs by aggregating the neighbouring slice features into a feature map.

At the baseline, the anchors' ratios and scales are set to $\{0.5, 1, 2\}$ and $\{2, 3, 4, 6, 12\}$, respectively. In our model, The length dictionary is set as $\{8, 16, 32, 64, 128, 256, 512\}$. As shown in Table 1, our model with nine slices outperforms the baseline with 21 slices in terms of sensitivity and inference time. As shown in Fig. 3, region proposals composed by anchor strings are flexible in shape.

In the training phase, we collect the region proposals made by the baseline and our model. The foreground ratio of the region proposals of our method is twice that of the baseline, as shown in Fig. 5. The foreground means the region proposal contains the object. Region proposals are usually imbalanced[14]. Our model overcomes IoU imbalance in the level of region proposals. In addition, our mechanism is more accurate and dynamic than RPN with the scale-sensitive loss function[12].

Faster RCNN[12] is also evaluated on DeepLesion by adding three fully-connected layers and removing pool4 and pool5 in VGG-16. Faster RCNN and improved R-FCN are trained with 3 slices. Basing the more accurate region proposals, our model reduces inference time significantly than models with 3D context aggregation. Our model with 21 slices needs less time to predict than 3DCE with 9 slices. Inference time complexity is roughly linearly proportional to the number of CT slices. The number of slices determines the inference time. In addition, we rerun 3DCE network and obtain higher results. 3DCE_CS_Att[21] is a novel framework in DeepLesion; it introduces a dual attention mechanism to boost sensitivity on the test set of DeepLesion. With less parameters, our model achieves competitive detection sensitivity. In Fig. 6, the loss function shows a better convergence rate. See in Figs. 7e and 7h, our model attaches importance to lesion with small scale.

Despite the overall detection on the test set, we further explore the detection results in three aspects. We split the test set according to lesion type, size, and slice intervals. The lesion types contain lung (LU), mediastinum (ME), liver (LV), soft tissue (ST), pelvis (PV), abdomen (AB), kidney (KD), and bone (BN). In addition, the dataset is split into some subsets according to lesion diameters and slice intervals. The diameter is the data provided in DeepLesion. To enrich the information, we also interpolate in CT slices to generate all slice intervals to 2 mm.

As shown in Table 2, our model boosts the detection sensitivity in all types of lesions. The ME type mainly consists of lymph nodes in the chest. AB lesions are miscellaneous ones that are not in the LV or KD. The

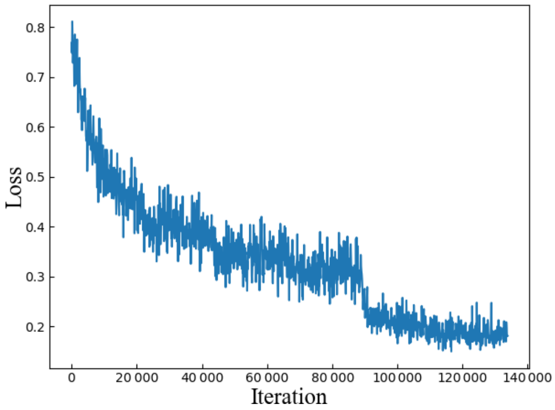
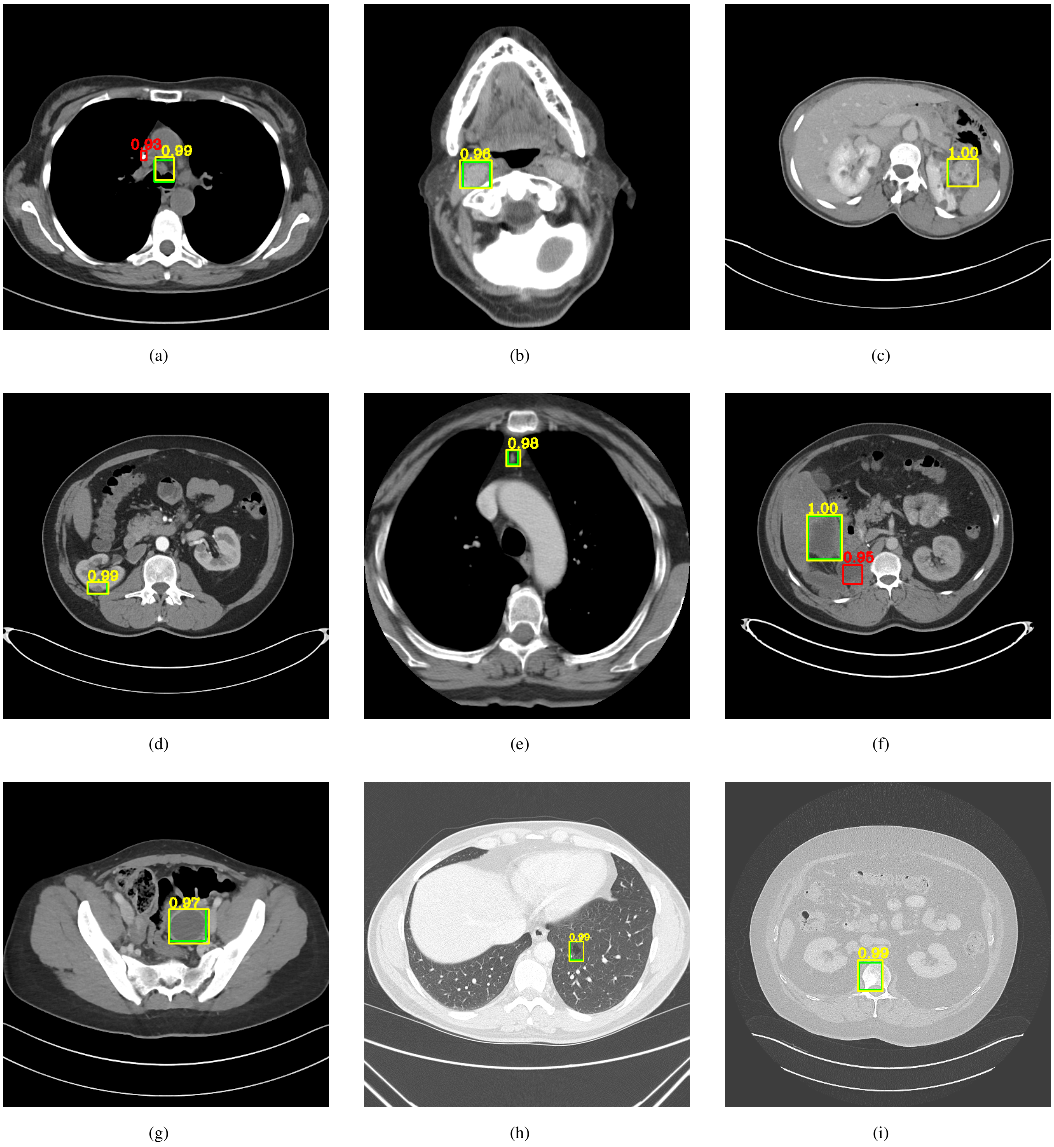


Fig. 6 Mini-batch loss schema of our model.

Table 1 Sensitivity at different FPs per image and inference time on the official test set (the bold text means the best-performed data).

Method	Sensitivity (%)						Inference time (ms)
	FP = 0.5	FP = 1	FP = 2	FP = 4	FP = 8	FP = 16	
Faster RCNN[12]	55.5	66.3	74.9	83.8	85.0	88.9	32
Improved R-FCN	56.5	67.7	76.9	82.8	87.0	89.8	**27**
3DCE[8], slices=9	61.7	71.9	79.2	84.3	87.8	89.7	56
3DCE[8], slices=15	63.0	73.1	80.2	85.2	87.8	89.7	74
3DCE[8], slices=21	63.2	73.4	80.9	85.6	88.4	90.2	98
3DCE_CS_Att[21], slices=9	67.8	76.3	82.9	86.6	89.3	90.7	–
3DCE_CS_Att[21], slices=15	70.8	78.6	83.9	87.5	89.9	91.4	–
3DCE_CS_Att[21], slices=21	**71.4**	**78.5**	**84.0**	**87.6**	**90.2**	**91.4**	–
Proposed, slices=9	63.3	73.4	80.3	85.0	88.7	90.6	**27**
Proposed, slices=15	66.2	76.0	81.9	86.9	88.4	91.0	38
Proposed, slices=21	69.2	78.3	83.3	87.1	89.9	91.0	50



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 7 Selected detection results in test set. Green, yellow, and red boxes represent ground truth, TPs, and FPs boxes, respectively.

Table 2 Sensitivity of 4 FPs on official test set of DeepLesion. The detection sensitivity is separately counted by lesion type, diameter, and slice interval.

(%)

Method	Lesion type								Lesion diameter			Slice interval	
	LU	ME	LV	ST	PV	AB	KD	BN	$\leqslant$ 10 (mm)	10–30 (mm)	$\geqslant$ 30 (mm)	$\leqslant$ 2.5 (mm)	$>$ 2.5 (mm)
Faster RCNN[12]	88.0	84.0	80.0	76.0	76.0	75.0	72.0	55.0	72.0	83.0	80.0	80.0	80.0
Improved R-FCN	88.9	84.1	80.2	75.8	77.3	74.3	72.1	56.6	72.4	83.9	81.1	80.8	81.0
3DCE[8]	90.9	88.1	90.4	73.6	82.1	81.3	82.1	75.0	81.0	87.9	83.0	85.8	85.0
3DCE_CS_Att[21]	**92.0**	88.5	**91.4**	**80.3**	**85.0**	**84.4**	**84.3**	75.0	82.3	90.0	**85.0**	**87.6**	**87.6**
Proposed	91.5	**88.7**	91.0	79.3	84.7	84.1	84.0	**75.6**	**84.2**	**91.0**	83.9	86.5	86.3

ST type contains lesions in the muscle, skin, and fats. The best improvements of sensitivity are the detection of LU, ME, and LV lesions, as shown in Table 2. On the basis of the conclusion of the LU lesion, a tissue with a smooth edge, focal fat, or fat alternating with calcific foci (popcorn calcification) can be easily detected as an LU lesion[31], supporting our inference. The foreground of these organs is more discriminative than others due to the abnormal intensity and representative appearance in normal backgrounds. To the best of our knowledge, the above-mentioned organs are normally easy to detect, thus benefiting from discriminative backgrounds.

In Table 2, the small lesions benefit mostly from the dynamic region proposal mechanism. The sensitivity of lesions whose diameters are less than 30 mm is promoted. As shown in Figs. 7f and 7h, our model attaches importance to lesion with a small scale. As shown in Fig. 5, our model normally generates accurate region proposals. With the help of Eq. (5), lesions with small shapes are detected properly. With the small slice intervals, 3D context is much more precise than that learned from loose intervals. Therefore, finer slices provide more representative features generated by the intermediate slices compared with the ones generated by interpolated slices. Faster RCNN[12] is trained with three slices, whereas 3DCE[8] and our model are trained with 15 slices, as shown in Table 1.

In summary, our model has better properties than the baseline, and the foreground ratio of region proposals is improved. In addition, the accurate region proposals boost the inference time without decreasing the sensitivity at FPs, as shown in Table 1. The lesions whose diameters are less than 30 mm benefit mostly from our model. The attention mechanism in 3DCE_CS_Att[21] acquires a considerable result. Our model generates competitive detection results and achieves a valid improvement for CT images with different slice intervals and diameters. Moreover, the training parameters of our model are less than those of 3DCE_CS_Att[21], and some hard negative samples are avoided with our model. In the experiments, our model is also restricted by the manually set length dictionary. When the length of lesion is not in the range we set, the bounding boxes predicted is not as tight as we expected. The mechanism is also restricted when the intensity of lesions is almost the same with the background[31].

5 Conclusion

In this study, we propose a new method to detect lesions. Our system is based on a new region proposal network, which is used to generate more accurate region proposals than existing RPN[12]. We also integrate the 3D context into the learning procedure. The aggregating features extracted from multislices result in high sensitivity at different FPs. Experiments show that our model constantly improves the accuracy of region proposals. Compared with the baseline, our model promotes the accuracy and remarkably reduces the inference time. Detection accuracy of lesions whose diameters are less than 30 mm is improved by approximately 3% in relation to the baseline. In the future, we will try to import additional medical knowledge and some exciting technologies, such as attention mechanism, into our studies to improve sensitivity at FPs continuously.

Acknowledgment

The paper was supported by the National Natural Science Foundation of China (Nos. 62072135, 61672181).

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, Lake Jahoe, NV, USA, 2012, pp. 1097–1105.

[2] C. Lam, C. Yu, L. Huang, and D. Rubin, Retinal lesion detection with deep learning using image patches, *Investigative Ophthalmology & Visual Science*, vol. 59, no. 1, pp. 590–596, 2018.

[3] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. Li, and F.-F. Li, Thoracic disease identification and localization with limited supervision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8290–8299.

[4] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection, *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2016.

[5] J. Wang, J. H. Noble, and B. M. Dawant, Metal artifact reduction for the segmentation of the intra cochlear anatomy in CT images of the ear with 3D-conditional GANs, *Medical Image Analysis*, vol. 58, p. 101553, 2019.

[6] J. Park, J. Yun, N. Kim, B. Park, Y. Cho, H. J. Park, M. Song, M. Lee, and J. B. Seo, Fully automated lung lobe segmentation in volumetric chest CT with 3D U-net: Validation with intra-and extra-datasets, *Journal of Digital Imaging*, vol. 33, no. 1, pp. 221–230, 2020.

[7] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, Evaluate the malignancy of pulmonar nodules using the 3D deep leaky noisy-or network, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3484–3495, 2019.

[8] K. Yan, M. Bagheri, and R. M. Summers, 3D context enhanced region-based convolutional neural network for end-to-end lesion detection, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, 2018, pp. 511–519.

[9] Q. Tao, Z. Ge, J. Cai, J. Yin, and S. See, Improving deep lesion detection using 3D contextual and spatial attention, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Shenzhen, China, 2019, pp. 185–193.

[10] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, Selective search for object recognition, *International Journal of Computer Vision,* vol. 104, no. 2, pp. 154–171, 2013.

[11] 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*, Columbus, OH, USA, 2014, pp. 580–587.

[12] S. Ren, K. He, R. B. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 91–99.

[13] J. Dai, Y. Li, K. He, and J. Sun, R-FCN: Object detection via region-based fully convolutional networks, in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 379–387.

[14] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, Libra R-CNN: Towards balanced learning for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 821–830.

[15] S. Bae, Object detection based on region decomposition and assembly, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 8094–8101.

[16] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, Free anchor: Learning to match anchors for visual object detection, in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 147–155.

[17] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 7263–7271.

[19] Z. Tian, C. Shen, H. Chen, and T. He, FCOS: Fully convolutional one-stage object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 9627–9636.

[20] C. L. Zitnick, and P. Dollár, Edge boxes: Locating object proposals from edges, in *Proc. of European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 391–405.

[21] Y. Li, Detecting lesion bounding ellipses with Gaussian proposal networks, in *Proc. of International Workshop on Machine Learning in Medical Imaging*, Shenzhen, China, 2019, pp. 337–344.

[22] J. Ribera, D. Guera, Y. Chen, and E. J. Delp, Locating objects without bounding boxes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6479–6489.

[23] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, Centernet: Keypoint triplets for object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 6568–6577.

[24] S. Liu, D. Huang, and Y. Wang, Adaptive NMS: Refining pedestrian detection in a crowd, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6459–6468.

[25] L. Cai, B. Zhao, Z. Wang, J. Lin, C. S. Foo, M. S. Aly, and V. Chandrasekhar, Maxpool NMS: Getting rid of NMS bottlenecks in two-stage object detectors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 9356–9364.

[26] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, Distance-IoU Loss: Faster and better learning for bounding box regression, in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 2020, pp. 12 993–13 000.

[27] K. Simonyan and Z. Andrew, Very deep convolutional networks for large-scale image recognition, https://arxiv.org/pdf/1409.1556, 2014.

[28] L. Xie, Y. Liu, L. Jin, and Z. Xie, DeRPN: Taking a further step toward more general object detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 2019, pp. 9046–9053.

[29] K. Yan, X. Wang, L. Lu, and R. M. Summers, Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations, https://arxiv.org/abs/1710.01766, 2017.

[30] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and F.-F. Li, ImageNet: A large-scale hierarchical image database, in *Proc. of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009, pp. 248–255.

[31] S. Suut, A. Zeid, A. Carolyn, and R. Prabhakar, Pictorial essay of radiological features of benign intrathoracic masses, *Annals of Thoracic Medicine*, vol. 10, no. 4, pp. 231–242, 2015.

Jiacheng Jiao received the BS degree in naval architecture and marine engineering from Harbin Engineering University in 2018. He is a master student at the College of Computer Science and Technology, Harbin Engineering University. His research interests include computer vision, computer-aided-diagnosis, medical-image detection, and deep learning.

Chunling Chen received the BS degree from Heilongjiang University in 2015 and the MS degree from Harbin Engineering University in 2018. She is a PhD candidate of Harbin Engineering University. Her research focuses on medical image processing, computer-aided-diagnosis, and deep learning.

Haiwei Pan received the PhD degree from Harbin Institute of Technology, Harbin, China in 2006. He is currently a professor and doctoral supervisor at Harbin Engineering University. His current research interests include big data, artificial intelligence, and medical image mining. He has authored or coauthored more than 60 publications in related areas. He has got more than 10 scientific research projects at the national level, and provincial and ministerial-level.

Tao Jin received the BS degree from Inner Mongolia University of Science and Technology, China in 2002, the MS degree from Tsinghua University, China in 2008, and the PhD degree from Tsinghua University, China in 2013, and currently he is a postdoctoral researcher at the School of Software, Tsinghua University, China. His research focuses on business process model management, including process model retrieval, process model refactoring, process model difference, behavior computing, and so on.

Yang Dong received the BS degree in computer science and technology from Harbin Engineering University in 2018. He is now a master student at the College of Computer Science and Technology, Harbin Engineering University. His research interests include computer vision and medical images processing.

Jingyi Chen received the BS degree in computer science and technology from Harbin Engineering University in 2018. She is now a master student at the College of Computer Science and Technology, Harbin Engineering University. Her research interests include object detection and deep learning.