Multi-scale Convolutional Neural Network Based on 3D Context Fusion for Lesion Detection

Zebiao Wu[1], Jinshan Chen[1], Zongyue Wang[1], Jinhe Su[1], and Guorong Cai[1,2(✉)]

[1] Computer Engineering College, Jimei University, Xiamen 360121, China
guorongcai.jmu@gmail.com
[2] Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350003, China

Abstract. Lesion detection is an essential technique in medical diagnostic systems. Since there are great differences in intensity and appearance within a same lesion category, lesion detection from computed tomography (CT) scans is still a challenging task. Sufficiently using 3D context information become the research hotpot in lesion detection area, since algorithms can benefit from geometry and texture of lesions. Motivated by this trend, we propose a multi-scale CNN based on 3D context fusion, called M3DCF, for extracting lesion area from CT scans. In order to speed up the algorithm, the one-stage regression-based detector, rather than region proposal network, is adopted. Specifically, we employ 3D context fusion strategy that allows M3DCF fusing features from neighboring slices. Finally, we use a multi-scale scheme to combine low-level and high-level features. This strategy allows us to get more meaningful semantic information. The experimental results conducted on DeepLesion dataset indicates that the proposed method outperformed state-of-the-arts, including RetinaNet, Faster R-CNN, and 3DCE. The source code is available on https://github.com/JMUAIA/M3DCF.

Keywords: Lesion detection · One-stage neural network · 3D context

1 Introduction

Computed tomography (CT) scans are the most frequently used medical images for high-precision X-ray measurement. Hence, lesion can be identified by a specialist based on CT images' density. Different from common images, CT images have more details and self-similar texture. So far, machine learning based automatic diagnosis from CT scans cannot make a breakthrough due to the fact that there is lack of large-scale annotated lesion database. Recently, NIHCC

The first author is a student. This work is supported by the National Natural Science Foundation of China under Grant No. 61702251, the Key Technical Project of Fujian Province under Grant No. 2017H6015.

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11857, pp. 573–585, 2019.
https://doi.org/10.1007/978-3-030-31654-9_49

released a large-scale medicine database called DeepLesion [1] on July 20, 2018. The DeepLesion dataset contains 32,735 lesions on 32,120 axial slices, which are annotated from 10,594 CT studies of 4,427 unique patients. Different from existing datasets that typically focus on one type of lesion, DeepLesion contains a variety of lesions including those in *Abdomen*, *Liver*, *Mediastinum*, etc. Basically, DeepLesion provides a good chance to handle the task of general lesion extraction, since mainstream object detection algorithms are mainly depended on large scale annotated data.

In the past five years, algorithms of object detection can be divided into two classes. The first one is based on two-stages scheme, which is based on proposal driven mechanism. The second scheme employs an end-to-end CNN [2] framework to regress the coordinates and the confidences of each target.

In the development of the two-stages network structures, R-CNN [3] (Region CNN) was the first use of deep learning for object detection. The author of RCNN has won awards in PASCAL VOC [4] at that time. R-CNN adopted selective search [5] to generate possible ROIs (Regions of Interest), then the proposals are classified by standard convolutional neural networks. Fast R-CNN [6] is an improved version of R-CNN that adopted ROI Pooling (Region of Interest Pooling) to share parameters. Faster R-CNN [7] integrates object proposal generation with classifier into a single convolution network, which is faster than Fast R-CNN since Region Proposal Network (RPN) shares full-image convolutional features. Mainstream two-stage algorithms including R-CNN, SPP-Net [8], Fast R-CNN, Faster R-CNN, R-FCN [9], etc.

OverFeat [10] is a first one-stage object detector based on CNNs. Recently, YOLO [13–15], SSD [11], DSSD [16] and RetinaNet [12] have innovative behavior in one-stage methods. SSD adopted multi-scale feature maps for object detection, 3×3 convolution kernels are used for fewer parameters. Besides this, anchor boxes are used for easier training, which are popular used in two-stages method. Since the featues SSD extracted are not as great as two-stages' features, DSSD adopted better backbone and deconvolution layers for feature extraction, therefore more expressive features are obtained. In addition, since imbalanced proportion of positive and negative samples have serious impacts on network training, a novel loss and well-designed network are used by RetinaNet and which achieves promising results.

Although deep learning achieves enduring greatness on the field of object detection. Most object detectors are designed for 2D images. Different from traditional 2D images, lesions in CT images have self-similar texture, which means non-lesions and true lesions areas may have similar appearances. Under such circumstances, using 3D context information to improve the performance of lesion detection becomes the research hotspot. To this end, Dou et al. [17] attempted to predict microbleeds in brain MRI by using a modified 3D CNN. Hwang and Kim [18] adopted weakly-supervised deep learning to detect nodules in chest radiographs and lesions in mammography. Besides, Teramoto et al. [19] proposed a novel CNNs for handling multi-model fusion task.

Recently, the DeepLesion team proposed a 3D Context Enhanced Region-based Convolutional Neural Network [20] (3DCE) for extracting lesion areas from CT scans. It is worth noting that this 3DCE adopts 3D context fusion strategy in the pipeline. The neighbor slices of target image are used to be the input to CNNs. Then the Region Proposal Network (RPN) of 3DCE generates a sparse set of candidate lesions. Finally, the feature maps of each candidate are used to lesion type classification. Although 3DCE CNN is a cost-efficient solution, the algorithm suffers from too much false positives. Since the proposed method is motivated by 3DCE, their results on Deeplesion are regarded as the baseline.

In this paper, we propose a novel method called Multi-scale Convolutional Neural Network based on 3D Context Fusion for lesion detection (M3DCF). First, we adopt one-stage framework rather than region proposal network. The purpose is to speed up the algorithm. Second, we designed multi-scale prediction scheme to fuse 3D context information. Third, our method performs bounding box regression directly from the last layer of our neural network. The experimental results show that the proposed method achieves state-of-the-art accuracy with faster speed than the mainstream algorithms such as RetinaNet, Faster R-CNN and 3DCE.

2 The Proposed Method

2.1 3D Context Features Fusion Network

Figure 1 describes the pipeline of the proposed method. First, we used three neighboring axial slices to compose a three-channel image. Second, 3*M

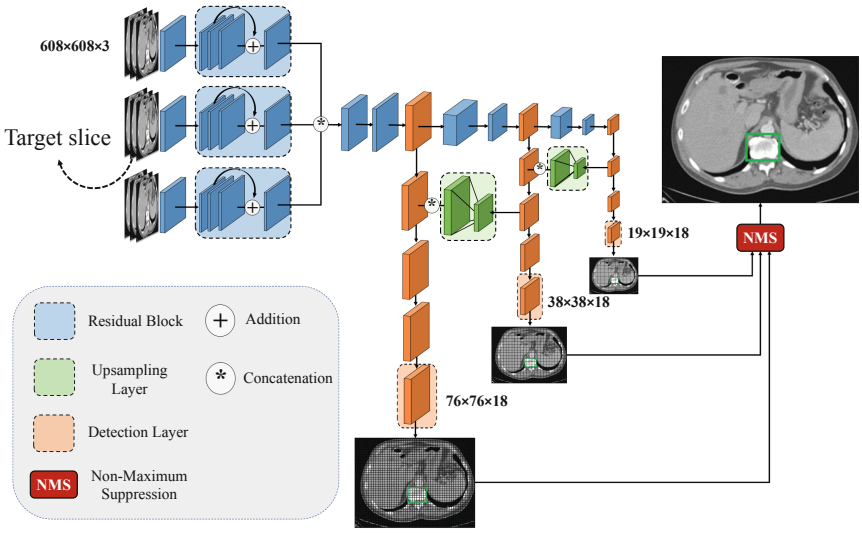

Fig. 1. Structure of the neural network we proposed

sequential slices have been divided into M images as the input of M3DCF. Note that the middlemost slice is the target slice with annotation. The main advantage of our fusion strategy is that it makes full use of shape information through 3D context, thus can enhance the discrimination ability of features.

Third, each image is fed into a respective convolutional neural network for performing feature extraction. M3DCF uses the Darknet-53 as its backbone network, since Darknet-53 is an efficient architecture that suitable for one stage object detection. In particular, Batch Normalization [21] (BN) is used in the backbone with each convolution layer. Moreover, we adopt the Rectified Linear Unit [22] (Relu) as our activation function after BN. The Convolution + Batch Norm + LeakyRelu (CBL) is a fundamental module in the pipeline of our framework. Additionally, residual modules [23] are proposed to improve the training of a deeper neural network. To ensure efficient training, we adopt residual modules of layer 1, 2 and 8. As shown in Fig. 1, we concatenate feature maps which are generated after several residual modules to construct a 3D context features. The concatenation of extracted features is a key component of M3DCF, which ensure more expressive features can be obtained. Different from simple addition, the concatenation in our method will expand the dimension of feature.

Most worthy of mention is that we adopt a multi-scale prediction strategy. The design of multi-scale anchors is an essential component for sharing features. Besides, the resolution of feature maps can be easily adjusted by changing the stride of convolution layer. In this way, we can transform the dimensions of tensor without any pooling layer and fully connected layer. As shown in Fig. 1, the input images are reduced to 1/32 of its original size in the first branch. For example, using stride as (2, 2) can halve the width and the height of the input size, then an input image with the size of 608 $\times$ 608 will be resized into 19$\times$19.

Moreover, with the goal of multi-scale prediction, we need to combine low-level and high-level features. As shown in Fig. 1, feature map from former layers will be merged with upsampled features from latter layers. After that, convolutional layers are used to process combined feature maps. It is worth noting that our method predicts 3 boxes at different scales. The output tensor encoding the coordinates of a bounding box, the confidence of an object, and the class predictions. Since the only category is 'lesion' in our task, the dimension of the output tensor is N $\times$ N $\times$ [3 $\times$ (4 + 1 + 1)]. Finally, we perform non-maximum suppression (NMS) on the detection confidence maps to obtain detection results.

2.2 Loss Function

Loss function has the important guiding significance of neural network during the training process. The neural network benefits a lot from a well-designed loss function and achieves competitive performance. There are many reliable loss functions and most of them have widespread adoption. Traditional loss function such as Mean Square Error [24] (MSE), Cross-Entropy [25] (CE), Root means squared error [26] (RMSE) and Sum Square Error [27] (SSE) are successfully used in different situations. Therefore, using a well-designed loss function for the practical problem may produce better results. As for object detection, there are

4 attributions need to be considered, such as location, size, class, and confidence. It's worth noting that there is only one category in the DeepLesion database and thus we haven't taken the category error into consideration.

2.2.1 Bounding Box Location Loss

First, we use binary cross-entropy loss for the location predictions. Thus, this formulation will be helpful when we move to more complex domains like the area of *Pelvis*. Since lesion detection has distinct difference of probability distribution, cross-entropy loss function is more suitable when there is large error between the predicted bounding boxes and the ground truths. Cross-entropy loss function makes the weights update faster, which accelerates the training process of CNNs. The location loss is defined as the following equation.

$$Loss_{location} = \sum_{l=1}^{L} \alpha_l \sum_{i=0}^{S_l^2} \sum_{j=0}^{B} \tau_{ij}^{obj} CrossEntropy[(x_i, y_i), (\hat{x}_i, \hat{y}_i)] \tag{1}$$

where $l = 1, 2..., L$ denotes the number of different resolutions of the output layer. x and y represent the upper-left coordinates of predicted bounding boxes, respectively. $\hat{x}$ and $\hat{y}$ stand for the ground truths. S_l^2 denotes the number of cells in each resolution, B indicates the number of boxes predicted in each cell. α_l is a weighting factor of the l^{th} resolution loss term. Additionally, τ_{ij}^{obj} stand for whether the Intersection over Union (IoU) of j^{th} window predicted by the i^{th} cell is higher than a specific threshold, which value is set to be 0 or 1. Generally, the IoU threshold is set to be 0.5.

2.2.2 Bounding Box Size Loss

As sum of squared error loss has been widely used in the object detection domain. We adopt sum of squared error loss as our bounding box size loss, and the formulation of bounding box size loss is shown as follow.

$$Loss_{size} = \sum_{l=1}^{L} \alpha_l \sum_{i=0}^{S_l^2} \sum_{j=0}^{B} \tau_{ij}^{obj} \cdot [(\sqrt{w}_i - \sqrt{\hat{w}_i})^2 + (\sqrt{h}_i - \sqrt{\hat{h}_i})^2] \tag{2}$$

where w and h denote the size of predicted bounding boxes, respectively. $\hat{w}$ and $\hat{h}$ respectively stand for the ground truths. The other symbols are defined in Sect. 2.2.1.

2.2.3 Confidence Loss

The last component represents whether a predicted bounding box contains a lesion or not. Since confidence describes the terms of probability distribution, we adopt the cross entropy to construct the confidence loss. Since the values of confidence ranges from 0 to 1, there is not weighting factor in the confidence loss for different resolutions. Hence, if regressed bounding box contains a lesion, the confidence loss is given as the following equation.

$$Loss_{object} = \sum_{l=1}^{L}\sum_{i=0}^{S_l^2}\sum_{j=0}^{B}\tau_{ij}^{obj} \cdot CrossEntropy(C_i, \hat{C}_i) \tag{3}$$

The definitions of L, S_l^2, B, τ_{ij}^{obj} are given in Sect. 2.2.1. As shown in Eq. 3, the loss combines the confidence errors in all proposal bounding boxes from different scales. On the contrary, if a regressed bounding box does not contain a lesion, the confidence loss is defined as Eq. 4.

$$Loss_{noobj} = \sum_{l=1}^{L}\sum_{i=0}^{S_l^2}\sum_{j=0}^{B}\tau_{ij}^{noobj} \cdot CrossEntropy(C_i, \hat{C}_i) \tag{4}$$

where τ_{ij}^{noobj} denotes whether the j^{th} bounding box that predicted by i^{th} cell contains a lesion or not. If the intersection over union (IoU) of prediction and the ground truth is below a well-setting threshold, τ_{ij}^{noobj} is set to be 0. As is mentioned in Sect. 2.2.1, we set $IoU = 0.5$ in the confidence loss function.

2.2.4 Overall Loss Function

In summary, the loss function of M3DCF is a combination of the mentioned components, and which is defined as the following equation.

$$Loss = \lambda_{coord}L_{location} + \lambda_{coord}L_{size} + L_{object} + \lambda_{noobj}L_{noobj} \tag{5}$$

where λ_{coord} and λ_{noobj} are respectively the weighting factors, which are designed for addressing imbalance of coordinate and confidence loss. For the reason that the ranges of location, size and confidence are different, the values of λ_{coord} and λ_{noobj} should depend on the ratio of their range. Generally, YOLO v3 recommends that the weights of location and the size can be the same, where the values should relate to the size of input object. Based on this scheme, we analysis the resolution of lesions in DeepLesion database. It's worth noting that the loss ranges of location, size and confidence will achieve stabilization during the training process when $\lambda_{coord} = 0.5$ and $\lambda_{noobj} = 0.5$.

3 Experiments

3.1 Experiment Setup

The experiments are conducted on DeepLesion dataset to evaluate the performance of the proposed algorithm. Specially, the training and the testing processes of all methods are performed on NVIDIA 1080Ti GPU. The DeepLesion is divided into training (70%, 22,901 lesions), validation (15%, 4887 lesions), and testing (15%, 4912 lesions) sets. There are several categories of DeepLesion, including *Bone*, *Mediastinum*, *Abdomen*, *Liver*, *Softtissue*, *Lung*, *Kidney*, and *Pelvis*. With the goal of encoding 3D information, we used three axial slices to compose three-channel images, and then each image is used to be the

input to CNN. Note that only the target CT slice that contains the ground truth bounding box. As for state-of-the-arts, we chose 3DCE, Faster R-CNN and RetinaNet as the baselines. As for Faster R-CNN, we use VGG-16 [28] pretrained on ImageNet as the backbone. The learning rate is set to be 0.001, and the batch is set to be 1 with the momentum is 0.9. RetinaNet is trained with Resnet-50 [23] where the input image size is $600 \times 1000 \times 3$. The initial learning rate of RetinaNet is set to be 0.00001, where the batch and the momentum is respectively as 1 and 0.9. As for 3DCE, we used VGG-16 as its pretrained backbone. The input image is with sized of $512 \times 512 \times 3$ (or $512 \times 512 \times 9$), the batch and the gradient descent momentum is set to be 1 and 0.9, respectively.

3.2 Results and Discussions

3.2.1 Overall Evaluation

Most existing works use sensitivity as their evaluative criteria, since the sensitivity measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of lesions which are correctly identified as having the condition). The 3DCE adopts "sensitivity at 4 false positives per image" as their assessment criteria. In 3DCE, they drew the sensitivity curve with average false positives per image, then the sensitivity rise as more false predictions allowed. In order to obtain more general evaluation, we considered this general object detection evaluation metric, namely mean average precision, to measure the performance of each algorithm. Specifically, we draw precision-recall curve to visualize the results of all methods.

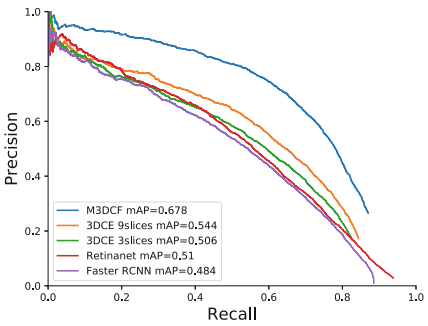

Fig. 2. Precision-recall curve and mAPs of different algorithms.

Figure 2 shows the precision-recall curve and the Mean Average Precision of all methods, respectively. It is obvious that the proposed algorithm significantly outperforms the other methods. Noteworthy that our method increases at least 0.13 mAP than state-of-the-art methods, such as 3DCE 9 slices. The results of 3DCE 9 slices indicates 3D context feature fusion drastically improve the discriminant ability of lesion features that extracted from different scales. Moreover,

one can see that the accuracy of all algorithms dropped significantly when recall reaches a high level such as recall > 0.8. This phenomenon means that there will be a large proportion of false positives with high confidence. It is worth noting that our method has significant produces a better result while M3DCF maintains the precision with 0.4. M3DCF obviously achieves state-of-the-art accuracy with faster than the other existing lesion detectors.

Table 1. The mAP and AP of the experimented algorithms.

Method	mAP	AP							
		Bone	*Abdomen*	*Mediastinum*	*Liver*	*Lung*	*Kidney*	*Softtissue*	*Pelvis*
RetinaNet	0.510	0.510	0.418	0.543	0.515	0.603	0.415	0.440	0.407
Faster R-CNN	0.484	0.530	0.369	0.495	0.541	0.574	0.411	0.422	0.362
3DCE 9slices	0.544	0.475	0.451	0.560	0.553	0.648	0.468	0.426	0.451
3DCE 3slices	0.506	0.423	0.408	0.507	0.537	0.617	0.415	0.413	404
3DCFF	**0.678**	**0.671**	**0.629**	**0.761**	**0.722**	**0.689**	**0.591**	**0.524**	**0.681**

As shown in Table 1, M3DCF has higher mAP and APs in all categories, which indicates that the feature fusion scheme M3DCF has significant benefits over the other lesion detection algorithms. Actually, M3DCF achieves an approximately 0.13 higher mAP than the other methods. Noteworthy that we achieve at least 0.2 mAP higher than the other lesion detectors in the categories of *Pelvis* and *Mediastinum*. Furthermore, our method also produces better qualitative results in *Bone*, *Abdomen*, *Liver*, *Kidney*, etc. General speaking, 3D feature fusion schemes are always better than detectors that only depend on one target image as the input. The exceptional thing is that RetinaNet and Faster R-CNN obtain better AP in category of *Bone* than 3DCE 9 slices.

3.2.2 Evaluation of Each Part

We plot the precision-recall curves of all algorithms shown in Fig. 3a, which was experimented in the area of *Mediastinum*. We can empirically observe that the average precision of most algorithms is above 0.5 while our method achieves an excellent result for $AP = 0.761$. Specially, the result of our method is close to the standard of practical application and which is 0.20 higher than 3DCE 9 slices. In addition, when the recall reaches 0.8, our method maintains the precision above 0.6 while the other methods drop rapidly. Figure 4 shows the detection results of experiments. It's shown that the texture of lesion is similar with surrounding areas and the lesion entirely blend with adjacent tissues. The 3DCE and Faster R-CNN which is based on Region Proposal Network generate a lot of candidate boxes, however, it's an extremely challenging task to recognize lesions under such conditions. Therefore, 3DCE and Faster R-CNN generate many useless false positives. Also, RetinaNet cannot handle this problem. On the contrary, the method we proposed successfully extract the lesion without any false positives. Meanwhile, we can infer that the fusion and multi-scale prediction strategy we designed can regress lesions precisely.

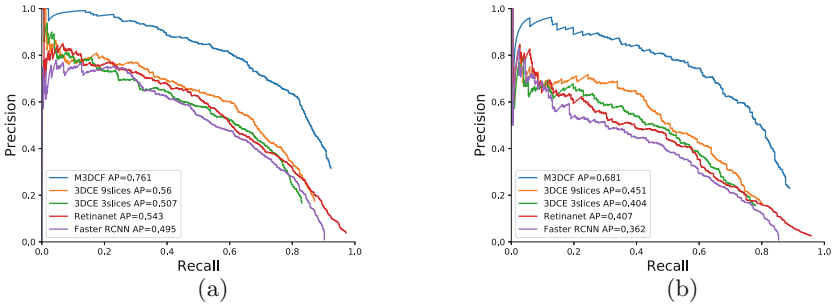

Fig. 3. Precision-recall curve of *Mediastinum* (a) and *Pelvis* (b)

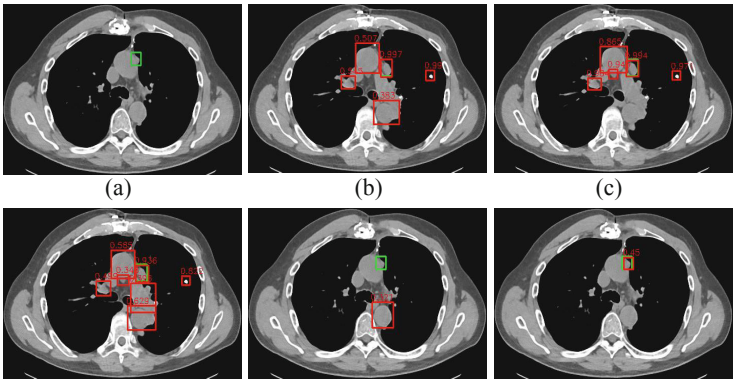

Fig. 4. The detection results on the position of *Mediastinum* (No. 000238 10 01 130), with (a) Ground truth; (b) 3DCE 3slices; (c) 3DCE 9slices; (d) Faster R-CNN; (e) RetinaNet; (f) M3DCF.

Figure 5 shows a typical slice of the area of *Pelvis*, the lesion located on the top left corner. It's an extremely challenging task to extract lesion since surrounding areas environmentally similar to the target area. Under such circumstances, current algorithms have poor performance on detection. It's worth noting that the average precision of 3DCE 9 slices is below 0.46 while Faster RCNN at 0.362 AP. However, our method achieves $AP = 0.681$ and which exceeds 0.2 accuracy. Besides, M3DCF has a competitive balance in the precision-recall trade off since the other algorithms generate a large proportion of false positives in a high recall. Obviously the strategy adopted is surprisingly effective.

3.2.3 Typical Fail Condition

We show representative failure cases in Fig. 6. CT slices' low-resolution, similar texture with surrounding, making the association of tissues complex. The results in Fig. 6 are difficult to regress precisely for M3DCF. As we can observe, the lesion

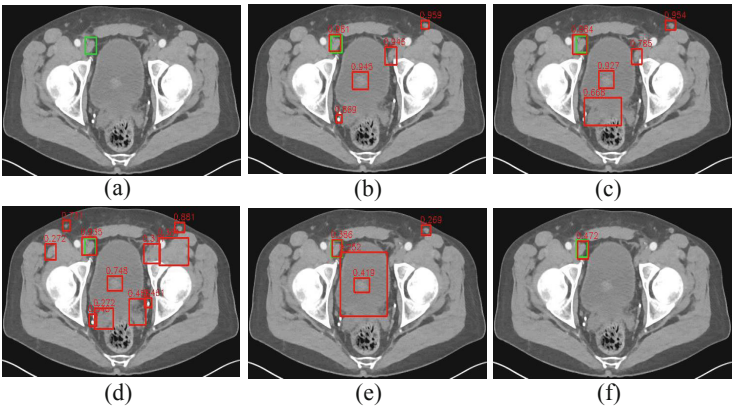

Fig. 5. The detection results on the position of *Pelvis* (No. 000268 03 01 127), with (a) Ground truth; (b) 3DCE 3slices; (c) 3DCE 9slices; (d) Faster R-CNN; (e) RetinaNet; (f) M3DCF.

of Fig. 6a and c located in the marginal region of the *Lung*, and there is plenty of similar tissues which make the performance worse. Since the lesion in Fig. 6b surrounded by the tissues and affected by light intensity, the method we proposed cannot extract the lesion well and makes a poor performance. In such cases, the algorithm is required to have strong edge sensitivity. Figure 6d and e are the slices of *Pelvis*, compared with other regions, more complex information like composition and texture bring more interference. Figure 6f shows the *Softtissue* area, it's difficult to recognize the lesion since the target is wrapped by tissues. In general, it is intractable for our method to detect the lesion especially for complicated shape, low-resolution, wrapped by surrounding tissues, etc.

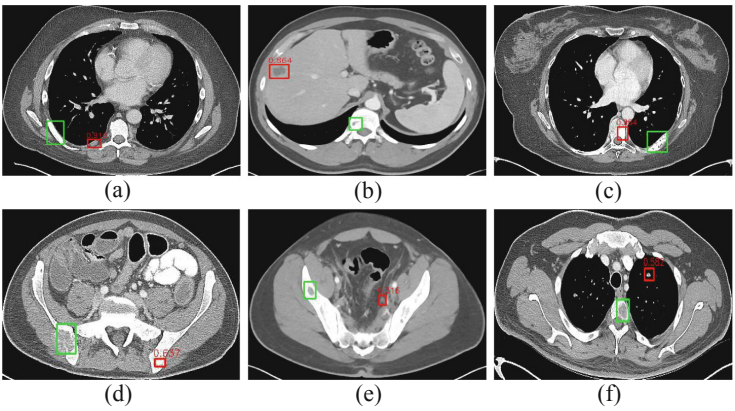

Fig. 6. The detection results on the position of *Pelvis* (No. 000268 03 01 127), with (a) Ground truth; (b) 3DCE 3slices; (c) 3DCE 9slices; (d) Faster R-CNN; (e) RetinaNet; (f) M3DCF.

Table 2. The average runtime of experimented algorithms.

Method	Average runtime (ms)
Faster R-CNN	159
RetinaNet	154
3DCE - 3 slices	79
3DCE - 9 slices	148
M3DCF	49

4 Runtime Evaluation

Runtime results for lesion detection are shown in Table 2. And runtimes are measured on a NVIDIA 1080 Ti. As shown in the experiment, since mainstream algorithms such as Faster RCNN and 3DCE adopt Region Proposal Network (RPN) for proposal generation proceeding. In addition, the backbone of Faster R-CNN for feature extraction is a relatively complicated network and which cost a lot of time. As shown in Table 2, the average processing time of Faster R-CNN is 159 ms. It is worth noting that the processing time of RetinaNet is 154 ms closed to Faster R-CNN. 3DCE 3 slices takes the half time of 3DCE 9 slices, which obviously proved that method based on 3DCE has not made a tradeoff of speed and accuracy. Most worthy of mention is that the extra time cost can be reduced by using Darknet-53 for its significant time efficiency. At 608×608 our method runs 49 ms at 67.8 mAP with three times faster than 3DCE 9 slices.

5 Conclusion

In this paper, we propose a new method called M3DCF that acts as a more effective alternative to previous approaches for lesion detection. First, we adopt a one-stage detector to enable the network is faster and simpler. And which gives rise to effective results. Second, for characteristics of computed tomography (CT) scans, we propose a novel feature fusion strategy based on 3D context to sufficiently fuse the extracted feature of multiple neighboring slices. Third, based on our 3D context fusion strategy, the multi-scale strategy allows us to get more meaningful semantic information, and it obviously improves the performance on the test set of DeepLesion database. The experimental results conducted on DeepLesion dataset indicates that the proposed method surpass the accuracy of the mainstream algorithms such as Faster R-CNN, RetinaNet, and 3DCE, while still being faster.

References

1. Yan, K., et al.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging **5**(3), 036501 (2018)
2. Krizhevsky, A., Sutskever, I. , Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
3. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
4. Everingham, M., et al.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
5. Uijlings, J.R.R., et al.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
6. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
7. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)
8. He, K., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
9. Dai, J., et al.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems (2016)
10. Sermanet, P., et al.: OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
11. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
12. Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
13. Redmon, J., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
14. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
15. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
16. Fu, C.-Y., et al.: DSSD: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
17. Dou, Q., Chen, H., Yu, L., et al.: Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans. Med. Imaging **35**(5), 1182–1195 (2016)
18. Hwang, S., Kim, H.E.: Self-transfer learning for fully weakly supervised object localization. arXiv preprint arXiv:1602.01625 (2016)
19. Teramoto, A., Fujita, H., Yamamuro, O., et al.: Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. Med. phys. **43**(6Part1), 2821–2827 (2016)

20. Yan, K., Bagheri, M., Summers, R.M.: 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 511–519. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_58
21. Courbariaux, M., et al.: Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or −1. arXiv preprint arXiv:1602.02830 (2016)
22. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10) (2010)
23. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
24. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
25. De Boer, P.-T., et al.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
26. Levinson, N.: The Wiener (root mean square) error criterion in filter design and prediction. J. Math. Phys. **25**(1–4), 261–278 (1946)
27. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res **30**(1), 79–82 (2005)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)