



ParallelNet: multiple backbone network for detection tasks on thigh bone fracture

Mengxuan Wang¹ · Jinkun Yao² · Guoshan Zhang¹ · Bin Guan¹ · Xinbo Wang¹ · Yueming Zhang¹

Received: 8 January 2020 / Accepted: 19 March 2021 / Published online: 12 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

In this paper, a novel two-stage R-CNN network called ParallelNet is proposed for thigh fracture detection task. In the proposed method, multiple parallel backbone networks and a feature fusion connection structure are designed, which can extract features with different reception fields. Specifically, the first backbone network is denoted as main network, which adopted normal convolution to detect small fractures, the rest backbone networks are denoted as sub-networks which adopted dilated convolution to detect large fractures. We evaluated the proposed method on a thigh fracture dataset containing 3842 X-ray radiographs, 3484 of which is assigned as a training dataset and 358 as a testing dataset. The experiments compare the proposed method with other state-of-the-art deep learning frameworks, including Faster R-CNN, FPN, Cascade R-CNN and RetinaNet, especially DCFPN which focus on thighbone fracture detection task. Our framework achieved 87.8% AP50 and 49.3% AP75 which outperformed other state-of-the-art frameworks. Moreover, ablation experiments on the backbone numbers, connection styles, different dilation rates and the position of dilated convolution have been attempted, and the function of each hyperparameter is analyzed.

Keywords Thigh fracture detection · Deep learning · Parallel backbone networks · Dilated convolution

1 Introduction

Computer-aided diagnosis (CAD) is a computer-based technology to reduce the workload and promote the efficiency of the clinicians. With the development of computer science and the upgrade of computer hardware technology, CAD has been applied in many aspects of the medical domain. In medical image analysis, CAD assists the clinicians with the suggestions predicted by the networks in classification, detection and segmentation tasks. With the development of deep learning methods, multiple frameworks have been implemented using deep convolutional neural networks (CNNs) to analysis different diseases. Some deep learning frameworks have achieved physician-level accuracy at

variety of diagnostic tasks [1]. Kooi et al. [2] adopted convolution neural network in breast lesion detection in mammograms and compared CNN with two certified radiologists, results show that CNN and human radiologists have similar detection accuracy. Esteva et al. [3] identified moles from melanomas by using Google Inception-v3 network, the authors evaluated the network with two dermatologists in two ways, the results indicate that the network achieved 72.1% and 55.4% accuracy which outperformed two dermatologists in both validate ways. Recently, deep learning-based network have been studied on many other kinds of diseases. Utilizing Faster R-CNN network, Liu et al. [4] achieved the detection task on the colitis dataset, and the mean average precision reaches 50.9%. Kermany et al. [5] proposed an image-based deep learning method to identify medical diagnoses with CNNs which can distinguish three diagnoses, including choroidal neovascularization diabetic macular edema and drusen. Drozdal et al. [6] established a network called FC-ResNet to achieve segmentation tasks on electron microscopy, magnetic resonance imaging, and computed tomography images. Gibson et al. [7] build a NiftyNet infrastructure which provides a deep learning pipeline for medical imaging applications including segmentation,

Communicated by B. Prabhakaran.

✉ Guoshan Zhang
zhanggs@tju.edu.cn

¹ School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

² Department of Radiology, Linyi People's Hospital, Linyi 276000, China

regression, image generation and representation learning applications. Wang et al. [8] proposed a deep CNN-based interactive framework for 2D and 3D medical image segmentation. In [9], Zhao et al. reviewed the most commonly used vessel segmentation methods, and indicates that deep neural networks dramatically improved the accuracy and robustness of vessel segmentation. Xia et al. [10] proposed oriented grouping-constrained spectral clustering (OGCSC) to deal efficiently with medical image segmentation problems. Using multi-scale structure prior, Xi et al. [11] proposed a CNV segmentation method in OCT images. These works indicate that deep learning is feasible in medical image analysis and capable of implementing multiple disease diagnosis.

As for fracture image analysis, several studies have been presented that related to deep learning methods. Lindsey et al. [12] proposed a deep neural network by extending the U-Net architecture on wrist fracture detection task. Cheng et al. [13] used DensNet-121 to extract the features from the hip fracture images to classify if the fracture exists and adopted transfer learning by pretraining the network on a limb dataset and the result reaches an accuracy of 91%. Badgeley et al. [14] used the inception-v3 CNN architecture to perform the recognition task on hip fracture images, by removing the final classification layer and computing the feature scores in the penultimate layer. Gale et al. [15] proposed a new algorithm and refined by a recurrent neural network (RNN) with two long short-term memory (LSTM) layers. Guan et al. [16] proposed a Dilated Feature Pyramid Network (DCFPN) by introducing dilated convolution in the original Feature Pyramid Network structure, and established a thigh bone dataset including 3484 training data and 358 testing data to evaluate the detection performance. The result reached 82.1% average precision and outperformed some well-known networks.

In fact, most of above-mentioned the deep learning networks are initially applied in generic objects. Recently, generic object detection networks are uniformed in two types: one-stage networks and two-stage networks. One-stage networks adopted uniform sampling from the feature maps to generate a large number of prior boxes and based on these prior boxes to classify and localize the objects. Redmon et al. [17] proposed a real-time detection network called YOLO, including 24 convolutional layers followed by 2 fully connected layers and achieved 57.9% mean average precision on PASCAL VOC 2012 dataset. YOLO v2 [18] proposed a new method to harness the large amount of classification data and use it to expand the scope of the detection systems, which improved the performance of YOLO. Lin et al. [19] proposed RetinaNet, which adopted FPN as its backbone network and designed a focal loss function to solve the superfluous negative sample problems. Two-stage networks first extract features from the images and

generate the proposals by the Region Proposal Network (RPN) presented in Faster R-CNN [20], then classify and localize the proposals in the second stage. Feature Pyramid Network (FPN) [21] introduced a pyramid-shaped network structure to promote the performance on detecting objects at vastly different scales, by combining the low-level high-resolution feature map with the high-level low-semantical feature maps. Mask R-CNN [22] based on FPN and Faster R-CNN proposed quantization-free layer called RoIAlign to preserve exact spatial locations. Cascade R-CNN [23] improved Faster R-CNN by adding extra detectors to promote the quantity of RPN introduced a learnable anchor in RPN, which shape is changing to fit the ground truths during training. Moreover, Guided Anchor [24], Libra R-CNN [25] and TridentNet [26] proposed new structures focusing on different purposes, respectively. Overall, two-stage methods are more time consuming due to the region proposal procedure, but the accuracy is higher than one-stage method. In our thigh bone detection task, real-time detection is not required, a higher detection accuracy can better help clinicians in diagnosis. Thus, two-stage networks are more suitable than one-stage networks in this task.

In this paper, we proposed a two-stage network with multiple parallel backbone networks called ParallelNet for thigh fracture detection task. The first backbone is denoted as the main network which adopted normal convolution to extract features from tiny fractures, the other backbones are denoted as sub-networks which adopted dilated convolution with different dilation rates to enrich the reception fields for detecting large fractures. A feature fusion connection structures named backward connection is designed between each individual backbone to fuse the feature maps. We assessed this novel network on the thigh fracture dataset established in [16], including 3842 thigh fracture X-ray images, 3484 of which are assigned as training data and the remainder are used as test data. The results show relatively higher accuracy of 87.8% AP50 and 49.3% AP75 which outperformed previous thigh bone detection network, DCFPN [16] by 5.7% in AP50 and 4.2% in AP75. We also compared the proposed network with other state-of-the-art detection networks, and outperformed the next best AP50 by 2.5% (RetinaNet) and AP75 by 0.6% (Cascade R-CNN).

2 Approach

The overview of the proposed ParallelNet is demonstrated as Fig. 1. First, original X-ray fracture images are fed into the multiple backbone networks, all networks consist of deep residual blocks in five stages and share the same stage 1. Each backbone network generates feature maps with different reception fields. Backward connection is designed to connect feature maps from different stages. Second, feature

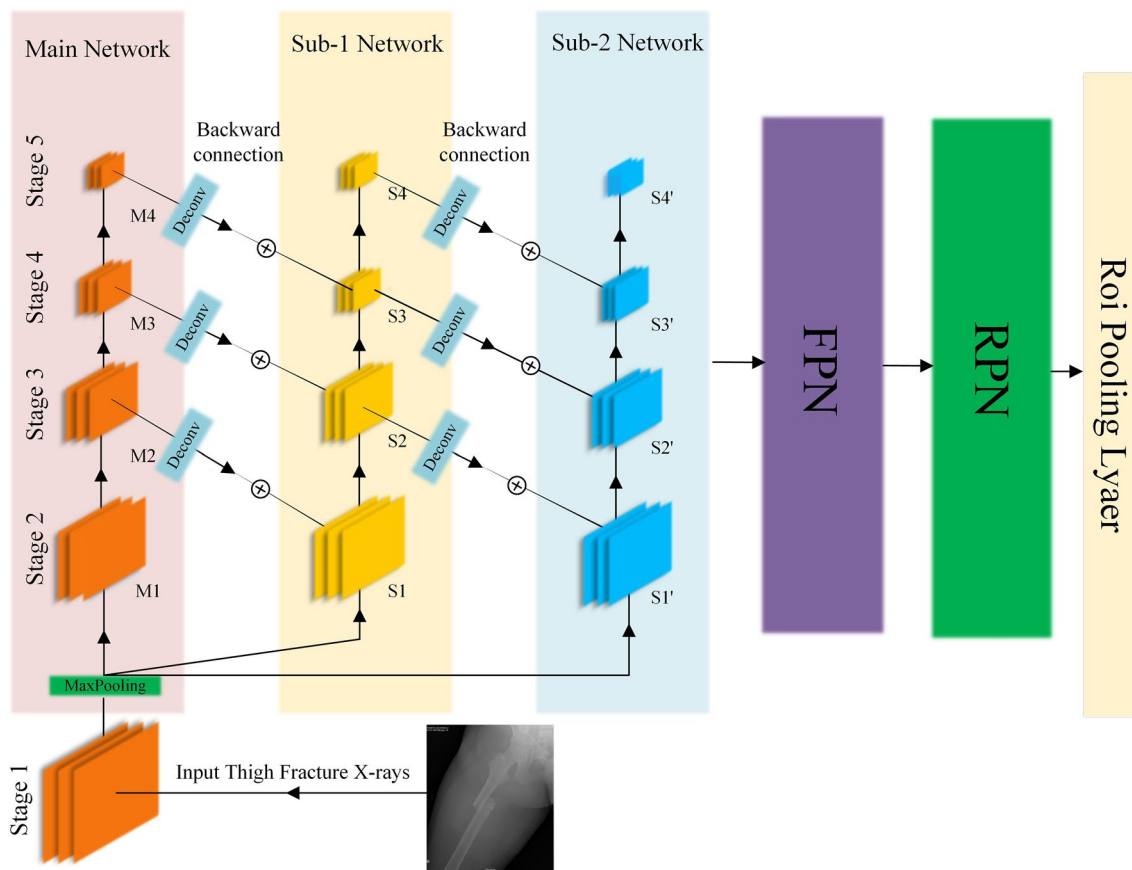


Fig. 1 The overview of ParallelNet

pyramid network structure generates feature maps with high-resolution information to detect fractures in large-various scales. Third, region proposal network (RPN) [20] RPN is employed to generate the region of interest (RoI) indicating which region contains the fracture. These RoIs will be converted into same size by RoI pooling layer to calculate their bounding-box regression and classification loss. Finally, focal loss function [19] evaluated the difference between the proposals and ground truths, both classification loss and localization loss are calculated, then Stochastic Gradient Descent (SGD) is adopted to update weights of the network.

2.1 Backbone network

Backbone network is the fundamental structure, it extracted learnable features from the original input images for the network. Previous work on thigh bone detection task, DCFPN [16] introduced dilated convolution in the backbone network and proved that enriching the reception field is conducive to improve the detection performance. However, we found the gridding effect of dilated convolution harms the performance of detecting small objects. As shown in Fig. 2, dilated convolution skipped some pixels during the calculation, this

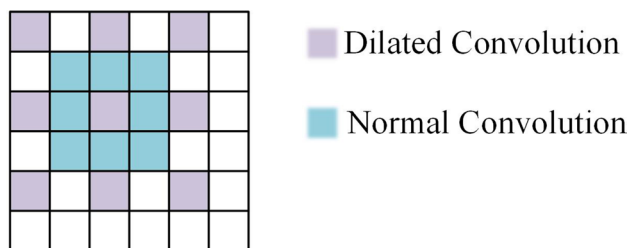


Fig. 2 Graphic illustration of normal convolution and dilated convolution

could be regarded as the normal convolution with holes, which means discrete dilated convolution sampling leads to lack of correlation in the long spatial distance sample. This indicates that, dilated convolution ignores vital information for detecting tiny objects while enriching the reception field. The larger dilation rates, the more pixels it skips. To overcome gridding effect, we designed a backbone network in three pathways with different dilation rates. Shown in Fig. 3, the first pathway is denoted as the main network, the second pathway is denoted as the sub-1 network and the last one is denoted as sub-2 network.

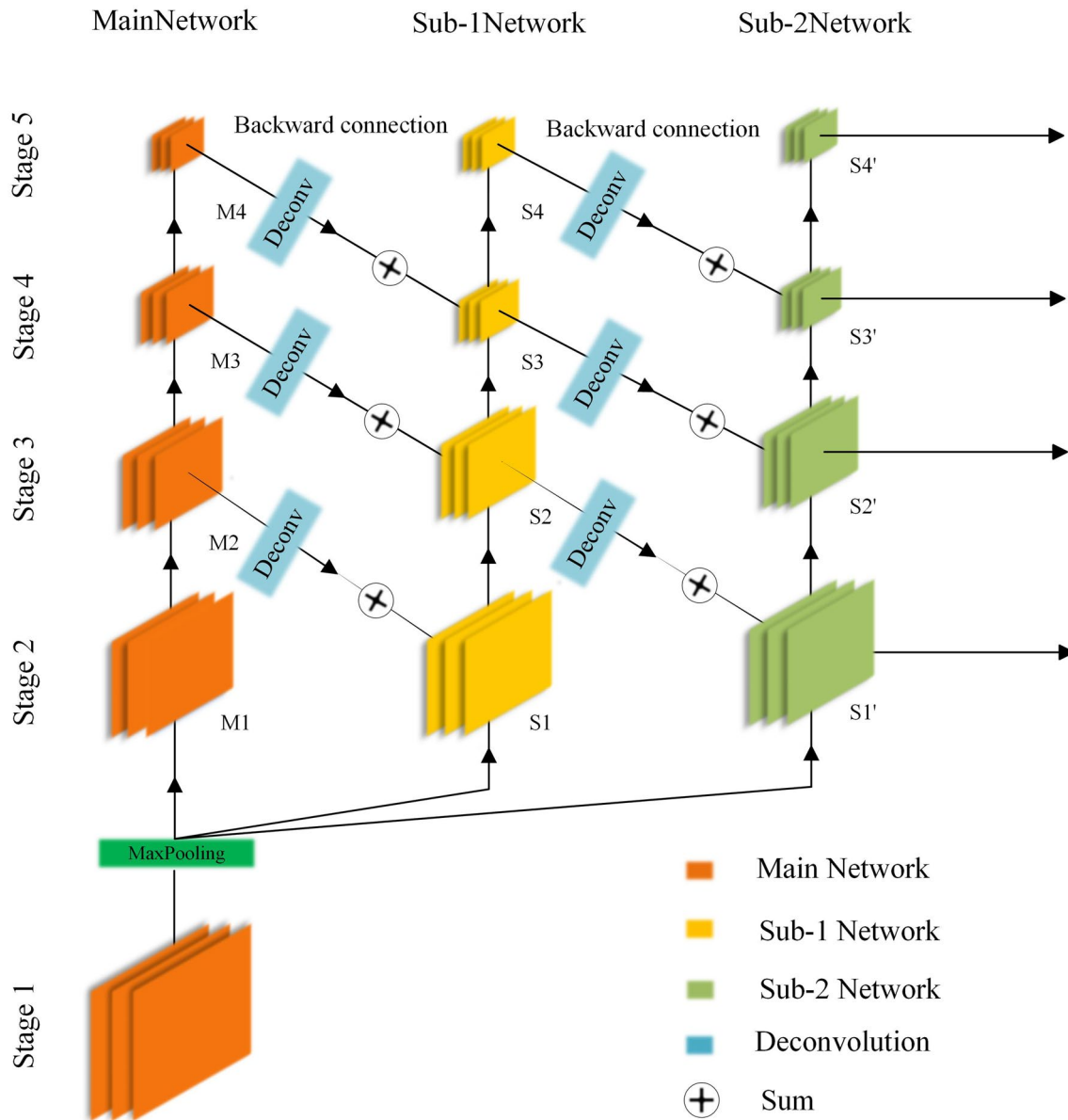
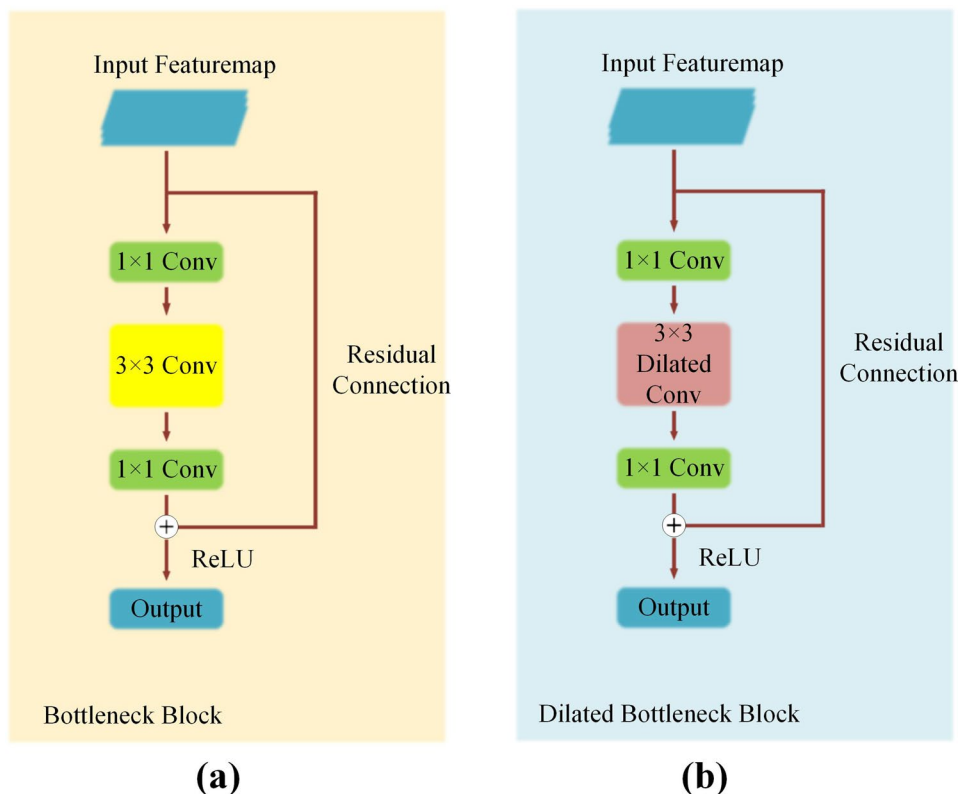


Fig. 3 The structure of proposed multiple pathway backbone network

The main network is composed of five stages and shares stage 1 with other pathways, and each stage includes several bottleneck blocks. Stage 1 composed of a 7×7 convolution kernel and a 3×3 max-pooling layer. The max-pooling layer is used to reduce the size of the feature map. Stage 2 to stage 5 contains 3, 4, 23, and 3 bottleneck blocks, respectively. These blocks extract features from the thigh fracture images and update the weights in training. The bottleneck block employed in the main network is demonstrated in Fig. 4a. The block includes three convolution kernels: the first 1×1 convolution kernel is used to decrease the input channels to $1/4$ to decrease the number of calculations required, the 3×3 convolution kernel is employed to extract features for

the network, then the final 1×1 kernel increase the channels to the input amounts. Here, the residual connection between the input and the output is considered as the identity mapping, which is used to ensure the network to train deeply. Parameters in stage 1 is frozen, because features learned in stage 1 are relatively elementary, usually some simple curves and edges. These primary features show similarity in different tasks, hence we frozen the parameters in stage 1 to save calculation resources. Parameters in stages 2–5 are trained, respectively, and did not share weights. At the end of each stage, backbone networks output a feature map with different reception fields, denoted as $\{M1, M2, M3, M4\}$. Feature maps reduce by 2 in size.

Fig. 4 The architecture of Bottleneck Block (a) and Dilated Bottleneck block (b)



The sub-1 network and sub-2 network are established with dilated convolution to obtain multi-scale features. Both of the sub-networks are composed of 3, 4, 23 and 3 dilated blocks in stages 2–5 and generate feature maps denoted as {S1, S2, S3, S4} and {S1', S2', S3', S4'}. The structure of dilated blocks is demonstrated in Fig. 4b. The dilated bottleneck block utilizes the dilated convolution kernel by replacing the 3 × 3 convolution kernel to extract more information.

The other problem of building multiple pathway networks is parameter redundancy. In our triple pathway network, parameters used are three times than single backbone networks, features from the same stage have strong similarity which brought difficulty to train the network. Hence, we designed a backward connection which fused the output feature map with the feature map of the previous stage in the second backbone. For instance, the output of the main-network M2 is fed back to fuse with S1 in the sub-network. Two types of feature maps are composed of different sizes and different channels. We used a deconvolution kernel to correct the channels and sizes of the feature maps. With this structure, feature maps are fused with different stages enriched the complexity of features, and parameters can be fully utilized.

2.2 Feature pyramid structure

Fractures have very distinguished spatial scales in different images, feature pyramid network (FPN) [21] is built to generating a series of high-resolution feature maps by

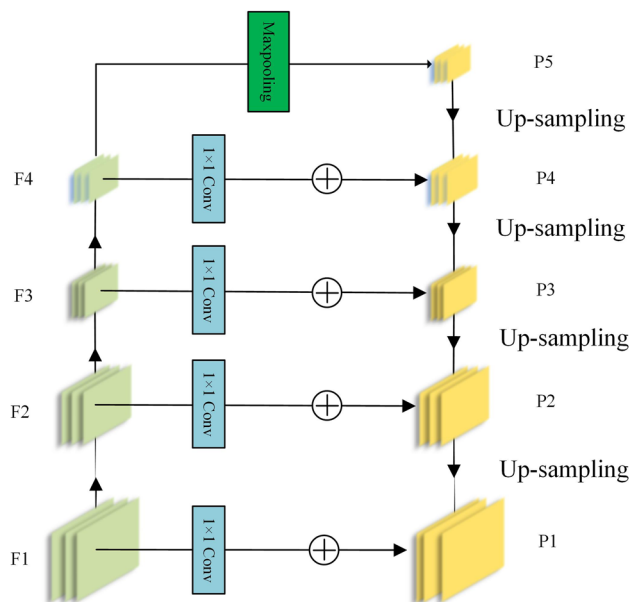


Fig. 5 Structure of feature pyramid network

up-sampling, and combined these high-resolution feature maps with the output feature maps to detect objects in vast different scales. Specifically, in Fig. 5, feature maps output by the backbones, denoted as $\{F1, F2, F3, F4\}$, are semantically high. FPN first generate feature map P5 as top layer with Maxpooling and uses up-sampling to generate high-resolution feature maps at each spatial size, denoted as $\{P1, P2, P3, P4, P5\}$. These feature maps are enhanced with features from feature map F1 to F4 via a 1×1 convolution kernel. The enhanced feature maps are combined both high-semantic information and high-resolution information which is friendly to detect vast object scales.

2.3 Region proposal network

Region proposal network (RPN) is first proposed in faster region-based convolution neural network (Faster R-CNN) [20]. This algorithm proposed a network to generate the Region of Interest (RoI), which indicates the region with the fractures. By creating anchors on every spatial pixel in feature map with three scales and three aspect ratios (0.5, 1, 2), the network will generate $W \times H \times 9$ anchors, where W and H are the width and height of the feature map. These anchors are fed to calculate the intersection over union (IoU) with the ground truths and assigned as a positive sample if IoU is over 0.6, a negative sample if IoU less than 0.3, and ignores the RoI when it is IoU is between 0.6 and 0.3. RPN predicts 128 RoIs including all positive samples if positive samples are fewer than 128, the network pads with the negative samples. The following RoI pooling layer converts the RoIs into fixed spatial size and then fed to two individual fully connected layers to turn the fixed RoI into two 4096-dimension vectors, one of which sent into the softmax classifier to label the positive and negative RoIs, the other sent to calculate the bounding-box regression with ground truths. In our network, five output feature maps in different depth and size represent the same fracture position in the original images, thus, we employed RPN in a stride of $\{4, 8, 16, 32, 64\}$ corresponding to the size to generate RoIs in each feature maps simultaneously. At last, the loss function adopted is the Fast R-CNN multi-task loss function, and gradient descent is employed with Stochastic Gradient Decent (SGD) in backward propagation.

3 Experiments

We first trained ParallelNet with the thigh fracture dataset established in [16], including 3842 thigh fracture X-ray images. 3484 images are used as training data to train the network, and 358 images assigned to test the result of the model. The dataset is annotated by Linyi People's Hospital in China and set in the same format as COCO dataset.

Considering the fracture dataset is relatively small and lacks information to teach the network, we employed transfer learning in our work using the ResNet-101 pretrained model to initial the weight of our network. In our ParallelNet, backbone network is composed of multiple pathways, hence we load the pretrained model of ResNet-101 in all individual pathways.

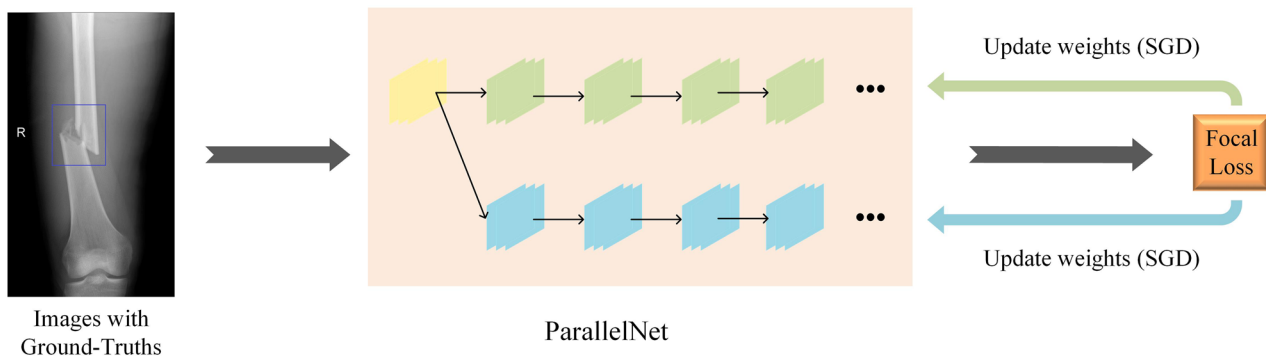
Training and testing pipeline are illustrated in Fig. 6. The input images with ground truths are fed into the network to update the weights in each backbone networks, respectively. The output trained model will be used in the testing procedure. In testing process, original images are sent into the model and output images with proposals. These proposals will calculate the IoU with ground truths. The result is evaluated with two metrics, AP50 and AP75. These two standards indicate different IoU between the predicted bounding boxes and the ground truths, if the IoU is over 0.5 in the test procedures, the prediction will be considered as a correct prediction in AP50, and if over 0.75 will be considered as a correct prediction in AP75.

Backbone network is the fundamental structure, it extracted learnable features from the original input images for the network. Previous work on thigh bone detection task, DCFPN [16] introduced dilated convolution in the backbone network and proved that enriching the reception field is conducive to improve the detection performance. However, we found the gridding effect of dilated convolution harms the performance of detecting small objects. As shown in Fig. 2, dilated convolution skipped some pixels during the calculation, this could be regarded as the normal convolution with holes, which means discrete dilated convolution sampling leads to lack of correlation in the long spatial distance sample. This indicates that, dilated convolution ignores vital information for detecting tiny objects while enriching the reception field. The larger dilation rates, the more pixels it skips. To overcome gridding effect, we designed a backbone network in three pathways with different dilation rates. Shown in Fig. 3, the first pathway is denoted as the main network, the second pathway is denoted as the sub-1 network and the last one is denoted as sub-2 network.

3.1 Implementation details

The method is exploited with Pytorch v1.1.0, CUDA 9.0 and cuDNN 7.1.4. and adopted with transfer learning of the backbone of ResNet-101 pretrained model. Data augmentation with a horizontal flip is used to enhance the quantities of our dataset. 4 NVIDIA GeForce GTX 1080TI GPU and each process 1 image with the learning rate of 0.005, the momentum of 0.8 and weight decay of 0.005, in 20 total epochs.

Training data



Testing data

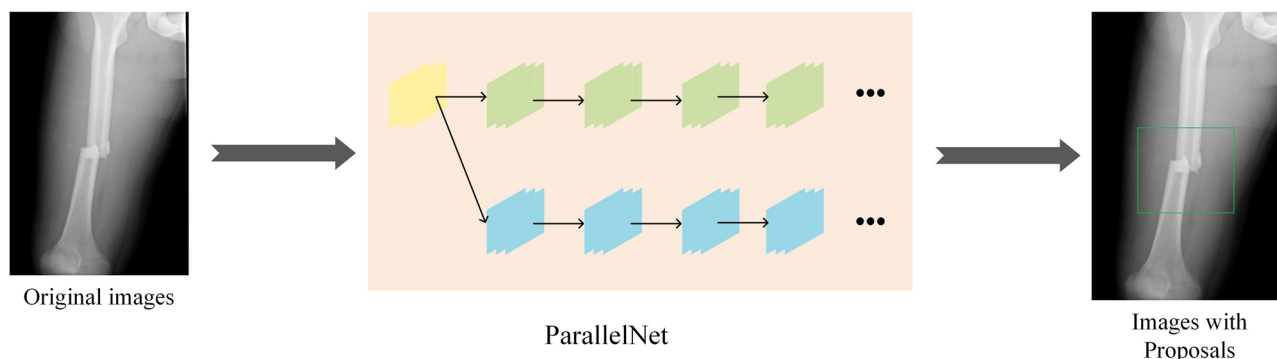


Fig. 6 Pipeline of training and testing the network

3.2 Results

The 3483 thigh fracture X-ray images are adopted in training. In test procedures, two metrics in the COCO style has been adopted to evaluate the results, including AP50 and AP75. We compared some of the state-of-art generic object detection methods with our network on the same thigh fracture dataset. To be fair, all of the networks are finetuned and chose the best detection results. Table 1 shows the result of different methods, including Faster R-CNN and feature pyramid network, etc. Previous work on thigh bone detection task, dilated convolution feature pyramid network (DCFPN) [16] reaches 82.1% AP50 and 45.1% AP75. It can be seen that our method is relatively higher in the final result, with the dilation rates of 2, 2, 1, 3 in stages 2–5, the results reach 87.8% AP50 and 49.3% AP75 which outperformed DCFPN by 5.7% AP50 and 4.2% AP70, and outperformed the next best AP50 by 2.5% (RetinaNet) and AP75 by 0.6% (Cascade R-CNN). Some samples of our detection results are illustrated in Fig. 7.

Table 1 Comparison of ParallelNet with other state-of-the-art methods

Method	Backbone	AP50 (%)	AP75 (%)
DCFPN	ResNet-101	82.1	45.1
Faster R-CNN	ResNet-50	78.1	46.3
FPN	ResNet-101	85.2	45.2
Cascade R-CNN	ResNet-50	85.0	48.7
FPN (guided anchoring)	ResNet-101	85.1	45.1
RetinaNet (with FPN)	ResNet-101	85.3	46.7
ParallelNet (ours)	TripleNet (ours)	87.8	49.3

3.3 Ablation experiments

3.3.1 Experiments on dilated convolution rates

Dilation rates are very vital hyperparameter for the network. We attempted several experiments on different dilation rates to evaluate the best baseline on the proposed network. Each backbone network is composed of five stages and shared a

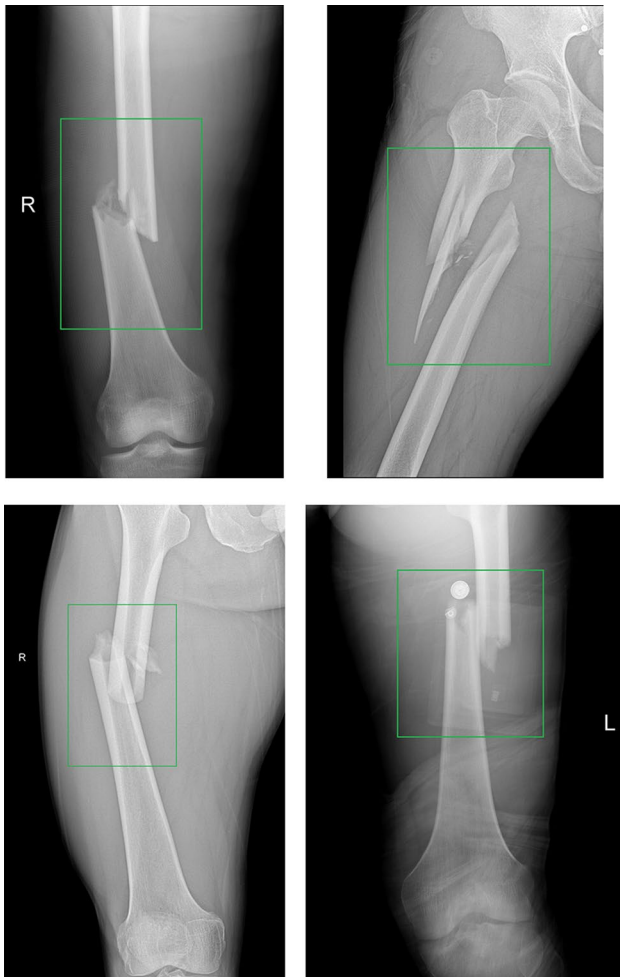


Fig. 7 Samples of detection results

frozen first stage, so dilation rates in stages 2–5 are denoted in a bracket with four numbers, each number representing the dilation rates in stages 2–5. This presentation of dilation rates is also applied in the following experiments.

In this experiment, the dilation rates of the main network are set as (1, 1, 1, 1) and different dilated rates are attempted in the sub-1 network and sub-2 network to evaluate the best setting of dilation rates. Table 2 presents the result of different dilation rates in each sub-network. In the experiments,

Table 2 Ablation experiments on dilation rates

	Sub-1 net Sub-2 net	AP50 (%)	AP75 (%)
1	(1, 1, 1, 1) (1, 1, 1, 1)	86.2	45.2
2	(1, 1, 1, 1) (2, 2, 2, 2)	87.3	47.6
3	(1, 1, 1, 1) (3, 3, 3, 3)	85.6	47.3
4	(2, 2, 2, 2) (1, 1, 1, 1)	86.0	46.9
5	(2, 2, 2, 2) (2, 2, 2, 2)	86.1	47.5
6	(2, 2, 2, 2) (3, 3, 3, 3)	85.4	46.1

we found that enlarging the dilation in sub-2 network can improved the detection accuracy, the results achieve 87.3% AP50 and 47.6% AP75 in experiment 2. However, it can also cause negative effect on the detection accuracy in some cases. Comparing experiment 2 and 3, the dilation rates of sub-2 network increased from (2, 2, 2, 2) to (3, 3, 3, 3) yet the result decreased significantly. Because larger dilation rates skipped more pixels, these skipped pixels contain vital information for detecting small objects. Also, irrelevant features, such as background information, are collected into the feature maps and harmed the final detection performance. Thus, we infer that the best setting of dilation rates is (1, 1, 1, 1) in sub-1 network and (2, 2, 2, 2) in sub-2 network.

3.3.2 Experiments on number of backbone pathways

In this section, we attempted dual backbone network and compared with the proposed triple backbone network. Dual backbone network composed of a main network and a sub-network. The dilation rate of the stages in the main network is frozen in (1, 1, 1, 1). Experiments on different dilation rates on individual sub-networks are demonstrated in Table 3. The results indicate that decreasing the backbone numbers greatly harmed the detection performance. It is because triple backbone network provides more learnable features due to their different dilation rates. This led the network to learn more complex features in the training process and contributes to the optimization of network parameters.

3.3.3 Experiments on different feature fusion connection structures

In this section, we compared backward connection structure with lateral connection structure which connected feature maps from the same stages of the backbones. Results are presented in Table 4. We test the performance of connection styles in two dilation rates baselines. The results indicate that backward connection outperformed lateral connection in both two baselines. The reason is that, in our ParallelNet, parameters used are three times than those in single backbone networks. When building lateral connection structures, features extracted from the same stage of backbones have strong similarity which causes parameter redundancy. This represents the network cannot learn more semantic

Table 3 Ablation experiments on number of backbone pathways

	Sub-1 net Sub-2 net	AP50 (%)	AP75 (%)
1	(1, 1, 1, 1) None	84.7	44.5
2	(2, 2, 2, 2) None	85.6	44.6
3	(1, 1, 1, 1) (1, 1, 1, 1)	86.2	45.2
4	(1, 1, 1, 1) (2, 2, 2, 2)	87.3	47.6

Table 4 Ablation experiments on number of backbone pathways

	Dilation rates		Connection style		Average precision	
	Sub-1 net	Sub-2 net	Lateral	Backward	AP50 (%)	AP75 (%)
1	(1, 1, 1, 1)	(1, 1, 1, 1)	√		85.1	44.2
2	(1, 1, 1, 1)	(1, 1, 1, 1)		√	86.2	45.2
3	(1, 1, 1, 1)	(2, 2, 2, 2)	√		84.6	44.8
4	(1, 1, 1, 1)	(2, 2, 2, 2)		√	87.3	47.6

Table 5 Ablation experiments on the employment of dilated convolution

Dilation rates			Average precision	
Main net	Sub-1 net	Sub-2 net	AP50 (%)	AP75 (%)
(1, 1, 1, 1)	(1, 1, 1, 1)	(2, 2, 2, 2)	87.3	47.3
(1, 1, 1, 1)	(2, 2, 2, 2)	(1, 1, 1, 1)	86.0	46.9
(2, 2, 2, 2)	(1, 1, 1, 1)	(1, 1, 1, 1)	81.3	38.9

information and parameters cannot fully utilized. Backward connection connected features from different semantic levels which enriched the complexity of features, parameters are fully trained and enhanced the detection accuracy.

3.3.4 The position of dilated convolution

In the experiments, we found that the position of dilated convolution is a vital hyperparameter to the final accuracy. The position could be regarded as which backbone is adopted the largest dilation rate. The following table 5 shows several experiments on the position of dilated convolution. All the following experiments are adopted with a triple backbone network and using a backward connection method. When employed dilation rates of (2, 2, 2, 2) in the sub-2 network, which is the final backbone network, the result achieves 87.3% in AP50. However, when applying the same dilation rates to the other two backbone networks, the result dropped significantly. The reason of this is because the latter backbone inherits the features extracted from previous one, if dilated convolution employed in previous backbones, gridding effects will lead to ignoring vital information of the small fractures and collect irrelevant information, this will further infect the performance of the detection accuracy. Thus, the last sub-network is the best employment position of dilated convolution (Table 5).

4 Conclusion

This work presented a novel network structure called ParallelNet to detect thigh bone fracture from X-ray images. ParallelNet is designed with several individual backbone pathways in the same shape with different dilated convolution

rates. A backward connection structure is proposed to connect each backbone networks to connect feature maps with different reception fields. Feature pyramid structure is adopted in the proposed network to detect fractures in large different scales. Furthermore, RPN is applied in the network to propose the candidate region, these candidate regions representing the specific fracture positions in the X-rays. Several experiments have been implemented based on the thigh bone fracture dataset established in DCFPN [16]. Specifically, the result achieved 87.8% AP50 and 49.3% AP75 which shows relatively higher results than the other state-of-the-art algorithms. Ablation experiments on the number of backbone networks, different types of feature fusion methods, different dilation rates in individual backbone networks and the employment of dilated convolution have been analyzed.

Acknowledgements The authors would like to thank Doctor Jinliang Wang and Fuzhou Li for their devotion in annotating and collecting the thigh fracture dataset; the radiologists in Department of Linyi People's Hospital for their devotion on evaluating the results of the experiments; Doctor Wanquan Liu from Curtin University for his pieces of advice on revising the paper. This work is under the support of the National Natural Science Foundation of China under Grants 62073237.

References

1. Esteva, A., Robicquet, A., Ramsundar, B., et al.: A guide to deep learning in healthcare. *Nat. Med.* **25**(1), 24–29 (2017)
2. Kooi, T., Litjens, G., Van Ginneken, B., et al.: Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017)
3. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017)
4. Liu, J., Wang, D., Lu, L., et al.: Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Med. Phys.* **44**, 4630–4642 (2017)
5. Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018). <https://doi.org/10.1016/j.cell.2018.02.010>
6. Drozdal, M., Chartrand, G., Vorontsov, E., et al.: Learning normalized inputs for iterative estimation in medical image segmentation. *Med. Image Anal.* **44**, 1–13 (2018)

7. Gibson, E., Li, W., Sudre, C., et al.: NiftyNet: a deep-learning platform for medical imaging. *Comput. Methods Progr. Biomed.* **158**, 113–122 (2018)
8. Wang, G., Zuluaga, M.A., Li, W., et al.: DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1559–1572 (2019)
9. Zhao, F., Chen, Y., Hou, Y., et al.: Segmentation of blood vessels using rule-based and machine-learning-based methods: a review. *Multimed. Syst.* **25**, 109–118 (2019)
10. Xia, K., Gu, X., Zhang, Y.: Oriented grouping-constrained spectral clustering for medical imaging segmentation. *Multimed. Syst.* **26**, 27–36 (2020)
11. Xi, X., Meng, X., Yang, L., et al.: Automated segmentation of choroidal neovascularization in optical coherence tomography images using multi-scale convolutional neural networks with structure prior. *Multimed. Syst.* **25**, 95–102 (2019)
12. Lindsey, R., Daluiski, A., Chopra, S., et al.: Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci.* **115**(45), 11591–11596 (2018)
13. Cheng, C., Ho, T., Lee, T., et al.: Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *Eur. Radiol.* **29**, 5469–5477 (2019)
14. Badgeley, M., Zech, J., Oakden-Rayner, L., et al.: Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2**, 1 (2019)
15. Gale, W., Oakden-Rayner, L., Carneiro, G., et al.: Producing radiologist-quality reports for interpretable deep learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging, pp. 1275–1279 (2019)
16. Guan, B., Yao, J., Zhang, G., Wang, X., et al.: Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network. *Pattern Recognit. Lett.* **125**, 521–526 (2019)
17. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
18. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517–6525 (2017)
19. Lin, T., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2999–3007 (2017)
20. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural. Inf. Process. Syst.* **39**, 1137–1149 (2017)
21. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 936–944 (2017)
22. He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
23. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
24. Wang, J., Chen, K., Yang, S., et al.: Region proposal by guided anchoring. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2960–2969 (2019)
25. Pang, J., Chen, K., Shi, J., et al.: Libra R-CNN: towards balanced learning for object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
26. Li, Y., Chen, Y., Wang, N., et al.: Scale-aware trident networks for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6054–6063 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.