Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Robust 3D Convolutional Neural Networks For Pulmonary Nodule Detection

Zhipeng Ding, Wenqiang Zhang and Kangzheng Gu
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
shanghai,chiana
{*zpding17,wqzhang&kzgu17*}*@fudan.edu.cn*

***Abstract*— Computer-aided Diagnosis(CADx) plays an important role in interpretation of medical images. In various kinds of medical images, Computed Tomography(CT) images are able to provide pivotal evidence for diagnosis of lung cancer. Especially, Early-stage lung cancer can be discovered by detecting small pulmonary nodules in patient's CT slices. Thus, it is worthwhile to develop pulmonary nodules detection algorithm. In this paper, we introduce a pulmonary nodules detection algorithm based on object detection algorithm, with three improved techniques that focus on characteristics of pulmonary nodules detection algorithm. First, we use 3D-convolutional layers to process CT images instead of 2D-convolutional layers. Second, We designed a powerful encoder to extract image features, which can train with a small batch for active learning or retraining by hospital with limited computing resources. Third, we propose filterRPN for filtering a large number of invalid anchors and accelerate the training process. Finally, we use some experiments to show that our model got the start-of-the-art results.**

***Index Terms*— 3D object detection, pulmonary nodules detection, filterRPN, data balance**

I. Introduction

Lung cancer is one of the most serious cancer in the world. High mortality of advanced lung cancer prompts medical researchers to develop methods to diagnose and treat lung cancer as early as possible. Computed tomography(CT) is one of the techniques that widely used to recognize the lung cancer. Early-stage lung cancer, even premalignant lesion, can be discovered by analyze patient's CT slices. In most cases, doctors use Computer-Aided Diagnosis(CADx) systems to help themselves detect pulmonary nodules fast and precisely. With the increasing use of CT, huge amount of CT slices data make applying deep learning methods on CT data processing become feasible. One of the methods to solve such image processing problems is called Convolutional Neural Networks(CNNs). In order to get better performance, we propose three improved techniques.

First, we designed a new 3D object detection algorithm and abandoned the traditional approach in 2D. Different from normal 2D images, CT slices contain full 3D information of a patient. Although regarding each slice as a single image is feasible, it is better to process the 3D slices at once. Traditional CADx systems detect pulmonary nodules based on some simple assumptions(e.g. nodules have some specific shapes), and check some simple low-level features[2] to decide if there are nodules in slices. Those methods are not accurate enough. So in this paper, we propose a 3D CNN framework to process 3D CT slices directly, inspired by[1][3][4].

Secondly, we design a module called Wide Bottleneck Block(abbreviated as WBB) that can be trained well by small batch size. Meanwhilethe encoder consisting of WBBs can generate adaptive features based on the size of the lung nodules. Since each hospital has a special data distribution, it is important to allow the hospital to train itself or to update the model online. Due to the limited computing resource of the hospital, they can only train the model by a small batch. It is well known that the scale of training batch affects training efficiency significantly. In general, if the batch is very small like 1 or 2, the model will be severely degraded. However, bigger batch size means that the model would cost large amount of computational resources. For 2D images, setting batch size to 16 or even 32 is acceptable. But when things come to 3D images, using such big batch size is not realistic, especially for hospitals. In addition, the encoder adopts a special supervision mechanism, which can independently select the size of the receptive field according to the size of the lung nodules, thereby increasing the recall rate of the pulmonary nodules. We will show our solution in section III. And in section IV, we will show the advantage of our model.

Thirdly, in RPN[5], there is no intersection between most of the anchors and ground truth. In 2D, The number of anchors is relatively small compared to 3D. A large number of invalid calculation points will seriously slow down the training and inference speed. We propose filterRPN module for filtering a lot of invalid anchors and accelerate the training process.

In the following part, we first introduce some previous works relevant to our research topic. Then, we propose our methods to overcome some shortcomings of general detection architectures on pulmonary detection problem. Finally, we arrange some experiments to show our design is effective and powerful. Our model gets the start-of-the-art results on

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

LUNA16[8] dataset.

II. RELATED WORK

A. Classic Object Detection Methods

The dominant object detection paradigm is two stage. The pioneered work source from RCNN[7] which is based on the Selective Search work[16], and Fast RCNN extend the RCNN work by updating the second-stage classier to convolutional network yielding large gains in accuracy and ushering in the modern era of object detection. In [5], the paper replace the Selective Search with a subconvolution network (simplied as RPN) to generalize region proposals and the architecture is faster and more accurate than Fast-Rcnn. MaskRcnn[17] add a subnetwork to the backbone network to predict the object mask and replace the ROI pooling layer with ROI align layer, performance of which is excellent. Later,numerous works have extended the architecture,such as [18,19,20,21,22,23]. In [6], the work exploited two-stage Faster-RCNN architecture, the backbone network of which is VGG5 and followed by a deconvolution layer. After that they connected two modules on the feature extractor: a region proposal network(RPN) aim to propose potential regions of nodules (also called ROI); a ROI classier to recognize whether ROIs are nodule or not. According to [3], The first weak point is that after four times downsampling, they upsample the feature maps by one deconvolution layer to recover enough size to propose regions, but they lose spatial information which would offset RPN network. The second defect is that for each axial slice in CT slices, the only concatenate its two neighboring slices in axial direction, and then rescale it into 600*600*3 pixels. In [6] and our work, the slice thickness is resampled into 1mm. So if the nodules is 3mm or smaller than 3mm in axial direction, the above architecture can ”see” all information, but the variance of nodules is very high(such as greater than 10mm nodules in axial direction) and the nodule information is lost a lot. Moreover, they only concatenate 3 slices together and don’t consider whether each of slices include a nodule, which will generate a lot of negatives and false positives. So they build another DCNN network to reduce false positives. Our input is 128*128*128 cubes and each cube include at least one pulmonary nodule which can contain enough information in axial information. And we exploit ensemble methods in network and build encoderdecoder architecture in which the decoder’s spatial information was supplemented from encoder by lateral connection inspired by [18,3,4] and have got excellent performance

III. METHODS

A. 3D object detection Neural Networks

In pulmonary nodules detection problem, there is no need to require the method to achieve real-time detection. In contrast, we would like to see that the algorithm detects pulmonary nodules more precisely even if it costs longer time. So, we follow the two-stage framework to construct our model. Since the LUNA16[8] dataset which does not include the categories of pulmonary nodules, we do not need to classify pulmonary nodules. As a result, after Region Proposal Networks(RPNs) in [5], the model does not have a classification branch.

Our model is shown in the figure 1. It takes a $128 * 128 * 128$ 3D grid from the rescaled CT slices as input. Then we use several convolution blocks, like ResBlock[27] or Wide group shuffle Bottleneck Block(WBB) proposed by us, to extract features from input data. And in order to make use of high-level semantic information, we learn from U-Net[28], linkNet[3]and design a pyramid structure to combine the high-level and low-level feature maps together in concatenation manner. Finally, our model outputs a 15-channel $32 * 32 * 32$ feature map, containing information of three types of anchors. Each anchor includes the coordinates, diameter and probability of a nodule. All the convolutional and transposed convolutional layers are 3D ones. The RPN contains the filterRPN and prediction branches.

B. Wide Bottleneck Block of the encoder

Due to the use of 3D convolutional layers, the model needs more memory to save parameters and calculate intermediate results. Thus, large batch size would no longer be affordable. Meanwhile, with special data distribution and limited computing resources, it is very important for the hospital to retrain the model. However, if the model is trained with very small batch size like one or two, the final true positive rate will be lower than models trained by larger batches. In order to overcome this problem, we design a Wide Bottleneck Block(WBB) as shown in figure 2.

In the WBB,in order to prevent feature collapse due to downsampling and to preserve effective features in small batches of data, feature maps will first pass a convolutional layer to expand the channel size for dispersing feature into high dimensional space.

$$\alpha = \frac{out}{\gamma} \tag{1}$$

$$T = \alpha BG \tag{2}$$

T represents the dimension after transferring and the γ represents widen factor. B is the basic group filters and G is group number.

According to Inception Networks[29,30], using multiple branches in neural networks could avoid bottlenecks of the information flow. And such structures could increase the diversity of gradient(different gradients’ scales and directions for convolution kernels in different branches). Besides, we discover that introducing such branches in CNNs could significantly enhance the model’s robustness of batch size. Moreover, several group branches also reduce the amount of extra parameters brought about by expanding dimension. And

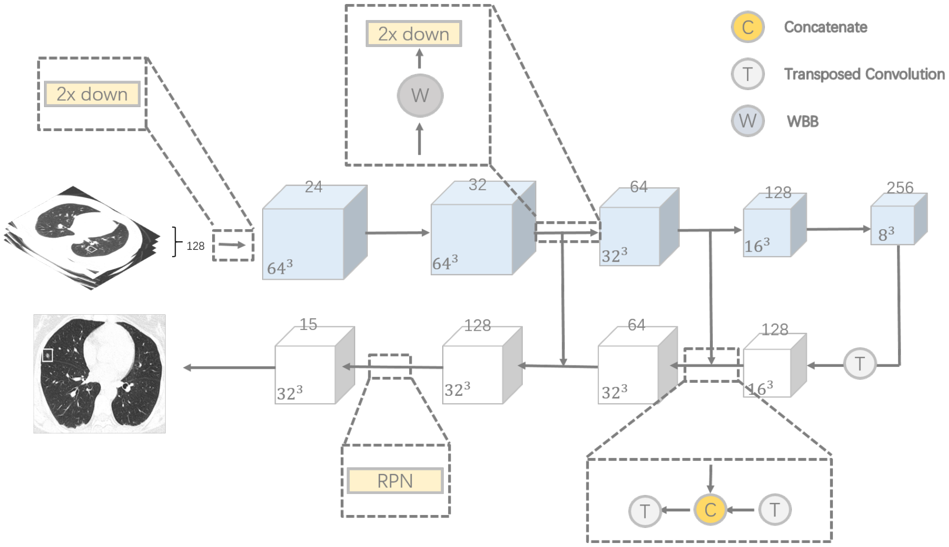
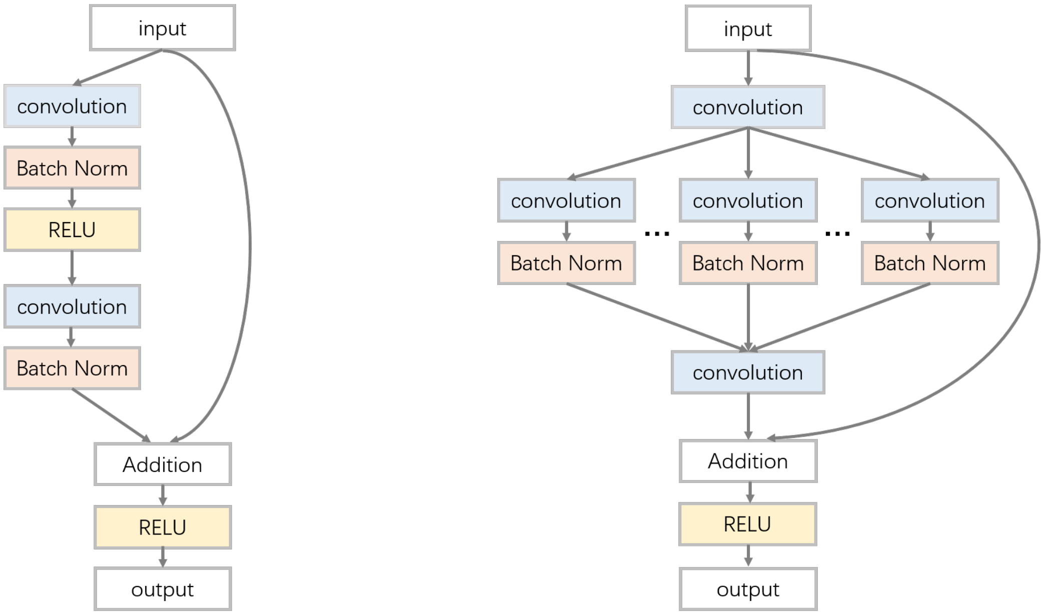


Fig. 1. **Network architecture.** Letter 'W' with a circle represents our Wide Bottleneck Block, which will be introduced in the next section. Letter 'C' with a circle represents concatenating layers. And letter 'T' with a circle represents a transposed convolutional layers used for upsampling.



(a) ResBlock

(b) Wide Bottleneck Block

Fig. 2. **ResBlock and Wide Bottleneck Block(WBB).** The left one is ResBlock[**?**] and the right one is Wide Bottleneck Block. Wide Bottleneck Block increases the bottleneck of information flow.

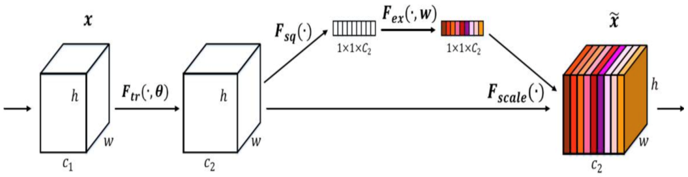


Fig. 3. **attention layer.** The attention layer[] for automatically selecting adaptive receptive field feature.

every branch has different dilation rate for capturing adaptive receptive field feature. We also split the branches into two groups. The one use 1×1 dilation rate and the other use 3×3 dilation rate. Since the batch is very small and the batch may only contain some special size pulmonary nodules, we have adopted extra attention layer[31] for automatically selecting adaptive receptive field feature. Also, we use a shortcut path to add inputs to outputs, like ResBlock[27].

C. filterRPN of the RPN

In liao[4]'s RPN branches, the total number of anchors that need to be calculated reach 100,000, which will slow down the speed of training and inference. Moreover, most of the anchors have no intersection with ground truth. So we add one filter branch to filter these invalid anchors. Through the branch of filterRPN, the probability of each position containing the pulmonary nodule is obtained. If the probability that a certain position exceeds the certain threshold, the corresponding anchors of the position are calculated. As a result, the training and inference process is accelerated by removing a lot of invalid calculations. The groundTruth is a binary cube in which 1 reprensents the position is located in pulmonary nodule. We adopted focal loss[7] to solve the imbalance between the positives and negatives.

The encoder and decoder can be seen in tableI. As for coordinate regression, we simply use the Smooth L1 loss[32] and coordinates are encoded for parameter learning.

$$d_x = \frac{G_x - A_x}{A_r} \tag{3}$$

$$d_y = \frac{G_y - A_y}{A_r} \tag{4}$$

$$d_z = \frac{G_z - A_z}{A_r} \tag{5}$$

$$d_r = \log \frac{G_r}{A_r} \tag{6}$$

$$Loss = L_{reg} + \lambda L_{cls} + \gamma L_{filterRPN} \tag{7}$$

$$L_{reg} = \begin{cases} 0.5x^2 & if \quad |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{8}$$

IV. Experiments

A. Datasets

Our experiments are based on Lung Nodule Analysis 2016[8], referred to as LUNA16, and Data Science Bowl 2017,referred to as DSB2017. LUNA16 includes 888 patients and 1188 pulmonary nodules. Each nodule is represented by its coordinates and diameter. The original DSB2017 is a lung cancer detection dataset. So it has no information of nodules. Previous work[4] supplied the manual labels of nodules of DSB2017. Thus we use DSB2017 as a supplementary to train the model.

In pulmonary nodule detection, we hope that we could find out all nodules without missing any of them, even with quantities of false positive samples. So, the sensitivity(Eq. 9) and accuracy(Eq.10) are commonly used to evaluate the performance of models.

TABLE I

The component of encoder

stage	output	encoder	
preblock	$64 \times 64 \times 64$	$3 \times 3 \times 3$,24 depthwise conv,s2	
stage1	$64 \times 64 \times 64$	1×1 D 3×3 $group\&dilate$ 1×1 32	$\times 2$
stage1_2	$64 \times 64 \times 64$	attention layer	
stage2	$32 \times 32 \times 32$	1×1 D 3×3 $group\&dilate$ 1×1 64	$\times 2$
stage2_3	$32 \times 32 \times 32$	attention layer	
stage3	$16 \times 16 \times 16$	1×1 D 3×3 $group\&dilate$ 1×1 128	$\times 3$
stage3_4	$16 \times 16 \times 16$	attention layer	
stage4	$8 \times 8 \times 8$	1×1 D 3×3 $group\&dilate$ 1×1 256	$\times 3$
stage4_deconv	$8 \times 8 \times 8$	attention layer	

TABLE II

The component of decoder

stage	output	decoder	
deconv_stage5	$8 \times 8 \times 8$	$2 \times 2 \times 2$,64 deconv depthwise,s2	
deconv_stage4	$16 \times 16 \times 16$	1×1 D 3×3 $group\&dilate$ 1×1 64	$\times 2$
deconv_stage3	$32 \times 32 \times 32$	1×1 D 3×3 $group\&dilate$ 1×1 64	$\times 2$
RPN	$32 \times 32 \times 32$	3×3 64 1×1 $1\&filterRPN$ 1×1 15	$\times 1$

$$sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{9}$$

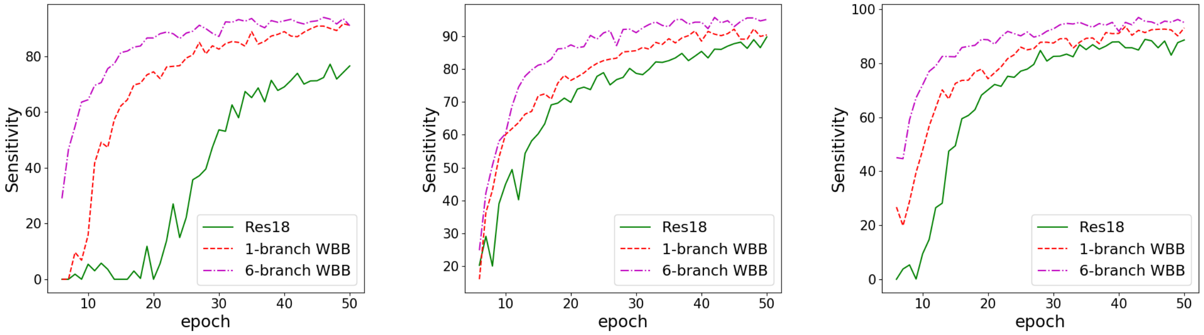
$$accuracy = \frac{TruePositives + TrueNegatives}{DataSamples} \tag{10}$$

B. Branch Size and Batch Size

First, we would like to show the advantage of our Wide Bottleneck Block. The model of [4] learns slowly under the really small batch size like 1. In opposite, our model could be trained effectively on small batches. And Wide Bottleneck Block with multiple branches makes the model training increasingly faster. Figure 4 shows the training process.

Then, we conduct several experiments to find the best branch size of Wide Bottleneck Block. Table IV-B shows the results. In this experiment, we set batch size as 4.

According to the results, increasing the branch size could improve the performance obviously. However, using too large branch size would not get better performance because of the



(a) $BatchSize = 1$

(b) $BatchSize = 2$

(c) $BatchSize = 4$

Fig. 4. **Sensitivity during Training.** The first one is the training process of batch size 1, while the second and third ones are 2 and 4. The three types of lines represent different models.

TABLE III

RESULTS OF DIFFERENT BRANCH SIZE

branch size	SEN	ACC
1	0.945	0.956
2	0.949	0.960
4	0.959	0.966
6	**0.972**	**0.975**
8	0.951	0.963

TABLE IV

RESULTS OF DIFFERENT LOSS FUNCTIONS

loss function	param	SEN	TNR	ACC
BCE	-	0.919	0.928	0.926
WFL	$\alpha = 0.51$	0.942	0.934	0.936
	$\alpha = 0.52$	**0.953**	**0.945**	**0.950**
	$\alpha = 0.53$	0.963	0.933	0.941
	$\alpha = 0.54$	0.977	0.913	0.930
	$\alpha = 0.55$	0.983	0.891	0.915

training difficulty. With our experiment, the best branch size is 6, which would be used in our best model.

C. filterRPN Focal Loss

We compared the performance of different loss functions, including Binary Cross Entropy (BCE) and our filerRPN weighted Focal Loss (WFL). Table IV shows the results.

The results indicate that bigger α we use, higher sensitivity we will get. This reflects the advantage of our Weighted Focal Loss to pay more attention to positive samples. If we consider sensitivity, TNR and accuracy synthetically, Weighted Focal Loss with $\alpha = 0.52$ is the best choice. It provides high sensitivity and the best TNR 94.5% and accuracy 95.0%, which is evidently better than BCE loss and original Focal Loss.

TABLE V

RESULTS OF DIFFERENT METHODS

Method	SEN	ACC
ISICAD[6]	0.856	-
SubsolidCAD[6]	0.361	-
LargeCAD[6]	0.318	-
M5L[6]	0.768	-
ETROCAD[6]	0.929	-
FasterRcnnCAD[6]	0.946	-
Res18CAD	0.965	0.967
ours	**0.972**	**0.975**

D. Compared with other methods

Table IV-D shows the performance of different methods. Row 'ours' represents the performance of our best model. And some methods do not provide their accuracy. Compared to all previous methods, our method achieves the state-of-the-art results, whose sensitivity is 97.2% and accuracy is 97.5%.

TABLE VI

THE SENSITIVE OF DIFFERENT FPS/SCAN

FPs/scan	ours
0.125	0.611
0.25	0.730
0.5	0.853
1	0.885
2	0.934
4	0.951
8	0.972

V. CONCLUSION

In this paper, we analyze the difficulties in pulmonary nodules detection and propose some improved techniques to overcome them. First, we use 3D convolutional encoder-decoder object detection model to extract features from CT slices. Second, we design a wide bottleneck block to increase the robustness of training by small batches and as the nodule size changes, the encoder can get adaptive feature maps. Third, we propose the filterRPN for removing invalid anchors and accelerating the process of training and inference. Next, our experiments indicate that our model is effective and robust. Finally, we get the state-of-the-art results on LUNA16 dataset. Furthermore, we are going to expand our model to more tasks on CT slices in the future, such as pulmonary nodules classification and cancer recognition.

Acknowledgment

We would like to show our gratitude to researchers from PET Center of Huashan Hospital and Changhai Hospital for their suggestions and feedbacks.

References

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[2] Lopez Torres, E., Fiorina, E., Pennazio, F., Peroni, C., Saletta, M., Camarlinghi, N., ... & Cerello, P. (2015). Large scale validation of the M5L lung CAD on heterogeneous CT datasets. Medical physics, 42(4), 1477-1489.

[3] Chaurasia, A., & Culurciello, E. (2017, December). Linknet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE Visual Communications and Image Processing (VCIP) (pp. 1-4). IEEE.

[4] Liao, F., Liang, M., Li, Z., Hu, X., & Song, S. (2019). Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-or Network. IEEE transactions on neural networks and learning systems.

[5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[6] Ding, Jia, et al. ”Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017.

[7] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[8] Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., van den Bogaard, C., Cerello, P., ... & van der Gugten, R. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical image analysis, 42, 1-13.

[9] Sermanet, Pierre, et al. ”Overfeat: Integrated recognition, localization and detection using convolutional networks.” arXiv preprint arXiv:1312.6229 (2013). Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

[10] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

[11] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. CVPR (1), 1(511-518), 3.

[12] Vaillant, R., Monrocq, C., & Le Cun, Y. (1994). Original approach for the localisation of objects in images. IEE Proceedings-Vision, Image and Signal Processing, 141(4), 245-250.

[13] elzenszwalb, P. F., Girshick, R. B., & McAllester, D. (2010, June). Cascade object detection with deformable part models. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 2241-2248). IEEE.

[14] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[16] Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

[17] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[18] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[19] Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 761-769).

[20] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., ... & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7310-7311).

[21] Shrivastava, A., Sukthankar, R., Malik, J., & Gupta, A. (2016). Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851.

[22] Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).

[23] Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-NMS–Improving Object Detection With One Line of Code. In Proceedings of the IEEE international conference on computer vision (pp. 5561-5569).

[24] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

[25] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[26] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Springer, Cham.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[28] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[29] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[30] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

[31] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[32] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).