VOLUME R-CNN: UNIFIED FRAMEWORK FOR CT OBJECT DETECTION AND INSTANCE SEGMENTATION

Yun Chen[1,2], *Junxuan Chen*[1] , *Bo Xiao*[2], *Zhengfang Wu*[2], *Ying Chi*[2], *Xuansong Xie*[2], *Xiansheng Hua*[2]

[1] Alibaba Group, Hangzhou, China
[2] Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

As a fundamental task in computer vision, object detection methods for the 2D image such as Faster R-CNN and SSD can be efficiently trained end-to-end. However, current methods for volumetric data like computed tomography (CT) usually contain two steps to do region proposal and classification separately. In this work, we present a unified framework called Volume R-CNN for object detection in volumetric data. Volume R-CNN is an end-to-end method that could perform region proposal, classification and instance segmentation all in one model, which dramatically reduces computational overhead and parameter numbers. These tasks are joined using a key component named RoIAlign3D that extracts features of RoIs smoothly and works superiorly well for small objects in the 3D image. To the best of our knowledge, Volume R-CNN is the first common end-to-end framework for both object detection and instance segmentation in CT. Without bells and whistles, our single model achieves remarkable results in LUNA16. Ablation experiments are conducted to analyze the effectiveness of our method.

Index Terms— object detection, computed tomography (CT), LUNA16.

1. INTRODUCTION

Different from the 2D image field, it is very challenging to fulfill the task of detection on CT due to its characteristic. The target in CT is much tinier than normal objects and it needs several experienced radiologists each spending tens of minutes to draw a convincing conclusion, which makes the CT annotation precious and rare. With tiny target, lack of data and high data dimension, the research on CT is easy to fail due to overfitting, especially when no pretrained models are available because of either commercial confidentiality or diverse data distribution.

Traditional CT diagnosis usually involves hand-designed features or descriptors requiring domain expertise [1, 2]. After the large-scale LIDC-IDRI [3] and LUNA16 [4] dataset became publicly available, deep learning-based methods have become the dominant framework for nodule research. Current leading methods for CT detection mainly contain two separate steps: propose candidates first and then perform false positive reduction on these candidates with a 3D convolutional neural network (CNN). [5] first established a 3D fully convolutional network (FCN) to screen the candidates from volumetric CT scans, and then a 3D ConvNet classification network is designed to move the false positive candidates [5]. [6] improved the first stage by introducing 2D RPN to extract proposals in individual 2D images then combine them to generate 3D proposals [6]. However, these methods are inefficient for both training and inference because candidate proposal and false positive reduction are performed in two separate steps. Worse still, they require sophisticated processing pipeline within the two steps, leading to low efficiency.

We address that candidate proposal and false positive reduction could be joined using RoI Pool methods to reduce the number of parameters and computational overhead by sharing convolutional feature maps. We further add mask prediction support by introducing a light mask head. The whole system is named Volume R-CNN (see Figure 1), which generate R-CNN family [7, 8, 9, 10] to 3D CT. In contrast to previous works that rely heavily on handcraft features, specialized knowledge or require complex multi-stage processing, Volume R-CNN is an end-to-end framework could perform object detection and instance segmentation simultaneously and efficiently. To the best of our knowledge, it is the first unified and common framework for object detection and instance segmentation in volume. We expected the proposed method could be applied to a wide range of volumetric data and serve as a meta-algorithm for further research in volumetric data. Experimental results have confirmed the effectiveness of our methods. Without bells and whistles, our method could gain competitive results in LUNA16 directly with one single model. Ablation experiments are conducted to investigate the behavior of Volume R-CNN, especially the key component RoIAlign3D.

2. METHODS

The proposed method consists of five components, as illustrated in Figure 1. The input is cuboid of size $D \times H \times W$, depth, height, width along the Z, Y, X axes respectively. The backbone is a 3D U-[11] extracting features of CT, from

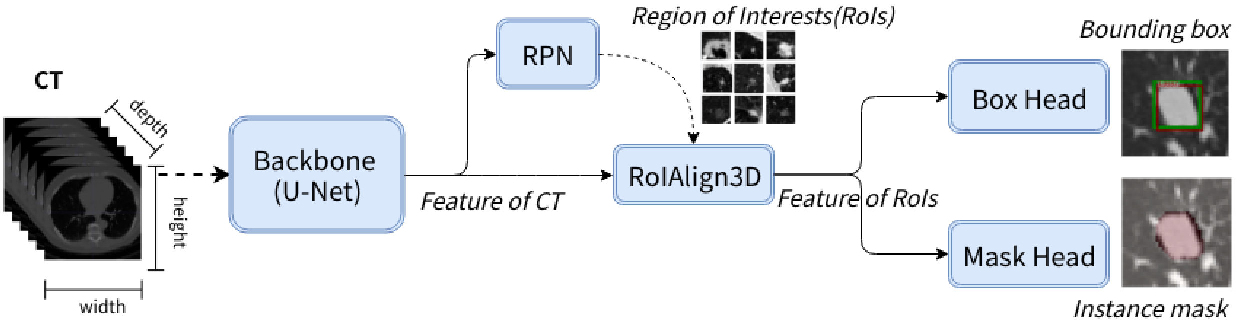

Fig. 1. The Volume R-CNN framework for object detection and instance segmentation. RoIs, bounding box and mask are all in 3D space, simplified for visualization herein. Loss from RPN, box head and mask head sum as final train objective. RoIs are seen as input data and the dotted line means no gradient during backward. RoIAlign are they key operation that joins other 4 modules and accelerates the whole process by directly extracting feature of RoIs on the feature map of CT.

which Region Proposal Network (RPN) proposes candidate bounding boxes called region of interests (RoI) — cuboid boxes of different shapes on different locations. The feature of RoIs is extracted using RoIAlign 3D — an efficient module that converts the features inside any valid RoIs with different size into a small feature map with a fixed spatial size. The feature of RoIs is further sent to two relatively independent head to parallelly predict bounding box (*Box Head*) and instance segmentation mask (*Mask Head*) for the target.

2.1. Region Proposal Network (RPN)

A Region Proposal Network (RPN) outputs a set of cuboid object proposals, each with a confidence score. This process is modeled with a fully convolutional network. For training RPNs, a binary class label (of being an object or not) is assigned to each anchor. We assign a positive label to two kinds of anchors: (i) anchors with the highest Intersection-over-Union (IoU) overlap with a ground-truth box, or (ii) anchors with an IoU overlap higher than 0.5 with any ground-truth box. We assign a negative label to a non-positive anchor if its IoU ratio is lower than 0.02 for all ground-truth boxes. Anchors that are neither positive nor negative do not contribute to the training objective. Only one positive anchor is randomly chosen as the target, and the others do not contribute to the training objective. There are much more negative anchors than positive ones. Hard negative mining [12] is used to deal with this problem. The N negative samples with highest classification confidence scores are selected as the hard negatives. The others are discarded and not included in the computation of loss. We adopt $N = 2$ in our experiments.

RPN produces a prediction for each anchor. Then a non-maximum suppression (NMS) operation with an IoU thresh of 0.2 is performed to rule out the overlapping proposals. The selected location-refined anchors are called Region of Interests (RoI). RoIs are seen as input data to be sent to RoIAlign, and gradient does not backward through them.

2.2. RoIAlign 3D

The RoIAlign 3D operation uses trilinear interpolation [13] to convert the features inside any valid RoIs into a small feature map with a fixed spatial extent of (oD, oH, oW) (e.g., $4\times 4\times 4$), where oD, oH and oW are layer hyper-parameters that are independent of any particular RoI. Each RoI is defined by a six-tuple (z, y, x, d, h, w) that specifies its center coordinates and shape.

RoIAlign brings better forward output and backward gradient, because of the way to computing the feature map in roi bins. For each target voxel in the bin, 8 nearest voxels of the feature map are used to calculate the interpolated value, while for RoIPool, only one voxel is selected after comparison, which results in the bottleneck of gradient backward. For an intuitive understanding of the strength of RoIAlign, we conduct simple experiments and show the results in Figure 2. The output of RoIAlign is much clearer than RoIPool under the same resolution. Also, the gradient of RoIPool tends to be noisy and fuzzy, while the gradient of RoIAlign is much more smooth, balanced and well-proportioned, which indicates that the RoIAlign has superior performance in both forward and backward period. RoIAlign leads to considerable improvements for both box and mask prediction which will be elucidated in the experiments.

2.3. Box Head and Mask Head

The RPN emits region proposals without category and the main purpose of the box head is to predict the categories of given RoIs and refine the RoIs to give more accurate location and shape prediction. We take 8 RoIs from region proposals that have IoU with a ground-truth bounding box of at least 0.3. These RoIs comprise the examples labeled with a foreground object class. 24 RoIs are sampled that have a maximum IoU with ground-truth in the interval $[0.0, 0.001)$, following [9]. These are the background examples and are labeled with 0. The sampled RoIs are also used as training target in mask

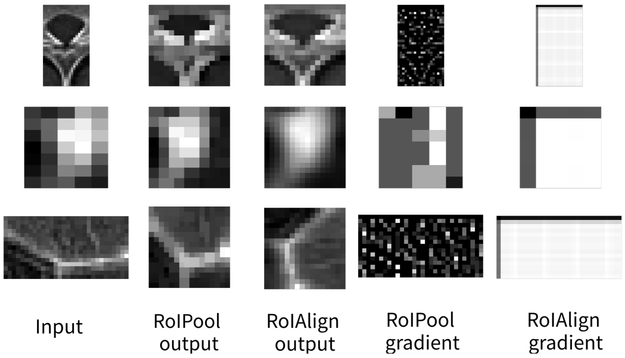

Fig. 2. RoIAlign *vs* RoIPool. RoIAlign give better forward output and the gradient is more balanced and well-propotioned in backward. Origin results are 3D cube, center slice is adopted for easy visualization and better understanding.

head.

Mask head gives mask prediction for every RoI. A similar strategy is used for mask representation and training objective as mask R-CNN [10] except that Volume R-CNN works in the 3D space. For every RoI, mask head gives a mask of $m \times m \times m$, double size of the RoI feature. The ground-truth mask within the bounding box is resampled to the same size. The prediction procedure is addressed naturally by the pixel-to-pixel correspondence provided by convolutions. Specifically, the mask from each RoI is predicted using an FCN. This allows each layer in the mask branch to maintain the explicit object spatial layout without collapsing it into a vector representation that lacks spatial dimensions.

3. EXPERIMENTS AND RESULTS

3.1. Main experiments on LUNA16

We perform a thorough comparison of Volume R-CNN to other methods along with detail ablation experiments. LUNA16 [4] is adopted in the experiments for comparison and analysis. LUNA16 contains 888 chest CT scans and 1186 pulmonary nodules. Each scan, with a slice size of 512×512 voxels, around 0.6 $mm/voxel$, and was annotated during a two-phase procedure by four experienced radiologists. Participants are required to perform 10-fold cross-validation when they use the provided data both as training and as test data. Results are evaluated using the Free-Response Receiver Operating Characteristic (FROC) analysis [14] which is defined as the average of the sensitivity at seven predefined false positive rates: $1/8, 1/4, 1/2, 1, 2, 4$, and 8 FPs per scan.

The result on box head is used as the final prediction but mask head is also included in the training objective. In Table 1, we compare Volume R-CNN to other methods reported in the official conclusion [4] of LUNA16 and some claimed leading results in the website (https://luna16.grand-challenge.org/results/). For those publicly available methods, our method gets very competitive results with one single model without bells and whistles. It is notable that other leading methods on the website do not offer the detailed description due to commercial confidentiality and intellectual property, and it may not be a fair comparison.

Table 1. Results in LUNA16

method	FROC
Our Single model (Volume R-CNN)	0.884
DeepLung [15]	0.842
3D FCN+CNN [5]	0.839
2D R-CNN+3D CNN [6]	0.891
2D SSD [16]	0.649
PAtech	0.951
JianpeiCAD	0.950
iFLYTEK-MIG	0.941
iDST-VC	0.897
AIDENce	0.807

Volume R-CNN outputs are visualized in Figure 3. The left is a nodule in the CT visualized in the 3D view. Detecting such a target is as hard as looking for a needle in a sea. We crop the CT at a side length of 36 with the nodule in the center for further visualization. It is notable that each mask is annotated by 4 radiologists and they are averaged as the ground-truth mask. The last row shows some unsatisfying results. One is missed while the other is the false positive. But it is also found that for the false positive, the mask prediction could work as a false positive reduction by stay inactivated on the false positive bounding boxes. But we do not use mask prediction to refine the box results in this work.

3.2. Ablation Experiments

We run several ablations experiments to analyze Volume R-CNN. Results are shown in Table 2 and discussed in detail next. We use 10-fold validation for the fair comparison with other methods in the previous subsection, however, in the ablation experiments, the results are compared within our method, so we used subset 0 as the test set and train on subset 1–9 to accelerate experiments.

As can be inferred from Table 2, box head (*Box* vs. *RPN*) and mask head (*Mask* vs. *Box*) can both give a promotion to the performance, compared to RPN. This can be interpreted

Table 2. Ablation Comparison

method	resolution	Box	Mask	RoI Layer	FROC
RPN 3D	1 mm	-	-	-	0.870
Box with Pool	1 mm	✓	-	RoIPool	0.875
Box with Align	1 mm	✓	-	RoIAlign	0.891
Mask with Pool	1 mm	✓	✓	RoIPool	0.880
Mask With Align	1 mm	✓	✓	RoIAlign	0.905
Box in 0.5 mm	0.5 mm	✓	-	RoIAlign	**0.915**

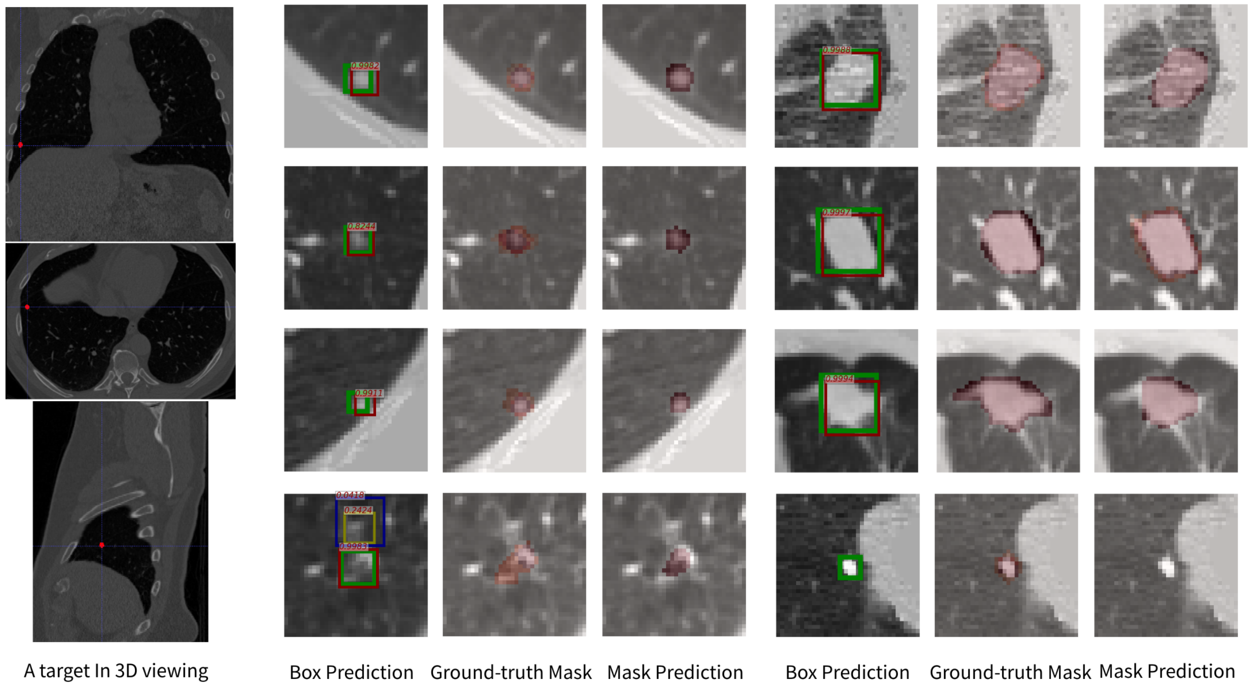

Fig. 3. Selected results of Volume R-CNN on LUNA16. Left shows a nodule mask in 3D view (better viewed in color) and others are results of detection. Origin results are the 3D cube. For easy visualization and better understanding, the target is cropped in the center with a side length of 36 voxels and the center slice is visualized. The ground-truth bounding box is drawn with a green box with a thicker edge without probability. The last row shows some unsatisfying results.

as that box head adds another procedure of classifying to give a more accurate prediction (the same as false positive reduction). While the mask head mainly benefits from more information (mask data) added to guide the training procedure.

RoIAlign gives significant improvement to the model, which can be seen obviously from the comparison of (*Box with Pool* vs. *Box with Align* and *Mask with Pool* vs. *Mask with Align*). Another proof is that *Box With RoIAlign* outperform *Mask with RoIPool*, which demonstrate that even with more data (mask data), RoIPool could not fully utilize them as RoIAlign.

Data resolution has a great impact on the performance (0.891 *vs.* 0.915), which seems straightforward. The original data resolution of CT is around 0.6 $mm/voxel$ and after resampled to 1 mm, some information is inevitably to lose, which has a crucial impact on the small targets. For example, a nodule with a diameter of 5 mm would only occupy less than 125 voxels (cube with side length of 5 pixels) with resolution 1 $mm/voxels$. If it is resampled to 0.5 $mm/voxel$, it would be a cube with side length 10, occupying around 1000 voxels. Even though higher data resolution gives a great promotion, it is not used online since it greatly slows down the processing (around 8x slower).

4. CONCLUSION

Most of the existing methods in volumetric CT detection require hand-crafted feature, multi-step processing or are confined to specific data. We novelly propose a unified detection framework named Volume R-CNN that joins region proposal, classification and instance segmentation using RoIAlign. It dramatically reduces computational overhead and number of parameter and could be trained end-to-end. Without bells and whistles, our single model gains competitive results on LUNA16. Ablation experiments have been conducted to detailedly analyze the effectiveness of our method.

5. REFERENCES

[1] Ayman El-Baz, Matthew Nitzken, Fahmi Khalifa, Ahmed Elnakib, Georgy Gimel'farb, Robert Falk, and Mohammed Abo El-Ghar, "3d shape analysis for early diagnosis of malignant lung nodules," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2011, pp. 772–783.

[2] Keelin Murphy, Bram van Ginneken, Arnold MR Schilham, BJ De Hoop, HA Gietema, and Mathias Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features

and k-nearest-neighbour classification," *Medical image analysis*, vol. 13, no. 5, pp. 757–770, 2009.

[3] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[4] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[5] Qi Dou, Hao Chen, Yueming Jin, Huangjing Lin, Jing Qin, and Pheng-Ann Heng, "Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 630–638.

[6] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 559–567.

[7] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[8] Leonard Berlin, "Faster reporting speed and interpretation errors: Conjecture, evidence, and malpractice implications," *Journal of the American College of Radiology*, vol. 12, no. 9, pp. 894–896, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[13] Paul Bourke, "Interpolation methods," *Miscellaneous: projection, modelling, rendering.*, , no. 1, 1999.

[14] International Commission on Radiation Units and Measurements, "Receiver operating characteristic analysis in medical imaging," *ICRU Report n 79*, p. 79, 2008.

[15] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and c lassification," *arXiv preprint arXiv:1801.09555*, 2018.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.