Trans-QWin-Bag: A Transfer Learning and Quantile Window Bagging Assisted Object Detection Framework for Chest X-ray (CXR) Nodule Detection

Di Xu [a, b], Fabien Scalzo [c] and Ke Sheng [b, *]

[a]Computer Science, University of California, Los Angeles, CA 90035, USA

[b]Radiation Oncology, University of California, Los Angeles, CA 90035, USA

[c]Computer Science, Pepperdine University, 24255 Pacific Coast Hwy, Los Angeles, CA 90263, USA

Di Xu: dixu@mednet.ucla.edu

Ke Sheng: KSheng@mednet.ucla.edu

1. Modelling Pipeline

As shown in **Figure 1**, the complete design of our Trans-QWin-Bag workflow includes three stages, the initial pretraining on DeepLesion[1] (full body tumors on CT) and Luna16 [2] (lung tumors on CT) datasets, the further sub-model transfer learning on our targeted NODE21 [3, p. 21] (lung nodule on Chest X-ray) dataset, and the final one- and two-stage detection network "quantile window bagging" for the optimal achievement of sensitivity. Both pretraining and transfer learning are implemented with Mask-RCNN [4] and RetinaNet [5] on detectron2 project[1]. Our rationales are elaborated as follows.

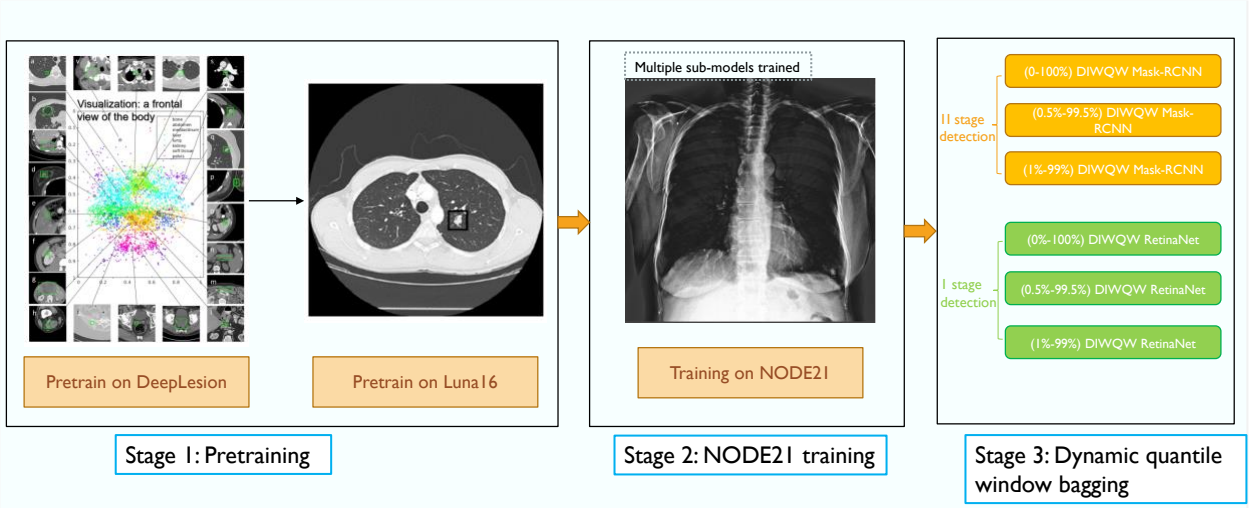

Figure 1: The pipeline for our proposed Trans-QWin-Bag detection mechanism

1.1 Data Preprocessing

Out-of-network Data Preprocessing: All of the datasets used, including DeepLesion, Luna16, and Node21, were processed into MS COCO annotation format [6] to abide by the input requirement of detectron2. Since we aim to roughly grasp some general domain knowledge in the stage of model pretraining, only the images with positive cases (positive images) were kept in processed DeepLesion and Luna16 datasets. Vice versa, for NODE21, we kept both positive and negative images. For data inputing into Mask-RCNN network, we further oversampled the positive images to balance out the entire dataset, whereas for that into RetinaNet, we did not conduct

[1] https://github.com/facebookresearch/detectron2

oversampling of positive cases since the focal loss function of RetinaNet is designed for handling imbalanced scenarios [5] .

In-network Data Preprocessing: Due to the intrinsic characteristics of CT imaging, absolute full body Hounsfield Unit (HU) window – (-1024, 3071) as well as lung HU window – (-1000, 400) are applied to the threshold and uniformly normalize the DeepLesion and Luna16 datasets before feeding into the neural network (NN).

Nevertheless, uniform normalization with various dynamic image-wise quantile windowing (DIWQW) was used for processing the NODE21 CXR dataset. Specifically, DIWQW was conducted by first finding and thresholding out the absolute values with respect to the given upper and lower quantile bounds in the target image and then conducting uniform normalization to rescale its pixel intensities. The ranges of DIWQW thresholds are selected to gradually and symmetrically exclude some extreme image pixel values at the left and right tails of the distribution (**Figure 2**). See **Table 1** for details of the thresholds used in DIWQW.

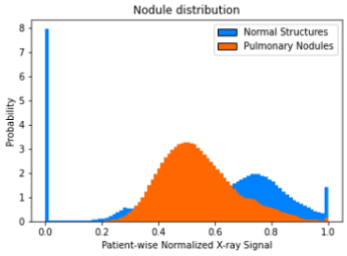

Figure 2: Uniformly normalized signal distribution visualization of CXR images and lung nodule

Quantile Range
(0% - 100%)
(0.5% - 99.5%)
(1% - 99%)

Table 1: Threshold ranges used for DIWQW strategies

1.2 Object Detection Network Architecture

As seen in Stage 3 of **Figure 1**, we selected two types of object detection framework – Mask-RCNN and RetinaNet – as modeling components for the final stage of quantile window bagging. Mask-RCNN and RetinaNet are two typical robust networks of two- and one-stage detection structures. In general, two-stage networks achieve higher accuracy, whereas one-stage networks generate more potential bounding box (bbox) proposals. In this task, to achieve the highest true positive (TP) predictions while minimizing false positive (FP) cases, we would like to combine and benefit from those two networks as mentioned above. Detailed model ensemble strategies will be elaborated in the quantile window bagging section.

For the model hyperparameters of RetinaNet, we set ResNet50 [7] as the backbone with training strategies of synchronized batch normalization (SyncBN) [8] to reduce internal covariate shift of backbone weights. The intersection over union (IoU) threshold for classification of negative and postive anchors was set as [0.3, 0.7] with an IoU score lower than 0.3 classifying to negative cases and vice versa.

For the model hyperparameters of Mask-RCNN, we also used the backbone of ResNet50 with SyncBN training and IoU threshold of [0.3, 0.7]. Since we only focused on the prediction of the bounding box in the Node21 Challenge, we turned off the mask branch of Mask-RCNN and only kept the bbox regression and classification branches during model training. Notably, we did not choose faster-RCNN [9] as our candidate for 2-stage model because we believe that the additional ROI align features of Mask-RCNN will better help the alignment of feature maps to original images and so as to improve the overall detection accuracy [4].

1.3 Transfer Learning

As mentioned in stage 1 of **Figure 1**, before officially training the NODE21 dataset, we conducted pretraining first on DeepLesion and then on Luna16 datasets to achieve better model weight initialization for both Mask-RCNN and RetinaNet, respectively. Specifically, we trained the DeepLesion dataset using weights of ResNet50 pretrained on ImageNet [9], and next, we trained

the Luna16 dataset with weights initialized from the ones pretrained on DeepLesion. The rationale for using DeepLesion plus Luna16 pretrained weights is to better prevent overfitting and ideally help escape local optima by providing more relevant background knowledge from similar datasets via the step of weight initialization.

1.4 Model Training

All the training was performed on a GPU cluster with 4×RTXA6000 and batch size of 8.

Pretraining: For Mask-RCNN and RetinaNet on DeepLesion and Luna16, we trained each model on the individual dataset for 10k iteration with a base learning rate (LR) of 0.01 and 10% LR decay on 8k and 9k iteration respectively. All the training has no backpropagation freezing on backbone residual blocks.

Fine-tuning: The data augmentation used in Mask-RCNN and RetinaNet includes random resizing in the scale of 0.8 to 2 (step 0.2), random rotation with an angle of 0 to 180, and random gaussian blur with a sigma of 0.3 to 3 and kernel size of (5, 5). Additionally, we ran Mask-RCNN fine-tuning for 10k iteration with base LR of 0.005 and 10% LR decay on 8k and 9k, and RetinaNet fine-tuning for 10k iteration with base LR of 0.001 and 10% LR decay on 8k and 9k. L2-norm gradient clips were both applied to Mask-RCNN and RetinaNet training.

1.5 Dynamic Quantile Window Bagging

Shown from **Figure 1**, dynamic quantile window bagging (DQWB) is the final stage before we output our final inferences. One noteworthy point of DQWB is that we did not conduct dataset bootstrapping for each sub-model since we did not bag up a great bunch of voters as traditional random forest or boosting [10] would do, and also we believe that quantile windowing could well diversify the input information for each sub-model. And thus, multicollinearity in the traditional ensemble model is not a major consideration here in our DQWB. The procedures for DQWB could break into the following two sub-stages.

Firstly, observing the fact that although RetinaNet could achieve higher AUC scores than Mask-RCNN, it was prone to generate more FP proposals. We grouped the bbox inferences from RetinaNet trained with various DIWQW only when at least two DIWQW sub-model predictions with area IoU of 0.2 to reduce the number of FP cases without scarifying its sensitivity on TP cases (see left-hand side of **Figure 3**).

Secondly, we bagged the bbox proposals from the RetinaNet bag as well as all Mask-RCNN inferences trained with different DIWQW and then applied non-maximum suppression (NMS) with an IoU threshold of 0.2 to reduce the heavily overlapped bboxes, which is also the last step of our entire detection pipeline (see right-hand side of **Figure 3**).

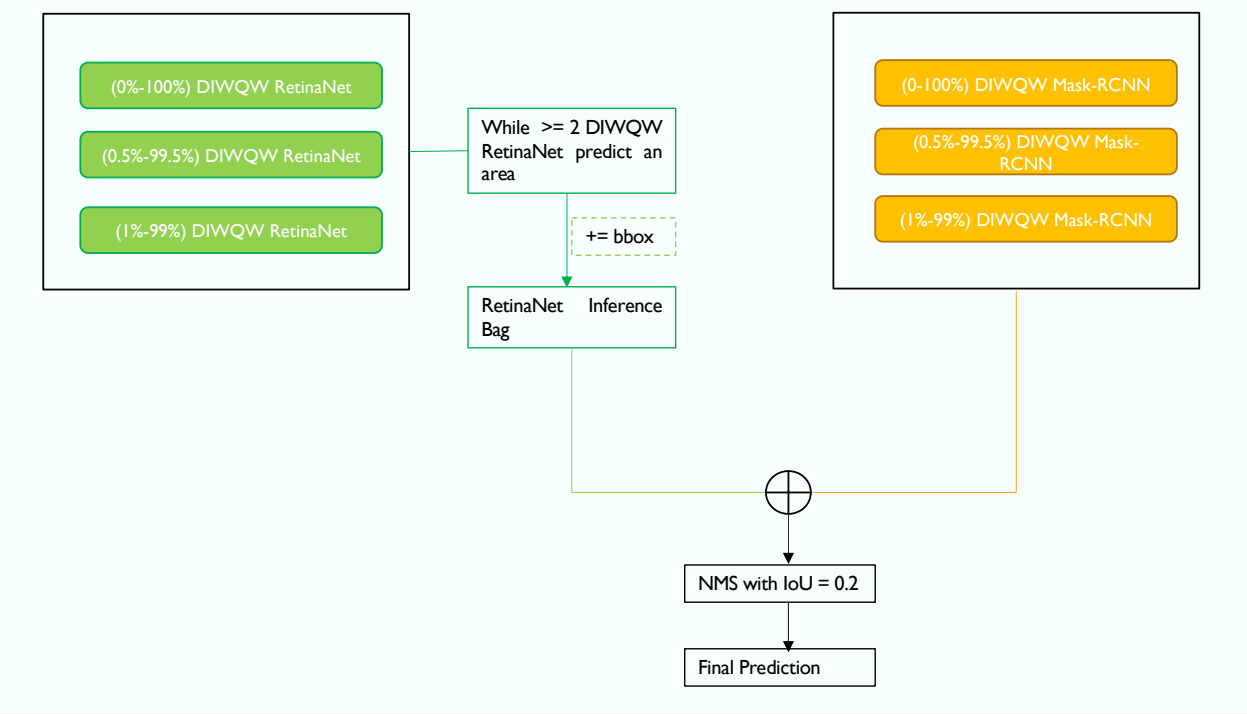

Figure 3: Pipeline for dynamic quantile window bagging

2. Experiment Results

As shown by the AUC score of RetinaNet in **Table 2**, the one-stage detection network is better at predicting more TP cases, whereas it also suffers from lots of FP predictions and therefore drags

down its sensitivity scores and overall final ranking. Different from RetinaNet, although Mask-RCNN lead to relatively more FN cases, it is more robust on sensitivity scores owning to smaller number of FP predictions.

The ablation study of transfer learning listed in **Table 3** shows that in compareson to pretrained ImageNet, datasets with similar background information could better boost model learning. Except that the combination of DeepLesion and Luna16 pretraining achieved the best inference on experimental test set, we could also observe that pretraining solely on DeepLesion is slightly more effective than that of only on Luna16.

Table 4 shows the effectiveness of our Trans-QWin-Bag strategies. By aggregating Mask-RCNN with different DIWQW, we observed an approximately 1% improvement on AUC, however its sensitivity on 0.25 FP per images drops a little compared to that of single Mask-RCNN model. In terms of Trans-QWin-Bag 1 and Trans-QWin-Bag 2, we could observe a around 2% rise on every metrics of Trans-QWin-Bag 2.

Modality	AUC	Sensitivity_5	Sensitivity_25	Sensitivity_125	Final_Ranking
Mask-RCNN	84.42%	58.47%	51.23%	39.57%	76.12%
RetinaNet	87.19%	48.89%	42.81%	31.73%	73.33%

Table 2: Ablation study of one- and two-stage prediction model. Results shown in the table were evaluated on the experimental test set on grand challenge. Both models were trained with (0.5%-99.5%) DIWQW.

Pretraining	Modality	AUC	Sensitivity_5	Sensitivity_25	Sensitivity_125	Final_Ranking
ImageNet	Mask-RCNN	81.82%	53.23%	49.19%	42.74%	73.66%
DeepLesion		83.38%	54.98%	50.07%	41.93%	75.04%
Luna16		82.28%	53.23%	49.60%	40.55%	74.50%
DeepLesion + Luna16		84.42%	58.47%	51.23%	39.57%	76.12%

Table 3: Ablation study of transfer learning from different datasets. Results shown in the table were evaluated on the experimental test set on grand challenge. Mask-RCNN results presented were trained with (0.5%-99.5%) DIWQW.

Modality	AUC	Sensitivity_5	Sensitivity_25	Sensitivity_125	Final_Ranking

Mask-RCNN	84.42%	58.47%	51.23%	39.57%	76.12%
Trans-QWin-Bag 1	85.37%	58.47%	47.58%	39.97%	75.92%
Trans-QWin-Bag 2	86.19%	60.89%	50.81%	40.73%	77.35%

Table 4: Results evaluated on the experimental test set on grand challenge. All three models were fine-tuned on DeepLesion+Luna16 dataset. Trans-QWin-Bag 1 contained three Mask-RCNN voters with different DIWQW. Trans-QWin-Bag2 contained three Mask-RCNN and three RetinaNet voters with different DIWQW.

3. Discussion

Chest X-rays (CXR) are commonly used for early-stage lung cancer. Compared with CT, CXR is more cost-effective, assessable, and delivers a lower imaging dose. However, it is more difficult to discriminate small lesions in CXR due to superimposing anatomies including ribs, clavicles, hilar, major vessels, and mediastinal structures. It has been reported that 19% of the non-small cell lung cancer was missed by radiologists in CXR examination [11]. Human observers tend to be inconsistent in performance, influenced by factors including experience and fatigue. Besides human radiologist performance, there is a shortage of trained radiologists to read CXR images, particularly in the medium-to-low income countries. Therefore, automated lung nodule detection is necessary to improve the process and bring lung cancer screening to various settings, including the low-resource ones.

Explicit image features including shape, surface, gray-level intensity, and gradient have been used in conventional machine learning methods to discriminate lung nodules in CXR [12], [13]. With the exponential growth of computational power, convolutional neural networks have shown superior detection capability in various diagnostic tasks, including lung nodule detection on CXR [14]. In this study, we further improved the CNN-based detection methods in the following aspects.

First, we combined and benefits of Mask-RCNN, which is a two-stage network for higher detection accuracy, and RetinaNet, which is a one-stage network generating more potential bounding box (bbox) proposals. The combined strengths of the two networks are shown in the dynamic quantile window bagging results.

Second, we pretrained both Mask-RNN and RetinaNet on DeepLesion and then on Luna16 datasets to achieve better model weight initialization. As shown in our result, pertaining on these datasets helps prevent overfitting and escape local optima by providing more relevant background knowledge from similar datasets via weight initialization.

Note that different sets of evaluation metrics may be used in clinical applications. For example, in the context of computer-assisted diagnosis (CAD), the detection is not fully automated. As such, high detection sensitivity is required with reasonable false positivity for the human observers to screen. Our network is fully adaptive to such tasks.

Additional improvement of the network may require better suppression of the interfering anatomical structures in the CXR images. For instance, dual-energy X-ray images may suppress the bone signals [15]. Suppression of anatomical structures via machine learning [16] may be possible pending future investigation.

References

[1] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imaging Bellingham Wash*, vol. 5, no. 3, p. 036501, Jul. 2018, doi: 10.1117/1.JMI.5.3.036501.

[2] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017, doi: 10.1016/j.media.2017.06.015.

[3] E. Sogancioglu, K. Murphy, and B. Van Ginneken, "NODE21." Zenodo, Apr. 28, 2021. doi: 10.5281/ZENODO.5548363.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *ArXiv170306870 Cs*, Jan. 2018, Accessed: Nov. 30, 2021. [Online]. Available: http://arxiv.org/abs/1703.06870

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *ArXiv170802002 Cs*, Feb. 2018, Accessed: Jan. 23, 2022. [Online]. Available: http://arxiv.org/abs/1708.02002

[6] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *ArXiv14050312 Cs*, Feb. 2015, Accessed: Jan. 21, 2022. [Online]. Available: http://arxiv.org/abs/1405.0312

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv151203385 Cs*, Dec. 2015, Accessed: Jan. 23, 2022. [Online]. Available: http://arxiv.org/abs/1512.03385

[8] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *ArXiv150203167 Cs*, Mar. 2015, Accessed: Jan. 23, 2022. [Online]. Available: http://arxiv.org/abs/1502.03167

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[10] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, Berlin, Heidelberg, 2000, pp. 1–15.

[11] L. G. Quekel, A. G. Kessels, R. Goei, and J. M. van Engelshoven, "Miss rate of lung cancer on the chest radiograph in clinical practice," *Chest*, vol. 115, no. 3, pp. 720–724, Mar. 1999, doi: 10.1378/chest.115.3.720.

[12] R. C. Hardie, S. K. Rogers, T. Wilson, and A. Rogers, "Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs," *Med. Image Anal.*, vol. 12, no. 3, pp. 240–258, Jun. 2008, doi: 10.1016/j.media.2007.10.004.

[13] S. Chen, K. Suzuki, and H. MacMahon, "Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification," *Med. Phys.*, vol. 38, no. 4, pp. 1844–1858, Apr. 2011, doi: 10.1118/1.3561504.

[14] X. Li *et al.*, "Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection," *Artif. Intell. Med.*, vol. 103, p. 101744, Mar. 2020, doi: 10.1016/j.artmed.2019.101744.

[15] F. Li, R. Engelmann, L. L. Pesce, K. Doi, C. E. Metz, and H. Macmahon, "Small lung cancers: improved detection by use of bone suppression imaging--comparison with dual-energy

subtraction chest radiography," *Radiology*, vol. 261, no. 3, pp. 937–949, Dec. 2011, doi: 10.1148/radiol.11110192.

[16] S. Oda *et al.*, "Performance of radiologists in detection of small pulmonary nodules on chest radiographs: effect of rib suppression with a massive-training artificial neural network," *AJR Am. J. Roentgenol.*, vol. 193, no. 5, pp. W397-402, Nov. 2009, doi: 10.2214/AJR.09.2431.