Journal Pre-proof

Global-Local Attention Network with Multi-task Uncertainty Loss for Abnormal Lymph Node Detection in MR Images

Shuai Wang, Yingying Zhu, Sungwon Lee, Daniel C. Elton, Thomas C. Shen, Youbao Tang, Yifan Peng, Zhiyong Lu, Ronald M. Summers

PII: S1361-8415(21)00390-X
DOI: https://doi.org/10.1016/j.media.2021.102345
Reference: MEDIMA 102345

To appear in: *Medical Image Analysis*

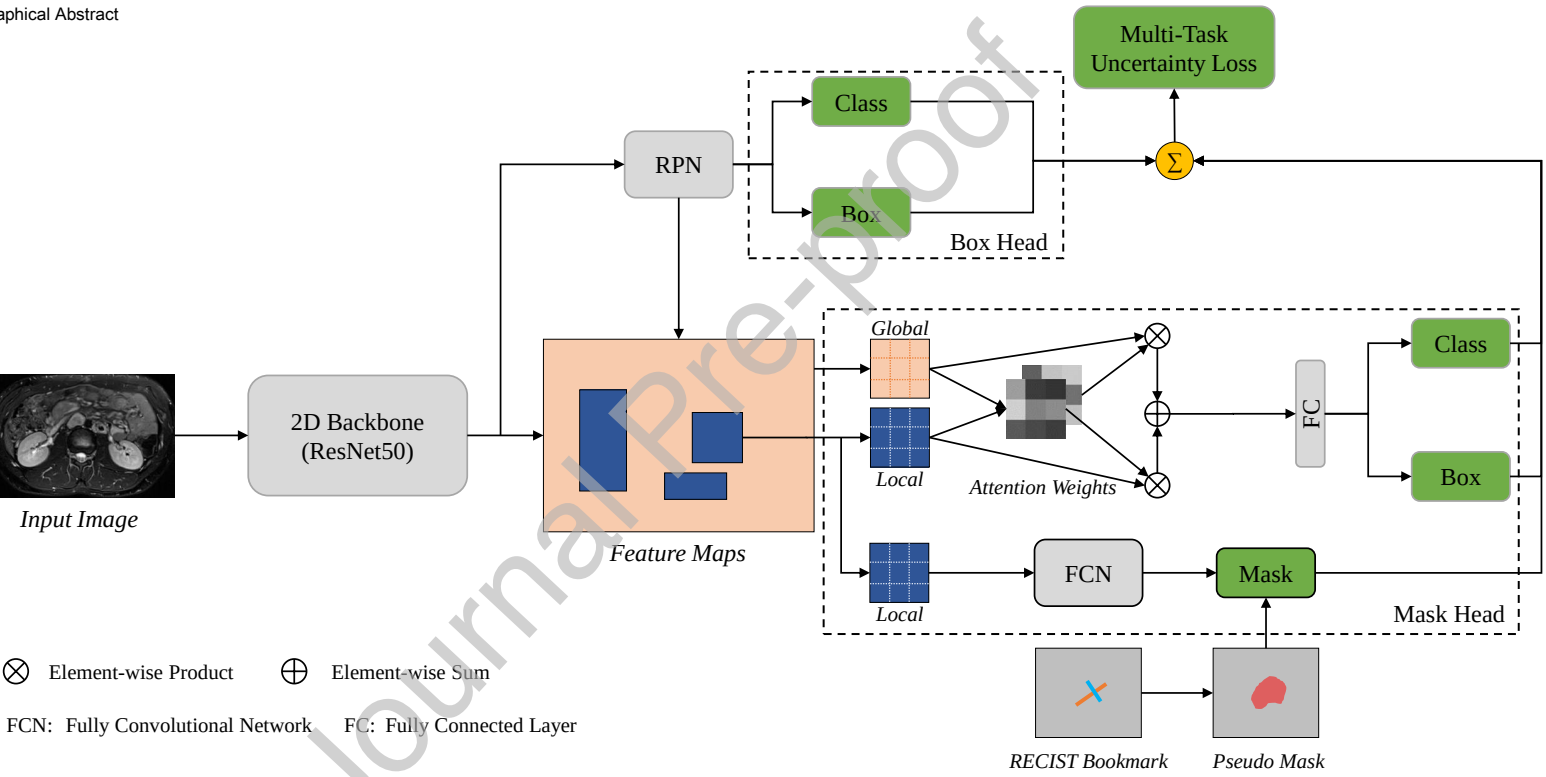
Received date: 19 March 2021
Revised date: 27 December 2021
Accepted date: 28 December 2021

Please cite this article as: Shuai Wang, Yingying Zhu, Sungwon Lee, Daniel C. Elton, Thomas C. Shen, Youbao Tang, Yifan Peng, Zhiyong Lu, Ronald M. Summers, Global-Local Attention Network with Multi-task Uncertainty Loss for Abnormal Lymph Node Detection in MR Images, *Medical Image Analysis* (2021), doi: https://doi.org/10.1016/j.media.2021.102345

Highlights

- We propose a novel network for the universal abnormal lymph node detection in MR images, which has great clinical value for the diagnosis of numerous diseases.
- We design a global-local context module to encode the image global and local scale context information for the detection and utilize the channel attention mechanism to weight different contexts.
- We introduce a multi-task uncertainty loss to adaptively balance the losses of different tasks, which can effectively alleviate the burden for tuning the loss weights by hand.
- We build a large-scale MRI abnormal lymph node dataset, which includes a total of 821 abnormal abdominal lymph nodes of 41 types from 584 different patients. Moreover, 123 images with complete 3D volume annotations are delineated by an experienced radiologist.

aphical Abstract



Medical Image Analysis (2022)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Global-Local Attention Network with Multi-task Uncertainty Loss for Abnormal Lymph Node Detection in MR Images

Shuai Wang[a,c], Yingying Zhu[a], Sungwon Lee[a], Daniel C. Elton[a], Thomas C. Shen[a], Youbao Tang[a], Yifan Peng[b], Zhiyong Lu[b], Ronald M. Summers[a,*]

[a]*Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA*
[b]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
[c]*School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, PR China*

ARTICLE INFO

Article history:

Keywords: Image Detection, Magnetic Resonance Imaging, Lymph Node, Deep Learning

ABSTRACT

Accurate and reliable detection of abnormal lymph nodes in magnetic resonance (MR) images is very helpful for the diagnosis and treatment of numerous diseases. However, it is still a challenging task due to similar appearances between abnormal lymph nodes and other tissues. In this paper, we propose a novel network based on an improved Mask R-CNN framework for the detection of abnormal lymph nodes in MR images. Instead of laboriously collecting large-scale pixel-wise annotated training data, pseudo masks generated from RECIST bookmarks on hand are utilized as the supervision. Different from the standard Mask R-CNN architecture, there are two main innovations in our proposed network: 1) global-local attention which encodes the global and local scale context for detection and utilizes the channel attention mechanism to extract more discriminative features and 2) multi-task uncertainty loss which adaptively weights multiple objective loss functions based on the uncertainty of each task to automatically search the optimal solution. For the experiments, we built a new abnormal lymph node dataset with 821 RECIST bookmarks of 41 different types of abnormal abdominal lymph nodes from 584 different patients. The experimental results showed the superior performance of our algorithm over compared state-of-the-art approaches.

1. Introduction

Detecting the location and size of abnormal lymph nodes in magnetic resonance (MR) images is a crucial step in the staging and treatment of cancer and lymphoproliferative disorders Amin et al. (2017). However, this task is time-consuming and burdensome, easily creating room for human errors. Therefore, automatic detection and segmentation of abnormal lymph nodes are highly desired for increasing work efficiency and reducing inter-observer variability.

There are many studies devoted to the detection of abnormal lymph nodes, but most are designed for specific regions, such as sentinel Liu et al. (2019); Kuwahata et al. (2020), axillary Barbu et al. (2010); Ha et al. (2018); Kitaizumi et al. (2020), and mediastinal lymph nodes Oda et al. (2018). However, in daily practice, radiologists need to identify all abnormal lymph nodes appearing in the entire scan to make an accurate diagnosis. Therefore, a universal detection method to find all abnormal lymph nodes in MR images will have great clinical value. To build an accurate universal abnormal lymph node detection model, there are many challenges to deal with such as 1) low contrast and

*This work was supported in part by the Intramural Research Programs of the National Institutes of Health Clinical Center and National Library of Medicine and the Qilu Youth Scholar Discipline Construction Funding from Shandong University (Corresponding author: R. Summers).

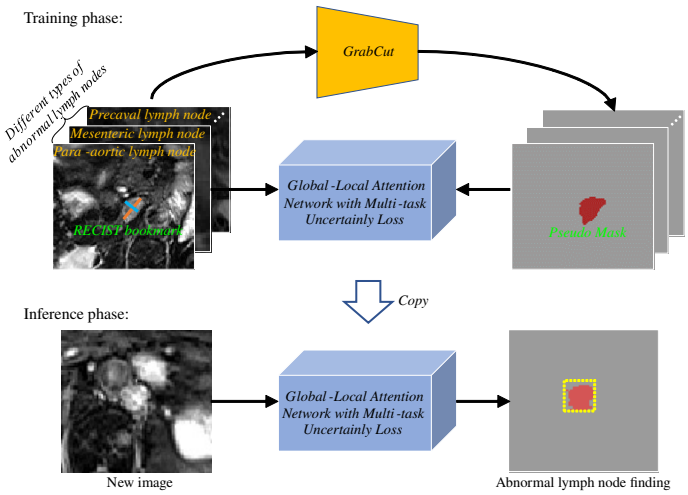


Fig. 1. The framework of our proposed method for abnormal lymph node detection starting from RECIST bookmarks. For better observation, the original images have been cropped around the bookmarks.

non-uniform intensity distribution due to MR imaging characteristics, 2) large intra-class variability between different lymph node types making it hard to train a robust model, and 3) high cost to collect large-scale accurate annotations for training.

In their routine work, radiologists generally annotate abnormal lymph nodes using some type of bookmarks, such as arrows, lines, diameters, or segmentations Yan et al. (2018b). In these types, lesion diameter is a good balance between cost and accuracy. As part of the RECIST guidelines Eisenhauer et al. (2009), lesion diameters consist of two lines for each abnormal finding, one measuring the longest diameter of the finding and the other one measuring its longest perpendicular diameter. For conciseness, we refer to the lesion diameter annotations as RECIST bookmarks. In this paper, we propose a novel network that is trained starting from RECIST bookmarks instead of full pixel-wise annotations, which eliminates the need for additional manual annotation. Our framework is shown in Fig. 1. To make full use of the RECIST bookmarks to supervise the network optimization, we first generate pseudo masks using a GrabCut-based method Rother et al. (2012) in the training phase. Then, we design a global-local attention network with multi-task uncertainty loss to detect all abnormal lymph nodes in MR images. Here, Mask R-CNN He et al. (2017) is taken as the base model but our proposed global-local module and multi-task uncertainty loss can be easily introduced into any type of detector with little effort.

The main contributions of this work can be summarized as follows. First, we propose a novel network for the universal abnormal lymph node detection in MR images, which has great clinical value for the diagnosis of numerous diseases. Second, we design a global-local context module to encode the image global and local scale context information for the detection and utilize the channel attention mechanism to weight different contexts. Third, we introduce a multi-task uncertainty loss to adaptively balance the losses of different tasks, which can effectively alleviate the burden for tuning the loss weights by hand. Finally, we build a large-scale MRI abnormal lymph node dataset, which includes a total of 821 abnormal abdominal lymph nodes of 41 types from 584 different patients. Moreover, 123 images with complete 3D volume annotations are delineated by an experienced radiologist.

2. Related Work

Object detection is a hot topic that has attracted much attention. Much effort in this field has been made in tasks such as face, object, and medical lesion detection Shen et al. (2017); Zhao et al. (2019); Oksuz et al. (2020); Kong et al. (2020); Fan et al. (2020). While our work concerns lymph node detection from RECIST bookmarks, we will first introduce the related works on general object detection methods, then specific approaches for lymph node detection, and finally lesion detection approaches based on RECIST bookmarks.

2.1. General Object Detection

The goal of object detection is to determine the categories of instances in the image and indicate their location with bounding boxes. Instead of extracting carefully designed hand-crafted features, deep learning has shown excellent performance in object detection, which combines feature extraction and model optimization in a unified end-to-end manner Shin et al. (2016). The existing deep learning-based detectors can be divided into two classes: two-stage and one-stage networks Chiang et al. (2018); Liu et al. (2020); Cao et al. (2020). The main difference between them is that the two-stage network extracts a set of regions of interest (ROIs) before making the detection while the one-stage network directly performs detection based on dense samplings. Compared with the two-stage network, the one-stage network achieves a higher inference speed.

The first two-stage detector was R-CNN Girshick et al. (2014), which combined region proposals with a convolutional neural network (CNN). Based on R-CNN Girshick et al. (2014), Fast R-CNN Girshick (2015) sped up training and testing time by jointly extracting object proposal features, training for classification, and bounding box regression. Faster R-CNN Ren et al. (2017) further reduced the running time by learning the region proposals instead of using the selective search algorithm. To effectively detect objects while simultaneously generating semantic segmentation masks, Mask R-CNN He et al. (2017) extended Faster R-CNN Ren et al. (2017) by adding an object segmentation branch to predict pixel-level labels. In object detection, detectors are usually subject to overfitting at training and quality mismatch at inference related to the used intersection over union (IoU) threshold. To address this problem, Cascade R-CNN Cai and Vasconcelos (2018, 2019) trained a multi-stage detector by sequentially increasing IoU thresholds.

The most representative one-stage detector was YOLO Redmon et al. (2016), which achieved high detection accuracy while also being able to run in real-time. After YOLO Redmon et al. (2016), YOLOv2 Redmon and Farhadi (2017), and YOLOv3 Redmon and Farhadi (2018) continuously improved the detection accuracy while still achieving high inference speed. The other representative one-stage detector was SSD

(Single Shot MultiBox Detector) Liu et al. (2016b), which covered objects using multiple feature maps at different resolutions and scales. To address the foreground-background class imbalance problem, Focal Loss Lin et al. (2017) was designed for one-stage detectors by down-weighting the loss assigned to well-classified samples. To make Focal Loss applicable to the continuous form, Generalized Focal Loss Li et al. (2020) was further proposed.

To introduce different scale context information into detection, multi-scale mechanism is a popular method Alansary et al. (2019); Wang et al. (2020b). The two common choices for multi-scale information extraction are the atrous spatial pyramid pooling (ASPP) Chen et al. (2017) and the pyramid pooling module (PPM) Zhao et al. (2017), which have many variants based on them, such as Wang et al. (2018); Piao et al. (2019); Guo et al. (2020); Wang et al. (2020a). To further improve the robustness and discrimination of learned features from different scales, an attention-based fusion module is usually used. For example, Shao et al. (2019) designed a new block that used dilated convolution operations to learn multi-scale features and followed channel attention and spatial module to pay attention to more important features. Cui et al. (2019) proposed a dense attention pyramid network for the ship detection in SAR images which densely connected convolutional block attention module to each concatenated feature map from top to bottom of the pyramid network.

2.2. Lymph Node Detection

Abnormalities of lymph nodes in different body parts often alert us to different diseases, such as infection, lymphoma, and metastatic cancer Roth et al. (2014); Liu et al. (2016a). Therefore, it is of great value to detect abnormal lymph nodes. Unfortunately, the great variations of lymph nodes in quantity, size, shape, and property make themselves difficult to identify and also cause the bottleneck of constructing lymph node datasets, leading to fewer studies on lymph node detection Roth et al. (2014); Wang et al. (2021). Instead of using deep learning-based detectors, most existing methods usually involve multiple separate stage methods for lymph node detection. For example, Ma and Peng (2020) first detected lymph node candidates using two preliminary computer-aided detection systems and then used a CNN model to identify true lymph nodes. Debats et al. (2019) fused the information from multi-view input to reduce the false-positive predictions after extracting some pre-defined patches. Bouget et al. (2019) combined U-net Ronneberger et al. (2015) and Mask R-CNN to make semantic segmentation and detection of mediastinal lymph nodes for lung cancer staging. Carolus et al. (2020) first employed image patches of different resolutions as candidates, and then trained a 3D CNN to detect enlarged potentially malignant lymph nodes. Instead of multi-stage methods, Zhu et al. (2020a) transformed the detection problem into a segmentation problem and designed a multi-branch detection-by-segmentation network for lymph node gross tumor volume detection. However, the limited works are mostly designed for lymph node detection in special regions because of the paucity of universal lymph node datasets. It is more challenging to design a universal detector to detect abnormal lymph nodes from the entire body.

2.3. RECIST-based Lesion Detection

To reduce the annotation cost, there are some works proposed to achieve lesion detection based on RECIST bookmarks. In Yan et al. (2018a); Shao et al. (2019); Tao et al. (2019), they directly utilized the bounding boxes drawn based on RECIST bookmarks to achieve universal lesion detection. Compared with bounding boxes, pixel/voxel-level pseudo mask may represent the lesion more accurately, and an additional segmentation task under its supervision may be beneficial to the improvement of detection performance. Therefore, Tang et al. (2019) and Yan et al. (2019) adopted GrabCut to generate pseudo masks and taken Mask R-CNN as the backbone to implement the detection and also segmentation. Zlocha et al. (2019) also used GrabCut to generate the pseudo masks from RECIST bookmarks by defaulting that most lesions are convex outlines and proposed an improved RetinaNet for CT lesion detection. Instead of generating bounding boxes or pseudo masks, Xie et al. (2021) directly detected the four extreme points and a center point of the RECIST bookmarks to achieve detection.

3. Methodology

The architecture of our network is shown in Fig. 2, which utilizes 2D Mask R-CNN He et al. (2017) as the base model and consists of three main components: pseudo mask generation from RECIST bookmarks, global-local attention, and multi-task uncertainty loss.

3.1. Pseudo Mask Generation

For each RECIST bookmark, there are two lines to delineate the corresponding abnormal lymph node, one measuring the longest diameter of the lymph node and the other one measuring its longest perpendicular diameter. For training the network, we can directly generate the bounding box annotation from these two lines, namely, by using their minimum bounding rectangle. However, due to the irregular shape variations of lymph nodes, this simple manner may lead to inaccurate bounding box representation and will degrade the detection performance. Therefore, we first generate pseudo masks from RECIST bookmarks to provide more accurate supervision for the model optimization in the training stage.

Here, we utilize GrabCut Rother et al. (2012); Tang et al. (2018); Wang et al. (2020c) to generate abnormal lymph node masks from the RECIST bookmarks. To reduce the complexity, we first indicate a square ROI centered on the midpoint of the longer diameter, with a side length twice the length of that diameter. Next, we perform GrabCut only inside the square ROI. In GrabCut, four classes of pixels should be initialized: obvious foreground pixels, obvious background pixels, possible foreground pixels, and possible background pixels. The pixels on the two lines belong to obvious foreground pixels. To initialize possible foreground pixels, we connect two adjacent endpoints of two lines in turn with a straight line. The pixels inside the resulting diamond are treated as possible foreground pixels. As it is difficult to determine the boundary between foreground and background, we set all other pixels as the possible background pixels. After initialization, we perform GrabCut to

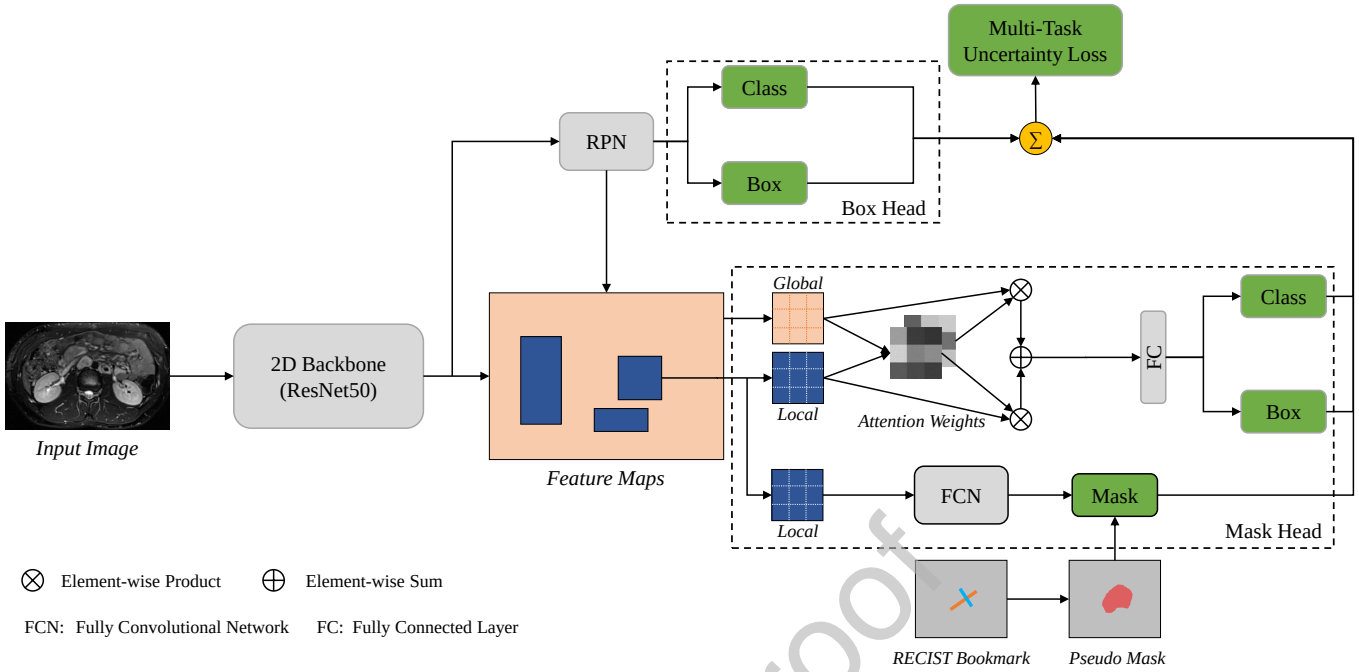


Fig. 2. Architecture of Global-Local Attention Network with Multi-task Uncertainty Loss. Since initial RECIST bookmarks can only provide coarse bounding box annotations, pseudo masks are generated to provide more accurate supervision information in the training stage and make the segmentation task available. For abnormal lymph node detection, a global-local attention module is introduced to improve the representation ability for better accuracy. To optimize the whole network, a multi-task uncertainty loss to adaptively weigh different tasks is designed to alleviate the burden of tuning these weights by hand.

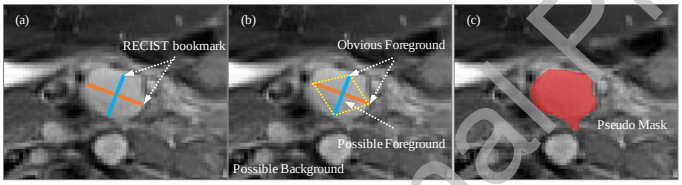


Fig. 3. An illustration showing how to generate pseudo masks from RECIST bookmarks. (a) Two lines of one RECIST bookmark for the corresponding abnormal lymph node. (b) Three types of initialized pixels used in GrabCut. (c) Generated pseudo mask (red).

generate pseudo masks. In Fig. 3, a pseudo mask generation example is given.

3.2. Global-Local Attention

In our study, our goal is to achieve universal abnormal lymph node detection in MR images which may cover different parts across the whole body. The large intra-class variability between different types of abnormal lymph nodes and the non-uniform intensity distribution between MR images present challenges to train a powerful detection model. In practice, a radiologist may make decisions based not only on the appearance of particular lymph nodes but also on broad and contextual information. To mimic this capability of considering context, we propose a global-local attention module for abnormal lymph node detection.

As shown in Fig. 2, our global-local attention module is performed for the final box classification and regression tasks of the mask head, not for the semantic segmentation task. The main reason for this is that the local scale itself has provided the broad and contextual information for each pixel inside the box and the global scale information is difficult to align with the local scale information for each pixel, which may weaken the segmentation performance. For each target ROI, both the fixed-size features of the target ROI (local, f^l) and the whole image (global, f^g) are extracted using the RoIAlign layer of Mask R-CNN. Instead of directly concatenating these local and global features together, we use an attention mechanism inspired by Chen et al. (2016) to weigh the information from different scales to improve the feature robustness. The detailed attention implementation is shown in Fig. 4. First, the local and global feature maps (e.g. 7×7) are concatenated and go through two convolutional layers. The first one is with kernel size 3×3 and 256 output channels followed by a Batch Normalization layer and a ReLU activation function while the second one is with kernel size 3×3 and 2 output channels. After that, the resulting feature maps (r) go through a Softmax layer to generate the weights:

$$w_i^s = \frac{exp(r_i^s)}{\sum_{t=1}^{S} exp(r_i^t)}, \tag{1}$$

where r_i^s is the resulting feature map of scale s at position i and S is the total number of scales. Here, $S = 2$ for the case of one global scale and one local scale. The final feature maps h for box regression and classification of the mask head are the weighted sum of f^l and f^g:

$$h_{i,c} = w_i^l \cdot f_{i,c}^l + w_i^g \cdot f_{i,c}^g \tag{2}$$

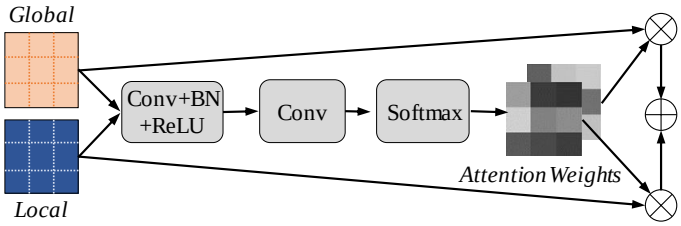


Fig. 4. Detailed global-local attention implementation.

where c corresponds to the feature channel and w_i^s is shared across all channels.

For the global context, we can adopt other sizes instead of the whole image to encode the global information and also easily extend single-scale implementation to multi-scale implementation without any cost. The global size and the number of global scales can be adaptively set according to the used different datasets. In the experimental section, we conducted comprehensive studies to verify the settings in our implementation.

3.3. Multi-task Uncertainty Loss

Regardless of whether the detector is a one-stage method or a two-stage method, it is usually a multi-task architecture, for example, to perform both box regression and classification tasks. To optimize the whole network, a grid search method is often adopted to tune the loss weight of each task for the best performance. However, the grid search method means that the corresponding network needs to be repeatedly trained many times. Such a tuning strategy is expensive in practice, as the cost grows exponentially with the number of tasks. Moreover, the grid search is only performed on a limited set of points so the optimal weights may not be achieved.

In our implementation, our network consists of five tasks similar to the original Mask R-CNN He et al. (2017). To alleviate the burden of tuning the loss weights by hand, we introduce an adaptive strategy to learn the loss weights in terms of task uncertainty. As described in Kendall et al. (2018), task uncertainty can capture the relative confidence between tasks. Let $f^\theta(x)$ be the output of a network with weights θ on input x. In Kendall et al. (2018), the regression likelihood is adapted as a Gaussian with mean given by the model output:

$$p(y|f^\theta(x)) = \mathcal{N}(f^\theta(x), \sigma^2) \tag{3}$$

with an observation noise scalar σ. And the classification likelihood is computed by squashing a scaled version of the model output through a softmax function:

$$p(y|f^\theta(x), \sigma) = \text{Softmax}(\frac{1}{\sigma^2} f^\theta(x)) \tag{4}$$

with a positive scalar σ.

In the case of multiple model outputs, the multi-task likelihood can factorise over the outputs by defining $f^\theta(x)$ as sufficient statistics:

$$p(y_1, ..., y_K|f^\theta(x)) = p(y_1|f^\theta(x))...p(y_K|f^\theta(x)), \tag{5}$$

in which y_1,...,y_K are the model outputs with respect to regression, classification, etc.

For the original Mask R-CNN, there are five tasks, two for regression and three for classification, modeled with Gaussian likelihoods and softmax likelihoods, respectively. And we denote y^r_{cls} and y^r_{box} as the outputs of the classification and regression tasks of the box head respectively, y^m_{cls}, y^m_{box}, and y^m_{mask} as the outputs of the classification, regression, and segmentation tasks of the mask head respectively. For optimization, we can minimize the negative log-likelihood of the model, and based on Eq. 3, 4, and 5, the overall loss can be written as:

$$\begin{aligned}
\mathcal{L} &= -\,log p(y^r_{cls} = c_r, y^r_{box}, y^m_{cls} = c_m, y^m_{box}, y^m_{mask} = s_m|f^\theta(x)) \\
&= -\,log p(y^r_{cls} = c_r|f^\theta(x)) \cdot p(y^r_{box}|f^\theta(x)) \\
&\quad \cdot p(y^m_{cls} = c_m|f^\theta(x)) \cdot p(y^m_{box}|f^\theta(x)) \cdot p(y^m_{mask} = s_m|f^\theta(x)) \\
&= -\,log p(y^r_{cls} = c_r|f^\theta(x)) - log p(y^r_{box}|f^\theta(x)) \\
&\quad - log p(y^m_{cls} = c_m|f^\theta(x)) - log p(y^m_{box}|f^\theta(x)) \\
&\quad - log p(y^m_{mask} = s_m|f^\theta(x)) \\
&= -\frac{1}{\sigma_1^2} log\text{Softmax}(y^r_{cls}, f^\theta(x)) + log\frac{\sum_{c'_r} \exp(\frac{1}{\sigma_1^2} f^\theta_{c'_r}(x))}{(\sum_{c'_r} \exp(f^\theta_{c'_r}(x)))^{\frac{1}{\sigma_1^2}}} \\
&\quad + \frac{1}{2\sigma_2^2}||y^r_{box} - f^\theta(x)||^2 + log\sigma_2 \\
&\quad - \frac{1}{\sigma_3^2} log\text{Softmax}(y^m_{cls}, f^\theta(x)) + log\frac{\sum_{c'_m} \exp(\frac{1}{\sigma_3^2} f^\theta_{c'_m}(x))}{(\sum_{c'_m} \exp(f^\theta_{c'_m}(x)))^{\frac{1}{\sigma_3^2}}} \\
&\quad + \frac{1}{2\sigma_4^2}||y^m_{box} - f^\theta(x)||^2 + log\sigma_4 \\
&\quad - \frac{1}{\sigma_5^2} log\text{Softmax}(y^m_{mask}, f^\theta(x)) + log\frac{\sum_{s'_m} \exp(\frac{1}{\sigma_5^2} f^\theta_{s'_m}(x))}{(\sum_{s'_m} \exp(f^\theta_{s'_m}(x)))^{\frac{1}{\sigma_5^2}}} \\
&\approx \frac{1}{\sigma_1^2}\mathcal{L}^r_{cls}(\theta) + \frac{1}{2\sigma_2^2}\mathcal{L}^r_{box}(\theta) + \frac{1}{\sigma_3^2}\mathcal{L}^m_{cls}(\theta) + \frac{1}{2\sigma_4^2}\mathcal{L}^m_{box}(\theta) + \\
&\quad \frac{1}{\sigma_5^2}\mathcal{L}^m_{mask}(\theta) + log\sigma_1 + log\sigma_2 + log\sigma_3 + log\sigma_4 + log\sigma_5 \\
&= \frac{1}{\sigma_1^2}\mathcal{L}^r_{cls}(\theta) + \frac{1}{\sigma_2^2}\mathcal{L}^r_{box}(\theta) + \frac{1}{\sigma_3^2}\mathcal{L}^m_{cls}(\theta) + \frac{1}{\sigma_4^2}\mathcal{L}^m_{box}(\theta) + \\
&\quad \frac{1}{\sigma_5^2}\mathcal{L}^m_{mask}(\theta) + log\sigma_1 + log\sigma_2 + log\sigma_3 + log\sigma_4 + log\sigma_5 - log2
\end{aligned} \tag{6}$$

in which $\frac{1}{\sigma}\sum_{c'} \exp(\frac{1}{\sigma^2} f^\theta_{c'}(x)) \approx (\sum_{c'} \exp(f^\theta_{c'}(x)))^{\frac{1}{\sigma^2}}$ is assumed to simplify the optimization objective. $\mathcal{L}^r_{cls}$ and $\mathcal{L}^r_{box}$ are the classification and bounding box regression losses of the box head, respectively. $\mathcal{L}^m_{cls}$, $\mathcal{L}^m_{box}$, and $\mathcal{L}^m_{mask}$ are the classification, bounding box regression, and mask segmentation losses of the mask head, respectively.

Under Eq. 6, $\sigma_{1\sim5}$ can be seen as the relative weights of the losses for each output. Large σ means the output with large variance and the prediction has low confidence. Thus, the weight of the corresponding loss should be decreased. The $\log(\sigma)$ term can be seen to regularize the loss weight. In this

Table 1. State-of-the-art comparison on our *RECIST slice test set* for detection. / means that the corresponding method has no segmentation mask prediction. The best performance in each metric is indicated in bold.

Method	Mask				Box			
	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}
YOLOv3 Redmon and Farhadi (2018)	/	/	/	/	39.92	36.14	44.51	43.35
Faster R-CNN Ren et al. (2017)	/	/	/	/	49.55	42.02	53.18	46.24
Mask R-CNN w/o Pseudo Mask He et al. (2017)	/	/	/	/	42.77	37.05	47.40	42.77
DetectoRS Qiao et al. (2020)	/	/	/	/	**55.97**	47.14	58.38	50.87
Mask R-CNN He et al. (2017)	45.28	41.97	52.02	48.55	46.49	40.76	53.18	48.55
Cascade Mask R-CNN Cai and Vasconcelos (2019)	51.26	43.54	41.04	34.68	51.79	44.27	43.93	36.99
Ours	**54.57**	**48.54**	**69.36**	**60.12**	55.16	**47.82**	**72.25**	**60.12**

$AP_{25/50}$: Average Precision with the IoU threshold 0.25 or 0.50
$SEN_{25/50}$: Sensitivity with the IoU threshold 0.25 or 0.50 at 5 false-positives per image

way, the loss weights can be learned with the network parameters θ, which eliminates the high cost of adjusting the weights manually by the grid search method to achieve good performance.

4. Experiments

4.1. Dataset

Our dataset consists of abdominal MRI studies scanned between Jan 2015 to Sep 2019 and downloaded from the National Institutes of Health's Picture Archiving and Communication System. Based on the MRI radiology reports, we collected a total of 584 T2-weighted image volumes from 584 different patients in which there were 821 RECIST bookmarks from 41 types of abnormal abdominal lymph nodes extracted in axial slices. These images range in size from $256 \sim 640 \times 192 \sim 640$ pixels. The linear min-max normalization is performed on each image separately to normalize its intensity distribution to the range [0,1] due to the large study-to-study intensity variations between different images. Specifically, the top 1% intensity values in each image are first truncated to the maximum of the remaining values to handle outliers. Then, the values of each image get linearly transformed into a number between 0 and 1 as $[X_i\text{-min}(X)]/[\max(X)\text{-min}(X)]$, where X_i is the intensity value of the i-th voxel in X. The whole dataset is randomly divided into training (70%), validation (15%), and test (15%) sets at the patient-level. Since the complete 3D annotations are unavailable, only the axial slices with RECIST bookmarks participate in training. And for better evaluation, an experienced radiologist has helped to complete the accurate voxel-level 3D annotations of RECIST bookmarks in the test set. So we have two test sets: one of 165 axial slices with RECIST bookmarks (*RECIST slice test set*) and the other one of 123 volumes with accurate voxel-level annotations which contains a total of 346 axial slices with lymph nodes (*3D volume test set*). We mainly compare on *RECIST slice test set* if no special mention while the training and test sets are in the same annotation form.

4.2. Implementation Details

We use 2D Mask R-CNN with Resnet50 as the backbone and train our model starting from pre-trained COCO weights. The initial learning rate is set to 10^{-3} and the batch size is 4. Faster R-CNN Ren et al. (2017), YOLOv3 Redmon and Farhadi (2018), Mask R-CNN He et al. (2017), Cascade Mask R-CNN Cai and Vasconcelos (2019), and DetectoRS Qiao et al. (2020) are trained on our dataset for comparison since they are representative and outstanding detection methods. Moreover, all compared methods are trained starting from pre-trained COCO weights. For our method, the loss weights are learned adaptively in the training process and for other compared methods, the grid search in the range of [1,9] with an interval 2 for different loss weights is used for selecting the optimal values for each method. All methods are run 160 epochs on an Nvidia GeForce GTX TITAN X card with 12GB memory. Our method, Mask R-CNN, and Cascade Mask R-CNN are implemented based on the implementation of Abdulla (2017). Faster R-CNN and DetectoRS are implemented with MMDetection, an open-source object detection toolbox based on PyTorch Chen et al. (2019). YOLOv3 is based on the implementation of Huynh (2017).

To evaluate the detection performance, we report APs (Average Precision) with IoU thresholds 0.25 (AP_{25}) and 0.50 (AP_{50}) because lymph nodes are usually small in size. We also compute the sensitivities with IoU thresholds 0.25 (SEN_{25}) and 0.50 (SEN_{50}) at 5 false-positives (FPs) per image, which can reflect the balance between the detection accuracy and false detection findings well. Because some compared methods (YOLOv3, Faster R-CNN, and DetectoRS) only have bounding box predictions, AP evaluated using mask IoU and bounding box IoU are both reported.

4.3. Comparison with the State-of-the-Arts

In this subsection, we show the main comparison results between our method and the state-of-the-art methods. The detailed performance comparison is reported in Table 1. These methods can be divided into 2 groups: one-stage methods and two-stage methods. YOLOv3 is the only one-stage method while all other methods are two-stage methods. We mainly

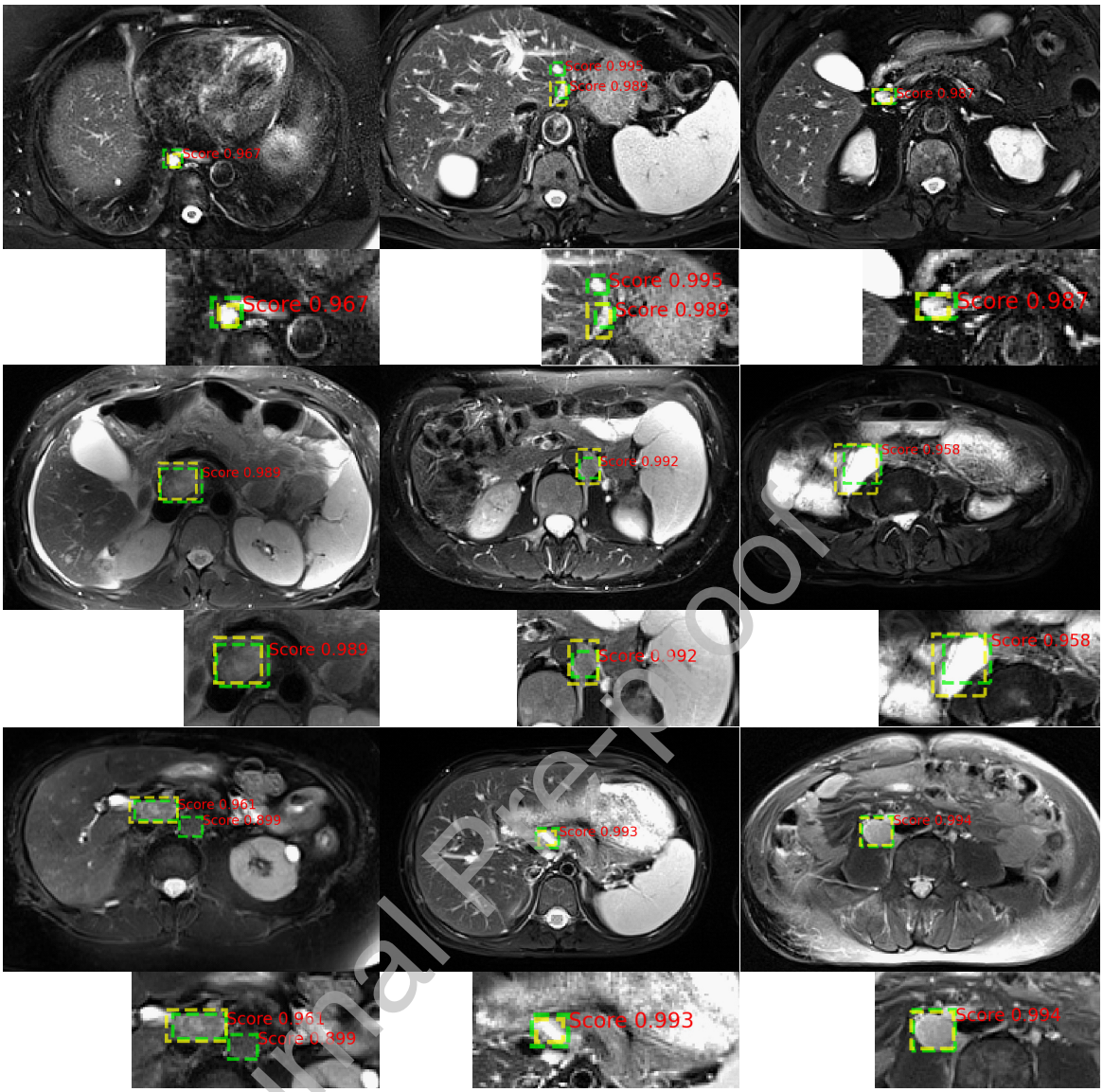


Fig. 5. Visualization of the detection results achieved by our method. Yellow box indicates the ground truth while green box indicates our prediction. Red number indicates the predicted score. The lesion regions are zoomed in for better observation.

compare the two-stage methods because two-stage detectors are usually more flexible and accurate than one-stage detectors. These methods can be also categorized as with and without the segmentation task. We report AP and sensitivity at 5 FPs per image evaluated using both mask IoU and bounding box IoU for detectors with the segmentation task and only using bounding box IoU for detectors without the segmentation task.

Our method achieved the best performance in seven of the eight metrics (as shown in Table 1). For AP_{25} evaluated using bounding box IoU, our method was only slightly worse than DetectoRS. For metrics evaluated using mask IoU, our method outperformed Mask R-CNN and Cascade Mask R-CNN by large margins. For example, for sensitivity at 5 FPs, Mask R-CNN and Cascade Mask R-CNN achieved poor performance while our method improved the sensitivity by more than 17% (IoU: 0.25) and 11% (IoU: 0.50). For metrics evaluated using bounding box IoU, we observed similar trends.

Fig. 5 provides the visualization of the results achieved by our method. The examples show the difficulty of detecting abnormal lymph nodes which are small relative to the field of view and quite similar in texture to "normal" regions. Nevertheless, our method can detect them reliably with few false-positive findings.

We also report the SEN_{25} and SEN_{50} values of all methods on the *3D volume test set* evaluated using both bounding box IoU and mask IoU as shown in Fig. 6. We can see that even if our method is only trained on slices with RECIST bookmarks, the performance of our model outperforms the comparison methods by large margins.

4.4. Ablation Study

In our network, there are three main innovative modules proposed to improve the detection ability: the pseudo mask, the global-local attention, and the multi-task uncertainty loss. The

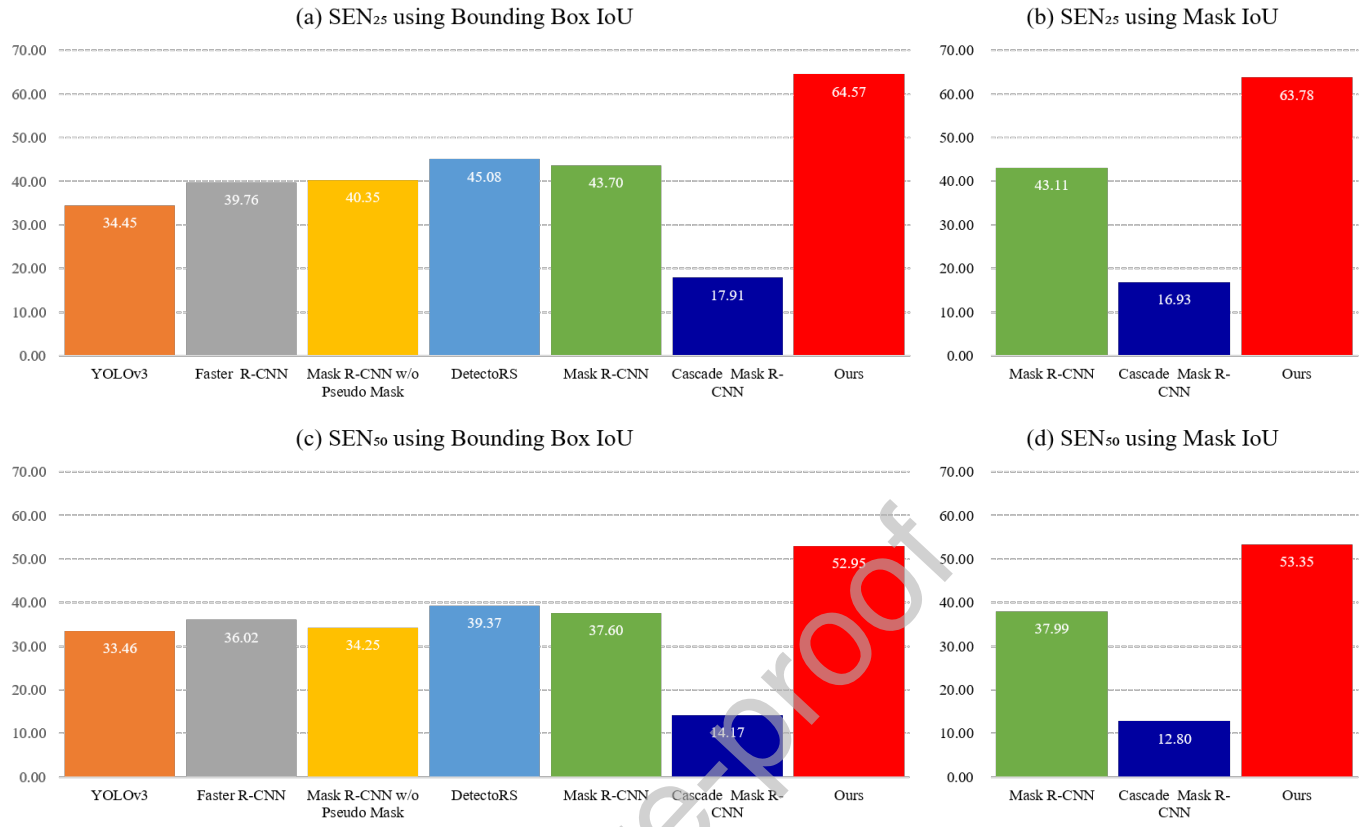


Fig. 6. State-of-the-art comparison on *3D volume test set* in terms of SEN_{25} and SEN_{50}. (a) and (c) show the results evaluated using bounding box IoU and (b) and (d) show the results evaluated using mask IoU. The number at the top of each bar indicates the corresponding metric value.

overall performance of our network has been verified by comparison with the state-of-the-art methods in the previous subsection. In this subsection, we study the effectiveness of each module by discarding one of them at a time in the network training. The corresponding results are reported in Table 2. The results of Mask R-CNN with and without the pseudo mask are also shown in Table 2 for reference.

By comparing Mask R-CNN with and without the pseudo mask, we see that using the pseudo mask instead of the bounding box makes the detection more effective, which is also verified by comparing our method with and without the pseudo mask. The effectiveness of global-local attention and multi-task uncertainty loss can be verified while the corresponding results achieve overall better performance than Mask R-CNN.

To further demonstrate the effectiveness of the pseudo masks, we reported the DSC (Dice Similarity Coefficient, %) and ASD (Average Symmetric Distance, mm) values by comparing the pseudo masks with ground truth in *3D volume test set* with accurate voxel-level annotations. Specifically, the average DSC value is 85.8% and the average ASD value is 0.82mm, which verify the effectiveness of the generated pseudo masks. In addition, we generate bounding boxes based on the pseudo masks and the ground truth masks respectively, and name them Pseudo bounding boxes and True bounding boxes respectively. Then, we compare RECIST bounding boxes (bounding boxes directly drawn from RECIST bookmarks) and Pseudo bounding boxes with True bounding boxes respectively and use average IoU values to evaluate the performance. In *3D volume test set*, the average IoU achieved by Pseudo bounding boxes is 0.68 while 0.63 is achieved by RECIST bounding boxes. The average IoU improvement of 0.05 proves that our generated pseudo masks can provide more accurate bounding boxes for lymph nodes and then can better supervise the detection task.

To further verify the effectiveness of our multi-task uncertainty loss, we train our methods by setting the loss weights manually. The detailed results are reported in Table 3. Since there are 5 tasks in our model, it is hard to traverse all possible weight settings and we show only several representative settings. Table 3 shows that the performance can benefit from adjusting the loss weights of different tasks on some metrics, which verifies the necessity of using the optimal loss weights. Compared with manual settings, our model with the multi-task uncertainty loss is overall better for all metrics, which can effectively avoid the cost of the grid search method to find the optimal loss weights.

4.5. Sensitivity to Global Scale Size

To extract the global context, we use the whole image as the receptive field. Here, we use different global sizes to verify whether our model is sensitive to this setting. To measure this sensitivity, we set the global size to 0.10, 0.25, and 0.50 of the whole image size separately. Different from using the whole image, the receptive field of the compared global sizes is centered on the corresponding target ROI. Fig. 7(a) and (b) show

Table 2. Ablation study of our three modules: pseudo mask, global-local attention, and multi-task uncertainty loss. / means that the corresponding method has no segmentation mask prediction.

Method	Mask				Box			
	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}
Ours	54.57	48.54	69.36	60.12	55.16	47.82	72.25	60.12
Mask R-CNN He et al. (2017)	45.28	41.97	52.02	48.55	46.49	40.76	53.18	48.55
Mask R-CNN w/o Pseudo Mask He et al. (2017)	/	/	/	/	42.77	37.05	47.40	42.77
Ours w/o Pseudo Mask	/	/	/	/	51.33	43.03	67.63	58.38
Ours w/o Global-Local Attention	52.33	47.36	65.90	58.96	52.94	46.71	66.47	59.54
Ours w/o Multi-task Uncertainty Loss	50.09	41.11	59.54	47.98	50.49	40.69	61.27	49.71

Table 3. Performance comparison between using our multi-task uncertainty loss and using the grid search method to set the loss weights. Mask R-CNN with equal weights for different tasks is used as a baseline, and the metric that outperforms the corresponding baseline achieved by other settings is marked in bold.

Setting	Mask				Box			
	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}	AP_{25}	AP_{50}	SEN_{25}	SEN_{50}
Multi-task Uncertainty Loss	**54.57**	**48.54**	**69.36**	**60.12**	**55.16**	**47.82**	**72.25**	**60.12**
$\lambda^r_{cls} = \lambda^r_{box} = \lambda^m_{cls} = \lambda^m_{box} = \lambda^m_{mask} = 1$	50.09	41.11	59.54	47.98	50.49	40.69	61.27	49.71
$\lambda^r_{cls} = 10, \lambda^r_{box} = \lambda^m_{cls} = \lambda^m_{box} = \lambda^m_{mask} = 1$	48.79	40.28	**61.85**	**49.71**	48.79	39.13	**62.43**	**50.29**
$\lambda^r_{box} = 10, \lambda^r_{cls} = \lambda^m_{cls} = \lambda^m_{box} = \lambda^m_{mask} = 1$	**50.16**	38.48	**64.16**	47.40	**51.88**	37.36	**68.21**	**50.29**
$\lambda^r_{cls} = \lambda^r_{box} = 10, \lambda^m_{cls} = \lambda^m_{box} = \lambda^m_{mask} = 1$	46.01	38.30	57.80	45.09	47.21	37.27	61.27	47.40
$\lambda^m_{mask} = 10, \lambda^r_{cls} = \lambda^r_{box} = \lambda^m_{cls} = \lambda^m_{box} = 1$	46.42	35.83	**60.69**	46.24	46.47	36.51	60.69	49.71
$\lambda^m_{cls} = 10, \lambda^r_{cls} = \lambda^r_{box} = \lambda^m_{box} = \lambda^m_{mask} = 1$	45.98	39.86	50.29	43.35	47.24	38.96	52.02	43.93
$\lambda^m_{box} = 10, \lambda^r_{cls} = \lambda^r_{box} = \lambda^m_{cls} = \lambda^m_{mask} = 1$	46.79	38.45	59.54	46.82	47.55	38.40	60.69	49.13
$\lambda^m_{cls} = \lambda^m_{box} = \lambda^m_{mask} = 10, \lambda^r_{cls} = \lambda^r_{box} = 1$	49.11	**41.36**	**62.43**	**52.02**	49.59	39.10	**63.01**	**52.60**

Table 4. Performance comparison on TCIA Lymph Node dataset. AP_{10} **evaluated using mask IoU is for evaluation.**

Method	Abdominal LN	Mediastinal LN
nnDetection	47.0	50.0
nnUNetPlus	31.1	34.2
Ours	51.9	52.2

the results under different global sizes, in which the performance is evaluated using mask IoU. We observed that our implementation is robust to different global sizes since all global sizes achieve similar performance. But using the whole image is more stable under different metrics. The main reason for this may be that the whole image can provide a more stable context compared with other scale sizes to identify the abnormal lymph node among different body parts.

4.6. Sensitivity to Multi-Global Scale

In this subsection, we extend the global context from single-scale implementation to multi-scale implementation by adding an additional global scale to check whether more global information represents better capability for detection. Similar to the settings before, the size of the additional global scale is set to 0.10, 0.25, and 0.50 of the whole image size, respectively. In Fig. 7(c) and (d), the performance evaluated using mask IoU with different sizes of additional global scales is reported. The results with an additional global size of 0.10 are overall better than the others. But the performance is slightly lower for the additional global size of 0.25 or 0.50 relative to the whole image scale. The main reason may be that the neighborhood context provided by the global sizes of 0.25 and 0.50 provide a larger neighborhood context which has a large overlap with the whole image context and may bring redundancy that weakens the effectiveness of the global-local attention module.

4.7. Comparison with Other Lymph Node Detection Methods

We further verify the effectiveness of our method by conducting another comparison on a public lymph node dataset, TCIA Lymph Nodes[1], marked by radiologists at the National Institutes of Health, including a total of 388 mediastinal lymph nodes in CT images of 90 patients and a total of 595 abdominal lymph nodes in 86 patients Roth et al. (2014). Two state-of-the-art methods are used for comparison, nnDetection and nnUNetPlus (a modified nnUNet for detection) Baumgartner et al. (2021), AP with the mask IoU threshold 0.10 ($AP_{0.10}$)

[1] https://wiki.cancerimagingarchive.net/display/Public/CT+Lymph+Nodes

Fig. 7. (a) AP_{25} and AP_{50} achieved by using different global sizes; (b) SEN_{25} and SEN_{50} achieved by using different global sizes; (c) AP_{25} and AP_{50} achieved by adding an additional global scale; (d) SEN_{25} and SEN_{50} achieved by adding an additional global scale. G_s denotes the implementation using global size s and AG_s denotes the implementation with an addition global scale of size s.

Table 5. Illustration of the results of other works using different datasets. #Patient and #LN denote the number of patients and lymph nodes, respectively. Recall@5.85FP and Recall@9FP mean the recall at 5.85 and 9 false-positives per image, respectively. mRecall@0.10-0.50 means the mean recall at a precision range of [0.10, 0.50] with 0.05 interval.

Method	#Patient	#LN	Recall@5.85FP	Recall@9FP	mRecall@0.10-0.50
Bouget et al. (2021)	120	1178	52.4	-	-
Bouget et al. (2019)	15	300	-	**74.2**	-
Yan et al. (2019)	141	651	-	-	72.5
Zhu et al. (2020a)	141	651	-	-	78.2
Zhu et al. (2020b)	141	651	-	-	74.7
Ours	584	821	**73.9**	73.9	**90.3**

is used to be consistent with the reported performance. The detailed results are shown in Table 4. The better performance of our method demonstrates the advantages of our method.

Moreover, we also illustrate our results along with the reported performance of other lymph node detection methods using different datasets, including Bouget et al. (2019); Yan et al. (2019); Zhu et al. (2020a,b); Bouget et al. (2021). Since these approaches used different metrics, we recalculate our results for a fair comparison and the corresponding performance is listed in Table 5. We can see that our method can achieve better performance than other methods, except that our Recall@9FP is slightly lower than Bouget et al. (2019). And the largest number of patients and the smallest number of lymph nodes per patient indicate that our dataset is more diverse and more challenging.

4.8. Comparison on DeepLesion

We also experimented on a large-scale and comprehensive dataset, DeepLesion Yan et al. (2019), which includes over 32K lesions on various body parts in CT scans. We used the official training set to train our model as in Yan et al. (2019, 2020) and evaluated the results on the 800 manually labeled sub-volumes in the test set of DeepLesion. To obtain the 3D volume detection results, we adopted the same strategy as in Yan et al. (2020), namely, stacking the predicted 2D boxes to 3D ones if the intersection over union (IOU) of two 2D boxes in consecutive slices is greater than θ (θ=0.1 in our implementation). The detailed comparison results are shown in Table 6. We can see that when the value of FPs per sub-volume is small (from 0.125 to 1), our method cannot achieve competitive results. And when this value keeps increasing, our method gets competitive results. The main reason is that our method is based on a 2D

backbone instead of a 2.5D backbone and does not design any mechanism for the 3D false prediction reductions. Therefore, when stacking the predicted 2D boxes to 3D ones, our method ignores the 3D context information and then generates many false positive predictions. Therefore, in future work, we will extend our method to 3D implementation to address this drawback.

5. Conclusion

In this paper, we propose a new network with Mask R-CNN as the base model for the detection of universal abnormal lymph nodes in MR images. Our implementation includes the pseudo mask generation, the global-local attention, and the multi-task uncertainty loss that are different from the standard Mask R-CNN. The generated pseudo mask provided more accurate supervision than the original RECIST bookmarks, which alleviates the burden for data annotation. Global-local attention mines more discriminative features for detection while multi-task uncertainty loss reduces the cost to tune the weights of different tasks. Our method was tested on our large-scale lymph node dataset and outperformed other state-of-the-art detectors.

References

Abdulla, W., 2017. Mask R-CNN for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

Alansary, A., Oktay, O., Li, Y., Le Folgoc, L., Hou, B., Vaillant, G., et al., 2019. Evaluating reinforcement learning agents for anatomical landmark detection. Medical image analysis 53, 156–164.

Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., et al., 2017. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more personalized approach to cancer staging. CA: a cancer journal for clinicians 67, 93–99.

Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D., 2010. Automatic detection and segmentation of axillary lymph nodes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 28–36.

Baumgartner, M., Jaeger, P.F., Isensee, F., Maier-Hein, K.H., 2021. nndetection: A self-configuring method for medical object detection. arXiv preprint arXiv:2106.00817 .

Bouget, D., Jørgensen, A., Kiss, G., Leira, H.O., Langø, T., 2019. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in ct data for lung cancer staging. International journal of computer assisted radiology and surgery 14, 977–986.

Bouget, D., Pedersen, A., Vanel, J., Leira, H.O., Langø, T., 2021. Mediastinal lymph nodes segmentation using 3d convolutional neural network ensembles and anatomical priors guiding. arXiv preprint arXiv:2102.06515 .

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.

Cai, Z., Vasconcelos, N., 2019. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence , 11URL: http://dx.doi.org/10.1109/tpami.2019.2956516, doi:10.1109/tpami.2019.2956516.

Cao, H., Liu, H., Song, E., Ma, G., Xu, X., Jin, R., et al., 2020. A two-stage convolutional neural networks for lung nodule detection. IEEE journal of biomedical and health informatics 24, 2006–2015.

Carolus, H., Iuga, A.I., Brosch, T., Wiemker, R., Thiele, F., Höink, A., et al., 2020. Automated detection and segmentation of mediastinal and axillary lymph nodes from ct using foveal fully convolutional networks, in: Medical Imaging 2020: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 113141B.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al., 2019. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 .

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016. Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3640–3649.

Chiang, T.C., Huang, Y.S., Chen, R.T., Huang, C.S., Chang, R.F., 2018. Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation. IEEE transactions on medical imaging 38, 240–249.

Cui, Z., Li, Q., Cao, Z., Liu, N., 2019. Dense attention pyramid networks for multi-scale ship detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 57, 8983–8997.

Debats, O.A., Litjens, G.J., Huisman, H.J., 2019. Lymph node detection in mr lymphography: false positive reduction using multi-view convolutional neural networks. PeerJ 7, e8052.

Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., et al., 2009. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). European journal of cancer 45, 228–247.

Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L., 2020. Camouflaged object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2777–2787.

Girshick, R., 2015. Fast R-CNN, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

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.

Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C., 2020. Augfpn: Improving multi-scale feature learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12595–12604.

Ha, R., Chang, P., Karcich, J., Mutasa, S., Fardanesh, R., Wynn, R.T., et al., 2018. Axillary lymph node evaluation utilizing convolutional neural networks using mri dataset. Journal of Digital Imaging 31, 851–856.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV) .

Huynh, N.A., 2017. Training and detecting objects with YOLO3. https://github.com/experiencor/keras-yolo3.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7482–7491.

Kitaizumi, T., Kuwahata, A., Saichi, K., Sato, T., Igarashi, R., Ohshima, T., et al., 2020. Magnetic field generation system of the magnetic probe with diamond quantum sensor and ferromagnetic materials for the detection of sentinel lymph nodes with magnetic nanoparticles. IEEE Transactions on Magnetics .

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J., 2020. Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing 29, 7389–7398.

Kuwahata, A., Taruno, K., Kurita, T., Makita, M., Chikaki, S., Saito, I., et al., 2020. Magnetic nanoparticle detection by utilizing nonlinear magnetization for sentinel lymph nodes of breast cancer patients. IEEE Transactions on Magnetics .

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., et al., 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. arXiv preprint arXiv:2006.04388 .

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.

Liu, J., Hoffman, J., Zhao, J., Yao, J., Lu, L., Kim, L., et al., 2016a. Mediastinal lymph node detection and station mapping on chest ct using spatial priors and random forest. Medical physics 43, 4362–4374.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., et al., 2020. Deep learning for generic object detection: A survey. International journal of computer vision 128, 261–318.

Liu, S., Wang, H., Zhang, C., Dong, J., Liu, S., Xu, R., Tian, C., 2019. In vivo photoacoustic sentinel lymph node imaging using clinically-approved carbon nanoparticles. IEEE Transactions on Biomedical Engineering .

Table 6. Sensitivity (%) at different FPs per sub-volume on the manually labeled test set of DeepLesion.

Method	AFP	Multi-dataset	Proposal fusion	MAM	FPR	0.125	0.25	0.5	1	2	4	8	Avg.
Yan et al. (2019)						11.2	16.3	24.3	32.8	41.6	50.9	60.1	33.9
Yan et al. (2020)	✓					15.8	21.4	27.9	35.9	43.4	52.0	60.9	36.8
	✓	✓				14.3	21.5	28.2	35.1	44.4	53.9	63.4	37.3
	✓	✓	✓			15.9	22.8	30.1	37.7	46.7	56.6	66.1	39.4
		✓	✓	✓		18.3	26.3	34.1	44.8	55.5	65.4	75.4	45.7
	✓			✓		22.0	28.4	36.6	45.2	55.0	65.5	75.0	46.8
	✓	✓		✓		21.3	28.3	37.1	46.7	55.5	66.2	75.9	47.3
	✓	✓	✓	✓		21.6	29.9	37.6	46.7	56.7	65.8	75.3	47.6
	✓	✓	✓	✓	✓	23.7	31.6	40.3	50.0	59.6	69.5	78.0	50.4
Ours						0.8	3.4	7.2	13.1	27.1	49.3	67.3	24.0

AFP, Multi-dataset, Proposal fusion, MAM, and FPR are the proposed modules in Yan et al. (2020). AFP: Anchor-free proposal network; MAM: Three missing annotation mining strategies; FPR: 3D false positive reduction network.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., et al., 2016b. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

Ma, Y., Peng, Y., 2020. Lymph node detection method based on multisource transfer learning and convolutional neural network. International Journal of Imaging Systems and Technology 30, 298–310.

Oda, H., Roth, H.R., Bhatia, K.K., Oda, M., Kitasaka, T., Iwano, S., et al., 2018. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest ct images, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 1057502.

Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E., 2020. Imbalance problems in object detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H., 2019. Depth-induced multi-scale recurrent attention network for saliency detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7254–7263.

Qiao, S., Chen, L.C., Yuille, A., 2020. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. arXiv preprint arXiv:2006.02334 .

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271.

Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., et al., 2014. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 520–527.

Rother, C., Kolmogorov, V., Blake, A., 2012. Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics 23, 3.

Shao, Q., Gong, L., Ma, K., Liu, H., Zheng, Y., 2019. Attentive ct lesion detection using deep pyramid inference with multi-scale booster, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 301–309.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35, 1285–1298.

Tang, Y., Harrison, A.P., Bagheri, M., Xiao, J., Summers, R.M., 2018. Semi-automatic recist labeling on ct scans with cascaded convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 405–413.

Tang, Y.B., Yan, K., Tang, Y.X., Liu, J., Xiao, J., Summers, R.M., 2019. Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 833–836.

Tao, Q., Ge, Z., Cai, J., Yin, J., See, S., 2019. Improving deep lesion detection using 3d contextual and spatial attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 185–193.

Wang, H., Huang, H., Wang, J., Wei, M., Yi, Z., Wang, Z., et al., 2021. An intelligent system of pelvic lymph node detection. International Journal of Intelligent Systems .

Wang, H., Wang, Q., Gao, M., Li, P., Zuo, W., 2018. Multi-scale location-aware kernel representation for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1248–1257.

Wang, R., Fan, J., Li, Y., 2020a. Deep multi-scale fusion neural network for multi-class arrhythmia detection. IEEE journal of biomedical and health informatics 24, 2461–2472.

Wang, S., Cong, Y., Zhu, H., Chen, X., Qu, L., Fan, H., Zhang, Q., Liu, M., 2020b. Multi-scale context-guided deep network for automated lesion segmentation with endoscopy images of gastrointestinal tract. IEEE Journal of Biomedical and Health Informatics 25, 514–525.

Wang, S., Wang, Q., Shao, Y., Qu, L., Lian, C., Lian, J., Shen, D., 2020c. Iterative label denoising network: Segmenting male pelvic organs in ct from 3d bounding box annotations. IEEE Transactions on Biomedical Engineering 67, 2710–2720.

Xie, C., Cao, S., Wei, D., Zhou, H., Ma, K., Zhang, X., et al., 2021. Recist-net: Lesion detection via grouping keypoints on recist-based annotation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 921–924.

Yan, K., Bagheri, M., Summers, R.M., 2018a. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 511–519.

Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y.b., Tang, Y.X., Huang, L., Xiao, J., Lu, L., 2020. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. IEEE Transactions on Medical Imaging .

Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., et al., 2019. Mulan: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 194–202.

Yan, K., Wang, X., Lu, L., Summers, R.M., 2018b. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging 5, 036501.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X., 2019. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems 30, 3212–3232.

Zhu, Z., Jin, D., Yan, K., Ho, T.Y., Ye, X., Guo, D., et al., 2020a. Lymph node gross tumor volume detection and segmentation via distance-based gating using 3d ct/pet imaging in radiotherapy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 753–762.

Zhu, Z., Yan, K., Jin, D., Cai, J., Ho, T.Y., Harrison, A.P., et al., 2020b. Detecting scatteredly-distributed, small, andcritically important objects in 3d oncologyimaging via decision stratification. arXiv preprint arXiv:2005.13705 .

Zlocha, M., Dou, Q., Glocker, B., 2019. Improving retinanet for ct lesion detection with dense masks from weak recist labels, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 402–410.

AUTHOR DECLARATION

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work entitled "Global-Local Attention Network with Multi-task Uncertainty Loss for Abnormal Lymph Node Detection in MR Images". We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We further confirm that any aspect of the work covered in this manuscript that has involved either experimental animals or human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

Credit Author Statement

Shuai Wang: Conceptualization, Methodology, Software, Investigation, Writing original draft

Yingying Zhu: Methodology, Writing - review & editing

Sungwon Lee: Data Annotation, Writing - review & editing

Daniel C. Elton: Software, Writing - review & editing

Thomas C. Shen: Data Collection, Writing - review & editing

Youbao Tang: Software, Writing - review & editing

Yifan Peng: Writing - review & editing

Zhiyong Lu: Writing - review & editing

Ronald M. Summers: Project administration, Writing - review & editing