Interpretable Heterogeneous Teacher-Student Learning Framework for Hybrid-Supervised Pulmonary Nodule Detection

Guangyu Huang, Yan Yan, *Senior Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* Wentao Zhu, and Xiongbiao Luo

***Abstract*—Existing pulmonary nodule detection methods often train models in a fully-supervised setting that requires strong labels (i.e., bounding box labels) as label information. However, manual annotation of bounding boxes in CT images is very time-consuming and labor-intensive. To alleviate the annotation burden, in this paper, we investigate pulmonary nodule detection by leveraging both strong labels and weak labels (i.e., center point labels) for training, and propose a novel hybrid-supervised pulmonary nodule detection (HND) method. The training of HND involves a heterogeneous teacher-student learning framework in two stages. In the first stage, we design a point-based consistency calibration network (PCC-Net) as a teacher, which is pre-trained to generate high-quality pseudo bounding box labels given point-augmented CT images as inputs. In the second stage, we develop an information bottleneck-guided pulmonary nodule detection network (IBD-Net) as a student to perform pulmonary nodule detection. In particular, we introduce information bottleneck to learn reliable pulmonary nodule-specific heatmaps under the guidance of PCC-Net, largely enhancing the model's interpretability and improving the final detection performance. Based on the above designs, our method can effectively detect pulmonary nodule regions with only a limited number of bounding box labels. Experimental results on the public pulmonary nodule detection dataset LUNA16 show that our HND method achieves an excellent balance between the annotation cost and the detection performance. The code will be released soon.**

***Index Terms*—Heatmap learning, Hybrid-supervised learning, Pseudo label generation, pulmonary nodule detection.**

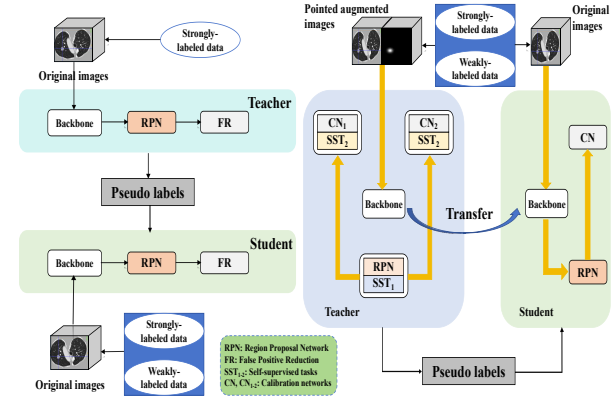
I. Introduction

LUNG cancer is the leading cause of cancer-related deaths worldwide over the past few years [1]. Computed tomography (CT) examination of pulmonary nodules often serves as a crucial indicator of lung cancer [2]. Early diagnosis of pulmonary nodules can significantly decrease the incidence of lung cancer. Accordingly, a variety of pulmonary nodule detection methods [3]–[5] have been developed and they often train models in the fully-supervised setting, which requires strongly-labeled CT images (i.e., the bounding box labels of all pulmonary nodules or tumors). However, accurately annotating CT images is very time-consuming and labor-intensive due to the significant diversity of the location, size, and appearance of pulmonary nodules.

G. Huang, Y. Yan, and X. Luo are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, and the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China (e-mail: guangyuhuang@stu.xmu.edu.cn; yanyan@xmu.edu.cn; xbluo@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

Wentao Zhu is with the Zhejiang Lab, Hangzhou 311121, China (email: wentao.zhu@zhejianglab.com).



(a) Homogeneous Teacher-Student Framework (b) Heterogeneous Teacher-Student Framework

Fig. 1: The comparison between (a) the homogeneous teacher-student framework and (b) our heterogeneous teacher-student framework.

To alleviate laborious annotations, several semi-supervised methods [6], [7], which leverage both strongly-labeled and unlabeled data for model training, have been proposed. Meanwhile, some recent efforts [2], [8], [9] have been made based on weakly-supervised learning with weak labels (such as electronic medical records (EMR), image-level labels, point labels, or scribble labels). Although the annotation cost is greatly reduced by using semi-supervised or weakly-supervised methods, their performance is still much worse than fully-supervised methods.

To balance the annotation cost and the detection performance, in this paper, we study a hybrid-supervised setting, which involves a relatively small amount of strong labels and a large amount of weak labels. Compared with the fully-supervised setting, the annotation cost in the hybrid-supervised setting is much smaller. Moreover, the hybrid-supervised setting can be easily implemented in practice since only a small amount of strongly-labeled data are required. In this paper, considering that the center point label takes less labeling cost than the scribble label and involves richer information than the image-level label, we choose the center point label as

the weak label. Generally, annotating the center point of a nodule is usually faster and simpler than annotating a bounding box. According to [10], annotating a bounding box consumes about 18 times more than clicking a center point of an object. Annotators can observe the nodule in the image or volume data and select a representative point as the center point of the nodule. In contrast, annotating the bounding box of a nodule requires more operations and time. Annotators need to draw a rectangular box around the nodule in the image or volume data, ensuring that the bounding box accurately encompasses the contour of the nodule [11]. This often involves adjustments based on the shape, size, and location of the nodule to obtain the best bounding box result.

Existing hybrid-supervised methods [12]–[14] mainly work on natural image segmentation or detection tasks. For example, Luo *et al.* [14] deal with strongly-labeled and weakly-labeled data separately by designing a strong-weak dual-branch network. Pan *et al.* [15] introduce a label-efficient hybrid-supervised framework to perform medical image segmentation. However, these methods only employ weak annotations of weakly-labeled data for training without generating strong labels for these data.

Recently, some methods [13], [16] leverage a homogeneous teacher-student framework, which takes the same form of inputs for both the teacher and student. This framework generates pseudo labels for weakly-labeled data by using the teacher model trained on strongly-labeled data and guides the learning of the student model with both strong labels and pseudo labels, as illustrated in Fig. 1 (a).

On the one hand, the pseudo labels (which are generated based on the model trained only with strongly-labeled data) may contain much noise, increasing the difficulty of learning an accurate student model. On the other hand, these methods often provide prediction results without justification, lacking interpretability and transparency. Although some recent methods [12] generate the class activation maps (CAM) from the teacher model as the guidance of the student model, the detection performance can be significantly affected if CAM fails to give reliable heatmaps (since CAM tends to give many false positives and overestimate the response regions [17]). Note that different from the objects in natural images, pulmonary nodule regions are small abnormal areas in the lungs, demanding a more dedicated and flexible way for hybrid-supervised learning. Hence, how to properly combine strongly-labeled and weakly-labeled CT images to generate *high-quality pseudo labels* and learn *reliable heatmaps* merits further investigation.

To address the above problems, we propose a hybrid-supervised pulmonary nodule detection (HND) method, which can generate high-quality pseudo bounding box labels and learn reliable pulmonary nodule-specific heatmaps based on a novel heterogeneous teacher-student learning framework, for pulmonary nodule detection, as shown in Fig. 1 (b). The training of HND involves two stages. In the first stage, we propose and pre-train a teacher model, called point-based consistency calibration network (PCC-Net), to obtain high-quality pseudo bounding box labels given point-augmented CT images as inputs. In the teacher model, we design two different self-supervised tasks (supervised by the consistency regression loss and the consistency classification loss, respectively) and apply them to the regression task and the classification task, respectively. Inspired by the multi-path detection calibration network (PDC-Net) [18], we also develop a two-path calibration network (TCN) to extract features from different layers for classification and regression. Different from PDC-Net which employs multi-path calibration, we incorporate self-supervised tasks with TCN to train a consistency calibration network. By integrating TCN into the teacher model, we leverage features from different layers of the model to formulate the consistency classification loss, leading to more accurate classification. In the second stage, based on the pre-trained PCC-Net, we develop and train a student model, called information bottleneck-guided pulmonary nodule detection network (IBD-Net), to detect pulmonary nodules of the input CT images by transferring the knowledge from the teacher model to the student model.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to study pulmonary nodule detection in the hybrid-supervised setting. We develop a novel HND method with a heterogeneous teacher-student learning framework for detecting pulmonary nodule regions. As a result, our method can obtain an effective student model with only a limited number of strong labels and a relatively large number of weak labels.
- We develop PCC-Net and integrate self-supervised tasks into its regression and classification networks, generating high-quality pseudo bounding box labels for weakly-labeled CT images. Furthermore, we design an information bottleneck-guided heatmap generation (IHG) module to learn reliable pulmonary nodule-specific heatmaps. Such a way provides intuitive explanations for the model, enhancing its transparency and interpretability.
- Our method performs favorably against state-of-the-art methods, demonstrating the great potential of hybrid-supervised learning for pulmonary nodule detection.

The remainder of this paper is organized as follows. We begin with a review of related work in Section II, followed by a detailed description of our proposed method in Section III. Then, in Section IV, we conduct extensive experiments on the representative pulmonary nodule detection dataset. Finally, we draw the conclusion in Section V.

II. Related Work

In this section, we review the methods closely related to our method. We first introduce pulmonary nodule detection in Section II-A. Then, we briefly review weakly-supervised and hybrid-supervised learning in Section II-B. Finally, we review the information bottleneck in Section II-C.

A. Pulmonary Nodule Detection

Pulmonary nodules have been extensively studied, including pulmonary nodule classification, detection, and segmentation. Zhu *et al.* [19] introduces a comprehensive framework that

includes both a classification network and an explanation network to distinguish between benign and malignant pulmonary nodules. Wang *et al.* [20] develop a novel DSNet architecture that incorporates a detailed representation transfer module and a soft mask-based adversarial training framework to perform accurate lung nodule segmentation. Existing automated pulmonary nodule detection methods typically involve two steps: nodule candidate detection and false positive reduction. Early methods [21] are based on the 2D image analysis. Recently, 3D-based methods have become very common. For example, Tang *et al.* [3] jointly train the detection and segmentation tasks and use decoupled feature maps to improve the accuracy of detecting pulmonary nodules. Wang *et al.* [22] utilize 3D Faster R-CNN for nodule detection and subsequently employ a deep 3D dual-path network to classify the nodules as benign or malignant. Mei *et al.* [4] introduce a slice-grouped non-local (SGNL) module to the encoder network, which can capture long-range dependencies of one slice group by considering cross-channel information, for pulmonary nodule detection. Based on SGNL, Xu *et al.* [23] propose to use short-distance slice grouping (SSG) and long-distance slice grouping (LSG) alternately. In this way, any similarities across multiple slices (regardless of their distance from each other) can be taken into account.

B. Weakly-Supervised and Hybrid-Supervised Learning

The annotation cost of fully-supervised learning is very high. To reduce the annotation cost, weakly-supervised learning [24]–[26] is developed by leveraging weak labels (such as image-level labels, point labels, scribble labels, electronic medical records (EMR), or response evaluation criteria in solid tumors (RECIST)) that are less laborious to be annotated than strong labels. Ibrahim *et al.* [13] train two models to generate pseudo labels of weakly-labeled data for image segmentation, and they employ a self-correction module to improve the quality of pseudo labels. Chen *et al.* [27] propose a casual CAM (C-CAM) method for weakly-supervised semantic segmentation on medical images. Based on image-level labels, two casual chains are designed to address the problem of ambiguous boundaries and co-occurrence. Note that Momoki *et al.* [2] address the challenge of limited annotated training data by using radiology reports for the automatic characterization of pulmonary nodules, facilitating radiologists in determining malignancy. Unlike [2], we annotate simple center points in the CT images to reduce the burden of annotation costs.

Zhu *et al.* [8] apply the expectation-maximization algorithm to exploit nodule-related information in EMR and propose a deep 3D convolutional neural network for pulmonary nodule detection. Yang *et al.* [9] leverage information (such as the presence of nodules in CT scans, the number of nodules, and related slice information) from EMR to address the problem of weakly-supervised pulmonary nodule detection. Shen *et al.* [28] introduce a two-stage weakly-supervised lung cancer detection and diagnosis network WS-LungNet. WS-LungNet consists of semi-supervised computer-aided detection and cross-nodule attention computer-aided diagnosis for segmenting nodules in CT images and performing patient-level diagnoses, respectively. Feng *et al.* [29] train a convolutional neural network for weakly-supervised segmentation of pulmonary nodules.

Despite the great progress achieved by weakly-supervised learning methods, their performance is still far from being satisfactory. Recently, hybrid-supervised learning, which jointly takes advantage of both strong and weak labels during training, has received much attention. Ning *et al.* [30] leverage a marco-micro framework to segment AS-OCT images. Pan *et al.* [15] propose a label-efficient hybrid-supervised framework for medical image segmentation.

In this paper, we study pulmonary nodule detection in a hybrid-supervised setting, which is an important but little-explored task. Notably, considering that pulmonary nodule regions are small abnormal areas in the lungs, we design a heterogeneous teacher-student learning framework to generate high-quality pseudo labels and learn reliable heatmaps. Such a way greatly reduces false positives and improves the detection performance.

C. Information Bottleneck

Information bottleneck [31], [32] characterizes a limit on the amount of mutual information between the original input and the latent representation obtained from the encoder. Li *et al.* [33] employ the information bottleneck theory to extract the minimal sufficient statistics of WSI. By leveraging the principle of information bottleneck, information bottleneck attribution (IBA) [34] has been proposed to provide interpretable visual explanations. IBA introduces noise into a feature map of a pretrained model to limit the information flow. In this way, an IBA heatmap can be obtained by summing along the channel axis of the Kullback-Leibler (KL) divergence term of the information loss. In addition to IBA, various other attribution methods (e.g., Grad-CAM [35], Layer-Wise Relevance Propagation (LRP) [36], IBA, and InputIBA [37]) have gained widespread use. Compared with Grad-CAM, IBA can give more accurate heatmaps, as validated in [34]. Demir *et al.* [17] design a visual attribution method using the IBA and show the superiority of the IBA over Grad-CAM in medical imaging diagnosis and prognosis. Wang *et al.* [38] employ the IBA heatmap obtained from the cancer classification branch to give location guidance to the tumor segmentation branch, leading to performance improvements. The IBA heatmap is directly adopted as a fixed weighting matrix for enhancing feature representations.

In this paper, instead of using fixed heatmaps, we design a module to learn reliable heatmaps without relying on extra models or point labels, enabling the network to capture effective information. By doing this, we can generate reliable heatmaps for predictions in the student model without using the teacher model. Such a way greatly reduces the computational cost for inference.

III. Proposed Method

In this section, we introduce our interpretable heterogeneous teacher-student learning method for hybrid-supervised pulmonary nodule detection. First, we give an overview of

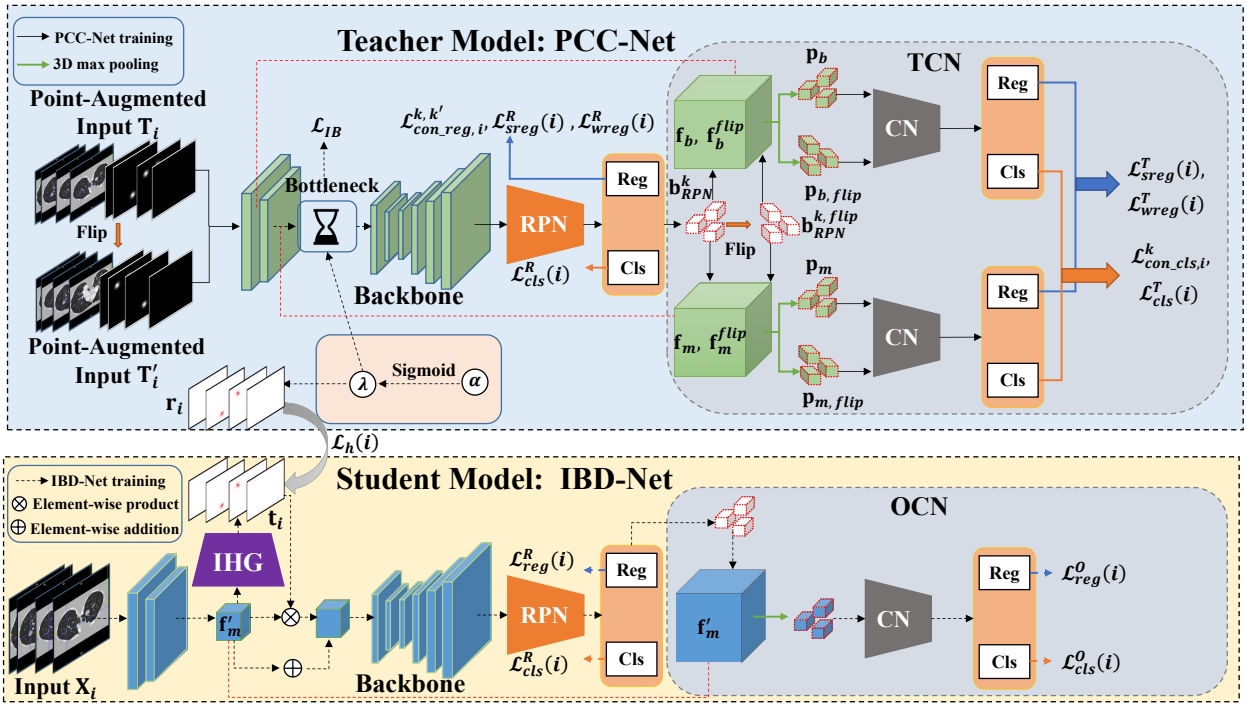


Fig. 2: The network architecture of HND. It consists of a two-stage heterogenous teacher-student learning framework (first pre-train PCC-Net (Sec. III-B) and then train IBD-Net (Sec. III-C)). In the first stage, a point-based consistency calibration network (PCC-Net) is pretrained as the teacher model to generate high-quality pseudo labels given point-augmented images as inputs. In the second stage, the student model (IBD-Net) leverages information bottleneck to learn reliable heatmaps and transfer knowledge from the teacher model. Details of symbols in the figure are given in Sec. III.

our method in Section III-A. Then, we introduce the key components (including the PCC-Net and the IBD-Net) of our proposed method in Sections III-B and III-C, respectively.

A. Overview

In this paper, we propose a novel hybrid-supervised pulmonary nodule detection (HND) method based on a heterogeneous teacher-student learning framework. The network architecture of HND is given in Fig. 2. The training of HND involves two stages. In the first stage, a teacher PCC-Net is pre-trained to generate high-quality pseudo bounding box labels given point-augmented CT images as inputs. In the second stage, a student IBD-Net is trained with both the original labels and generated pseudo bounding box labels for all CT images. In IBD-Net, we leverage information bottleneck to obtain pulmonary nodule-specific heatmaps from PCC-Net. Based on these heatmaps, an IHG module is designed to learn reliable heatmaps in IBD-Net. Note that different from the conventional homogeneous teacher-student framework (which usually considers the same inputs for both the teacher and student), our heterogeneous teacher-student learning framework leverages different inputs for the teacher and student while exploiting the rich knowledge in the learned teacher model to guide the training of the student model. Based on the framework, PCC-Net can generate high-quality pseudo labels and guide IBD-Net to learn reliable heatmaps, greatly enhancing the detection performance.

B. Point-Based Consistency Calibration Network (PCC-Net)

PCC-Net consists of a backbone, a 3D region proposal network (RPN), and a two-path calibration network (TCN). Following [4], 3D RPN consists of a $3\times3\times3$ convolutional layer followed by two parallel $1\times1\times1$ convolutional layers to predict classification probability and regression terms for each voxel. To make full use of center point labels, we incorporate the point representations into CT images and use them as the inputs of PCC-Net. Based on this, we apply self-supervised learning to the input and its flipped version, where we impose the consistency loss on both regression and classification networks, enhancing the network learning ability.

Specifically, we generate a Gaussian heatmap by fitting a Gaussian (with a fixed variance) to the center point of each nodule region in the i-th input CT image $\mathbf{X}_i$, so that the regions around the point are activated. In the Gaussian heatmap, the values near the center point are close to 1 while those far from the center point are close to 0. Then, the Gaussian heatmap is concatenated with the i-th CT image along the channel dimension and reshaped into a 2-channel input tensor (a point-augmented CT image) $\mathbf{T}_i$. Next, the input tensor $\mathbf{T}_i$ is flipped along the z-axis to obtain the flipped version $\mathbf{T}'_i$. Both $\mathbf{T}_i$ and

$\mathbf{T}'_i$, which involve the point representations, are used as the inputs of PCC-Net.

Consistency Regression Loss. Given $\mathbf{T}_i$ and $\mathbf{T}'_i$ as inputs, we denote the regression results (predicted by RPN) of the k-th candidate bounding box w.r.t. $\mathbf{T}_i$ and the k'-th candidate bounding box (the opposite location along the z-axis of the k-th candidate bounding box) w.r.t. $\mathbf{T}'_i$ as $\mathbf{d}^k_{RPN} = [\Delta cz^k, \Delta cy^k, \Delta cx^k, \Delta d^k, \Delta h^k, \Delta w^k]$ and $\mathbf{d}^{k'}_{RPN} = [\Delta \hat{c}z^{k'}, \Delta \hat{c}y^{k'}, \Delta \hat{c}x^{k'}, \Delta d^{k'}, \Delta h^{k'}, \Delta w^{k'}]$, respectively, where the elements in both $\mathbf{d}^k_{RPN}$ and $\mathbf{d}^{k'}_{RPN}$ represent the displacements of the center and scale coefficients of a candidate box. Thus, the consistency regression loss for the k-th and k'-th candidate boxes is

$$\mathcal{L}^{k,k'}_{con\_reg,i} = \frac{1}{6}\left(\left\|\Delta cz^k - \left(-\Delta \hat{c}z^{k'}\right)\right\|^2 + \left\|\Delta cy^k - \Delta \hat{c}y^{k'}\right\|^2 + \left\|\Delta cx^k - \Delta \hat{c}x^{k'}\right\|^2 + \left\|\Delta d^k - \Delta \hat{d}^{k'}\right\|^2 + \left\|\Delta h^k - \Delta \hat{h}^{k'}\right\|^2 + \left\|\Delta w^k - \Delta \hat{w}^{k'}\right\|^2\right). \quad (1)$$

where $\|\cdot\|$ denotes the L_2 norm.

Consistency Classification Loss. Given $\mathbf{T}_i$ and $\mathbf{T}'_i$ as inputs, we denote the extracted features at the middle layer of the backbone as $\mathbf{f}_m$ and $\mathbf{f}^{flip}_m$, respectively, while the features at the bottom layer of the backbone as $\mathbf{f}_b$ and $\mathbf{f}^{flip}_b$, respectively. The k-th bounding box $\mathbf{b}^k_{RPN}$ predicted by RPN is first flipped along the z-axis and we obtain $\mathbf{b}^{k,flip}_{RPN}$. Then, the features $\mathbf{f}_m$ and $\mathbf{f}^{flip}_m$ are respectively cropped by using $\mathbf{b}^k_{RPN}$ and $\mathbf{b}^{k,flip}_{RPN}$, followed by a 3D max pooling layer, to obtain the features $\mathbf{p}_m$ and $\mathbf{p}_{m,flip}$, respectively. Similarly, we repeat the same operations for $\mathbf{f}_b$ and $\mathbf{f}^{flip}_b$, and obtain $\mathbf{p}_b$ and $\mathbf{p}_{b,flip}$, respectively. Finally, $\mathbf{p}_m$ and $\mathbf{p}_{m,flip}$ are fed into one path of TCN (consisting of two fully-connected layers) to get the classification scores $s^k_{TCN_1}$ and $s^{k,flip}_{TCN_1}$, respectively, for the k-th candidate bounding box. Meanwhile, $\mathbf{p}_b$ and $\mathbf{p}_{b,flip}$ are fed into another path of TCN (consisting of two fully-connected layers) to get the classification scores $s^k_{TCN_2}$ and $s^{k,flip}_{TCN_2}$, respectively. Hence, the consistency classification loss for the k-th candidate bounding box is defined as

$$\mathcal{L}^k_{con\_cls,i} = \mathrm{Q}\left(s^k_{TCN_1}, s^{k,flip}_{TCN_1}\right) + \mathrm{Q}\left(s^k_{TCN_2}, s^{k,flip}_{TCN_2}\right), \quad (2)$$

where $\mathrm{Q}(\cdot)$ denotes the Jensen-Shannon Divergence.

Then, the overall consistency loss for the i-th point-augmented CT image is obtained by the average of loss values from all bounding box pairs:

$$\mathcal{L}_{cons}(i) = \mathbb{E}(\mathcal{L}^{k,k'}_{con\_reg,i} + \mathcal{L}^k_{con\_cls,i}). \quad (3)$$

where '$\mathbb{E}$' denotes the expectation operation.

Joint Loss. The joint loss for PCC-Net is defined as

$$\mathcal{L}_{PCC} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{cons}(i) + \mathcal{L}^{R+T}_{cls}(i)) + \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}^{R+T}_{sreg}(i) + \frac{1}{N_w}\sum_{i=1}^{N_w}\mathcal{L}^{R+T}_{wreg}(i), \quad (4)$$

where $\mathcal{L}^{R+T}_{cls}(i) = \mathcal{L}^{R}_{cls}(i) + \mathcal{L}^{T}_{cls}(i)$, $\mathcal{L}^{R+T}_{sreg}(i) = \mathcal{L}^{R}_{sreg}(i) + \mathcal{L}^{T}_{sreg}(i)$, and $\mathcal{L}^{R+T}_{wreg}(i) = \mathcal{L}^{R}_{wreg}(i) + \mathcal{L}^{T}_{wreg}(i)$. N, N_s, and N_w represent the numbers of total training images, strongly-labeled images, and weakly-labeled images, respectively. $\mathcal{L}^{R+T}_{cls}(i)$ denotes the classification loss on RPN and TCN for the i-th image. $\mathcal{L}^{R+T}_{sreg}(i)$ and $\mathcal{L}^{R+T}_{wreg}(i)$ denote the regression losses on RPN and TCN for the i-th strongly-labeled image and the i-th weakly-labeled image, respectively. Note that weak annotations only consist of point coordinates (i.e., the center point of a bounding box). Hence, we can only calculate the regression losses based on these coordinates.

We exploit consistency learning on a point-augmented CT image and its flipped version, where we impose constraints on both the regression and classification of candidate bounding boxes. Note that the flipping transformation is used due to its simplicity and effectiveness in augmenting training data. In particular, the flipping can help capture variations in the nodule orientation and improve the model's learning ability. Moreover, the predicted bounding boxes are comprised of the regression results of RPN and the ensemble of the classification results of RPN and TCN. RPN mainly focuses on localization while TCN focuses on classification. Hence, we apply the consistency regression loss and the consistency classification losses to different modules (RPN and TCN). This greatly facilitates the model to learn the scale knowledge. In this way, PCC-Net can generate high-quality pseudo bounding box labels for weakly-labeled CT images. Note that CSD [39] introduces the consistency constraints on the final classifier only. Different from CSD, we impose the consistency regression loss on RPN and the consistency classification loss on TCN. This enables RPN and TCN to focus on their respective tasks.

C. Information Bottleneck-Guided Pulmonary Nodule Detection Network (IBD-Net)

To ensure the generalization capability, IBD-Net adopts a simple structure, which consists of a backbone, an RPN, and a one-path calibration network (OCN). The structures of the backbone and RPN is the same as those of PCC-Net. One simple way to train IBD-Net is to directly use all the CT images (with the original labels and generated pseudo bounding box labels) as training data. However, this way does not well exploit the pre-trained teacher PCC-Net that incorporates the point supervision information. Therefore, it is desirable to train the student IBD-Net under the guidance of PCC-Net. To achieve this, knowledge distillation, which uses the features of the teacher to guide the learning of the student, is a natural choice. Unfortunately, naive knowledge distillation is not applicable to our task, since PCC-Net and IBD-Net use different inputs. Here, we introduce information bottleneck (IB) to generate pulmonary nodule-specific heatmaps from PCC-Net. Based on this, we develop an IHG module to learn reliable heatmaps (rather than directly using the IB-generated heatmap as a weighting matrix in MIB-Net [38] which requires the annotated information for model learning).

Information Bottleneck Loss. We insert an information bottleneck layer into the middle layer of the backbone in the pre-trained PCC-Net. Suppose that the input of the information

bottleneck layer is denoted as $\mathbf{F}_i$ and its predicted label is denoted as $\mathbf{Y}_i$. Generally, all available information in $\mathbf{F}_i$ is used to predict $\mathbf{Y}_i$. The information bottleneck can depict a limitation of available information. It introduces a variable $\mathbf{M}_i$ to limit the information used to predict the label $\mathbf{Y}_i$. Mathematically, the information bottleneck maximizes the shared information between the variable $\mathbf{M}_i$ and the label $\mathbf{Y}_i$ while minimizing the shared information between the variable $\mathbf{M}_i$ and the input $\mathbf{F}_i$, that is,

$$\max I\left[\mathbf{Y}_i;\mathbf{M}_i\right]-\beta I\left[\mathbf{F}_i;\mathbf{M}_i\right], \tag{5}$$

where $I[;]$ represents the mutual information and the parameter β determines the trade-off between accurate label prediction and minimal utilization of information from $\mathbf{F}_i$. Typically, $\mathbf{M}_i$ is the feature representation calculated by adding noise to $\mathbf{F}_i$, i.e.,

$$\mathbf{M}_i=\gamma\mathbf{F}_i+(1-\gamma)\,\epsilon, \tag{6}$$

where the tensor γ has the same dimension as $\mathbf{F}_i$, and it controls the signal reduction and noise injection. ϵ is the noise that has the same mean and variance as $\mathbf{F}_i$. Usually, γ is set to $\gamma=\mathrm{sigmoid}(\alpha)$ and α is the learnable parameters of the information bottleneck (initialized as 5 as done in [34]). Hence, the information loss is defined as

$$\mathcal{L}_{info}=\mathbb{E}_{\mathbf{F}_i}\left[D_{\mathrm{KL}}\left[P\left(\mathbf{M}_i|\mathbf{F}_i\right)||\,Q\left(\mathbf{M}_i\right)\right]\right], \tag{7}$$

where $P\left(\mathbf{M}_i|\mathbf{F}_i\right)$ represents the probability distribution and $Q(\mathbf{M}_i)=\mathcal{N}\left(\mu_{\mathrm{F}_i},\sigma_{\mathrm{F}_i}\right)$. μ_{F_i} and σ_{F_i} are the mean and variance of $\mathbf{F}_i$, respectively.

Based on the above, we optimize the information bottleneck by using the following IB loss:

$$\mathcal{L}_{IB}=\mathcal{L}_s^{PCC}+\chi\mathcal{L}_{info}, \tag{8}$$

where $\mathcal{L}_s^{PCC}$ denotes the detection loss (including the classification and regression losses) for PCC-Net. χ is a parameter empirically set to 10 as done in [34].

Accordingly, we calculate the three-dimensional heatmap $\mathbf{r}_i$ by performing a summation along the channel dimension:

$$\mathbf{r}_i=\sum_{j=0}^{c}D_{\mathrm{KL}}\left[P\left(\mathbf{M}_{[j,d,h,w]}|\mathbf{F}_{[j,d,h,w]}\right)||\,Q\left(\mathbf{M}_{[j,d,h,w]}\right)\right], \tag{9}$$

where c, d, h, and w denote the channel, depth, height, and width of the features, respectively.

IHG Module. Unlike Grad-CAM [35], IB tends to give more accurate heatmaps [17], [34]. Once IB is trained, we can obtain a pulmonary nodule-specific heatmap $\mathbf{r}_i$ from PCC-Net for $\mathbf{X}_i$. Based on $\mathbf{r}_i$ from PCC-Net, IHG is trained to learn a reliable heatmap $\mathbf{t}_i$ for the feature $\mathbf{f}_m^{'}$ (extracted from the middle layer of the backbone in IBD-Net). Technically, the IHG module contains a $1\times1\times$ 1 convolutional block and a softmax layer. We use a distance loss to enforce $\mathbf{t}_i$ to be similar to $\mathbf{r}_i$:

$$\mathcal{L}_h=\left\|\frac{\mathbf{t}_i}{\|\mathbf{t}_i\|_F}-\frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_F}\right\|_F, \tag{10}$$

where $\|\cdot\|_F$ represents the Frobenius norm. The learned heatmap $\mathbf{t}_i$ is used to activate pulmonary nodule-specific regions in $\mathbf{f}_m^{'}$ (see Fig. 2).

Then, we combine the heatmap $\mathbf{t}_i$ with the feature $\mathbf{f}_m^{'}$ by

$$\mathbf{f}^p=\mathbf{f}_m^{'}\oplus\left(\mathbf{f}_{\mathbf{m}}^{'}\otimes\mathbf{t}_i\right), \tag{11}$$

where $\mathbf{f}^p$ is the enhanced feature by incorporating the reliable heatmap. $\oplus$ and $\otimes$ represent element-wise addition and element-wise multiplication, respectively.

Joint Loss. The joint loss for IBD-Net is defined as

$$\begin{aligned}\mathcal{L}_{IBD}&=\sum_{i=1}^{N}(\mathcal{L}_{reg}^{R+O}(i)+\mathcal{L}_{cls}^{R+O}(i))+\sum_{i=1}^{N}\mathcal{L}_h(i),\\ \mathcal{L}_{reg}^{R+O}(i)&=\mathcal{L}_{reg}^{R}(i)+\mathcal{L}_{reg}^{O}(i),\\ \mathcal{L}_{cls}^{R+O}(i)&=\mathcal{L}_{cls}^{R}(i)+\mathcal{L}_{cls}^{O}(i),\end{aligned} \tag{12}$$

where $\mathcal{L}_{reg}^{R}(i)$ and $\mathcal{L}_{reg}^{O}(i)$ are the regression losses on RPN and OCN for the i-th CT image $\mathbf{X}_i$, respectively. $\mathcal{L}_{cls}^{R}(i)$ and $\mathcal{L}_{cls}^{O}(i)$ are the classification losses on RPN and OCN for the i-th CT image $\mathbf{X}_i$, respectively.

IV. Experiments

In this section, we first introduce the datasets and evaluation metrics in Section IV-A. Then, we give the implementation details in Section IV-B. Next, we conduct ablation studies in Section IV-C and give some visualization results in Section-D. Finally, we compare our method with several state-of-the-art methods in Section IV-E.

A. Datasets and Evaluation Metrics

We conduct extensive experiments on the public LUNA16 dataset [40] to evaluate the detection performance. The LUNA16 dataset is collected based on the LIDC-LDRI dataset [41], which is a widely used public dataset for studying pulmonary nodules.

For the LUNA16 dataset, similar to NoduleNet [3], we use 583 CT images with 1,131 nodules in our experiments. Specifically, we randomly select 483 CT images and 100 CT images for training and testing, respectively. For each round of training, we randomly select 48 ($\sim$ 10% of the training data) and 435 ($\sim$ 90% of the training data) CT images as strongly-labeled and weakly-labeled images, respectively, from the whole dataset. The rest is used for testing.

As done in [40], Free-Response Receiver Operating Characteristic (FROC) and the competition performance metric (CPM) are used as the evaluation metric. The horizontal axis of FROC represents the false positive rate per scan (FPs/scan) while the vertical axis represents the sensitivity. Mathematically, the sensitivity is calculated as

$$Sensitivity=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{13}$$

where TP and FN represent the true positive and the false negative, respectively.

For calculating CPM, several values of FPs/scan are first selected (such as 0.125, 0.25, 0.5, 1.0, 2.0, 4.0, and 8.0), and then the corresponding sensitivity values are calculated. Finally, CPM is the average of these sensitivity values.

TABLE I: Ablation studies for different variants of our HND method on the LUNA16 dataset. CPM is used for performance evaluation. The values from the second to eighth columns are nodule detection sensitivities (unit:%) under a specific false positive rate per scan (FPs/scan). The values in the last column are the average sensitivities. The best results are marked in **bold**.

Method	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Avg.
Baseline	52.63	57.89	66.08	78.36	80.12	84.21	86.55	72.26
HND w/o point	52.04	62.57	72.51	81.29	84.21	88.30	90.64	75.94
HND w/o CL	47.37	59.06	73.10	80.12	**87.72**	**90.64**	92.40	75.77
HND w/o CRL	56.73	63.16	74.85	83.04	**87.72**	88.89	91.81	78.03
HND_OCN	57.31	61.99	73.10	81.87	85.96	89.47	**92.98**	77.53
HND w/o IHG	**60.23**	64.33	73.68	80.12	84.80	90.06	90.06	77.61
HND	**60.23**	**67.25**	**77.19**	**84.80**	**87.72**	**90.64**	92.40	**80.03**

B. Implementation Details

Each CT image is pre-processed and cropped into a 3D patch with the size of $128\times128\times128$ for LUNA16. The models are trained for 200 epochs and 300 epochs in two stages, respectively, by using SGD with an initial learning rate of 0.001. The learning rate is divided by 10 after 100 and 160 epochs. We iteratively feed the strongly-labeled images (with a batch size of 8) and weakly-labeled images (with a batch size of 1) to the network in each learning stage. In PCC-Net and IBD-Net, 3D ResNet-50 [42] is adopted as the backbone. In IBD-Net, the information bottleneck is trained every ten epochs. We adopt IBD-Net, which is trained based on only the regression and classification losses with bounding box labels and center point labels, as our baseline method. For strongly-labeled data, we utilize their bounding box labels and category labels in the classification and regression losses. For weakly-labeled data, we employ their center point labels and category labels in the classification and regression losses. During the test phase, we only use the student model for inference, where the original CT images are taken as inputs.

C. Ablation Studies

We perform extensive ablation studies to evaluate the key components of our method. The results obtained by the baseline and several variants of our HND method are given in Table I. The LUNA16 dataset is used for ablation studies. We use 10% of the training data and 90% of the training data as strongly-labeled and weakly-labeled data, respectively.

Effectiveness of the Two-Stage Learning framework. In this paper, we develop a heterogeneous two-stage teacher-student learning framework. We evaluate the effectiveness of the two-stage learning framework. The baseline method is trained based on the one-stage learning framework.

From Table I, the baseline method performs much worse than the other variants (trained with the two-stage learning framework). Notably, HND outperforms the baseline method by a large margin (7.77% improvements in terms of the average sensitivity). This shows the superiority of the two-stage learning framework, which generates pseudo labels in the first stage and trains the pulmonary nodule detection model in the second stage.

Effectiveness of Point-Augmented Representations. We propose to use point-augmented representations as inputs of PCC-Net to fully exploit both strong labels and weak labels. To demonstrate the effectiveness of the heterogeneous teacher-student framework, we also evaluate the homogeneous teacher-student framework, where both the teacher and student networks adopt the same inputs (without any point augmentation). The homogeneous teacher-student framework is HND that uses only the original CT images as the inputs of PCC-Net and is denoted as "HND w/o point".

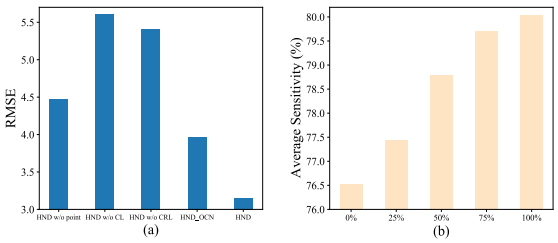


Fig. 3: Performance comparisons of (a) the quality of pseudo labels and (b) different data proportions.

As shown in Table I, by comparing the "HND w/o point" method with HND, HND obtains better performance than the "HND w/o point" method. Specifically, the average sensitivity obtained by the HND method is improved by 4.09% in comparison with that obtained by the "HND w/o point" method. This indicates the importance of enhancing input representations by the point information in the first stage. Such a way is beneficial to generate high-quality pseudo labels of CT images.

Effectiveness of the Consistency Loss. We evaluate the performance of our HND without the overall consistency loss (denoted as "HND w/o CL") and without only the consistency regression loss (denoted as "HND w/o CRL").

As shown in Table I, the "HND w/o CRL" method achieves better results than the "HND w/o CL" method. This is because that the consistency classification loss enables the model to better distinguish the nodules from non-nodules and further reduce false positives. Meanwhile, by adding the consistency regression loss, HND obtains better performance than the "HND w/o CRL" method. This indicates the importance of predicting the scale displacement of the bounding box by using the consistency regression loss. These two losses complement each other, leading to the improvement of the detection performance.

TABLE II: Ablation studies for the influence of different center points.

δ	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Avg.
0	60.23	67.25	**77.19**	**84.80**	**87.72**	90.64	92.40	**80.03**
1	60.82	65.50	76.02	80.70	85.38	**91.23**	**92.98**	78.95
2	57.31	68.42	76.61	81.87	85.38	90.06	92.40	78.86
3	**61.40**	**71.35**	76.61	81.29	84.21	87.72	89.47	78.86

Effectiveness of TCN. We evaluate the effectiveness of TCN. HND that uses a one-path calibration network (OCN) in PCC-Net is denoted as "HND_OCN".

As shown in Table I, HND outperforms HND_OCN in terms of different sensitives (about 2.5% improvements in terms of the average sensitivity). Therefore, the adoption of TCN in different layers of the teacher model is beneficial for extracting different levels of information, further reducing false positive rates for detecting pulmonary nodules.

Effectiveness of the IHG Module. We evaluate the effectiveness of the IHG module. HND trained without using the IHG module is denoted as "HND w/o IHG".

As given in Table I, HND achieves better performance than HND w/o IHG. The IHG module can effectively learn reliable heatmaps that encode pulmonary nodule-specific information, reducing false positives and highlighting response regions. The learned heatmaps can then be used to activate pulmonary nodule-relevant areas in the features. Therefore, the IHG module successfully facilitates the transfer of information between the teacher network and the student network, leading to enhanced detection performance.

Comparison of the Quality of Pseudo Labels. We compare the quality of pseudo labels obtained by different methods. We evaluate four variants of HND (including "HND w/o point", "HND w/o CL", "HND w/o CRL", and "HND_OCN"). Root Mean squared error (RMSE) is used to measure the difference between pseudo labels and ground-truth labels. The smaller the difference is, the higher the quality of pseudo labels is. Results are given in Fig. 3(a).

From Fig. 3(a), we can observe that HND achieves a smaller RMSE than HND w/o point, HND w/o CL, HND w/o CRL, and HND_OCN. This shows that the quality of pseudo labels is improved by using the point-augmented representations, the consistency loss, and TCN, respectively.

Comparison of Different Data Proportions. We evaluate HND with different proportions of weakly-labeled CT images (the number of strongly-labeled images is fixed). Results are given in Fig. 3(b).

The performance of HND is improved when the proportion of weakly-labeled CT images is increased from 0% to 100%. This can be ascribed to the fact that CT images with point annotations can provide useful weak information. When more annotation information is used, the performance can be effectively enhanced.

Comparison of Different Center Points. We impose variations to the original ground-truth center point by adding zero-mean Gaussian noise with standard deviation δ to each point. Table II shows the results obtained by our method with different levels of Gaussian noise.

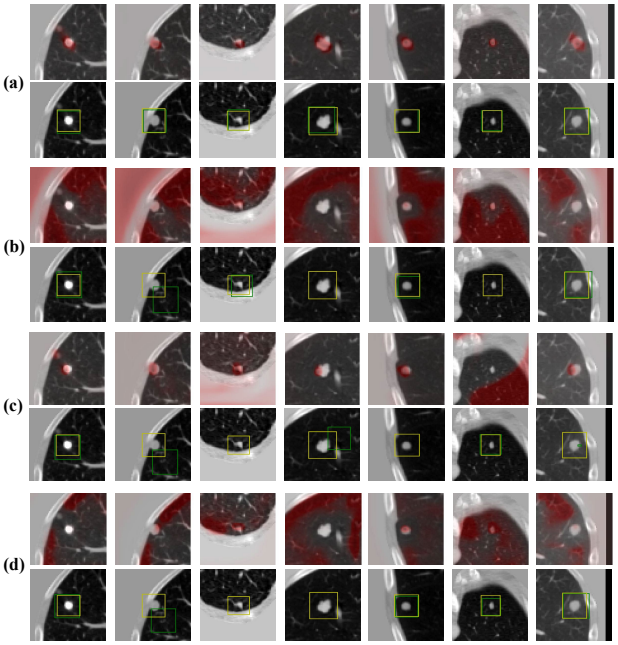


Fig. 4: Visualizations of some generated heatmaps (the first row), some detection results (the bounding boxes with green borders in the second row), and ground-truths (the bounding boxes with yellow borders in the second row) obtained by using (a) our IBA, (b) Saliency Maps [43], (c) GuidedBP [44], and (d) GradCAM [35]. The images without green bounding boxes indicate that no nodule is detected.

As shown in Table II, our method achieves the best results when $\delta = 0$. Meanwhile, the performance of our method only slightly changes when different values of δ are used. These results show the robustness of our method against small perturbations of center points.

Influence of the Layer of the Backbone to Insert the Information Bottleneck.

We insert the information bottleneck into different layers of the backbone and evaluate the final performance. We denote the four layers after the first four blocks of the backbone as "Layer 1", "Layer 2", "Layer 3", and "Layer 4", respectively. The performance of our method by inserting the information bottleneck into the four layers of the backbone is given in Table III.

We can see that our method achieves the best average sensitivity when the information bottleneck is inserted into Layer 2 of the backbone. The features from the middle layers

TABLE III: Ablation studies for the influence of different layers of the backbone that the information bottleneck is inserted into.

Layer	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Avg.
Layer 1	52.63	63.74	70.76	80.70	85.38	88.89	90.06	76.02
Layer 2	60.23	67.25	77.19	**84.80**	**87.72**	**90.64**	**92.40**	**80.03**
Layer 3	56.73	64.91	73.68	80.12	83.63	88.89	90.06	78.86
Layer 4	**60.82**	**69.59**	**78.36**	83.04	85.96	89.47	91.23	79.78

TABLE IV: Ablation studies for the influence of IBA and other interpretable strategies employed in our method when 10% of training data is used as strongly-labeled data.

Strategy	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Avg.
IBA [34]	**60.23**	**67.25**	**77.19**	**84.80**	**87.72**	**90.64**	**92.40**	**80.03**
GuidedBP [44]	49.71	55.56	64.91	70.76	77.19	80.12	83.63	68.84
GradCAM [35]	38.01	42.69	48.54	52.05	57.31	63.16	70.18	53.13
Saliency [43]	38.01	40.35	43.27	52.05	55.56	59.06	61.40	49.96

TABLE V: CPM (%) comparisons obtained by fully-supervised methods on the LUNA16 dataset.

Methods	False positive rate per scan								Data Type
	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Average	
Roth *et al.* [45]	49.90	59.80	66.60	70.50	75.60	80.30	82.90	69.40	2D
Lee *et al.* [46]	26.20	35.90	47.20	58.60	67.40	71.30	76.60	54.80	2D
Setio *et al.* [47]	63.60	72.70	79.20	84.40	87.60	90.50	91.60	81.40	2D
Liao *et al.* [48]	59.38	72.66	78.13	84.38	87.50	89.06	89.84	80.13	3D
Dou *et al.* [49]	67.70	73.70	81.50	84.80	87.90	90.70	92.20	82.70	2D
Tang *et al.* [3]	65.18	76.79	83.93	87.50	91.07	92.86	93.75	84.43	3D
Zhang *et al.* [50]	73.70	76.40	80.40	84.70	89.00	92.10	93.70	84.30	2D
Mei *et al.* [4]	71.17	80.18	86.49	90.09	93.69	94.59	95.50	87.39	3D
Lin *et al.* [51]	72.15	79.22	86.53	90.13	93.20	94.77	95.78	87.39	3D
Jian *et al.* [52]	76.43	82.14	85.71	89.29	92.86	94.29	95.71	88.08	3D
Luo *et al.* [5]	74.30	82.90	88.90	92.20	**93.90**	**95.80**	**96.40**	**89.20**	3D
HND	**77.78**	**84.21**	**89.47**	**92.40**	92.98	92.98	94.15	89.14	3D

contain both spatial and contextual information, which can be useful for learning reliable heatmaps.

D. Visualization Results

In this section, we visualize several examples of heatmaps generated by our IBA and other interpretable methods, including three gradient-based methods (Saliency maps [43], Guided Backpropagation (GuidedBP) [44] and GradCAM [35]) on LUNA16, as shown in Fig. 4. GradCAM [35], Saliency maps [43], and GuidedBP [44] are all gradient-based methods. GradCAM, Saliency maps, and GuidedBP all utilize the backpropagation algorithm to compute the relevance between image pixels and the model output, thereby generating heatmaps based on the gradient information. This enables them to visualize the input regions that the model focuses on. IBA introduces perturbations or masking to the input, evaluating the influence on the model's predictions. In this way, IBA can generate heatmaps based on the results calculated from using the information bottleneck theory.

For a fair comparison, we apply these methods to obtain activation maps (typically having the same size as the input images) from the pretrained PCC-Net. The activation maps extracted from the PCC-Net are resized to the same dimension as the feature $\mathbf{f}_m^{'}$ and used to guide the learning of the IHG module. Here, we employ 10% of training data in LUNA16 as strongly-labeled training data and visualize the heatmaps learned by the IHG module.

From Fig. 4, the heatmaps generated by our method contain much less noise than our method with other interpretable methods. By using our IBA, our method can generate dense and reliable heatmaps that cover all the targeted areas. The predicted bounding boxes and the ground-truth labels are similar. This demonstrates the effectiveness of our IBA in removing unrelated information. In this way, the pulmonary nodule-specific heatmaps obtained by our IBA can effectively teach the IBD-Net to learn reliable heatmaps and further improve the final performance. Saliency Map and GradCAM capture too much background information when locating nodules and they easily miss some nodules (e.g., the third column in Fig. 4(b) and the fourth column in Fig. 4(d)). Compared with Saliency Map and GradCAM, GuidedBP can depict the position of nodules more accurately but focuses only on a small portion of nodules, resulting in a significant deviation between the detection results and the ground-truth labels (e.g., the first and last examples in Fig. 4(c)).

The visualization results demonstrate that our method is capable of providing accurate predictions while offering intuitive explanations through heatmaps. Moreover, we also evaluate

TABLE VI: CPM (%) comparisons obtained by semi-supervised and hybrid-supervised methods on LUNA16.

Proportion		Method	Average
Strongly-Labeled	Weakly(Un)-Labeled		
100%	0%	NoduleNet (N2) [3]	84.43
		SANet [4]	87.39
		HS-SANet	83.29
		SS-ND	84.89
		HND	**89.14**
50%	50%	HS-SANet	81.12
		HS-LSSANet	74.94
		SS-N	84.02
		SS-N+	79.65
		SS-ND	81.58
		HND	**87.30**
25%	75%	HS-SANet	75.52
		HS-LSSANet	66.83
		SS-N	78.07
		SS-N+	80.80
		SS-ND	78.33
		HND	**84.46**
10%	90%	HS-SANet	49.96
		HS-LSSANet	39.77
		SS-N	64.16
		SS-N+	67.34
		SS-ND	77.44
		HND	**80.03**

the influence of our IBA and other interpretable methods on the performance of our method in Table IV. Our method achieves the best results across all metrics, consistent with the visualization results in Fig. 4.

E. Comparison with State-of-the-Art Methods

In this subsection, we evaluate our methods on the representative pulmonary nodule detection dataset LUNA16. Table V gives the comparison results obtained by fully-supervised methods on the LUNA16 dataset. Table VI gives the comparison results obtained by semi-supervised and hybrid-supervised methods on the LUNA16 dataset. The state-of-the-art methods include fully-supervised methods [3]–[5], [45]–[50], two hybrid-supervised methods (HS-SANet and HS-LSSANet), and three semi-supervised methods (SS-ND, SS-N, and SS-N+). Since most of the hybrid-supervised methods work on 2D image segmentation and natural image detection tasks, they cannot be directly applied to our task. Hence, for a fair comparison, we construct a simple hybrid-supervised method (denoted as "HS-SANet") based on SANet [4]. In this method, SANet is first pre-trained by strongly-labeled CT images to predict pseudo bounding box labels for weakly-labeled images. Then, SANet is fine-tuned with the original strong labels and generated pseudo labels. Similarly, we construct another hybrid-supervised method based on LSSANet [23] (denoted as "HS-LSSANet"). For the semi-supervised method, we remove the point-related operations in HND and denote this method as "SS-ND". Based on NoduleNet [3] "SS-N" is constructed by utilizing a similar two-stage learning strategy. In the first stage, "SS-N" generates pseudo labels for unlabeled images, whereas, in the second stage, both labeled images and unlabeled images are trained alternately. We extend "SS-N" to "SS-N+" by adding two self-supervised tasks designed by us.

From Table V and Table VI, our HND method achieves much better performance than other methods in different cases. This can be ascribed to the two-stage teacher-student learning framework, which involves the pre-training of PCC-Net and the training of IBD-Net under the guidance of PCC-Net. Compared with fully-supervised methods, SCPM-Net (average sensitivity: 89.20%) and SANet (average sensitivity: 87.39%) using 100% of strongly-labeled data, HND using only 50% of strongly-labeled data can achieve close performance (average sensitivity: 87.30%), significantly reducing annotation costs by only requiring half of strongly-labeled CT images. Moreover, as shown in Table VI, the average sensitivity (87.30%) of HND using 50% of strongly-labeled data outperforms those (84.43% and 83.29%) of fully-supervised methods (NoduleNet and HS-SANet). Note that the time of annotating 100% of the training data is significantly greater than that of annotating 50% of the training data. When only 10% and 25% of strongly-labeled data are available, HND still outperforms other hybrid-supervised and semi-supervised methods (HS-SANet and SS-ND) by a large margin.

In SANet, a SGNL module is introduced in the encoder network to capture long-term dependencies within a slice group. Based on the SGNL module, LSSANet further exploits SSG and LSG to detect pulmonary nodules. In this way, when the slices are far apart, the similarity between them can still be exploited for feature extraction. HS-LSSANet performs worse than HS-SANet. This is because SSG and LSG in LSSANet are easily affected by the different sizes of test images, making HS-LSSANet fail to perform well in LUNA16 containing images with different sizes.

In Table V, our method performs better than most fully-supervised methods. Most pulmonary nodule detection methods detect nodules in the form of 3D bounding boxes, while SCPM-Net detects pulmonary nodules in the form of 3D bounding spheres. When predicting bounding boxes, the model is required to predict scale offsets in three directions. In contrast, only the scale offsets in one direction (the radial direction) are predicted for bounding spheres. This simplifies the prediction of bounding spheres, improving the performance. Note that SCPM-Net is specifically designed under a fully-supervised setting, while our method focuses on a hybrid-supervised setting. Although the sensitivities of our method are slightly lower than that of SCPM-Net, our method exhibits a performance increase in comparison with SCPM-Net at low false positive rates per scan (5.23% and 1.31% improvements at 0.125 and 0.25, respectively). The above experiments validate the effectiveness of our method.

V. Conclusion and Future Work

In this paper, we develop a novel HND method for pulmonary nodule detection in a hybrid-supervised setting, which requires only a small number of labeled bounding boxes and a relatively larger number of labeled center points for training. HND is trained via an interpretable heterogeneous two-stage teacher-student learning framework. In the first stage,

PCC-Net is pre-trained as a teacher to generate high-quality pseudo bounding box labels. In the second stage, IBD-Net is trained as a student to detect pulmonary nodule regions under the guidance of PCC-Net. In IBD-Net, an IHG module is introduced to learn reliable heatmaps that closely resemble the pulmonary nodule-specific heatmaps extracted from PCC-Net. Experimental results on the public LUNA16 dataset show the superiority of our method against state-of-the-art methods.

Although our method can achieve excellent performance for hybrid-supervised pulmonary nodule detection, the training complexity of our method is relatively high. The training of our method involves a heterogeneous teacher-student framework, where a teacher model is trained in the first stage while a student model is trained in the second stage under the guidance of the teacher model. In future work, we plan to simplify the model architecture to reduce the training complexity.

REFERENCES

[1] H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, M. M. Coates *et al.*, "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1459–1544, 2016.

[2] Y. Momoki, A. Ichinose, Y. Shigeto, U. Honda, K. Nakamura, and Y. Matsumoto, "Characterization of pulmonary nodules in computed tomography images based on pseudo-labeling using radiology reports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2582–2591, 2022.

[3] H. Tang, C. Zhang, and X. Xie, "NoduleNet: Decoupled false positive reduction for pulmonary nodule detection and segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 266–274.

[4] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, and H. Zhang, "SANet: A slice-aware network for pulmonary nodule detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4374–4387, 2022.

[5] X. Luo, T. Song, G. Wang, J. Chen, Y. Chen, K. Li, D. N. Metaxas, and S. Zhang, "SCPM-Net: An anchor-free 3d lung nodule detection network using sphere representation and center points matching," *Medical Image Analysis*, vol. 75, p. 102287, 2022.

[6] D. Wang, Y. Zhang, K. Zhang, and L. Wang, "FocalMix: Semi-supervised learning for 3D medical image detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3951–3960.

[7] H.-Y. Zhou, C. Wang, H. Li, G. Wang, S. Zhang, W. Li, and Y. Yu, "SSMD: Semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation," *Medical Image Analysis*, vol. 72, p. 102117, 2021.

[8] W. Zhu, Y. S. Vang, Y. Huang, and X. Xie, "DeepEM: Deep 3D convnets with EM for weakly supervised pulmonary nodule detection," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 812–820.

[9] H.-H. Yang, F.-E. Wang, C. Sun, K.-C. Huang, H.-W. Chen, Y. Chen, H.-C. Chen, C.-Y. Liao, S.-H. Kao, Y.-C. F. Wang *et al.*, "Leveraging auxiliary information from EMR for weakly supervised pulmonary nodule detection," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 251–261.

[10] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Training object class detectors with click supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6374–6383.

[11] G. Kang, K. Liu, B. Hou, and N. Zhang, "3D multi-view convolutional neural networks for lung nodule classification," *PloS One*, vol. 12, no. 11, p. e0188290, 2017.

[12] L. Fang, H. Xu, Z. Liu, S. Parisot, and Z. Li, "EHSOD: CAM-guided end-to-end hybrid-supervised object detection with cascade refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 10 778–10 785.

[13] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 715–12 725.

[14] W. Luo and M. Yang, "Semi-supervised semantic segmentation via strong-weak dual-branch network," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 784–800.

[15] J. Pan, Q. Bi, Y. Yang, P. Zhu, and C. Bian, "Label-efficient hybrid-supervised learning for medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2026–2034.

[16] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang, "Interactive self-training with mean teachers for semi-supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5941–5950.

[17] U. Demir, I. Irmakci, E. Keles, A. Topcu, Z. Xu, C. Spampinato, S. Jambawalikar, E. Turkbey, B. Turkbey, and U. Bagci, "Information bottleneck attribution for visual explanations of diagnosis and prognosis," in *Machine Learning in Medical Imaging*, 2021, pp. 396–405.

[18] X. Chen, H. Li, Q. Wu, K. N. Ngan, and L. Xu, "High-quality R-CNN object detection using multi-path detection calibration network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 715–727, 2021.

[19] H. Zhu, W. Liu, Z. Gao, and H. Zhang, "Explainable classification of benign-malignant pulmonary nodules with neural networks and information bottleneck," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[20] C. Wang, R. Xu, S. Xu, W. Meng, J. Xiao, and X. Zhang, "Accurate lung nodule segmentation with detailed representation transfer and soft mask supervision," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[21] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[22] W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 673–681.

[23] R. Xu, Y. Luo, B. Du, K. Kuang, and J. Yang, "LSSANet: A long short slice-aware network for pulmonary nodule detection," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 664–674.

[24] J. Dong, Y. Cong, G. Sun, Y. Yang, X. Xu, and Z. Ding, "Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2020–2033, 2021.

[25] X. Wang, D. Cai, S. Yang, Y. Cui, J. Zhu, K. Wang, and J. Zhao, "SAC-Net: Enhancing spatiotemporal aggregation in cervical histological image classification via label-efficient weakly supervised learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[26] Y. Zhao, Q. Ye, W. Wu, C. Shen, and F. Wan, "Generative prompt model for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 6351–6361.

[27] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 676–11 685.

[28] Z. Shen, P. Cao, J. Yang, and O. R. Zaiane, "WS-LungNet: a two-stage weakly-supervised lung cancer detection and diagnosis network," *Computers in Biology and Medicine*, vol. 154, p. 106587, 2023.

[29] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 568–576.

[30] M. Ning, C. Bian, D. Lu, H.-Y. Zhou, S. Yu, C. Yuan, Y. Guo, Y. Wang, K. Ma, and Y. Zheng, "A macro-micro weakly-supervised framework for AS-OCT tissue segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 725–734.

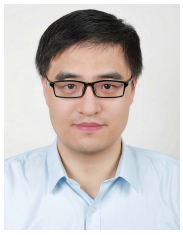
[31] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[32] Y. Lyu, Y. Jiang, B. Peng, and J. Dong, "Infostyler: Disentanglement information bottleneck for artistic style transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2070–2082, 2024.

[33] H. Li, C. Zhu, Y. Zhang, Y. Sun, Z. Shui, W. Kuang, S. Zheng, and L. Yang, "Task-specific fine-tuning via variational information bottle-

neck for weakly-supervised pathology whole slide image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7454–7463.
[34] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," *arXiv preprint arXiv:2001.00396*, 2020.
[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
[36] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
[37] Y. Zhang, A. Khakzar, Y. Li, A. Farshad, S. T. Kim, and N. Navab, "Fine-grained neural network explanation by identifying input features with predictive information," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 20 040–20 051.
[38] J. Wang, Y. Zheng, J. Ma, X. Li, C. Wang, J. Gee, H. Wang, and W. Huang, "Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation," *Medical Image Analysis*, vol. 83, p. 102687, 2023.
[39] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *Proceedings of the Advances in Neural Information Processing Systems*, 2019, pp. 10 759–10 768.
[40] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. 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.
[41] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. 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.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[44] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
[45] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1170–1181, 2016.
[46] H. Lee, H. Lee, M. Park, and J. Kim, "Contextual convolutional neural networks for lung nodule classification using gaussian-weighted average image patches," in *Medical Imaging 2017: Computer-Aided Diagnosis*, 2017, pp. 544–550.
[47] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. Van Ginneken, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.
[48] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3484–3495, 2019.
[49] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.
[50] Z. Zhang, X. Li, Q. You, and X. Luo, "Multicontext 3D residual CNN for false positive reduction of pulmonary nodule detection," *International Journal of Imaging Systems and Technology*, vol. 29, no. 1, pp. 42–49, 2019.
[51] J. Lin, Q. She, and Y. Chen, "Pulmonary nodule detection based on ir-unet++," *Medical & Biological Engineering & Computing*, vol. 61, no. 2, pp. 485–495, 2023.
[52] M. Jian, L. Zhang, H. Jin, and X. Li, "3DAGNet: 3D deep attention and global search network for pulmonary nodule detection," *Electronics*, vol. 12, no. 10, p. 2333, 2023.

Guangyu Huang is currently pursuing the master's degree with the School of informatics, Xiamen University, China. Her research interests include deep learning, computer vision and medical image analysis.



Yan Yan (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2009. He worked as a Research Engineer with the Nokia Japan Research and Development Center from 2009 to 2010. He worked as a Project Leader with the Panasonic Singapore Laboratory in 2011. He is currently a Full Professor with the School of Informatics, Xiamen University, China. He has published around 100 papers in the international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, CVPR, ICCV, ECCV, AAAI, and ACM MM. His research interests include computer vision and pattern recognition.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Professor with the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the Outstanding Associate Editor Award of 2022 from the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Wentao Zhu received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2014. He worked in the U.S. Subsidiary of United Imaging, Houston, TX, USA, as a Manager and a Senior Scientist from 2014 and 2019. He is currently a Professor with Zhejiang Lab, Hangzhou, China. His research interests are image reconstruction and machine learning in the field of medical imaging. He has been granted more than ten U.S. Patents and more than 30 China Patents.

Xiongbiao Luo (Senior Member, IEEE) received the Ph.D. degree in information science from Nagoya University, Japan, in 2011. He was a Postdoctoral Fellow and an Assistant Professor with Nagoya University, a Postdoctoral Fellow with the University of Western Ontario, Canada, and a Senior Researcher with the French National Institute of Health and Medical Research, France. He is currently a Full Professor with the Department of Computer Science and Technology, National Institute for Data Science in Health and Medicine, and the Director of the XMU Center for Surgery and Engineering, Xiamen University. He has edited seven books and published more than 150 peer-reviewed articles. His current research interests include artificial intelligence in healthcare, surgical vision, autonomous navigation and robotic, and computational photography.