IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 4, 2020, accepted August 20, 2020, date of publication August 24, 2020, date of current version September 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019104

Towards Accurate Pulmonary Nodule Detection by Representing Nodules as Points With High-Resolution Network

ZEHUI GONG[1], DONG LI[1], JIATAI LIN[1], YUN ZHANG[1], AND KIN-MAN LAM[2], (Member, IEEE)

[1]School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[2]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Dong Li (dong.li@gdut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61503084 and Grant U1501251.

ABSTRACT Almost all successful nodule detectors rely heavily on a fixed set of anchor boxes. In this paper, inspired by the success of the keypoint estimation method in natural image detection, we propose an anchor-free framework for accurate pulmonary nodule detection. We first present a novel representation for detecting nodules, in terms of their 3D center locations, which reduces the number of hyper-parameters and the corresponding computation related to anchors, thus making the nodule detection pipeline much simpler. Then, an effective two-stream network is introduced to reduce the false positive nodule candidates, by aggregating information from the image stream and motion-history stream. Experiments show that the proposed approach achieves a sensitivity of 96.1%, with 8 false positives per scan, and a CPM score of 90.6%, on the publicly available LUNA16 dataset, which outperforms other state-of-the-art methods. By testing on the SPIE-AAPM dataset with models pre-trained on the LUNA16, our proposed method yields 92.8% sensitivity with 8 false positives per scan. This demonstrates the effectiveness and generalization ability of our method.

INDEX TERMS Lung nodule detection, 3D convolution neural network, keypoint estimation.

I. INTRODUCTION

Lung cancer accounts for the vast majority of cancer-related death worldwide [1], [2]. However, most of these deaths could have been avoided, if lung cancer could have been diagnosed at an early stage. Therefore, the research on lung cancer diagnosis is extremely essential and urgent. Computed-aided diagnosis (CAD) systems have been widely used to assist radiologists in accelerating diagnosing process and therapy planning [3], [4].

Many algorithms [5]–[10] have been proposed for lung nodule detection on low-dose computed tomography (CT) scans, by using powerful deep convolutional neural networks (CNNs) with either 2D or 3D detection frameworks. 2D CNNs [11], [12] process each slice of the scan separately, without taking the inter-slice relations into account. On the other hand, most of the 3D CNNs [5], [13], [14] simply generalize the recent deep 2D CNN-based detection models in computer vision, such as Faster RCNN [15]. These methods take 3D patches (in general, the patch size is $96 \times 96 \times 96$) as input, and exhaustively and inefficiently iterate through the entire CT volume in a sliding window manner, to output probability maps over the patches.

All current mainstream nodule detectors [5], [11], [14], [16] rely heavily on a fixed set of anchor boxes to achieve high sensitivity (recall rate), where each anchor box is sampled uniformly over the spatial positions with a pre-defined set of scales and aspect ratios. Despite the great success achieved by anchor-based nodule detectors, it is worth emphasizing that they suffer from some shortcomings. (1) These approaches require a large number of anchor boxes placed on each position of the input images (up to 100K) to ensure a high sensitivity, which results in high computational cost especially when the model adopts a heavy classifier during candidate generation. (2) A large number of anchor boxes will cause a significant imbalance between positive and negative samples because most of the anchors are assigned with negative labels. (3) The scales and aspect ratios of these anchor boxes have to be manually designed for different problems, and an incorrect setting of these parameters may greatly impair the accuracy of

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

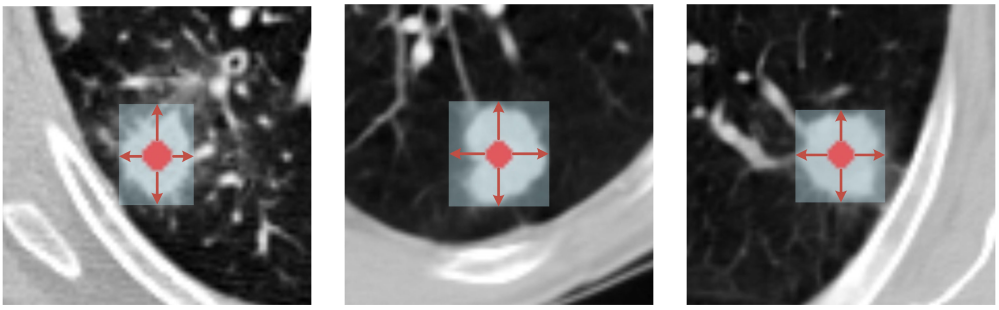


FIGURE 1. We model a nodule as a center point, in terms of its 3D spatial position, and regress the size of the nodule with the feature maps at the center. Note that nodules and corresponding centers are three-dimensional, of which we only show central slices of nodules in 2D images for better visualization. Best viewed in color. The images are randomly sampled from LUNA16.

the detector. (4) Even with careful design, fixed anchors may hardly cope with nodules with a wide range of variations in shape and size.

Another challenge we may face for accurate nodule detection is the similar appearance between normal tissues and nodules. This may make the detector incapable of distinguishing them correctly, *i.e.*, wrongly detecting normal tissues as nodules, thus resulting in high false positive rates. Therefore, it is very essential to devise an approach that can differentiate between tissues and nodules, to reduce false positives for the nodule detection pipeline.

A natural question is: *Can we detect nodules in a simpler formulation, e.g., using keypoint estimation, and bypasses the need for fixed anchor boxes*? In this paper, we demonstrate that the answer is affirmative. Furthermore, it is showed that the much simpler center-point-based nodule detector even achieves higher performance than the anchor-based counterparts. More specifically, we first present a novel framework for lung nodule detection. This framework detects the center point of each potential nodule and, regresses the size of each detected nodule using image features from the center location (see Fig. 1). Therefore, nodule detection is transferred into a three-dimension keypoint estimation problem [17], [18]. Furthermore, a two-stream network for false positive reduction is proposed, which integrates the information from the image stream and motion-history stream. The proposed two-stream network reduces the false positive nodule candidates considerably, while still achieving high sensitivity.

The main contributions of this paper are summarized as follows:

1. We propose a novel framework for nodule detection, termed 3D-CenterNet, which represents nodules in a much simpler manner, in terms of their 3D spatial center locations.
2. We propose an anchor-free nodule detector without requiring any anchor boxes, which reduces the number of hyper-parameters that need heuristic tuning to achieve good performance, thus making it much simpler in training and inference.
3. We propose a two-stream network for false positive reduction, which reduces the false positives considerably (more than 90%), while still obtaining high sensitivity.
4. Comprehensive experiments conducted on the LUNA16 dataset demonstrate the effectiveness of our proposed network, which achieves state-of-the-art performance compared to other lung nodule detection methods.

In Section II, we review the related work of keypoint estimation methods in 2D object detection and that of nodule detection in the medical image. Section III describes the details of our proposed method, and in Section IV, we describe the experimental setup and analyse the experimental results. Finally, we make a brief conclusion of our proposed method in Section V.

II. RELATED WORK

A. OBJECT DETECTION USING KEYPOINT ESTIMATION

Before recently proposed anchor-free keypoint estimation object detection methods [19], [21]–[23], 2D object detection has been dominated by generating pre-defined anchors in a sliding window manner, *e.g.*, [15], [24], [25]. CornerNet [21] is the pioneering work that tries to adopt the keypoint estimation method for detecting objects. It first detects a pair of corners of an object. Then, the detected corners are grouped using distances in embedding space to produce the final detected bounding box. Discovering that CornerNet generates some incorrect bounding boxes due to wrong grouping of corners, Duan *et al.* [22] employ additional center locations of objects to filter out those false positives, *i.e.*, detecting objects using keypoint triplets. Besides, Zhou *et al.* [23] detect four extreme points and one center point, and group the five keypoints into a bounding box using the geometric method. Moreover, Zhou *et al.* [19] further simplify the formulation, by representing objects as their center locations of the 2D bounding boxes, and regressing object size using image features from the center location directly. The design of our work for nodule detection shares the same spirits with Zhou *et al.* [19], detecting the 3D spatial center locations of nodules and predicting corresponding nodule size using the features at each center location.

B. NODULE DETECTION IN 3D MEDICAL IMAGES

The hand-craft features, *e.g.*, spherical filter and local binary feature [26], are adopted by earlier lung nodule detectors [27]–[31], which achieve inferior performance compared with deep learning-based methods [5], [8], [11], [32].

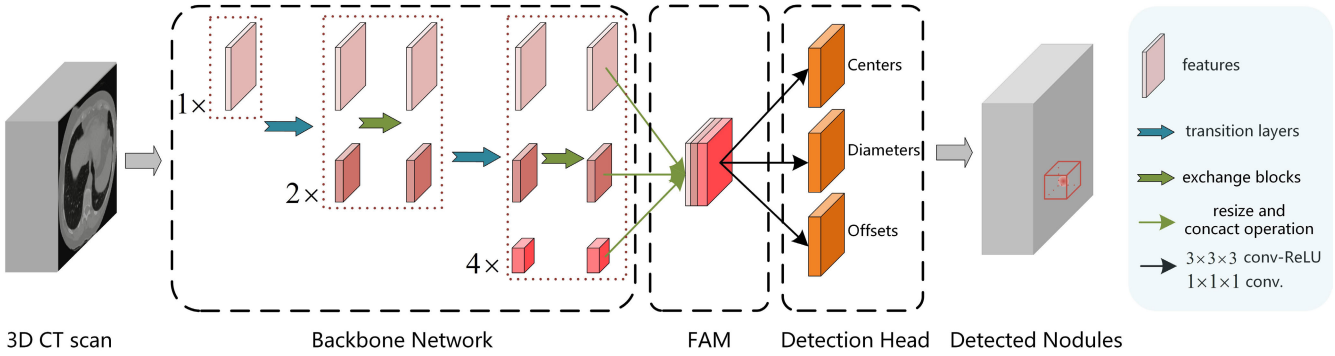


FIGURE 2. **The overall architecture of the proposed 3D-CenterNet. The backbone network is first adopted to encode the 3D CT scan. Then, we employ feature aggregation module (FAM) to aggregate representations from multi-level feature maps, following a detection head to predict center locations and sizes of potential nodules, along with corresponding offsets. Notably, we just show three stages with three parallel subnetworks of the backbone for simplicity. 'Transition layers' are adopted to add lower resolution subnetworks between different stages.**

TABLE 1. **Comparison between anchor-based nodule detectors and our proposed center-point-based framework. Our algorithm removes the computation and all hyper-parameters related to anchors during training.**

Method	pre-defined scales	pre-defined aspect ratios	proposal stage	setting training IOU thresholds	training IOU computing
anchor-based	✓	✓	✓	✓	✓
Ours	No	No	No	No	No

Ding *et al.* [11] employ a 2D Faster R-CNN as a nodule detector and use 3D CNN for false positive reduction, which achieves 89.1% average sensitivity. Zhu *et al.* [5] devise a 3D Faster R-CNN to detect nodules, which can effectively learn rich features by combining 3D dual-path blocks and a U-net-like encoder-decoder structure. Li and Fan [14] propose using a 3D region proposal network (RPN) for nodule detection, which uses an encoder-decoder structure and, a squeeze-and-excitation structure to further enhance the performance. Khosravan and Bagci [8] design a 3D CNN with dense connection and propose adopting max-pooling throughout the network to achieve better performance. Another line of research [7], [10], [13], [32], [33] investigate multi-scale feature maps either in image or feature pyramid to cope with the variance of nodule size. All mentioned nodule detectors use anchors sampled uniformly over the spatial position as candidate nodules and classify each anchor to be a nodule or not as well as adjust their locations. The anchoring scheme introduces excessively many hyper-parameters that need to be carefully tuned across various problems (see Table 1). In this work, we show that even without pre-defined anchors, the much simpler nodule detector can still achieve better performance.

III. METHODOLOGY

In this section, we will describe the proposed 3D-CenterNet and the two-stream network for false positive reduction in detailed. As shown in Fig.2, there are two major components for 3D-CenterNet.

Backbone Network we employ HRNet [34], a highly efficient feature extractor, to obtain high-level semantic but also spatially finer features to facilitate the subsequent detection task. Moreover, we make some modifications to extract features from 3D images, which form the feature extractor of 3D-CenterNet.

Detection Head The detection head aims to predict the center locations of nodules along with size corresponding to each center. Furthermore, to generate more precise center locations, the detection head also predicts offset to slightly adjust the location of each center position.

A. BACKBONE NETWOK

1) 3D-HRNet

We adopt HRNet [34] as the backbone network and make some adjustments to extract features from the input 3D images. HRNet was originally proposed for the 2D human pose estimation task. A high-resolution subnetwork is used as its first stage. Then, the lower resolution subnetworks are added in series, which form the other stages of HRNet. Unlike other methods that first downsample the input data and then restore the high-resolution representation, HRNet maintains high-resolution representation throughout the entire network. Moreover, it introduces exchange blocks for efficient information exchange across different resolutions, enabling the network to gather richer semantic and spatial information. Therefore, HRNet is a preferred choice for 3D-CenterNet. The middle part of Fig. 2 shows an illustration of the architecture of HRNet.

Here we show an example of two exchange blocks in 3th stage, which is given as follows,

$$\begin{array}{ccccccc} \mathcal{C}_{31}^{1} & \searrow & & \nearrow & \mathcal{C}_{31}^{2} & \searrow & \\ \mathcal{C}_{32}^{1} & \rightarrow & \mathcal{U}_{3}^{1} & \rightarrow & \mathcal{C}_{32}^{2} & \rightarrow & \mathcal{U}_{3}^{2}, \\ \mathcal{C}_{33}^{1} & \nearrow & & \searrow & \mathcal{C}_{33}^{2} & \nearrow & \end{array} \tag{1}$$

where $\mathcal{C}_{sr}^{b}$ is the convolution unit of resolution r (the resolution is $1/2^{r-1}$ of the first subnetwork) in bth exchange block, in the sth stage. $\mathcal{U}_{s}^{b}$ represents the corresponding exchange unit.

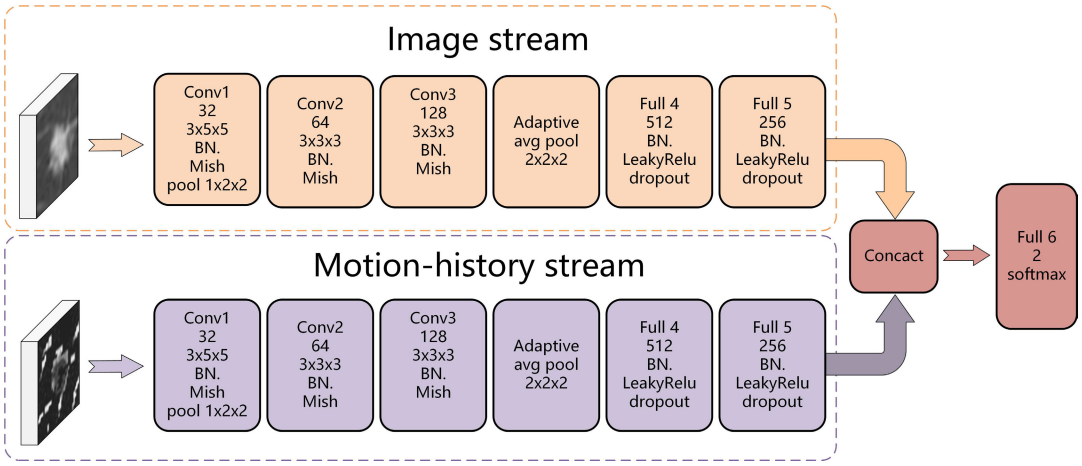


FIGURE 3. The overall framework of the proposed two-stream network for false positive reduction.

TABLE 2. Comparison between CenterNet [19] and our proposed 3D-CenterNet. Our model takes 3D volume as input and, employs 3D-HRNet as the backbone network. Besides, we aggregate multi-level features from the backbone for the subsequent detection task (see Fig. 2). Moreover, instead of applying L1 loss for object size regression, we propose using a version that is irrelevant to the object scale (see Eq. 6), to achieve more accurate 3D bounding boxes.

Method	Input 3D volume	Backbone	Aggregate multi-level features	Output 3D locations	Scale irrelevant regression loss
2DCenterNet	No	Hourglass-104 [20]	No	No	No
3D-CenterNet (Ours)	✓	3D-HRNet	✓	✓	✓

The exchage unit takes s feature maps in sth stage as input, denoted as $X = \{X_1, X_2, \ldots, X_s\}$. The subscript r is showed only, for description simplicity. After multi-scale feature fusion, the exchange unit outputs s feature maps, represented as $Y = \{Y_1, Y_2, \ldots, Y_s\}$, where the output dimensions are exactly the same as that of input. We obtain each output Y_j by aggregating input feature maps, $Y_j = \sum_{i=1}^{s} f(X_i, j)$, where $f(X_i, j)$ is downsampling or upsampling operation of X_i from resolution i to resolution j. If $i = j, f(X_i, j)$ is just an identity mapping.

Our backbone network is composed of four stages, and a total of four parallel subnetworks. 2D convolution kernels are all replaced by corresponding 3D kernels. The resolution of the subnetworks is gradually downsampled by $2\times$, while doubling the number of channels accordingly. The first stage contains two residual units, which are formed by a bottleneck [35] with 64 channels. Then a $3 \times 3 \times 3$ convolution reduces feature maps to 18 channels. There are 1, 4, and 3 exchange blocks in the 2nd, 3rd, and 4th stages, respectively. Specifically, one exchange block contains two residual convolution units for each of the resolution and one exchange unit across resolutions. Combining all stages gives rise to our backbone network, termed 3D-HRNet. The overall architecture of the backbone network is tabulated in Table 3.

TABLE 3. Detailed network architecture of 3D-HRNet. Input size for stem layer is 512 × 512 × 16. Output size is in $w \times h \times l$ format. '#Subnet' represents the number of parallel subnetworks in each layer.

Layer name	Operator	output size	#Subnet
stem layer	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}$	$128 \times 128 \times 8$	1
stage 1	$\begin{bmatrix} 1\times1\times1, 64 \\ 3\times3\times3, 64 \\ 1\times1\times1, 256 \end{bmatrix} \times 2$	$128 \times 128 \times 8$	1
stage 2	exchange block $\times1$	$128 \times 128 \times 8$ $64 \times 64 \times 4$	2
stage 3	exchange block $\times4$	$128 \times 128 \times 8$ $64 \times 64 \times 4$ $32 \times 32 \times 2$	3
stage 4	exchange block $\times3$	$128 \times 128 \times 8$ $64 \times 64 \times 4$ $32 \times 32 \times 2$ $16 \times 16 \times 1$	4

2) FEATURE AGGREGATION MODULE

Due to the significant variation in nodule scale (from 3mm to 30mm), it is crucial to aggregate information from multi-level features, to provide more powerful representation for the subsequent detection task. Therefore, we propose a feature aggregation module (FAM) to integrate features from various resolutions. Specifically, the output of 4th stage in the backbone network contains four different resolutions of feature maps, denoted as $\mathcal{O} = \{\mathrm{x}_l\}_{l=1}^{4}$, where x_l represents the feature maps at level l, whose resolution is $1/2^{l+1}$ of the input CT volume, and the depth is $1/2^l$ of the input.

For the feature maps at level l, we first use *trilinear* interpolation operation to resize x_l, with a scale of 2^{l-1}. Then, the resized feature maps are concatenated in the channel dimension. Finally, a convolution operation with a kernel size of $1 \times 1 \times 1$ is further applied to the concatenated features, to reduce the channels to 64.

With our proposed feature aggregation module, the information from different resolutions are aggregated, which

provides a richer representation for the following detection task.

B. DETECTION HEAD

1) DETECTING NODULE AS POINTS

Inspired by recently proposed keypoint-based approach [19], [21], [22] for 2D object detection, especially the work of Zhou *et al*. [19], we present a novel formulation for detecting nodules, regarding the nodule detection problem as the estimation of corresponding center locations. Moreover, we regress the nodule size based on the local features at the centers. The differences between the work of Zhou *et al*. [19] and our 3D-CenterNet are highlighted in Table 2.

For each 3D CT scan $I \in R^{D\times H\times W}$, we predict a binary keypoint heatmap $\hat{P} \in R^{(D/r_d)\times(H/r_s)\times(W/r_s)}$ to represent the likeliness of the center location of nodules, where r_d and r_s are the output stride in depth and in the 2D spatial directions, respectively. We set $r_d = 2$ and $r_s = 4$. A prediction $\hat{P}_{xyz} = 1$ means that a nodule is present at the current position, while $\hat{P}_{xyz} = 0$ represents the background.

For each center location $c \in R^3$, we first map it to the output space, according to the given stride. Then, we obtain $\widetilde{c} = (\lfloor c_x/r_s \rfloor, \lfloor c_y/r_s \rfloor, \lfloor c_z/r_d \rfloor)$, where c_x, c_y and c_z represent the corresponding coordinates along the x, y and z axis. With the center point c, the ground-truth heatmap $P \in [0, 1]^{(D/r_d)\times(H/r_s)\times(W/r_s)}$ can be computed using a Gaussian kernel, as follows:

$$P_{\widetilde{c}} = \exp\left(-\frac{(x-\widetilde{c}_x)^2 + (y-\widetilde{c}_y)^2 + (z-\widetilde{c}_z)^2}{2\sigma_c^2}\right) \tag{2}$$

where (x, y, z) is a point in the heatmap, and σ_c is the standard deviation of the Gaussian function, which is adjusted automatically according to the nodule's size [21], determining the amount of penalty reduction to a negative point, *i.e*., not a center point. The keypoint-estimation objective is a variant of the focal loss [36], as follows:

$$L_c = \frac{-1}{N}\sum_{i=1}^{D}\sum_{j=1}^{H}\sum_{k=1}^{W}\begin{cases}(1-\hat{P}_{ijk})^{\alpha}\log(\hat{P}_{ijk}), & \text{if } P_{ijk}=1\\ (1-P_{ijk})^{\beta}(\hat{P}_{ijk})^{\alpha} & \\ \cdot\log(1-\hat{P}_{ijk}), & \text{otherwise}\end{cases} \tag{3}$$

where N is the number of nodules in a CT scan I, and α and β are the parameters of the focal loss. We set α to 2 and β to 4 in all our experiments. With the $(1 - P_{ijk})$ term computed by the 3D Gaussian kernel function, false positives around a ground truth are greatly suppressed.

We quantify the coordinates of the center locations, during the mapping operation from the input space to the output feature space, with a given downsampling stride. This may cause some precision loss, which will affect the accuracy of the predicted center location, especially for small nodules. To address this issue and recover the quantization error, we predict the offsets for each center point, to slightly adjust its location. Furthermore, we apply the L1 loss to train the offsets. Note that only the offsets at the ground-truth center locations contribute to the loss during training, while all the other locations are ignored.

$$\epsilon = \left(\frac{c_x}{r_s} - \left\lfloor\frac{c_x}{r_s}\right\rfloor, \frac{c_y}{r_s} - \left\lfloor\frac{c_y}{r_s}\right\rfloor, \frac{c_z}{r_d} - \left\lfloor\frac{c_z}{r_d}\right\rfloor\right) \tag{4}$$

$$L_{off} = \frac{1}{N}\sum_{c}\left|\epsilon - \hat{\epsilon}\right| \tag{5}$$

where ϵ and $\hat{\epsilon}$ are the ground-truth and predicted offsets, respectively. In addition, our proposed model also regresses nodule size using the features at the center. Let d_n be the diameter of nth nodule. Then the loss function for nodule size regression can be formulated as follows:

$$L_{size} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_1\left(1 - \min\left(\frac{d_n}{\hat{d}_n}, \frac{\hat{d}_n}{d_n}\right)\right) \tag{6}$$

where $\mathcal{L}_1$ is smooth L1 loss, as follows:

$$\mathcal{L}_1(t) = \begin{cases}0.5t^2 & \text{if } |t| < 1\\ |t| - 0.5 & \text{otherwise}\end{cases} \tag{7}$$

We train L_{size} with the raw voxel coordinates and do not normalize the nodule diameter. The overall training objective is given as follows:

$$L = \lambda_c L_c + \lambda_{off} L_{off} + \lambda_{size} L_{size} \tag{8}$$

Unless otherwise specified, we set $\lambda_c = 1$, $\lambda_{off} = 1$ and $\lambda_{size} = 0.1$ in all our experiments. The center heatmap $\hat{P}$, predicted sizes $\hat{S}$ and offsets $\hat{E}$ can be inferred with a single network. Furthermore, all the outputs share the same convolutional backbone network. Once we obtain the features from an input CT scan through the backbone network, the features are used to predict each of the modalities with the detection head.

2) FROM POINTS TO 3D BOUNDING BOXES

Generating 3D boxes from network predictions is straightforward. Let $\hat{C} = \{(\hat{x}_m, \hat{y}_m, \hat{z}_m)\}_{m=1}^{M}$ be the set of M detected center points. Each center location is equipped with an offset and size property, denoted as $\hat{E} = \{(\Delta\hat{x}_m, \Delta\hat{y}_m, \Delta\hat{z}_m)\}_{m=1}^{M}$ and $\hat{S} = \{\hat{d}_m\}_{m=1}^{M}$, respectively. Then, the mth 3D box can be obtained as follows:

$$\mathcal{B}_m = \left(\hat{x}_m + \Delta\hat{x}_m - \frac{\hat{d}_m}{2}, \hat{y}_m + \Delta\hat{y}_m - \frac{\hat{d}_m}{2}, \hat{z}_m + \Delta\hat{z}_m - \frac{\hat{d}_m}{2}, \hat{x}_m + \Delta\hat{x}_m + \frac{\hat{d}_m}{2}, \hat{y}_m + \Delta\hat{y}_m + \frac{\hat{d}_m}{2}, \hat{z}_m + \Delta\hat{z}_m + \frac{\hat{d}_m}{2}\right) \tag{9}$$

C. TWO-STREAM NETWORK FOR FALSE POSITIVE REDCUTION

As shown in Table 4, we analyze the detection results generated by our proposed model and, observe that only 1.66% of the predicted candidates are true positives, however, 91.04% of them are false positives. Moreover, we visualize some of the real nodules and tissues from the LUNA16 dataset. As shown in Fig. 4, nodules and tissues are very similar in appearance, which causes the 3D-CenterNet incapable of distinguishing them correctly. Consequently, the model

TABLE 4. Statistic analysis of the detection results. TPs means true positives, FPs the false positives, ICs the ignore candidates, and DoD the double detection.

	TPs	FPs	ICs	DoD
Percentage (%)	1.66	91.04	7.21	0.09

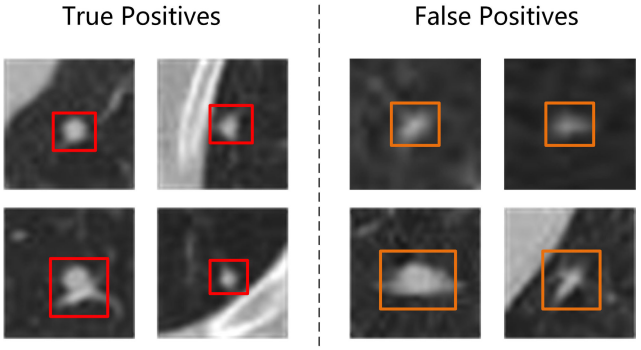


FIGURE 4. True positives (nodules) and false positives (*e.g.*, tissue) are very similar in appearance.

produces a large number of false positives, which affect the final performance considerably.

Therefore, it is crucial to design an effective method to filter out these false positives. We propose a novel two-stream network for false positive reduction (TSN-FPR), which integrates the information from the image stream and the motion-history stream (MHS), enhancing the discrimination between real nodules and false positives. The overall framework of the TSN-FPR is showed in Fig. 3.

Specifically, the idea of the motion-history stream is inspired by the work of [37]. The intensity value of MHS $M(s, y, x)$ within $(1, \tau)$ slice can be computed given the pixel position (x, y) on slice s. The detailed process of computing MHS is illustrated in Alg. 1. As shown in Fig. 5, the patterns between real nodules and tissues in the motion-history stream are different, which can act as an additional cue to differentiate between normal tissues and nodules.

Algorithm 1 Process of Computing Motion-History Stream

Input: Threshold ϵ, duration τ, CT scan $I \in R^{S\times H\times W}$
Output: The motion history image stream $M \in R^{S\times H\times W}$

initialize $M \leftarrow 0$
$D^{S\times H\times W} \leftarrow 0$ // variation between slices
$\Psi^{S\times H\times W} \leftarrow 0$. // update function
for each s in $[1, S]$ **do**
  for each position (x, y) in $I(s)$ **do**
    $D(s, y, x) \leftarrow |I(s, y, x) - I(s-1, y, x)|$
    if $D(s, y, x) \geq \epsilon$ **then**
      $\Psi(s, y, x) \leftarrow 1$
    end if
    if $\Psi(s, y, x) = 1$ **then**
      $M(s, y, x) \leftarrow \tau$
    else
      $M(s, y, x) \leftarrow \max(0, M(s-1, y, x) - 1)$
    end if
  end for
end for
return M

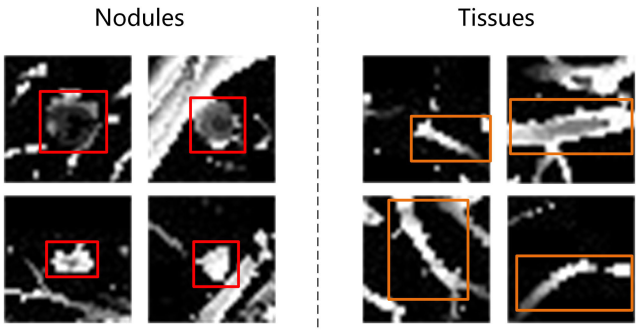


FIGURE 5. The patterns between nodules and tissues are different in the motion-history image stream. Specifically, nodules typically have a specific region, *e.g.*, circular, with either a brighter or darker center. However, the motion history images of tissues present an irregular appearance and, tend to extend in a specific direction.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. DATASETS

To evaluate the performance of the proposed 3D-CenterNet, we employ two publicly available datasets: LUNA16 and SPIE-AAPM. LUNA16 dataset is used for training and testing, and due to the limited data size of SPIE-AAPM, it is employed for evaluation only. The details of these two datasets are summarized in Table 5.

LUNA16 *Dataset*: LUNA16 challenge dataset is the most widely used dataset when developing nodule detection algorithms. It contains 888 CT scans and a total of 1186 lung nodules, with the nodule locations and diameters accepted by at least 3 out of 4 experienced radiologists. Nodules that are less than 3mm in diameter and annotated by only 1 or 2 radiologists are ignored during training. Formally, LUNA16 is divided into 10 subsets for cross-validation, alternately applying nine subsets as training and one subset as testing, and the average performance is reported.

TABLE 5. Detailed information of the datasets.

Dataset	Manufacturer	#Scans	#Nodules
LUNA16	GE Healthcare	888	1186
SPIE-AAPM	Philips Healthcare	70	83

SPIE-AAPM *Dataset*: Collected for the 'Grand Challenge', SPIE-AAPM dataset [40] aims at developing quantitative algorithms for medical image analysis, especially for classifying the lung nodules of being malignant or benign. It consists of 70 CT scans from 70 participants, with a total of 22489 images. The nodule locations and diagnoses of each CT scan are given as annotation information. We employ this dataset for evaluation to verify the generalization ability of our 3D-CenterNet.

B. DATA PREPROCESSING

Data preprocessing is critical in order to get a sufficiently accurate nodule detection model. The spacing (mm/pixel)

TABLE 6. Performance comparison with other methods on the LUNA16 dataset. We show the sensitivity (%) at 7FP/scan rates, including 1/8, 1/4, 1/2, 1, 2, 4, and 8. The performance measures of other methods were copied from the original publication.

Methods	1/8	1/4	1/2	1	2	4	8	CPM score
MOT-M5Lv1	59.7	67.0	71.8	75.9	78.8	81.6	84.3	74.2
Gupta *et al.* [32]	53.1	62.9	79.0	83.5	84.3	84.8	85.6	76.3
Xie *et al.* [38]	49.3	68.8	79.6	85.2	86.4	86.4	86.4	77.5
Dou *et al.* [7]	67.7	73.7	81.5	84.8	87.9	90.7	92.2	82.7
Dou *et al.* [39]	65.9	74.5	81.9	86.5	90.6	93.3	94.6	83.9
Zhu *et al.* [5]	69.2	76.9	82.4	86.5	89.3	91.7	93.3	84.2
Wang *et al.* [10]	67.6	77.6	87.9	94.9	**95.8**	95.8	95.8	87.8
Ding *et al.* [11]	74.8	**85.3**	88.7	92.2	93.8	94.4	94.6	89.1
Khosravan and Bagci [8]	70.9	83.6	**92.1**	**95.3**	95.3	95.3	95.3	89.7
3D-CenterNet (Ours)	71.3	80.1	86.7	91.7	95.0	**96.2**	**97.1**	88.3
3D-CenterNet (Ours)*	**78.4**	84.7	90.6	93.8	95.0	95.5	96.1	**90.6**

* shows the results after applying TSN-FPR for false positive reduction.

among CT scans is inconsistent and the resampling operation is adopted to ensure the spacing is 1 mm. To save computation, the black borders in images are clipped by setting a fixed threshold. Moreover, we transform each CT scan with an effective value of Hounsfield unit between $[-1200, 600]$ into the grey value of $[0, 255]$ using a linear transformation.

C. EXPERIMENTAL SETTINGS

1) 3D-CenterNet

a: TRAINING

During training, we randomly sample 3D volumes along the z-axis, extracting 16 slices in succession as the input of 3D-CenterNet. Then, the sampled volumes are rescaled to the resolution of $512 \times 512 \times 16$. For data augmentation, we flip the images about the x and z axis randomly, and use random rotation (choose from $\{0°, 90°, 180°, 270°\}$), random scaling (between 0.8 and 1.4) and random cropping on every slice. Adam [41] is adopted to optimize the training objective. The network is trained for a total of 230 epochs with a batch size of 8 on a V100 GPU. We use an initial learning rate of 1e-4, and drop it at 180 and 210 epochs with a ratio of 0.1, respectively. The parameters of our network are randomly initialized.

b: INFERENCE

In the inference time, we take 16 overlapped slices each time as the input volume. A 3D max pooling layer with a kernel of size $3 \times 3 \times 3$ is applied to the center heatmap to suppress 26 negative points in a $3 \times 3 \times 3$ region, and finally, the top 32 centers are reserved. Flipping testing is employed by default. Non-maximum suppression (NMS) is applied to the detected 3D bounding boxes to obtain the final detection results.

2) TSN-FPR

The cross-entropy loss is used for training the TSN-FPR. Image patches centered with each nodule candidate are cropped from the whole CT scan with a fixed scale of $15 \times 32 \times 32$. Then, each image patch is resized to 40×40 pixels, as the input of the TSN-FPR. Since the imbalance between positive and negative candidates (approximately $1/250$), each positive candidate is rotated by 0, 90, 180, 270 degrees. For each rotated candidate, the flipping operation is further applied on the x, y, and z axis, respectively, thus generating a total of 12 augmented versions for each of the positive candidates.

In training, the network is trained for a total of 24 epochs with a batch size of 128. The initial learning rate is set to 0.01 and decreases by a ratio of 0.1 at 16 and 22 epochs. The average prediction time for one candidate is about 21.0 milliseconds with one GTX 1070 GPU.

D. EVALUATION METRICS

Following [42], Free-Response Receiver Operating Characteristic (FROC) [12] and Competition Performance Metric (CPM) are employed to evaluate the performance of our method. For comparison with other nodule detectors, we perform 10-fold cross-validation and the final results are obtained by averaging the 10 experiments. Specifically, the FROC curve plots the nodule sensitivity and corresponding false positives. As recommended by the LUNA challenge organizers, CPM score is obtained by averaging the sensitivity at 7 FP/scan rates (*i.e.*, 0.125, 0.25, 0.5, 1, 2, 4, 8).

E. EXPERIMENTAL RESULTS AND ANALYSIS

1) RESULTS

a: COMPARISON WITH OTHER METHODS

Table 6 shows the results of our proposed 3D-CenterNet and other nodule detectors on the LUNA16 dataset, with false-positive rates at 1/8, 1/4, 1/2, 1, 2, 4, and 8 per scan, respectively. The best performance achieved by various methods for each of the false-positive rates is highlighted in bold. As shown in the table, our proposed method outperforms other state-of-the-art anchor-based algorithms, which implies that without pre-defined anchors, our simpler center-point-based formulation can still achieve good performance. We yield a sensitivity of 97.1% at 8 FPs/scan, outperforming all other methods, which is what the CAD

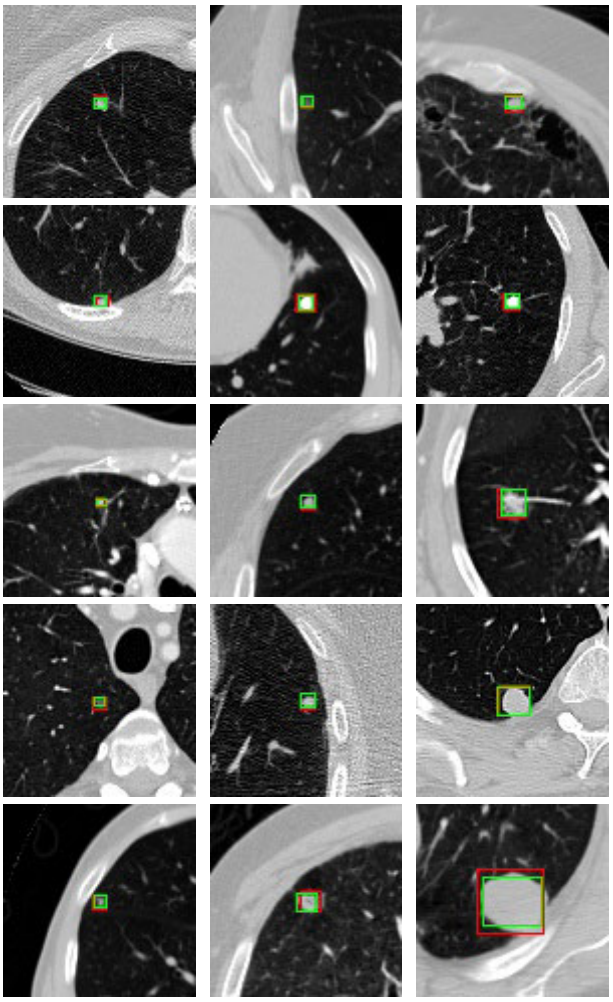


FIGURE 6. Visualization of some detected true positives with different sizes by using our proposed 3D-CenterNet. We only show the central slices of the nodule, and, for better visualization, we crop and zoom in on the regions around the nodule. Detection results are shown with green box, while the red box indicates the ground-truth nodules. The visualization results show that our 3D-CenterNet framework is capable of detecting nodules of different sizes, based on features with a single scale only.

systems require clinically (higher recall rate). By applying our proposed two-stream network for false positive reduction, 3D-CenterNet outperforms Zhu *et al*. [5] by 6.4%, and obtains an improvement of 0.9% compared to Khosravan and Bagci [8], demonstrating the effectiveness of our proposed detector. Some visual nodule detection results are illustrated in Fig. 6.

b: VERIFICATION OF GENERALIZATION ABILITY

Besides, we evaluate the performance of our model on SPIE-AAPM dataset. Notably, we only train on LUNA16 dataset and report the CPM score and the sensitivity at 7 FP/scan rates. As shown in Table 7, although the sensitivity at low false-positive levels, *e.g*., 1/8, 1/2, is quite low, 3D-CenterNet still yields 60.9% CPM score and a sensitivity of 92.8% at 8 false positives per scan. It is worth mentioning that the equipment manufacturers used to collect CT scans for these two datasets are different, which causes

TABLE 7. FROC performance of 3D-CenterNet on the SPIE-AAPM dataset. Note that we only train our model on the LUNA16 dataset. The CPM score (%) and the sensitivity (%) at 7 FP/scan rates are reported.

1/8	1/4	1/2	1	2	4	8	CPM score
12.0	37.3	51.8	69.9	77.1	85.5	92.8	60.9

remarkable differences on the scans, *e.g*., intensity, matching setting and image protocol, etc., and affects the performance of the model considerably. Even if the significant differences, our model can still achieve high sensitivity (recall) on SPIE-AAPM dataset, demonstrating the generalization ability of our proposed model.

2) EFFECTIVENESS OF TWO-STREAM NETWORK FOR FP REDUCTION

We conduct two experiments to demonstrate the effectiveness of the proposed two-stream network for false positive reduction. As shown in Fig. 7a, after applying the proposed TSN-FPR network, the sensitivities of low FP levels are boosted considerably, *e.g*., with an improvement of 7.1% at 1/8 FP level, and improvement of 3.6% at 1/4 FP level, respectively. However, due to the incorrect classification (filtering out the true positives), the sensitivity at high FP level drops a little bit (96.1% *vs* 97.1% at 8 FP level). Besides, the number of nodule candidates from a total of 67 CT scans is further showed in Fig. 7b. TSN-FPR filters out three true positives due to the incorrect classification. On the other hand, the number of false positives drops considerably by applying TSN-FPR, which reduces 97.3% of the false positives. This demonstrates the effectiveness of our proposed two-stream network for false positive reduction.

3) EFFECT OF INPUT RESOLUTION IN TRAINING AND TESTING

In the training phase, we adopt a fixed input resolution of $512 \times 512 \times 16$. During testing, we keep the original CT slice resolution and zero-pad each slice with up to 32 pixels. Besides, we test our model with fixed resolutions of 256, 384, and 512, respectively. Notably, all 3D input CT scans share the same depth of 16. Table 8 shows the results under various testing resolutions. Maintaining the original resolution achieves better performance than fixing testing resolution at all false-positive rates. Testing in lower resolution ($256 \times 256 \times 16$) runs 1.3 times faster but significantly drops the CPM score (0.889 *vs.* 0.463). The average inference time for one CT scan ranges from 10.7 seconds to 17.7 seconds, under various testing resolutions.

In addition, we evaluate the performance of different nodule sizes, under various testing resolutions. The diameters in the range from 3mm to 10mm are considered as small nodules, those with the diameter in the range from 10mm to 20mm belong to medium level, and those with a diameter greater than 20mm are large nodules. For convenience, we use CPM^s, CPM^m, and CPM^l to represent the CPM score for small, medium, and large nodules, respectively. As shown in Table 9, CPM^l is always equal to 100.0%,

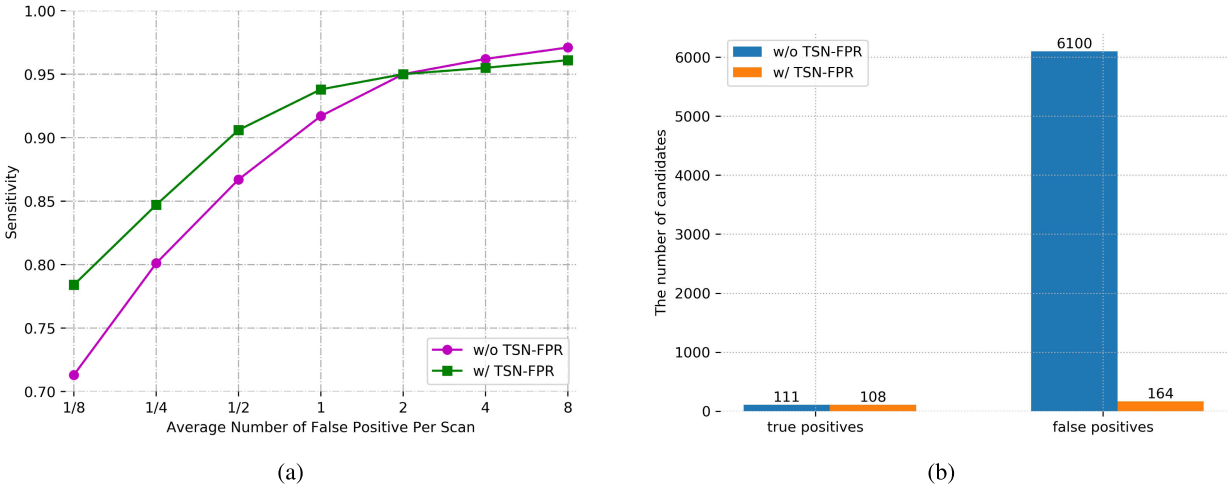


FIGURE 7. Comparison between the performance (left) and the number of candidates (right), with and without false positive reduction.

TABLE 8. Effect of testing resolution. Larger resolutions obtain higher performance but run slower. However, keeping the original resolution for evaluation is the best choice, which achieves the best balance between speed and performance. Resolution is the height and width of each CT slice and the 3D volume depth is fixed at 16. Time is in seconds/scan.

Resolution	1/8	1/4	1/2	1	2	4	8	CPM score	Time(s)	Params(M)	FLOPs(G)
256	23.2	31.2	36.6	49.1	53.6	63.4	67.0	46.3	**10.7**	21.2	426.0
384	54.5	67.9	73.2	79.5	84.8	87.5	92.9	77.2	17.0	21.2	958.5
512	70.5	75.0	83.0	**91.1**	**93.8**	96.4	**98.2**	86.9	17.7	21.2	1704.0
Original	**75.9**	**79.5**	**86.6**	**91.1**	**93.8**	**97.3**	**98.2**	**88.9**	14.3	21.2	665.6

TABLE 9. CPM scores for different nodule size, under various testing resolutions. A higher resolution is essential for performance improvement of small nodules.

Resolution	CPM^s	CPM^m	CPM^l
256	32.7	85.2	**100.0**
384	72.8	97.9	**100.0**
512	83.3	**98.4**	**100.0**
Original	**86.7**	96.3	**100.0**

which indicates that 3D-CenterNet is excellent in locating large nodules. The improvement of performance is mainly reflected in small and medium-sized nodules, with CPM^s and CPM^m increased by 54.0% and 11.1%, respectively, when the resolution changes from 256 to the original. This suggests that a higher resolution is essential for detecting small nodules accurately.

4) ANALYSIS OF VARIOUS REGRESSION LOSS

In order to achieve more accurate predictions for nodule diameter, we conduct experiments on three different regression losses: L1 loss, smooth L1, and our proposed L_{size} that is less sensitive to the nodule size. The experiments are tabulated in Table 10. L1 loss performs better than smooth L1 at low FP levels, *e.g.*, 1/8, 1/4, 1/2, etc. However, they yield the same sensitivity of 97.3% at 8 false positives per scan. L_{size} achieves the best performance compared to the other two regression losses, especially at high false-positive rates, with an increase of 0.9% at 8 FP/scan.

5) ANALYSIS OF NODULE SIZE REGRESSION WEIGHT

We analyze the impact of λ_{size} on our model. As is showed in Table 11, $\lambda_{size} = 0.1$ is a preferred choice because it works better than the other parameter settings, with improvement of 1.0% and 0.6% over $\lambda_{size} = 0.03$ and $\lambda_{size} = 0.3$, respectively. For lower values (< 0.1), the performance drops a little bit, since the value is too small to help the network properly learning to predict the nodule size. Although we train L_{size} with the raw coordinates, the $\min(d_n/\hat{d}_n, \hat{d}_n/d_n)$ term normalizes the loss value into a range of [0, 1]. Therefore, a larger value of λ_{size} still works well.

6) EFFECT OF TRAINING SCHEDULE

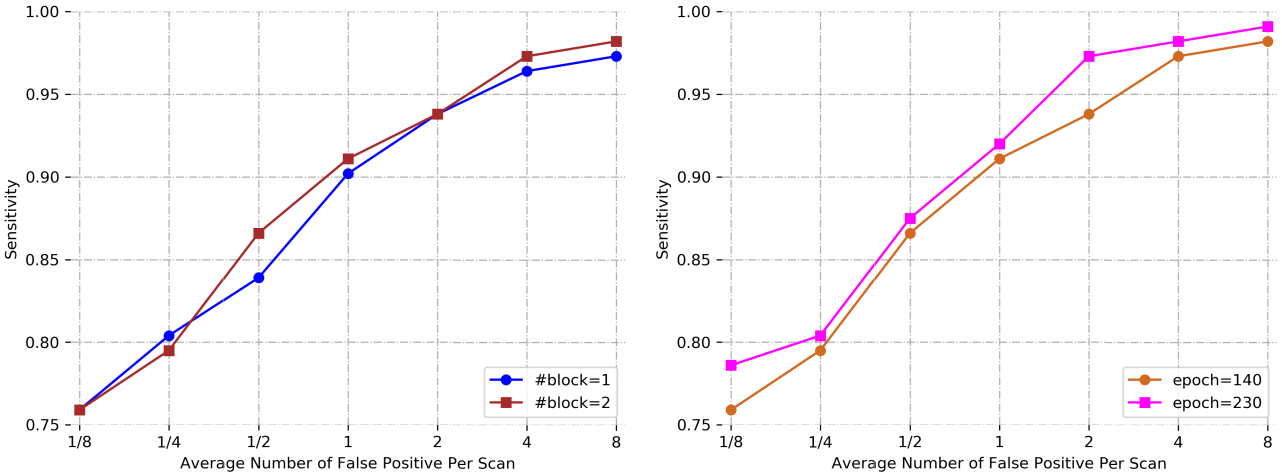
Moreover, we analyze the effect of the training schedule on our model by varying the total training epoch. Specifically, we first train 3D-CenterNet for 140 epochs and, decay the learning rate by a ratio of 0.1 at 90 epochs. Then, the training epoch is further extended to 230. As shown in Fig. 8b, with a longer training schedule, we further boost the performance at the cost of more computational resources. Therefore, we train all the models for 140 epochs in ablation experiments, while using 230 epochs training schedule when comparing to other methods.

7) IMPACT OF MODEL CAPACITY

By default, we employ two residual units in stage 1 to stage 4 in the backbone network, denoted as $\#block = 2$. In addition, we analyse the impact of model capacity by using only one residual unit, represented as $\#block = 1$. Fig. 8a shows that the larger model achieves slightly higher

TABLE 10. Analysis of various regression loss. Our proposed nodule size regression loss works better than Smooth L1 and L1 loss.

Loss	1/8	1/4	1/2	1	2	4	8	CPM score
Smooth L1	70.5	77.7	84.8	91.1	94.6	**97.3**	97.3	87.6
L1	71.4	78.6	**87.5**	**92.9**	**95.5**	96.4	97.3	88.5
L_{size}	**75.9**	**79.5**	86.6	91.1	93.8	**97.3**	**98.2**	**88.9**



(a) Impact of model capacity.

(b) Effect of training schedule.

FIGURE 8. Analysis of the model capacity and training schedule. (a) Larger model achieves better performance with stronger ability to learn the mapping function from input CT scan to corresponding nodule center locations. (b) Training longer performs better.

TABLE 11. Analysis of nodule size regression weight. $\lambda_{size} \geq 0.1$ achieves preferable results.

λ_{size}	CPM score
0.03	87.9
0.1	**88.9**
0.3	88.3

CPM score, implying that a larger model is capable of learning the nodule detection task better, because of the stronger ability.

V. CONCLUSION

In this work, a novel representation for detecting nodules is first proposed, in terms of their 3D spatial center locations. Based on this novel representation, an anchor-free framework for lung nodule detection, called 3D-CenterNet, is presented. The proposed nodule detector finds nodules' centers and, regresses their corresponding diameters. To reduce false positives generated by the detection model, we further propose a novel two-stream network, which aggregates information from image stream and motion-history stream, to enhance the discrimination between tissues and nodules. The resulting model is much simpler, while still obtaining state-of-the-art performance. In the future, we will continue to focus on predicting other properties, *e.g.*, nodule density, one of the important criteria for judging the benign and malignant nodules. This can provide radiologists with more information about lesions, easing their diagnosis process considerably.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA, A Cancer J. Clin.*, vol. 65, no. 1, pp. 5–29, Jan. 2015.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[3] R. H. Mak, M. G. Endres, J. H. Paik, R. A. Sergeev, H. J. W. L. Aerts, C. L. Williams, K. R. Lakhani, and E. C. Guinan, "Use of crowd innovation to develop an artificial intelligence-based solution for radiation therapy targeting," *JAMA Oncol.*, vol. 5, no. 5, pp. 654–661, 2019.

[4] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Rev. Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[5] W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 673–681.

[6] H. Cao, H. Liu, E. Song, G. Ma, X. Xu, R. Jin, T. Liu, and C.-C. Hung, "Two-stage convolutional neural network architecture for lung nodule detection," 2019, *arXiv:1905.03445*. [Online]. Available: http://arxiv.org/abs/1905.03445

[7] 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 Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[8] N. Khosravan and U. Bagci, "S4ND: Single-shot single-scale lung nodule detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 794–802.

[9] X. Huang, J. Shan, and V. Vaidya, "Lung nodule detection in CT using 3D convolutional neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 379–383.

[10] B. Wang, G. Qi, S. Tang, L. Zhang, L. Deng, and Y. Zhang, "Automated pulmonary nodule detection: High sensitivity with few candidates," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 759–767.

[11] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 559–567.

[12] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sanchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[13] J. Liu, L. Cao, O. Akin, and Y. Tian, "Accurate and robust pulmonary nodule detection by 3D feature pyramid network with self-supervised feature learning," 2019, *arXiv:1907.11704*. [Online]. Available: http://arxiv.org/abs/1907.11704

[14] Y. Li and Y. Fan, "DeepSEED: 3D squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1866–1869.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] H. Tang, C. Zhang, and X. Xie, "NoduleNet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation," 2019, *arXiv:1907.11320*. [Online]. Available: http://arxiv.org/abs/1907.11320

[17] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 17, 2019, doi: 10.1109/TPAMI.2019.2929257.

[18] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," 2016, *arXiv:1611.05424*. [Online]. Available: http://arxiv.org/abs/1611.05424

[19] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[21] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[22] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[23] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[26] B. Xiao, K. Wang, X. Bi, W. Li, and J. Han, "2D-LBP: An enhanced local binary feature for texture image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2796–2808, Sep. 2019.

[27] S. C. van de Leemput, F. Dorssers, and B. E. Bejnordi, "A novel spherical shell filter for reducing false positives in automatic detection of pulmonary nodules in thoracic ct scans," *Proc. SPIE*, vol. 9414, Mar. 2015, Art. no. 94142P.

[28] B. Chen, T. Kitasaka, H. Honma, H. Takabatake, M. Mori, H. Natori, and K. Mori, "Automatic segmentation of pulmonary blood vessels and nodules based on local intensity structure analysis and surface propagation in 3D chest CT images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 7, no. 3, pp. 465–482, May 2012.

[29] S. Akram, M. Y. Javed, M. U. Akram, U. Qamar, and A. Hassan, "Pulmonary nodules detection and classification using hybrid features from computerized tomographic images," *J. Med. Imag. Health Informat.*, vol. 6, no. 1, pp. 252–259, Feb. 2016.

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

[31] B. van Ginneken *et al.*, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study," *Med. Image Anal.*, vol. 14, no. 6, pp. 707–722, 2010.

[32] A. Gupta, T. Saar, O. Martens, and Y. L. Moullec, "Automatic detection of multisize pulmonary nodules in CT images: Large-scale validation of the false-positive reduction step," *Med. Phys.*, vol. 45, no. 3, pp. 1135–1149, Mar. 2018.

[33] B.-C. Kim, J.-S. Choi, and H.-I. Suk, "Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection," *Neural Netw., Off. J. Int. Neural Netw. Soc.*, vol. 115, pp. 1–10, Jul. 2019.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*. [Online]. Available: http://arxiv.org/abs/1902.09212

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[37] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *Proc. IEEE Workshop Detection Recognit. Events Video*, Jul. 2001, pp. 39–46.

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

[39] Q. Dou, H. Chen, Y. Jin, H. Lin, J. Qin, and P.-A. Heng, "Automated pulmonary nodule detection via 3D convnets with online sample filtering and hybrid-loss residual learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 630–638.

[40] S. G. Armato, L. M. Hadjiiski, G. D. Tourassi, K. Drukker, M. L. Giger, F. Li, G. Redmond, K. Farahani, J. S. Kirby, and L. P. Clarke, "Lungx challenge for computerized lung nodule classification: Reflections and lessons learned," *J. Med. Imag.*, vol. 2, no. 2, 2015, Art. no. 020103.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

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

ZEHUI GONG received the B.S. degree in electronic science and technology from the Guangdong University of Technology (GDUT), Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree in control science and engineering. His current research interests include computer vision, facial expression recognition, and lesion detection.

DONG LI received the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, under the supervision of Prof. K.-M. (Kenneth) Lam, in February 2014.

He is currently an Associate Professor with the Faculty of Automation, Guangdong University of Technology (GDUT). His work designed robust and distinctive features to describe pore-scale facial key points, such as pores, fine winkles, and hair. He focus on designing new pore-scale feature extraction algorithms and developing pore-scale facial feature applications, such as face verification and adapting existing algorithms to pore-scale application. His research interests include computer vision, pattern recognition, image analysis, color correction, and image restoration.

JIATAI LIN received the B.S. and M.S. degrees from the Guangdong University of Technology, Guangdong, China, in 2017 and 2020, respectively. His research interests include machine learning and medical image.

YUN ZHANG received the B.S. and M.S. degrees in automatic engineering from Hunan University, Changsha, China, in 1982 and 1986, respectively, and the Ph.D. degree in automatic engineering from the South China University of Science and Technology, Guangzhou, China, in 1998.

He is currently a Full Professor with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include intelligent control systems, multiagent systems, neural networks, and signal processing.

KIN-MAN LAM (Member, IEEE) received the Associateship degree (Hons.) in electronic engineering from The Hong Kong Polytechnic University (Hong Kong Polytechnic), in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, The University of Sydney, Sydney, NSW, Australia, in August 1996.

He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, as an Assistant Professor, in October 1996, where he was an Associate Professor in February 1999. He is actively involved in professional activities. His current research interests include human-face recognition, image and video processing, and computer vision.

Dr. Lam is a Treasurer of the IEEE Hong Kong Chapter of Signal Processing, a member of the Program Committee of the Advanced Concepts for Intelligent Vision Systems (ACIVS) in 2004, the Eighth International Conference on Control, Automation, Robotics, and Vision (ICARCV) in 2004, and the IASTED International Conference on Internet and Multimedia Systems and Applications (EuroIMSA) in 2005. He received the Australia Postgraduate Award, the IBM Australia Research Student Project Prize, and the S. L. POA Scholarship for overseas studies. He serves as the Technical Chair for the 2004 International Symposium on Intelligent Multimedia, Video, and Speech Processing (ISIMP) in 2004 and the Technical Co-Chair for the 2005 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) in 2005. He served as a Guest Editor for the Special Issue on Biometric Signal Processing of *EURASIP Journal on Applied Signal Processing*.