Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection

Qi Dou*, *Student Member, IEEE*, Hao Chen, *Student Member, IEEE*, Lequan Yu, *Student Member, IEEE*, Jing Qin, *Member, IEEE*, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*— *Objective*: False positive reduction is one of the most crucial components in an automated pulmonary nodule detection system, which plays an important role in lung cancer diagnosis and early treatment. The objective of this paper is to effectively address the challenges in this task and therefore to accurately discriminate the true nodules from a large number of candidates. *Methods*: We propose a novel method employing three-dimensional (3-D) convolutional neural networks (CNNs) for false positive reduction in automated pulmonary nodule detection from volumetric computed tomography (CT) scans. Compared with its 2-D counterparts, the 3-D CNNs can encode richer spatial information and extract more representative features via their hierarchical architecture trained with 3-D samples. More importantly, we further propose a simple yet effective strategy to encode multilevel contextual information to meet the challenges coming with the large variations and hard mimics of pulmonary nodules. *Results*: The proposed framework has been extensively validated in the LUNA16 challenge held in conjunction with ISBI 2016, where we achieved the highest competition performance metric (CPM) score in the false positive reduction track. *Conclusion*: Experimental results demonstrated the importance and effectiveness of integrating multilevel contextual information into 3-D CNN framework for automated pulmonary nodule detection in volumetric CT data. *Significance*: While our method is tailored for pulmonary nodule detection, the proposed framework is general and can be easily extended to many other 3-D object detection tasks from volumetric medical images, where the targeting objects have large variations and are accompanied by a number of hard mimics.**

***Index Terms*—Computer-aided diagnosis, deep learning, false positive reduction, pulmonary nodule detection, 3-D convolutional neural networks.**

Manuscript received May 21, 2016; revised August 4, 2016; accepted September 12, 2016. Date of publication September 26, 2016; date of current version June 15, 2017. This work was supported in part by the Research Grants Council of The Hong Kong Special Administrative Region under Project CUHK 412513, in part by the National Natural Science Foundation of China under Project 61233012, and in part by the Shenzhen-Hong Kong Innovation Circle Funding under Program SGLH20131010151755080 and Program GHP/002/13SZ. *Asterisk indicates corresponding author.*

*Q. Dou, H. Chen, L. Yu, and P. A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong(e-mail: qdou@cse.cuhk.edu.hk).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University.

Digital Object Identifier 10.1109/TBME.2016.2613502

I. Introduction

AUTOMATED detection of pulmonary nodules in volumetric thoracic computed tomography (CT) scans plays an important role in computer-aided lung cancer diagnosis and early treatment [1]–[3]. The pulmonary nodules are radiologically visible as small structures that are roughly spherical opacities within the pulmonary interstitium images [4]. They have been regarded as crucial indicators of primary lung cancer, which has been the leading cause of cancer death in recent years [5]. Based on reliable detection of lung nodules, radiologists and surgeons can perform size measurements and appearance characterizations for cancer malignancy diagnosis [6] and, if necessary, timely surgical intervention in order to increase the survival chances of patients [7], [8].

An automated pulmonary nodule detection system mainly consists of two steps: 1) candidate screening and 2) false positive reduction. In candidate screening, a considerable number of coarse candidates are rapidly screened throughout the whole volume using a variety of criteria, e.g., intensity thresholding, shape curvedness, and mathematical morphology [3], [9], [10]. In false positive reduction, effective classifiers together with discriminative features are developed to reduce a large number of false positive candidates. In order to maintain a high sensitivity in candidate screening, the criteria employed in this step are usually quite straightforward and lenient, and consequently a great number of candidates are selected out and forwarded to the second step. In this regard, the false positive reduction stands as the most crucial component of an automated pulmonary nodule detection system [1] and a lot of efforts have been dedicated to improving the performance of this step.

Automated identification of the pulmonary nodules from thoracic CT scans is, however, among the most challenging tasks in computer-aided chest radiograph analysis [11] for at least the following two reasons. First, the pulmonary nodules have large variations in sizes, shapes, and locations, as shown in the green rectangle in Fig. 1. Moreover, the contextual environments around them are often diversified for different categories of lung nodules, such as solitary nodules, ground-glass opacity nodules, cavity nodules, and pleural nodules [12]. Second, some false positive candidates carry quite similar morphological appearance to the true pulmonary nodules, as shown in the red

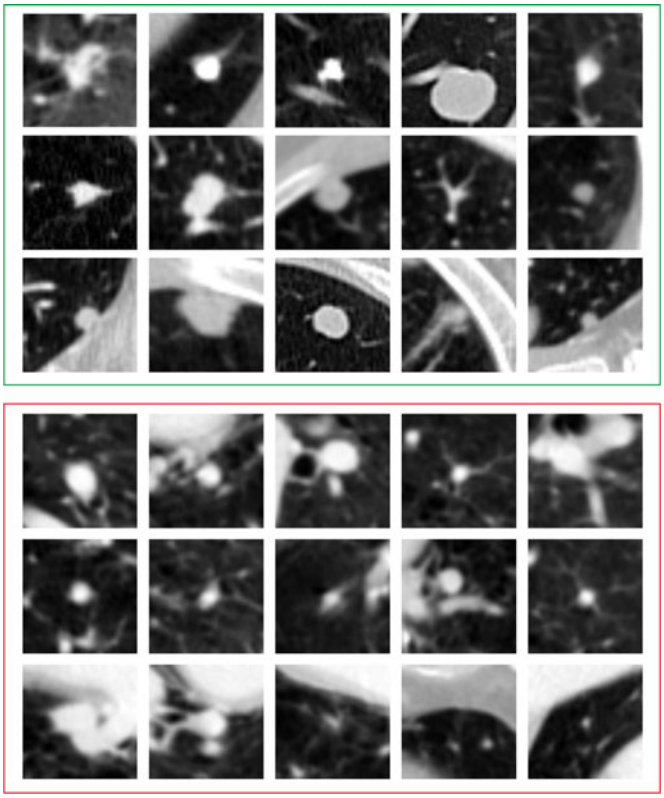

Fig. 1. Examples of the pulmonary nodules with various sizes, shapes, and locations (green rectangle), and the false positive candidates (red rectangle) which carry similar appearance and make the task challenging. Each example is a representative 2-D transverse plane extracted from a location.

rectangle in Fig. 1. The existence of these hard mimics would heavily hinder the detection process.

Many research works have devoted efforts in developing efficient and robust false positive reduction algorithms in order to meet the aforementioned challenges. Some of them endeavored to design representative features for pulmonary nodules by combining a set of discriminative characteristics of the nodules. For examples, Messay *et al.* [8] designed a set of shape, position, intensity, and gradient features from segmented nodule candidates, and achieved a detection sensitivity of 82.66% with an average of three false positives per scan. Jacobs *et al.* [3] employed features based on the intensity, shape, texture characteristics, and incorporated contextual information in respect to some surrounding anatomical structures. This method achieved a detection sensitivity of 80% with 1.0 false positive per scan. Unfortunately, these hand-crafted features tend to suffer from limited representation capability and are insufficient to deal with the large variations of lung nodules.

Recently, with the remarkable successes of deep convolutional neural networks (CNNs) in image and video processing [13]–[16], the representation capability of the high-level features which are learned from large amounts of training data has been broadly recognized. This also inspired some researchers to employ CNNs in automated pulmonary nodule detection. In a recent work, Setio *et al.* [17] proposed to employ two-dimensional (2-D) multiview convolutional networks to learn representative features for pulmonary nodule detection. This method can incorporate relatively wide volumetric spatial information for detection by extracting many 2-D patches from differently oriented planes. Superior to those works employing low-level hand-crafted features, this method achieved a state-of-the-art detection sensitivity of 85.4% at 1.0 false positive per subject on the benchmark of Lung Image Database Consortium–Image Database Resource Initiative (LIDC–IDRI) [11] dataset, demonstrating the effectiveness of convolutional networks on this task. However, this 2-D CNNs based solution still could not take full advantage of 3-D spatial contextual information of pulmonary nodules to single them out from hard mimics and complicated environments. After all, detecting pulmonary nodules from volumetric CT scans is, in essence, a 3-D object detection problem.

From a broader perspective, while 2-D CNNs have witnessed many fruitful applications in medical image analysis field in recent years [18]–[20], 3-D CNN is still in its infant stage in medical applications even though 3-D medical data are quite common and popular in clinical practice. Just a few 3-D variants of CNNs have been very lately proposed for medical image computing [21]–[25]. As alternatives, some variants of 2-D CNNs attempted to exploit sequentially adjacent slices [26]; orthogonal planes [27] or multiview planes [17] to aggregate more 3-D spatial information in the network. However, due to the nature of 2-D network architecture, it is difficult for these solutions to sufficiently encompass 3-D spatial information within volumetric data into the model.

In this paper, we propose a novel framework on top of 3-D CNNs for false positive reduction in automated pulmonary nodule detection from CT images. By taking full advantage of the 3-D spatial information, our method can learn representative features with higher discrimination capability than those learned from 2-D CNNs. To the best of our knowledge, this is a pioneer work that exploits 3-D CNNs for pulmonary nodule detection in volumetric CT scans. To deal with the large variations of pulmonary nodules and more robustly distinguish them from their hard mimics, we further propose to consider multilevel contextual information around pulmonary nodules by integrating a set of 3-D CNNs with different sizes of receptive field. Experiments performed on a large-scale benchmark dataset demonstrate the effectiveness of the 3-D CNNs as well as the multilevel contextual information integration strategy for improving the detection accuracy.

Our main contributions can be summarized as:

1) We propose a novel method to exploit 3-D CNNs for pulmonary nodule detection in volumetric CT scans; compared with their 2-D counterparts, the 3-D CNNs can encode richer spatial information and extract more discriminative representations via the hierarchical architecture trained with 3-D samples;
2) Considering the complicated anatomical surrounding environments of pulmonary nodules, we propose a simple yet effective strategy to encode multilevel contextual information to meet the challenges coming with the large variations and hard mimics of pulmonary nodules;
3) We validated our proposed framework on the LUNA16 challenge held in conjunction with ISBI 2016. Our team achieved the highest score in the false positive

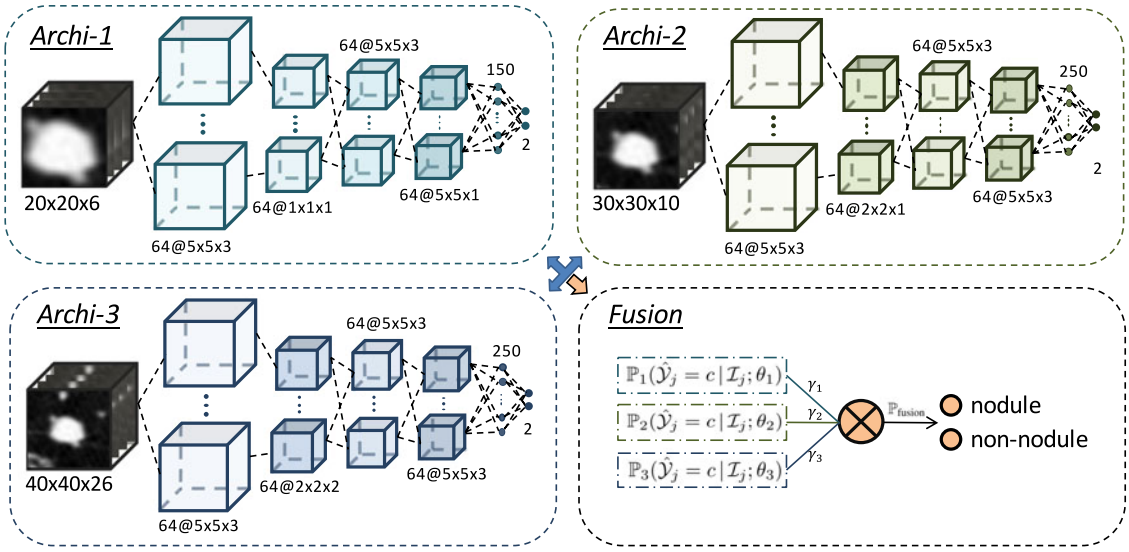

Fig. 2. Framework of the proposed method. We design three 3-D convolutional networks incorporating different levels of contextual information. The posterior predictions of these networks are fused to produce the final classification result.

reduction track, corroborating the outstanding efficacy of our method.

The remainder of this paper is organized as follows. We describe our method in Section II and report the experimental results in Section III. Section IV further discusses some key issues of the proposed method. The conclusions are drawn in Section V.

II. Methodology

The proposed multilevel contextual 3-D CNNs framework for false positive reduction in automated pulmonary nodule detection is illustrated in Fig. 2. We develop three 3-D convolutional networks, each encoding a specific level of contextual information. The final classification results are obtained by fusing the probability prediction outputs of these networks.

A. Construction of 3-D CNNs

In general, each 3-D convolutional network consists of 3-D convolutional, 3-D max-pooling, and fully-connected layers to hierarchically extract representations (also called *features*) and a softmax layer for the final regression to probabilities. Each layer contains a number of channels, and every channel encodes a different pattern. For 3-D CNN, each channel in the convolutional/max-pooling layer is actually a 3-D feature volume, rather than a 2-D feature map in conventional CNNs. The 3-D feature volume includes a group of neurons structured in a cubic manner.

1) 3-D Convolutional Layer: To construct a 3-D convolutional layer, we first establish a set of small 3-D feature extractors (or usually called *kernels*), which sweep over their input (i.e., the output of the previous layer) to extract a stack of higher-level representations. In order to generate a new feature volume, we use different 3-D kernels to convolve different input feature volumes (each feature volume corresponding to a unique 3-D kernel). Then, we add a bias term, and employ a nonlinear activation function. We formulate the 3-D convolutional layer in an element-wise manner as follows:

$$\boldsymbol{h}_i^l(x,y,z) = \boldsymbol{\sigma}\left(\boldsymbol{b}_i^l + \sum_k \sum_{u,v,w} \boldsymbol{h}_k^{l-1}(x-u, y-v, z-w)\boldsymbol{W}_{\text{ki}}^l(u,v,w)\right) \quad (1)$$

where $\boldsymbol{h}_i^l$ and $\boldsymbol{h}_k^{l-1}$ represent the ith 3-D feature volume in the lth layer and the k-th 3-D feature volume in the previous layer, respectively; $\boldsymbol{W}_{\text{ki}}^l \in \mathbb{R}^3$ is the 3-D convolutional kernel connecting $\boldsymbol{h}_i^l$ and $\boldsymbol{h}_k^{l-1}$; $\boldsymbol{h}_i^l(x,y,z)$, $\boldsymbol{h}_k^{l-1}(x-u, y-v, z-w)$, and $\boldsymbol{W}_{\text{ki}}^l(u,v,w)$ represent their element-wise values with (x,y,z) being the coordinates of $\boldsymbol{h}_i^l$ and (u,v,w) being the coordinates of the 3-D kernel $\boldsymbol{W}_{\text{ki}}^l$; the $\boldsymbol{b}_i^l$ is a bias term; and $\boldsymbol{\sigma}(\cdot)$ is the nonlinear activation function, i.e., the rectified linear units (ReLU) ($\boldsymbol{\sigma}(\boldsymbol{a}) = \max(\boldsymbol{0}, \boldsymbol{a})$) [28]. Note that activations from different 3-D kernels should be summed up before adding the bias term; the summation over k in (1) means the summation of activations from different 3-D kernels.

2) 3-D Max-Pooling Layer: In-between successive 3-D convolutional layers, we periodically insert 3-D max-pooling layers to subsample the 3-D feature volumes, and therefore acquiring invariance to local translations in 3-D space. Assuming that the lth layer is a convolutional layer and the $(l+1)$th layer is the 3-D max-pooling layer following it, the max-pooling layer accepts a 4-D tensor $\boldsymbol{T} = [\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, ..., \boldsymbol{h}_K^l] \in \mathbb{R}^{X\times Y\times Z\times K}$. For the max-pooling operation, it selects the maximum activation within a cubic neighborhood and generates an abstracted output $\boldsymbol{T}' \in \mathbb{R}^{X'\times Y'\times Z'\times K}$, where (X,Y,Z) and (X',Y',Z') are the sizes of feature volumes before and after the max-pooling operation, respectively; K denotes the number of feature volumes which remains unchanged during the pooling operation. Given the pooling kernel size of M and stride of S, the size of feature volumes is reduced as $X' = (X-M)/S + 1$ (same for Y' and Z').

3) Fully Connected Layer: In fully connected layers, the neurons have much denser connections than those of convolutional layers. Specifically, each neuron is connected with all neurons in adjacent layers. This is different from the local connection style equipped in the convolutional layers. These dense connections can benefit stronger representation capability of the extracted representations. To implement the fully connected layer, we first flatten the feature volumes into a neuron vector, next perform a vector-matrix multiplication, then add a bias term to it, and finally apply a nonlinear function to generate the activations as follows:

$$\boldsymbol{h}^f = \boldsymbol{\sigma}(\boldsymbol{b}^f + \boldsymbol{W}^f \boldsymbol{h}^{f-1}) \tag{2}$$

where the $\boldsymbol{h}^{f-1}$ is the input feature vector obtained by flattening the 3-D feature volumes of the $(f-1)$th layer; $\boldsymbol{h}^f$ is the output feature vector of the fth layer, which is a fully connected one; $\boldsymbol{W}^f$ is the weight matrix; $\boldsymbol{b}^f$ is the bias term; and $\boldsymbol{\sigma}(\cdot)$ is the ReLU [28].

4) Softmax Layer: The output layer of the 3-D CNN is the softmax layer. Denoting the neuron vector in the last layer by $\boldsymbol{h}^L$, and C is the number of target classes, we calculate the prediction probability for each class c via the softmax regression $p_c(\boldsymbol{h}^L) = \exp(\boldsymbol{h}_c^L)/\sum_{c=0}^{C-1}\exp(\boldsymbol{h}_c^L)$, where $\boldsymbol{h}_c^L$ is the cth element of the neuron vector. The output activations of the softmax layer are all positive values within the interval $(0, 1)$ and summed up to one. As a result, they can be interpreted as the estimated probability distribution predicted by the network.

5) Cost Function: Given a set $\{(\mathcal{I}^{(1)}, \mathcal{Y}^{(1)}), ..., (\mathcal{I}^{(N)}, \mathcal{Y}^{(N)})\}$ of N paired 3-D training samples, where $\mathcal{I}^{(j)}$ is an input cubic patch and $\mathcal{Y}^{(j)}$ is the corresponding ground-truth label, $\hat{\mathcal{Y}}^{(j)}$ is the predicted label; representing all the trainable parameters in 3-D CNNs by θ, we construct the following cost function:

$$\ell(\theta) = -\frac{1}{N}\sum_{j=1}^{N}\sum_{c=0}^{C-1}\mathbb{1}\{\mathcal{Y}^{(j)}=c\}\log\mathbb{P}(\hat{\mathcal{Y}}^{(j)}=c\,|\,\mathcal{I}^{(j)};\theta) \tag{3}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function; $\mathbb{P}(\hat{\mathcal{Y}}^{(j)}=c\,|\,\mathcal{I}^{(j)};\theta)$ is the estimated probability of sample $\mathcal{I}^{(j)}$ belonging to class c, which is exactly the output value $p_c(\boldsymbol{h}^L)$ from the softmax regression layer. The parameters in 3-D CNNs are optimized by minimizing the loss $\ell(\theta)$.

B. Importance of Receptive Field

The pulmonary nodules have large variations regarding the volume sizes (with diameter ranging from 3 to 30 mm), shapes, and many other characteristics, such as subtlety, solidity, internal structure, spiculation, sphericity, etc. [11]. In addition, the nodules often come with complicated contextual environments and hard mimics. To deal with these challenges, a batch of previous works employed features meticulously designed for a specific class of nodules for discrimination and detection [9], [29], [30]. Although these methods have achieved encouraging results in detecting specific lung nodules from CT scans, the flexibility and extensibility of these methods were quite limited, as the features tailored for a kind of nodules are often not suitable for other types of nodules with different characteristics and contextual environments. In addition, the discrimination capability of these hand-crafted features are usually insufficient for differentiating lung nodules from their hard mimics in complex environments. Leveraging the strong discrimination capability of 3-D CNNs, we propose to detect various types of nodules under a unified framework, which is able to learn the parameters in a data-driven way, and produce accurate detection results based on the learned high-level representations.

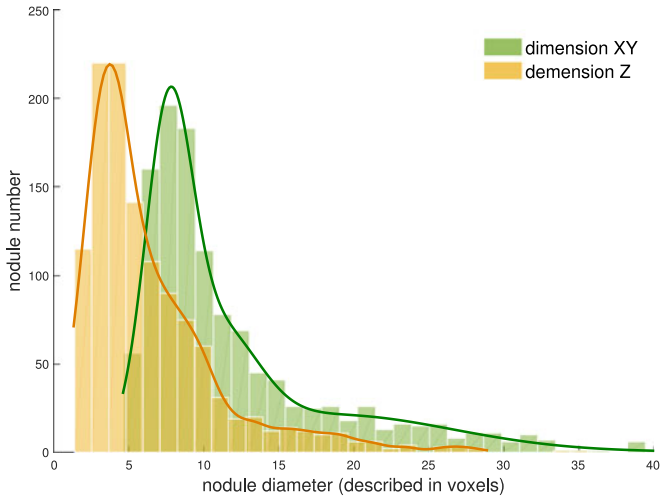

Fig. 3. Distribution analysis of the sizes of pulmonary nodules for determining receptive fields, with diameters measured in voxels across different dimensions. (Note: *X* and *Y* dimensions have the same resolution.)

In 3-D CNNs for pulmonary nodule detection, cubic samples centering on the interested candidate positions are input to the networks to train their discrimination capacity. The size of the cubic samples, i.e., the surrounding range of a target position, is called the *receptive field* of a network. The size of receptive field plays a crucial role for the recognition performance of a network. In other words, the amount of surrounding contextual information considered by the network will implicitly yet greatly influence the generated prediction probability distribution and, hence, the accuracy of the detection results. In principle, if the size of receptive field is too small, only limited contextual information will be exploited to train the networks and its discrimination capability should be deficient to handle large variations of detection targets. On the other hand, if the receptive field is too large, more redundant messages or even noises would be involved in the training, which would degrade the performance of the networks, especially when the number of training samples is quite limited. In this regard, it is difficult, if not impossible, to figure out a single optimal receptive field for a detection target with large variations. We propose to design a set of multilevel contextual 3-D CNNs to meet this challenge and improve the detection performance by fusing the results obtained from the networks learned from different levels of contextual information.

C. Multilevel Contextual Networks and Model Fusion

We determine the sizes of receptive fields employed in our framework by analyzing the size distribution of the pulmonary nodules. Based on the statistical analysis in Fig. 3, we carefully

TABLE I
ARCHITECTURES OF THE MULTILEVEL CONTEXTUAL 3-D CNNS

Archi-1			Archi-2			Archi-3		
Layer	Kernel	Channel	Layer	Kernel	Channel	Layer	Kernel	Channel
Input	–	1	Input	–	1	Input	–	1
C1	$5 \times 5 \times 3$	64	C1	$5 \times 5 \times 3$	64	C1	$5 \times 5 \times 3$	64
M1	$1 \times 1 \times 1$	64	M1	$2 \times 2 \times 1$	64	M1	$2 \times 2 \times 2$	64
C2	$5 \times 5 \times 3$	64	C2	$5 \times 5 \times 3$	64	C2	$5 \times 5 \times 3$	64
C3	$5 \times 5 \times 1$	64	C3	$5 \times 5 \times 3$	64	C3	$5 \times 5 \times 3$	64
FC1	–	150	FC1	–	250	FC1	–	250
FC2	–	2	FC2	–	2	FC2	–	2
Softmax	–	2	Softmax	–	2	Softmax	–	2

C: convolution, M: max-pooling, FC: fully connected.

design three networks that incorporate different levels of contextual information surrounding the pulmonary nodules. First, observing that the diameter density peak of small nodules lies in around nine voxels in dimension *X* and *Y*, and four voxels in dimension *Z*, we set the first network, namely *Archi-1*, with a receptive field of $20 \times 20 \times 6$ (voxels). This receptive field is able to encompass small-sized pulmonary nodules with proper amount of context, and it covers 58% of all the nodules in the dataset. Next, we design the model *Archi-2* with a larger receptive field as $30 \times 30 \times 10$. This size covers the majority (85%) of the annotated nodules, and thus it can perform well on the normal situations that most frequently happen among patients. This window size aims to provide rich context for small nodules and suitable amount of contextual information for the middle-sized lesions, while for some large nodules, it can usually include main parts of them with some marginal regions excluded. Finally, we construct the model *Archi-3* with a coverall receptive field of $40 \times 40 \times 26$. According to our statistical analysis, this model bounds over 99% of the nodules except for several outliers. Under this receptive field, rich contextual information could be provided for middle-sized lesions, taking the risk of bringing in noisy surrounding signals to some small-sized cases. Nevertheless, this architecture can better handle those nodules with extremely large sizes than the other two models. The detailed constructions of the three networks are shown in Table I. In addition, Fig. 4 presents the appearance of pulmonary nodules under different window sizes. It is observed that the amount of included contextual information for nodules can be diverse in subject to the various lesion sizes. For the small lung nodule in Fig. 4(a), the *Archi-1* is the most suitable, given that the large patches may include noisy backgrounds [see the last row of (a)]. For the nodule with diameter of 12 mm in Fig. 4(d), the receptive field of *Archi-2* is the best with proper amount of clear context included. For the case of Fig. 4(f), the large patch is more advisable for the extremely large lung nodule, especially considering that the small patch cannot even cover the whole lesion region.

After designing the networks with different receptive fields, given a testing nodule candidate $\mathcal{I}_j$, each model will assign a prediction probability for it. To aggregate the considered multilevel contextual information explored by different models for the final classification, we fuse the softmax regression outputs from all networks. Denoting the regressed probability of $\mathcal{I}_j$ belonging to the cth class from model *Archi-1* by $\mathbb{P}_1(\hat{\mathcal{Y}}_j = c|\mathcal{I}_j;\theta_1)$

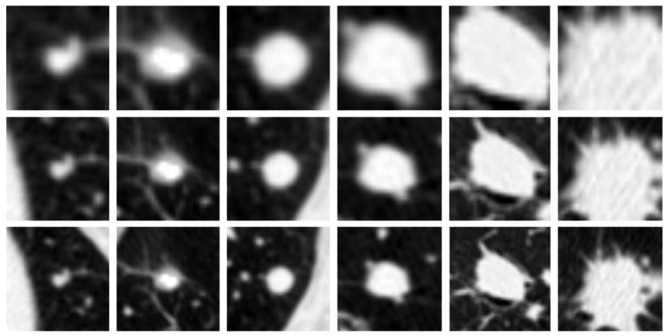

Fig. 4. Illustration of multilevel contextual information surrounding nodules. The patch sizes are $20\times 20 \times 6$, $30 \times 30 \times 10$, and $40 \times 40 \times 26$ for the first, second, and third row, respectively. We show the transverse plane only, and all patches are scaled to the same image resolution for clear visualization. The examples (a) and (b) are small nodules with diameter lower than 7 mm, (c–e) are middle-sized nodules with diameter between 9and 16 mm, (f) is a large nodule with a diameter of over 24 mm.

(analogous for *Archi-2* and *Archi-3*), the fused posterior probability $\mathbb{P}_{\text{fusion}}$ is estimated by weighted linear combination as follows:

$$\mathbb{P}_{\text{fusion}}(\hat{\mathcal{Y}}_j = c\,|\,\mathcal{I}_j) = \sum_{\varphi\in\{1,2,3\}} \gamma_\varphi \mathbb{P}_\varphi(\hat{\mathcal{Y}}_j = c\,|\,\mathcal{I}_j;\theta_\varphi) \quad (4)$$

where $\mathbb{P}_{\text{fusion}}(\hat{\mathcal{Y}}_j = c\,|\,\mathcal{I}_j)$ is the fused prediction probability of $\mathcal{I}_j$ belonging to class c output by the whole framework. The constant weights γ_φ were determined using grid search on a small subset of the training data in our experiments ($\gamma_1 = 0.3$, $\gamma_2 = 0.4$, $\gamma_3 = 0.3$).

D. Training Process

The weights θ were learned with stochastic gradient descent, i.e., each iteration of the parameter update was computed based on a mini-batch of training samples. The positive and negative samples were obtained according to the candidates with labels provided by the challenge. We extracted patches centering on the candidate locations with sizes of $20 \times 20 \times 6, 30 \times 30 \times 10$, and $40 \times 40 \times 26$, corresponding to the three architectures. To deal with the severe class imbalance between the false positive candidates and the true nodules (around 490:1 in this challenge), translation and rotation augmentations were conducted for the ground truth nodule positions. Specifically, we translated the

TABLE II
RESULTS OF THE FALSE POSITIVE REDUCTION TRACK IN ISBI LUNA16 CHALLENGE

Team	Team no.	CNN type	0.125	0.25	0.5	1	2	4	8	Score (CPM)
DIAG_CONVNET(arnaud.setio) [17]	T1	2-D	0.636	0.727	0.792	0.844	0.876	0.905	0.916	0.814
iitm03(subru1603)	T2	2-D	0.394	0.491	0.570	0.660	0.732	0.795	0.851	0.642
LUNA16CAD(hirokinakano)	T3	2-D	0.113	0.165	0.265	0.465	0.596	0.695	0.785	0.440
LUNA16CAD(mattdns100689)	T4	3-D	0.640	0.698	0.750	0.804	0.847	0.874	0.897	0.787
LungNess(bim_bam)	T5	2-D	0.453	0.535	0.591	0.635	0.696	0.741	0.797	0.635
UACNN(ccanoespinosa)	T6	2-D	0.655	0.745	0.807	0.849	0.880	0.907	0.925	0.824
CUMedVis(QiDou) **(Ours)**	T7	3-D	0.677	0.737	0.815	0.848	0.879	0.907	0.922	**0.827**

centroid coordinates by one voxel along each axis and rotated 90°, 180°, and 270° within the transverse plane. In total, we obtained 0.65 million samples to train the networks. We clipped the intensities into the interval $(-1000, 400$ Houndsfield Unit$)$ and normalized them to the range of $(0, 1)$. The mean gray-scale value was subtracted to adjust the distribution of training and testing data. During the training process, The weights were randomly initialized from the Gaussian distribution $\mathcal{N}(0, 0.01^2)$ and updated with standard backpropagation [31]. The learning rate was initialized as 0.3 and decayed by 5% every 5000 iterations. We set a relatively high learning rate at the beginning of the training process because we considered that the 3-D network is trained from scratch rather than fine tuned from a pretrained model. The mini-batch size was set to 200, the momentum [32] was set to 0.9, and the dropout [33] (rate = 0.2) strategy was utilized in convolutional and fully connected layers to improve the generalization capability of the model. The networks were implemented in Python based on the deep learning library of Theano [34]. The three network architectures were independently trained and validated. It took about 6 h to train each network using a GPU of NVIDIA TITAN Z.

III. EXPERIMENTS

A. Dataset and Candidate Generation

We evaluated the proposed approach on a large-scale benchmark dataset, which was released by the LUNA16 Challenge held in conjunction with ISBI 2016. We participated in the false positive reduction track, in which participants were asked to, given a set of candidate locations, assign each candidate a probability for being the pulmonary nodule.

The challenge filtered out 888 CT scans from the publicly available LIDC dataset [11]. The volumes were with resolution in the transverse plane as 512×512, element spacing as $0.74 \times 0.74\,\text{mm}^2$, and variable slice thickness but not larger than 2.5 mm. The annotations of lung nodules were collected with a two-phase manual labeling process conducted by four experienced thoracic radiologists. During the process, each radiologist marked the identified lesions as nonnodule, nodule < 3 mm, and nodules $>= 3$ mm. Then, the challenge selected a total of 1186 nodules $>= 3$ mm accepted by three or four radiologists as the reference standard (i.e., ground truth). Annotations that were not included in the reference standard (i.e., nonnodules, nodules < 3 mm, and nodules annotated by merely one or two radiologists) were referred as irrelevant findings.

In the challenge of false positive reduction track, the organizers provided a set of prescreened candidates to participants. The candidates were figured out by three existing candidate detection algorithms [3], [9], [10], and 1120 out of 1186 ground truth nodules (sensitivity of 94.4%) were detected with $551\,065$ candidates.

B. Evaluations Metrics

The challenge evaluated detection results by measuring the detection sensitivity and average false positive rate per scan. A predicted candidate location was counted as a true positive if it was located within the radius of a true nodule center. Detections of irrelevant findings were ignored (i.e., considered as neither false positives nor true positives) in the evaluation. The challenge organizers performed the free receiver operation characteristic (FROC) analysis by setting different thresholds on the raw prediction probabilities submitted by the participating teams. The evaluation also computed the 95% confidence interval using the bootstrapping [35]. A competition performance metric (CPM) score [36], which was calculated as the average sensitivity at seven predefined false positive rates: $1/8$, $1/4$, $1/2$, 1, 2, 4, and 8 false positives per scan, was produced for each algorithm. The tenfold cross validation on the dataset was specified.

C. Results of the Challenge

There were seven teams participating in the ISBI challenge, and the challenge results are listed in Table II. For the convenience of description, we assign each team a number as listed in the Table. All the seven teams employed deep CNNs for the challenge, demonstrating the enormous influence of deep CNNs on medical image analysis community nowadays. However, despite of the 3-D nature of this detection task, five of the seven teams utilized variants of 2-D CNNs based on multiview planes (T1), orthogonal planes (T5), adjacent planes along a specific direction (T6), or separate 2-D slices (T2 and T3). Only T4 and our team (T7) employed 3-D CNNs.

It is observed that the CMP scores of T2, T3, and T5 were far behind those of other competitors. T2 and T3 constructed their 2-D models based on separate 2-D slices, where volumetric contextual information cannot be sufficiently explored. The performance of T2 was much better than that of T3 because T2 attempted to ensemble two 2-D CNNs to reduce bias of a single network. On the other hand, T5 employed orthogonal planes to integrate volumetric spatial information in the 2-D model training.

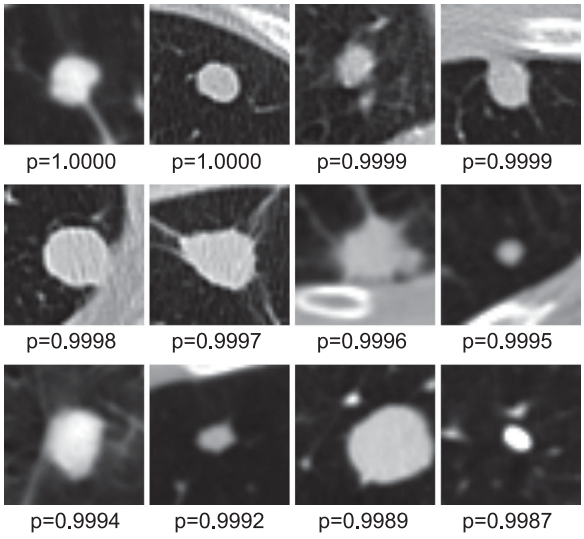

Fig. 5. Examples of pulmonary nodule detection results of our framework. Each patch is a representative transverse plane of one annotated nodule and the *p*-value below the patch is its prediction probability from our framework.

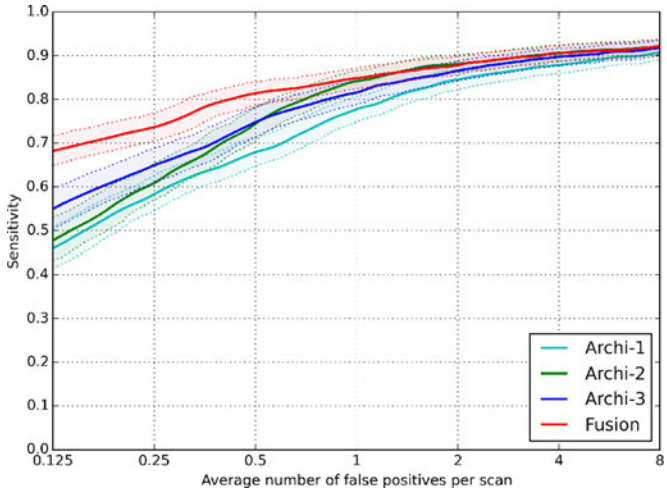

Fig. 6. FROC curves of different architectures and their fusion result. Dashed lines denote the 95% confidence interval estimated via bootstrapping [35].

T1 and T6 achieved quite good results based on 2-D CNNs. T1 trained multiple streams of 2-D CNNs with a set of patches extracted from differently oriented planes [17]. Although this scheme was still not able to fully leverage the 3-D spatial information, it was an effective variant of 2-D CNNs for volumetric image analysis, especially when concerning its computational and storage efficiency. T6 resampled the CT scans into a homogenous voxel size and combined three patches, including the plane centering on the candidate point as well as two patches located at 3 mm above and 3 mm below the candidate point. The success of this scheme was also attributed to the inclusion of more spatial contextual information in the trained model.

Like our approach, T4 also employed 3-D convolutional networks. They took 3-D patches of size $42 \times 42 \times 42$ with different resolutions as input, and constructed networks consisting of six convolutional layers with equal kernel size ($3 \times 3 \times 3$) and three max-pooling layers to down-sample feature volumes. Although 3-D CNNs are considered being able to encode more volumetric information for discriminating the true lung nodules, the performance of T4 was slightly lower than that of T1 and T6. The reasons for this could be that 1) they only used twofold cross validation, much less than the specification of tenfold cross validation, and hence the insufficient training data degraded the power of 3-D CNNs, and 2) they employed the same input size and kernel size in all three dimensions; however, as the third dimension (Z dimension) of the CT scans had a relatively lower resolution, if the input size and kernel size were the same in all dimensions, the actual receptive field in the world space could be incommensurate. Nevertheless, it outperformed the other three 2-D CNNs based methods (T2, T3, and T5) by a large margin, demonstrating the effectiveness of the 3-D variant of CNNs in volumetric detection tasks.

Our method achieved the highest CPM score in the challenge. Different from T4, we designed smaller input size and kernel sizes in the third dimension in order to proportionate the receptive field across all directions. More importantly, we carefully analyzed the diameter distribution of the nodules, and designed a framework that fused multilevel spatial contextual information to effectively resolve the conflicts between the large variations of pulmonary modules and the limited training dataset, which is one of the main challenges of employing deep CNNs in medical image analysis applications. Fig. 5 presents examples of successfully detected pulmonary nodules. It is observed that our framework accurately identifies nodules of various sizes, shapes, and locations with very high confidence.

TABLE III
SENSITIVITIES OF MODELS UNDER DIFFERENT FALSE POSITIVE RATES

FP/Scan	Archi-1	Archi-2	Archi-3	Fusion
8	0.909	0.920	0.918	0.922
4	0.879	0.905	0.897	0.907
2	0.846	0.881	0.866	0.879
1	0.777	0.842	0.814	0.848
0.5	0.681	0.747	0.750	0.815
0.25	0.580	0.604	0.651	0.737
0.125	0.459	0.473	0.546	0.677

D. Quantitative Analysis of Our Method

We further quantitatively analyzed the performance of the three network architectures (i.e., *Archi-1*, *Archi-2*, *Archi-3*) in our framework, which incorporated different levels of volumetric contextual information surrounding the pulmonary nodules. The FROC curves of each network as well as the fusion model are presented in Fig. 6. It is observed that, for all of the three individual networks, the detection sensitivities can reach beyond 90% under the false positive rate of 8 per scan, demonstrating that the 3-D CNNs are able to effectively extract discriminative representations from volumetric CT scans for pulmonary nodule detection.

Table III lists the detection sensitivities of different network architectures at different false positive rates specified by the challenge. All the three architectures can achieve a sensitivity

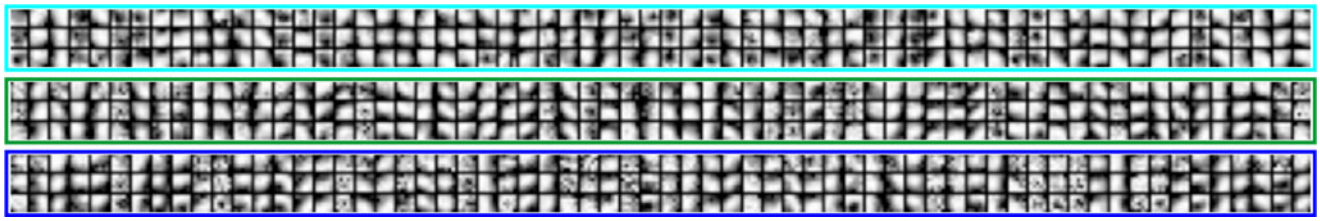

Fig. 7. Visualization of the learned 3-D kernels in the first layer of the networks incorporating different levels of contextual information. Each $5\times5\times3$ kernel is embedded as three 5×5 maps presented in a column. The rectangles with color cyan, green, and blue correspond to *Archi-1*, *Archi-2*, and *Archi-3*, respectively.

of over 87% with 4 false positives per scan. *Archi-2* even reached detection sensitivity of 84% at 1 false positive rate. In order to increase the difficulty of the challenge, several extremely low false positive rates (0.5, 0.25, 0.125 false positives per scan) were included in the challenge evaluation scheme, which is of significance as it determines if a system can identify an acceptable percentage of modules with very few false positives, and hence increase the automation level of current computer-assisted diagnosis systems. In such cases, our multilevel contextual fusion model demonstrated a strong capability in reducing false positives while maintaining a satisfactory sensitivity. For example, when confining 0.125 false positives per scan, *Archi-1*, *Archi-2*, and *Archi-3* obtained sensitivity of merely 45.9%, 47.3%, and 54.6%, respectively. Meanwhile, our fusion model achieved a sensitivity of 67.7%, which exceeded that of the *Archi-1*, *Archi-2*, and *Archi-3* by 21.8%, 20.4%, and 13.1%, respectively. It is worthwhile to note that, in Table II, although our method achieved similar sensitivities to T1 and T6 at the 1 or above false positives per scan, we achieved much better performance than these two teams under the relatively low false positive rates (for example, 0.677 versus 0.655 of T6 and 0.636 of T1 at 0.125 false positives per scan). These experiments demonstrated that the networks incorporating different levels of contextual information can be complementary with each other and the combination of them brings a boost in detection performance.

IV. Discussion

We present a novel 3-D CNNs based framework to effectively reduce the false positive candidates in pulmonary nodule detection from volumetric CT scans. The success of the proposed framework mainly lies in two aspects. First, compared with the 2-D CNNs, the 3-D CNNs, equipped with 3-D convolutions and max-poolings, are naturally suitable for volumetric medical image processing. The 3-D networks are more proficient in encoding 3-D spatial information, and therefore produce representations with higher discrimination capability. Second, taking advantages of the multilevel contextual networks, our method achieves a more promising detection accuracy for clinical application of automated pulmonary nodule detection system. The contextual information is vital in the lung nodule detection task, given their considerable variations in sizes, shapes, and locations. This assumption of relating detection performance to the amount of contextual information is also true for 2-D CNNs based methods. For example, the method of T1 [17] employed multiview planes to input more spatial contextual information to the networks and produced highlighted results among those 2-D CNNs based methods.

Instead of developing a whole pulmonary nodule detection system, which usually integrates a candidate detector and a false positive reducer, this paper emphasizes a special focus on the false positive reduction component. This means that the proposed approach is independent of the candidate screening methods, and therefore can be combined with any candidate detector. It is true that the final detection accuracy will also depend on the performance of the candidate screening methods. If the provided candidates come with a higher sensitivity, it is promising to achieve better results with our framework.

Note that all the participating teams employed deep convolutional networks in this challenge. We can see that the CNNs, being a dominant trend in natural image processing, pervade quickly in the medical image analysis community. Even though CNNs have been increasingly employed on medical imaging applications, most works to date have been built on top of 2-D CNNs [37]. It was also the same case in this challenge, where five out of seven participants utilized 2-D CNN variants. Successful training of 3-D CNNs is not easy, given its larger parameter scales compared with 2-D CNNs. The lack of sufficient training samples, due to expensive expert annotation and privacy issues, is one of the main obstacles hindering the applications of 3-D CNNs in medical image analysis. With the joint efforts of the whole community, recent challenges, including the LUNA16, have been providing large-scale benchmark datasets. This provides those data-driven methods an opportunity to present outstanding performance in medical applications. For this specific task of lung nodule detection, we totally extracted 0.65 million samples to train the 3-D CNNs. Fig. 7 visualizes the 64 learned 3-D convolutional kernels in the first layer of the three network architectures. It is observed that all the networks were effectively trained with filters presenting similar patterns of various orientations, as the early layer is normally responsible for common low-level features, such as edges, corners, and intensity gradients.

Some examples of pulmonary nodule detection results with relatively low confidence are shown in Fig. 8. The left group presents true nodules with either irregular shapes or ambiguous boundaries. Nevertheless, our framework was able to retrieve these challenging cases with a probability of higher than 0.75. The right group shows true nodules coming with extremely

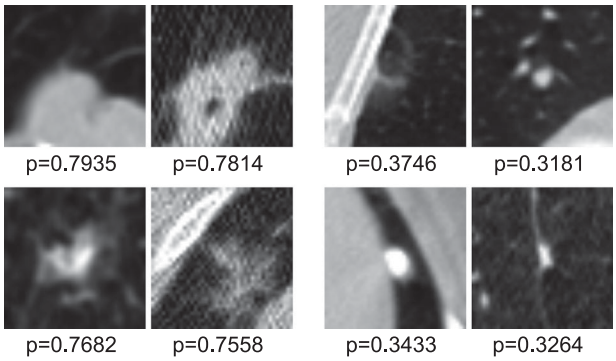

Fig. 8. Examples of detection results with relatively low confidence. Each patch is a representative transverse plane of one annotated nodule and the *p*-value below the patch is its degree of suspicion figured out by our framework.

small sizes and complex surrounding particularities. These cases were underrepresented in the dataset and therefore classified by our framework with a limited belief (between 0.3 and 0.4). Special augmentations targeting these outlier cases might bring a potential to increase the recognition performance.

V. Conclusion

In this paper, we present a 3-D CNNs based framework for computer-aided detection of pulmonary nodules from volumetric CT scans. We demonstrate the importance and effectiveness of leveraging 3-D multilevel contextual information when exploiting convolutional networks to detect lesions with large variations and hard mimics from volumetric medical data. Experimental results in the LUNA16 challenge demonstrated impressive efficacy of the proposed approach for the false positive reduction task. In principle, the proposed framework is general and can be easily extended to other object detection tasks in 3-D medical images. Further investigations include evaluating it on more clinical data and promoting it in clinical practice with the aid of radiologists and surgeons.

Acknowledgment

The authors would like to thank the LUNA16 challenge organizers for providing the dataset and evaluating our results.

References

[1] I. Sluimer *et al.*, "Computer analysis of computed tomography scans of the lung: A survey," *IEEE Trans. Med. Imag.*, vol. 25, no. 4, pp. 385–405, Apr. 2006.
[2] 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.
[3] C. Jacobs *et al.*, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Med. Image Anal.*, vol. 18, no. 2, pp. 374–384, 2014.
[4] M. Tan *et al.*, "A novel computer-aided lung nodule detection system for ct images," *Med. Phys.*, vol. 38, no. 10, pp. 5630–5645, 2011.
[5] American Cancer Society. Cancer Facts and Figures 2015. [Online]. Available: http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf
[6] H. MacMahon *et al.*, "Guidelines for management of small pulmonary nodules detected on ct scans: A statement from the fleischner society 1," *Radiology*, vol. 237, no. 2, pp. 395–400, 2005.
[7] C. I. Henschke *et al.*, "Early lung cancer action project: Overall design and findings from baseline screening," *Lancet*, vol. 354, no. 9173, pp. 99–105, 1999.
[8] T. Messay *et al.*, "A new computationally efficient cad system for pulmonary nodule detection in ct imagery," *Med. Image Anal.*, vol. 14, no. 3, pp. 390–406, 2010.
[9] A. A. Setio *et al.*, "Automatic detection of large pulmonary solid nodules in thoracic ct images," *Med. Phys.*, vol. 42, no. 10, pp. 5642–5653, 2015.
[10] K. Murphy *et al.*, "A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification," *Med. Image Anal.*, vol. 13, no. 5, pp. 757–770, 2009.
[11] S. G. Armato III *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.
[12] M. Firmino *et al.*, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects," *Biomed. Eng. Online*, vol. 13, pp. 1–16, 2014.
[13] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[14] K. He *et al.*, "Deep residual learning for image recognition," 2015. [Online]. Available: arXiv:1512.03385.
[15] X. Qi *et al.*, "Semantic segmentation with object clique potential," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2587–2595.
[16] S. Ji *et al.*, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[17] A. A. A. Setio *et al.*, "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, Mar. 2016.
[18] D. C. Cireşan *et al.*, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Berlin, Germany: Springer-Verlag, 2013, pp. 411–418.
[19] H. Chen *et al.*, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Berlin, Germany: Springer-Verlag, 2015, pp. 507–514.
[20] H. Chen *et al.*, "Dcan: Deep contour-aware networks for accurate gland segmentation," 2016. [Online]. Available: arXiv:1604.02677.
[21] K. Kamnitsas *et al.*, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," 2016. [Online]. Available: arXiv:1603.05959.
[22] Q. Dou *et al.*, "3d deeply supervised network for automatic liver segmentation from ct volumes," 2016. [Online]. Available: arXiv:1607.00582.
[23] Ö. Çiçek *et al.*, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016. [Online]. Available: arXiv:1606.06650.
[24] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.
[25] H. Chen *et al.*, "Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation," 2016. [Online]. Available: arXiv:1608.05895.
[26] H. Chen *et al.*, "Automatic detection of cerebral microbleeds via deep learning based 3d feature representation," in *Proc. 12th Int. Symp. Biomed. Imag.*, 2015, pp. 764–767.
[27] H. R. Roth *et al.*, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*. Berlin, Germany: Springer-Verlag, 2014, pp. 520–527.
[28] X. Glorot *et al.*, "Deep sparse rectifier neural networks," in *Proc. Fourteenth Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[29] F. Ciompi *et al.*, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, 2015.
[30] A. A. Enquobahrie *et al.*, "Automated detection of small pulmonary nodules in whole lung ct scans," *Acad. Radiol.*, vol. 14, no. 5, pp. 579–593, 2007.
[31] G. E. Hinton *et al.*, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[32] I. Sutskever *et al.*, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[33] G. E. Hinton *et al.*, "Improving neural networks by preventing co-adaptation of feature detectors," 2012. [Online]. Available: arXiv:1207.0580.
[34] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688
[35] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[36] M. Niemeijer *et al.*, "On combining computer-aided detection systems," *IEEE Trans. Med. Imag.*, vol. 30, no. 2, pp. 215–223, Feb. 2011.
[37] H. Greenspan *et al.*, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

Qi Dou (S'14) received the B.E. degree in biomedical engineering from Beihang University, Beijing, China, in 2014. She is currently working toward the Ph.D. degree in computer science from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

Her research interests include medical image computing, deep learning, computer-aided detection, so on.

Hao Chen (S'14) received the B.S. degree in information engineering from Beihang University, Beijing, China, in 2013. He is currently working toward the Ph.D. degree in computer science from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

His research interests include medical image analysis, deep learning, object detection, segmentation, so on.

Lequan Yu (S'16) received the B.S. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2015. He is currently working toward the Ph.D. degree in computer science and engineering at The Chinese University of Hong Kong, Shatin, Hong Kong.

His research interests include deep learning, medical image computing, and computer vision.

Jing Qin (M'16) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2009.

He is an Assistant Professor in School of Nursing, The Hong Kong Polytechnic University. He is also a Key Member in the Centre for Smart Health, SN, PolyU, HK. He has participated in more than 10 research projects and published more than 90 papers in major journals and conferences. His research interests include virtual/augmented reality for healthcare and medicine training, medical image processing, deep learning, visualization and human–computer interaction, and health informatics.

Pheng Ann Heng (M'92–SM'06) received the Ph.D. degree in computer science from Indiana University, Indianapolis, IN, USA.

He is currently a Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, where he is also the Director of the Virtual Reality, Visualization, and Imaging Research Centre. He is also the Director of the Research Center for Human–Computer Interaction, Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include virtual reality applications in medicine, visualization, medical imaging, human–computer interfaces, rendering and modeling, interactive graphics, and animation.