IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 30, 2019, accepted June 13, 2019, date of publication June 26, 2019, date of current version July 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925210

A Two-Stage Convolutional Neural Network for Pulmonary Embolism Detection From CTPA Images

XIN YANG[1], (Member, IEEE), YI LIN[1], JIANCHAO SU[1], XIANG WANG[2], XIANG LI[2], JINGEN LIN[3], AND KWANG-TING CHENG[4], (Fellow, IEEE)

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Department of Radiology, The Central Hospital of Wuhan, Wuhan 430074, China
[3]JD AI Research, Mountain View, CA 94039, USA
[4]Department of Electrical Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong

Corresponding author: Xin Yang (xinyang2014@hust.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700402, in part by the National Natural Science Foundation of China under Grant 61502188, in part by the Hubei Provincial Natural Science Foundation under Grant ZRMS2017000375, in part by the Wuhan Science and Technology Bureau under Award 2017010201010111, in part by the Fundamental Research Funds for the Central Universities under Grant 2019kfyRCPY118, and in part by the Program for HUST Acadamic Frontier Youth Team.

ABSTRACT This paper presents a two-stage convolutional neural network (CNN) for automated detection of pulmonary embolisms (PEs) on CT pulmonary angiography (CTPA) images. The first stage utilizes a novel 3D candidate proposal network that detects a set of cubes containing suspected PEs from the entire 3D CTPA volume. In the second stage, each candidate cube is transformed to be aligned to the direction of the affected vessel and the cross-sections of the vessel-aligned cubes are input to a 2D classification network for false positive elimination. We have evaluated our approach using both the test dataset from the PE challenge and our own dataset consisting of 129 CTPA data with a total of 269 embolisms. The experimental results demonstrate that our method achieves a sensitivity of 75.4% at two false positives per scan at 0 mm localization error, which is superior to the winning system in the literature (i.e., sensitivity of 60.8% at the same level of false positives and localization error). On our own dataset, our method achieves sensitivities of 76.3%, 78.9%, and 84.2% at two false positives per scan at 0, 2, and 5 mm localization error, respectively.

INDEX TERMS Convolutional neural network, pulmonary embolism detection, two-stage.

I. INTRODUCTION

Pulmonary Embolism (PE) refers to the situation when a blood clot becomes lodged in one of the arteries that go from the heart to the lungs. This blockage can obstruct the normal flow of blood and in turn causes low oxygen levels of the vital organs and becomes life-threatening [1]. Besides, it can influence the pulmonary arterial pressure and right heart pressure, yielding right-sided heart failure and ischemia. Early detection and treatment of PE could effectively decrease the morality rate [2] and computed tomography pulmonary angiography (CTPA) is the primary means for diagnosing PE in today's practice. However, manually interpreting a CTPA volume demands a radiologist to carefully trace each pulmonary artery across 300-500 slices for any suspected PEs, which is time consuming and error-prone due to lack of experience and eye fatigue.

Automated detection of PE is of high demand for improving the accuracy and efficiency of PE detection and diagnosis. Existing methods typically consist of two steps: 1) detecting a list of candidates from an entire CTPA volume based on voxel-level features, and 2) removing false positives from candidates based on region-level features and a classifier [3], [4]. For instance, Masutani *et al.* [5] extracted handcrafted features based on CT values, local contrast and the second derivatives of voxels for candidate detection and leveraged the volume, effective length and mean local contrast of grouped voxels as region-level features for false positive removal. In [6], Bouma *et al.* proposed to compute isophote curvature and circularity of the bright lumen as region-level features for false positive removal.

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

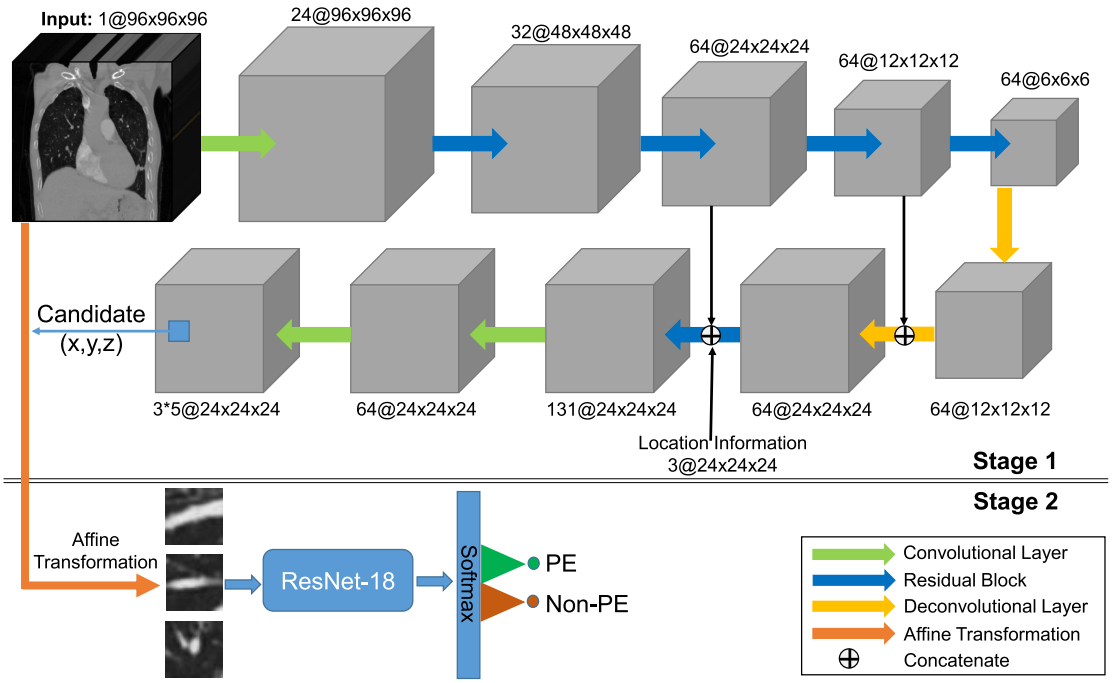

FIGURE 1. The framework of our two-stage PE detection network.

However, due to the limited representation ability of these handcrafted features, conventional methods often suffer from a high false positive rate in order to achieve an acceptable sensitivity. To address this problem, Nima *et al.* [7] investigated the feasibility of 2D CNN features for eliminating false positives in the task of PE detection. To reduce the appearance variations of PEs on 2D cross-section images, Nima *et al.* proposed to align each 3D candidate cube to the orientation of the affected vessel and then extracted the 2D cross-sections from the transformed cube for 2D CNN feature extraction. However, the authors in [7] still utilized handcrafted features [8] for detecting candidate PEs. As a result, a large number of false positives could be generated and in turn place heavy burden to the following step despite the usage of vessel-aligned feature representations. In this work, we for the first time implement all steps including PE candidate proposal, the generation of vessel-aligned image representation and FP removal into a two-stage CNN for a highly accurate PE detector.

As shown in Fig. 1, our PE detection network is a cascade of two stages: 1) a 3D candidate proposal subnet based on a 3D fully convolutional neural network (FCN), and 2) a false positive removal subnet based on vessel-aligned candidate transformation and a 2D classification network. Specifically, the first subnet extracts 3D feature hierarchies using 3D FCN which are then combined with location information to generate candidate cubes containing PEs via two 3D convolutional layers. The second subnet first transforms the original data within each candidate cube so as to align the suspected embolus with the orientation of the affected vessel segment. Then three cross-sections of the transformed candidate cube are extracted as input to a 2D classification network to output the candidate's probability of being a PE. We have evaluated our approach using the entire 20 CTPA test dataset from the PE challenge [9], achieving a sensitivity of 75.4% at 2 false positives per scan at 0mm localization error. This performance is superior to the winning system in the literature, which achieves a sensitivity of 60.8% at the same level of false positives. We have also evaluated our system on our own dataset consisting of 129 CTPA data. Our system achieves a sensitivity of 76.3%, 78.9% and 84.2% at 2 false positives per scan at 0mm, 2mm and 5mm localization error, respectively. A series of ablation study have been conducted to examine the impact of each component in the proposed system.

II. METHOD

Fig. 1 illustrates the framework of our two-stage PE detection network. The first stage is a 3D fully convolutional network that proposes candidate, and the second stage extracts vessel-aligned 3D candidate cubes and removes false positives based on 2D cross-sections of vessel-aligned cubes and a ResNet-18 classifier [10].

A. STAGE 1: CANDIDATE PROPOSAL SUBNET USING FCN

The first stage aims at a high sensitivity and a reasonable false positive rate. To fully exploit the 3D context information of a pulmonary CTPA volume, our candidate proposal subnet employs a 3D FCN to extract 3D feature hierarchies.

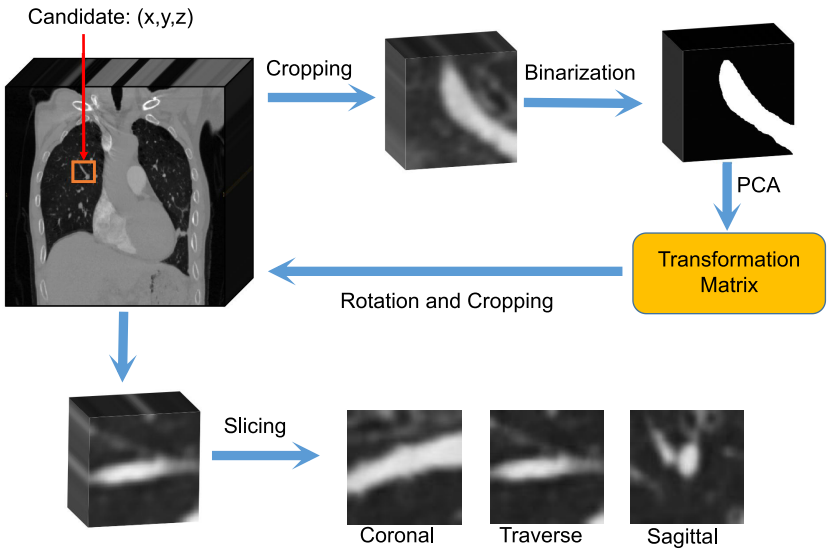

FIGURE 2. The pipeline of vessel-aligned slice generation.

As shown in the top part of Fig. 1, the 3D FCN utilizes an encoder-decoder network architecture with skip connections. The encoder starts with a 3D convolutional layer, followed by a max-pooling layer and another four residual blocks [10] to encode hierarchical feature maps. The decoder up-samples the feature maps by two deconvolutional layers, one residual block and two convolutional layers. Skip connections are utilized to connect the last two residual blocks in the encoder and the corresponding residual blocks in the decoder. In addition to visual cues, the location is also an important indicator for identifying PEs as PEs usually reside at some unique regions, i.e. bifurcations of the main pulmonary arteries or lobar branches [11]. Therefore, we also input the location information and combine it with the FCN feature maps in the decoder. Specifically, we form a 3-channel location map which has the same size as FCN feature maps at the second deconvolutional layer (i.e. $24 \times 24 \times 24$). Each voxel of the location map is a 3-dimensional vector indicating the x, y, z coordinates in the entire 3D volume. We directly concatenate [12] the 3-channel location map with the 64-channel FCN feature map, together with the 64-channel FCN feature map passed from the skip connection to form a 131-channel feature map. A residual block is then applied to the concatenation of maps for information fusion.

To detect candidate proposals from the fused 3D feature map, we incorporate anchor cubes into our candidate proposal subnet to facilitate accurate detection of variable size proposals as Faster R-CNN [13]. Specifically, the anchor cubes are pre-defined multiscale 3D windows centered at every voxel location of the feature map. For each voxel location, we specify $N = 3$ anchor cubes, each of which has a different scales (i.e. s = 10mm, 30mm and 60mm respectively). For each anchor cube of a scale s, we design five regressors to compute five values $(\Delta x_s, \Delta y_s, \Delta z_s, \Delta d_s, p_s)$ which indicate the location and size of the candidate cube in the entire pulmonary volume as well as the probability of this cube containing a PE. Similar as Faster R-CNN, we regress the offset values of the location $(\Delta x_s, \Delta y_s, \Delta z_s)$ and size (Δd_s) with respect to the anchor cube at scale s for easier and more accurate regression in training and inference. To this end, we first apply a 3D convolution layer with 64 kernels size of $1 \times 1 \times 1$ to the fused feature map, then we apply another 3D convolution layer with $5N$ kernels (the size of each kernel is $1 \times 1 \times 1$) to output a feature map size of $24 \times 24 \times 24$. Each voxel of the output feature map is a $5N$-dimensional vector indicating $(\Delta x_s, \Delta y_s, \Delta z_s, \Delta d_s, p_s)$, $s = \{1, \hat{a}\check{A}\c{e}, N\}$ and $N = 3$ in this study.

A typical size of a 3D pulmonary CTPA volume is $512 \times 512 \times 400$, which could be too large to be directly input into the network due to the constraint of GPU memory. To alleviate the memory cost, in the training phase, we divide the entire volume uniformly into overlapping cubes size of $96 \times 96 \times 96$ and input each cube into the network. During tesing, we divide the entire volume uniformly into overlapping sizes of $208 \times 208 \times 208$. Using different input sizes during training and testing is enabled by the benefits of fully convolutional network.

B. STAGE 2: FALSE POSITIVE REMOVAL SUBNET USING VESSEL-ALIGNED REPRESENTATION

The second stage aims to remove false positives as many as possible via a classifier and meanwhile maintain a high sensitivity. This is a very challenging task as the first stage could generate many false positive in order to achieve a satisfactory sensitivity, yielding a severe imbalance between the positive and negative samples. In addition, the appearance of all possible PEs could vary significantly on the three cross-sections

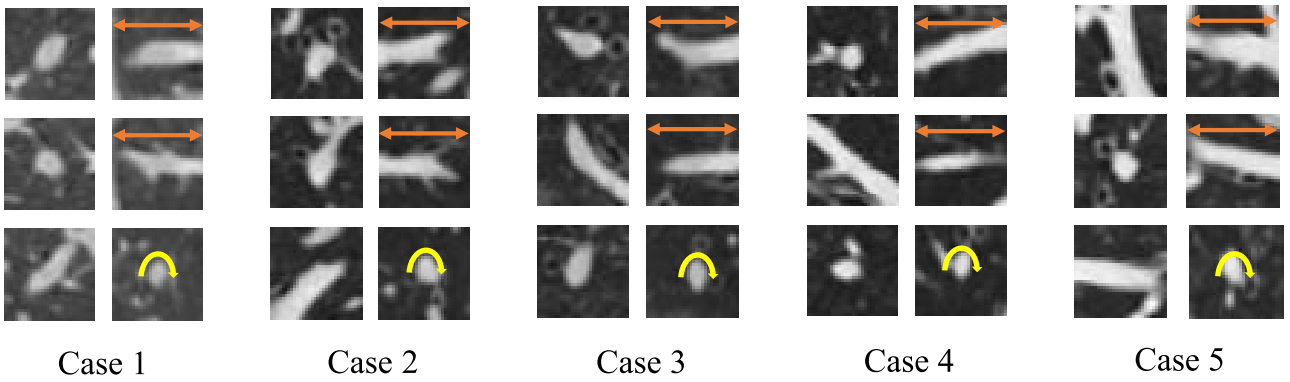

FIGURE 3. Coronal(top row), transverse (middle row) and sagittal (bottom row) slices of five exemplar PE cubes. The left column of each case denotes slices of the original cube and the right column indicates those after vessel-alignment transformation. The orange arrows denote the vessels' longitude directions and the yellow arch arrows indicate the vessels' cross-section view.

due to various orientations, sizes and shapes of PEs. Utilizing a 3D classifier [14] could alleviate the appearance variation problem to some extent while also lead to severe overfitting to limited training data due to the lack of sufficient 3D samples for training the 3D classifier. To address the above problem, we adopted the vessel-aligned 2.5D image representation proposed by Nima *et al.* [7], which aligns each candidate proposal to the orientation of the affected vessel to reduce the appearance variations of PEs in the three cross-section slices. We describe details of our false positive removal subnet based on vessel-aligned image representation in the following.

The bottom part of Fig. 2 illustrates the process of aligning a candidate cube to the orientation of the affected vessel segment. First, we crop the candidate cube from the original volume and binarize the cube via intensity thresholding. According to the radiologists' experience, typical vessel intensities are above 100 Hounsfield Units (HU) while the other tissues have intensity values below 100 HU. Considering the fact that a PE appears as a filling defect in CTPA (i.e. a dark region surrounded by the bright vessel lumen), thus the intensity values of PE are slightly lower than those of vessels, we empirically choose 70HU as our binarization threshold. Accordingly, voxels whose intensity values are greater than 70HU are labeled as vessels and others are set to zeros as non-vessels in the binarized cube. We then apply principal component analysis (PCA) [15] to the binarized cube to calculate the orientation of the vessel segment in the cube. We obtain three eigen vectors $(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$ and their corresponding eigen values $(\lambda_1, \lambda_2, \lambda_3)$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3$. According to the physical meaning of eigen vectors, $\boldsymbol{v}_1$ represents the direction in which the vessel elongates, and $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ represent two orthogonal directions in the plane vertical to $\boldsymbol{v}_1$. After that we apply a 3D rotation transformation to the original candidate cube based on the 3D affine transformation matrix A_θ defined by $(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$ as

$$\begin{bmatrix} x^s \\ y^s \\ z^s \end{bmatrix} = A_\theta \begin{bmatrix} x^t \\ y^t \\ z^t \\ 1 \end{bmatrix} = \begin{bmatrix} s_x \boldsymbol{e}_1^T \boldsymbol{v}_1 & s_y \boldsymbol{e}_1^T \boldsymbol{v}_2 & s_z \boldsymbol{e}_1^T \boldsymbol{v}_3 & t_x \\ s_x \boldsymbol{e}_2^T \boldsymbol{v}_1 & s_y \boldsymbol{e}_2^T \boldsymbol{v}_2 & s_z \boldsymbol{e}_2^T \boldsymbol{v}_3 & t_y \\ s_x \boldsymbol{e}_3^T \boldsymbol{v}_1 & s_y \boldsymbol{e}_3^T \boldsymbol{v}_2 & s_z \boldsymbol{e}_3^T \boldsymbol{v}_3 & t_z \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ z^t \\ 1 \end{bmatrix} \quad (1)$$

where (x^t, y^t, z^t) are the target coordinates of the regular grid of the vessel-aligned cube, (x^s, y^s, z^s) are the source coordinates of the original volume. All the coordinates are normalized to $[-1, 1]$ in our study, i.e. $-1 \leq x^t, y^t, z^t \leq 1$ and $-1 \leq x^s, y^s, z^s \leq 1$. t_x, t_y, t_z denote the offsets of the candidate cube with respect to the center of the entire volume, s_x, s_y, s_z denote the scaling ratio between the candidate cube and the entire volume, and $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ form a unit matrix. Regarding the derivation of A_θ, we refer readers to Appendix for details. For non-integer coordinates calculated by (1) we apply trilinear interpolation to get the intensity values from the original volume.

Once we obtain the vessel-aligned cube, we extract three cross-sections of the cube to form a 3-channel input to our 2D classification network based on ResNet-18. Fig. 3 illustrates the cross-sections (i.e. coronal, transverse and sagittal views) of five exemplar candidate cubes containing PE before and after the vessel-alignment transformation. By comparing the left and right column of each case we observe that, vessel alignment operations can effectively reduce appearance variations of PEs on cross-sections.

C. TRAINING OUR TWO-STAGE PE DETECTION NETWORK

We define the objective function of the candidate proposal subnet as:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (2)$$

where the classification loss L_{cls} is the binary cross entropy loss, and the regression loss L_{reg} is the smooth L_1 loss. The two terms are normalized by the mini-batch size N_{cls} and the number of anchor locations N_{reg}, and weighted

by λ. Note that L_{reg} only applies to the positive anchors. i denotes the i_{th} anchor in a mini-batch, p_i and p_i^* denote the predicted probability of being PE and the ground-truth label ($p_i^* = \{0, 1\}$), $\boldsymbol{t}_i$ and $\boldsymbol{t}_i^*$ denote the predicted position and ground-truth position associated with a positive anchor, which consists of 4 parameters:

$$\Delta x = (x - x_a)/d_a, \quad \Delta y = (y - y_a)/d_a,$$
$$\Delta z = (z - z_a)/d_a, \quad \Delta d = \log(d/d_a) \tag{3}$$

where (x, y, z, d) are predicted or ground-truth cube's center coordinates and its side length, and (x_a, y_a, z_a, d_a) are for anchor cube.

To collect training samples of the candidate proposal subnet, we assign a binary class label to each anchor. If an anchor overlaps with some ground-truth with intersection-over-Union (IoU) greater than 0.5, we label it as positive. If the anchor has an IoU smaller than 0.02 with all ground-truth, we label it as negative. Anchors that are neither labeled as positive nor negative are excluded for training. We use online hard sample mining [16] in training by randomly selecting M negative samples in each mini-batch and sorting them in a descending order based on their classification scores. The top k samples are selected as hard samples and contribute to the calculation of the objective function. The rest are abandoned by setting its loss to 0.

The objective function of the false positive (FP) removal subnet is cross entropy loss with softmax. We collect training data of the FP removal subnet based on the output of the first stage. If the center of a candidate cube generated by the proposal subnet does not reside on any ground-truth masks, the candidate cube is labeled as a negative sample for training. Otherwise, it is labeled as a positive sample. Such training data collection scheme leads to severe imbalance between positive and negative samples. To alleviate such problem, we perform data augmentation to positive samples by performing scaling, random translation and rotation to the original volume and then extracting the vessel-aligned candidate cubes from the transformed volumes. Specifically, we perform random scaling by N_s times within a range between 15mm to 35mm, random translation by N_t times within a range of -5mm to 5mm, and random rotation by N_r times around the axis of $\boldsymbol{v}_1$. Regarding the random translation, we also need to ensure that the center of the shifted candidate still resides on the ground-truth mask. Accordingly, we can finally augment each positive training sample by $N_p = N_s \times N_t \times N_r$ times for the FP removal subnet. Meanwhile, we randomly sample a similar number of negative samples as positives for training.

We train the two stages separately, i.e. train the 1^{st} stage till convergence and then train the 2^{nd} stage based on the output of the trained 1^{st} stage model.

III. EXPERIMENTS

A. DATA

We evaluated our method on two datasets:1) a public dataset from the PE challenge [9] which consists of 20 patients for training and another different 20 patient for testing, 2) a composition of two sets named PE129 dataset. PE129 contains 99 patients collected from our local hospital and another 30 patients from a public dataset [17]. PE129 contains a total of 269 embolisms.

B. IMPLEMENTATION DETAILS

We pre-processed every data for evaluation as follows before inputting it into our network. Firstly, we segmented the lung regions to exclude the background tissues based on the connected component labeling algorithm [18]. Secondly, we resampled the data to an isotropic resolution ($1mm \times 1mm \times 1mm$). Thirdly, we adjusted the contrast of each data by clipping its intensity values into [-1200, 600] HU and linearly transforming them into [-1, 1].

During training, the entire CT volume is uniformly divided into small cubes ($96 \times 96 \times 96$). In order to keep a sufficient number of negative samples, we selected from the small cubes to make sure 70% of our training samples contains a PE and other 30% do not contain any PE. To ease the problem of overfitting, we augment the dataset by randomly flipping, rotation and rescale the size of patches between 0.75 to 1.25.

The 3D FCN in the first stage was pre-trained on the largest publicly available dataset LUNA16 [19] for pulmonary nodule detection. We trained the first stage of our model for 100 epochs using the stochastic gradient descent (SGD) optimizer with a learning rate being 1e-3, a momentum being 0.9, and weight decay being 1e-4. The RestNet-18 in the second stage was trained on the output of the 1^{st} stage as described in Sec. II-C. To train ResNet-18, we used the SGD optimizer with a momentum 0.99 and set the initial learning rate as 1e-4. We decayed the learning rate by 10 times every 30 epochs. We trained the ResNet-18 for about 100 epochs till convergence.

For data augmentation in the second stage, we set N_s to 3 (i.e. 15mm, 25mm, and 35mm) and resized cross-sections of all candidate cubes to 32×32 for the convenience of being sent to the ResNet-18. For translation, we shift the candidate point in a random direction by $N_t = 4$ times within 5mm range. Regarding rotation, we set $N_r = 5$. Accordingly, we augmented positive training samples for the second stage by 60 times. To collect a similar number of negative samples as positives, we randomly sampled negative candidates from the output of the first stage without data augmentation. It is worth noting that the data augmentation and negative sample collection are only performed in training, not in testing.

C. EXPERIMENT RESULTS

1) PE CHALLENGE DATASET

We have compared our approach with the methods [7] using the entire 20 CTPA test dataset from the PE challenge [9]. As the ground truth labels are not available on the website, we asked a radiologist with over 10 years' reading experience to manually delineate PEs in each test scan. The manual annotations were further validated by a 2^{nd} observer. We evaluated

all the methods using the FROC curve per CTPA scan. A detection is counted as positive if it locates within a certain range (i.e. 0mm, 2mm and 5mm) to an embolus manual mask.

The network is trained on our PE129 dataset. In the test phase, the first stage of our network generates 3263 candidates in total, among which 459 are true PEs and 2804 are false positives. That is, the sensitivity and the number of false positives per patient of our candidate proposal subnet is 91.2% and 50.6 respectively. In comparison, Nima et al's method [7] achieves a similar sensitivity (i.e. 93%) and much greater FPs per patient (i.e. 65.8) in the first stage of their approach, demonstrating the superiority of our 3D CNN-based proposal subnet to the handcrafted-feature based proposal method in [11]. Further applying the FP removal subnet, we could significantly reduce the number of FPs. Figs. 4, 5 and 6 compare the performance among our method and all other methods evaluated on this challenge [9] at 0mm, 2mm, and 5mm localization error, respectively. A system that can achieve a high sensitivity while maintaining a relatively low number of false positives (i.e. 1 to 5 false positives per CTPA study) is desirable for radiologists. In our comparison,

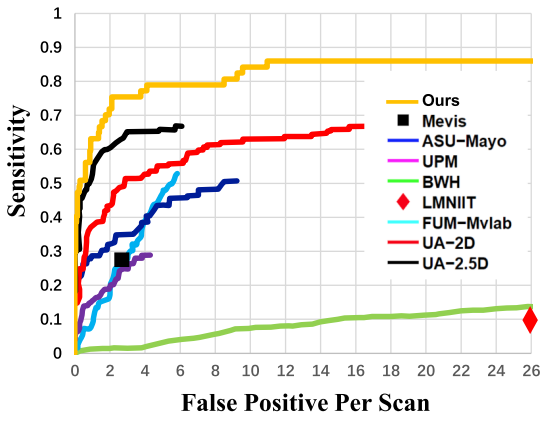

FIGURE 4. Comparison with the state-of-the-art methods on the PE challenge dataset. Localization error = 0mm.

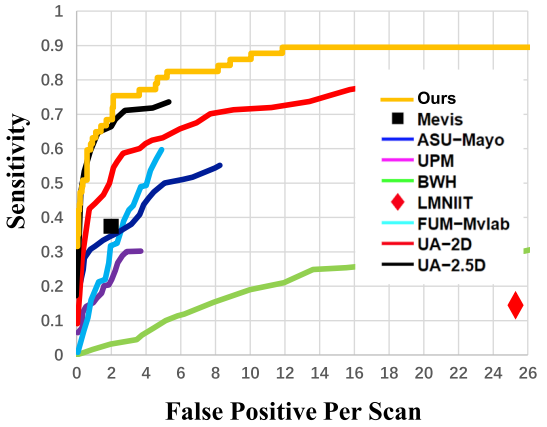

FIGURE 5. Comparison with the state-of-the-art methods on the PE challenge dataset. Localization error = 2mm.

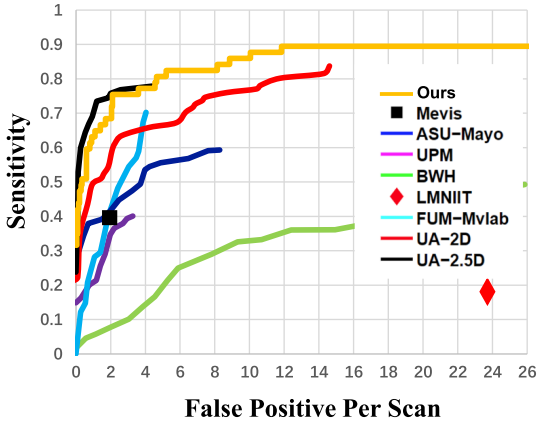

FIGURE 6. Comparison with the state-of-the-art methods on the PE challenge dataset. Localization error = 5mm.

we pay attention to the sensitivity at 2 false positive per scan. For our system, we achieve sensitivities of 75.4%, 75.4%, and 75.4% at 2 false positives per scan at 0mm, 2mm, and 5mm localization error, respectively. As shown in Fig. 4 and 5, this result outperforms the winning system UA-2.5D at which achieves 60.8% and 66.4% at 0mm and 2mm localization error, respectively. In fig. 6, our result is slightly inferior to its sensitivity of 75.8% at 5mm localization error. However, we believe the performance at lower localization error is more important. Usually, the highest sensitivity one method can achieve is also important. At 0mm localization error, our system can achieve 85.96% while producing 10.95 false positives per scan. And at 2mm and 5mm localization error, our system can achieve 89.47% while producing 11.85 false positives per scan. These results are also superior to the state-of-the art methods.

2) PE129 DATASET

We have further evaluated our system on our PE129 CTPA dataset. We randomly split our PE129 dataset to a training set with 100 scans and a test set with 29 scans. In the test phase, our proposal subnet generates 4014 candidates in total, among which 162 are true PEs and 3852 are false positives. That is, the sensitivity of our proposal subnet is 92.1% and the number of FPs per patient is 48.1. We further apply the FP removal subnet to all candidates and the final sensitivity achieved by our network is 76.3%, 78.9% and 84.2% at 2 false positives per scan at 0mm, 2mm and 5mm localization error, respectively. The top row of Fig. 7 shows ground-truth masks of PEs plotted on CT slices extracted from transverse view, denoted by red color. The middle row (blue dots) displays the centers of candidate cubes generated by our proposal subnet and the bottom row of Fig. 7 (blue dots) show the center points of the final detections after applying the FP removal subnet. If the center of a detected cube locates on the ground-truth mask, it is a true positive (TP), otherwise it is false positives. Clearly, the

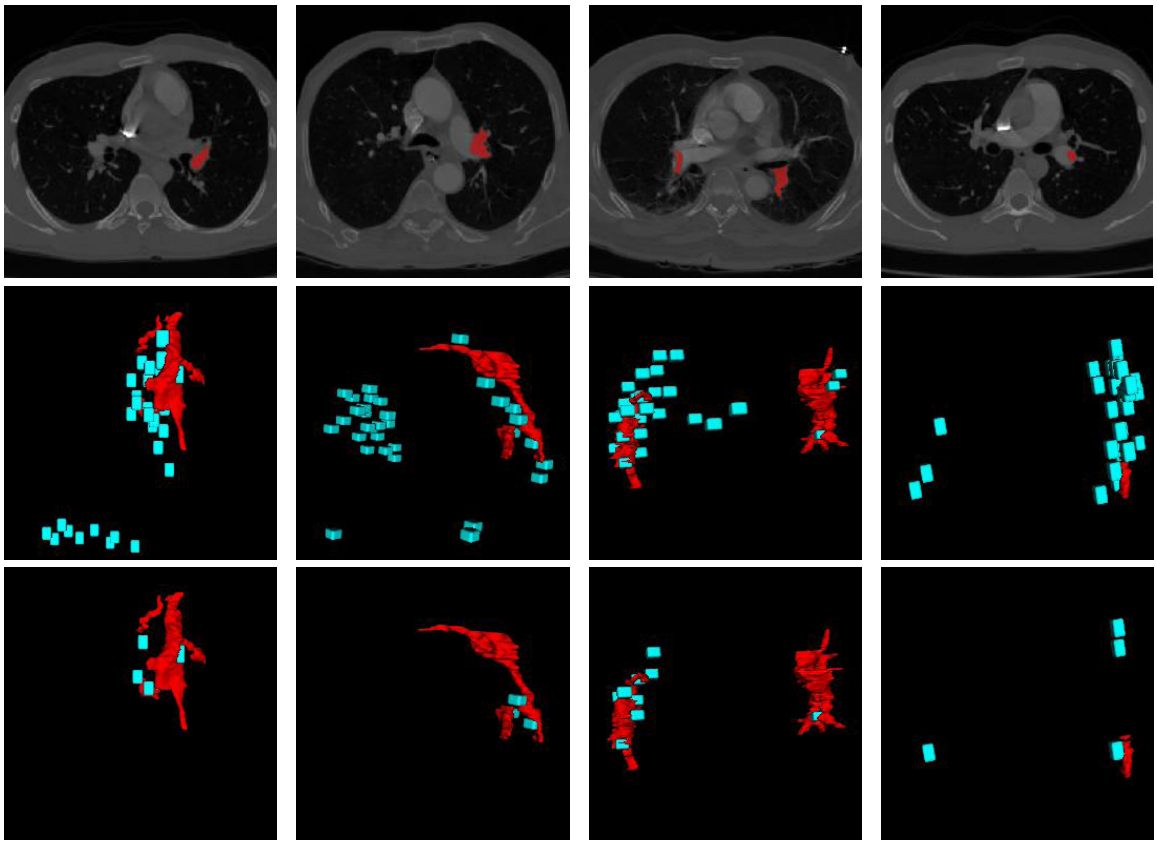

FIGURE 7. Candidates plotted on the ground-truth masks of four patients. Color red denotes the PE's ground-truth mask. All the candidates that have the top predicted probabilities on each patient are plotted as blue blocks. For each patient, the first row shows the ground truth PE masks plotted on CT slices. The second row shows the candidates that FCN generates and the thrid row shows the candidates after false positive reduction.

2^{nd} stage significantly removes FPs and meanwhile maintains sufficient TPs.

D. ABLATION STUDIES

We further examine the effectiveness of each component in our PE detection system and generate two variants of our method: 1) using only the 3D candidate proposal subnet of our approach (denoted as 1^{st} stage), and 2) using both stages but the input of the 2^{nd} stage is not the vessel-aligned 2.5D representation. Instead, we directly extracted the cross-sections of each candidate cube from the axial, coronal and sagittal views to form a 3-channel input to the FP removal subnet. We denote this variant as both stages with plain 2.5D. We keep all the network design and parameters identical for our method and the two variants in comparison evaluation. Quantitative results in terms of sensitivity at 2 false positives at 0mm are reported in Table. 1 for all the methods.

On the PE challenge test set, our proposal subnet achieves a sensitivity of 71.9% at 2FPs. Directly removing FPs using the 2nd stage with plain 2.5D representation could improve the sensitivity by 1.8%. Integrating the vessel-aligned representation could further improving the sensitivity by 3.5%. For PE129 dataset, the proposal subnet achieves 47.5% sensitivity at 2FPs, and the false positive removal subnet with the plain 2.5D representation and the vessel-aligned 2.5D representation could improve the sensitivity by 20.9% and 28.8%, respectively. Greater improvements achieved by stage 2 on our PE129 dataset than on the PE challenge might be due to many small emboli with various rotations in our dataset. Aligning them with the vessel orientation can effectively reduce the variations.

TABLE 1. Ablation Studies of Our Two-stage System. The sensitivities are at 2 false positives per scan at 0mm localization error.

Method	PE Challenge	PE129
1^{st} stage	71.9%	47.5%
Both stages with plain 2.5D	73.7%	68.4%
Our method	75.4%	76.3%

IV. CONCLUSION

This work presents a novel two-stage PE detection network. In the first stage, we establish a 3D fully convolutional

network (FCN) to efficiently propose candidates from the original CT scans. In the second stage, we rotate the 3D candidate cubes to make it align with the longitude direction of the affected vessel segment and input the cross-sections of the rotated cubes into the subsequent FP removal subnet. Extensive experiments on both public dataset and our own dataset demonstrate the superior performance of our method to the state-of-the-arts.

Future work includes investigate the performance of our system on cross-center CTPA images and data with small PEs in subsegmental pulmonary arteries.

APPENDIX
DERIVATION OF AFFINE TRANSFORMATION MATRIX

In order to get a vessel-aligned 3D cube according to the three eigen vectors, we apply a 3D affine transformation:

$$\begin{bmatrix} x^s \\ y^s \\ z^s \end{bmatrix} = A_\theta \begin{bmatrix} x^t \\ y^t \\ z^t \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ z^t \\ 1 \end{bmatrix} \tag{4}$$

where (x^t, y^t, z^t) are the target coordinates of the regular grid in the output crop, (x^s, y^s, z^s) are the source coordinates in the original CT that define the sample points, and A_θ is the affine transformation matrix. In general, A_θ is defined by three kind of transformation: translation, scaling, and rotation. We will describe them one by one.

First we consider translation. The last column of A_θ is given by:

$$\begin{bmatrix} \theta^{14} \\ \theta^{24} \\ \theta^{34} \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{5}$$

where t_x, t_y, t_z represent the offsets of the candidate center to the center of the original entire CTPA volume.

When considering only the scaling factors of the three axes, the relationship between source coordinates and target coordinates is given by:

$$\begin{bmatrix} x^s \\ y^s \\ z^s \end{bmatrix} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ z^t \end{bmatrix} \tag{6}$$

where s_x, s_y, s_z represent the scaling ratio between the candidate's length in x, y and z axes and the orginal CTPA volume's length in three axes.

When considering rotation only, the relationship between source coordinates and target coordinates is given by:

$$\begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \boldsymbol{e}_3 \end{bmatrix} \begin{bmatrix} x^s \\ y^s \\ z^s \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ z^t \end{bmatrix} \tag{7}$$

By combining (4) and (7) we have:

$$\begin{bmatrix} x^s \\ y^s \\ z^s \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1^T\boldsymbol{v}_1 & \boldsymbol{e}_1^T\boldsymbol{v}_2 & \boldsymbol{e}_1^T\boldsymbol{v}_3 \\ \boldsymbol{e}_2^T\boldsymbol{v}_1 & \boldsymbol{e}_2^T\boldsymbol{v}_2 & \boldsymbol{e}_2^T\boldsymbol{v}_3 \\ \boldsymbol{e}_3^T\boldsymbol{v}_1 & \boldsymbol{e}_3^T\boldsymbol{v}_2 & \boldsymbol{e}_3^T\boldsymbol{v}_3 \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ z^t \end{bmatrix} \tag{8}$$

where $[\boldsymbol{e}_1\ \boldsymbol{e}_2\ \boldsymbol{e}_3]$ represents a unit matirx and $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$ are exactly the eigen vectors computed before.

If we combine the results of scaling and rotation, we can compute the first three columns of A_θ as:

$$\begin{aligned} &\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{e}_1^T\boldsymbol{v}_1 & \boldsymbol{e}_1^T\boldsymbol{v}_2 & \boldsymbol{e}_1^T\boldsymbol{v}_3 \\ \boldsymbol{e}_2^T\boldsymbol{v}_1 & \boldsymbol{e}_2^T\boldsymbol{v}_2 & \boldsymbol{e}_2^T\boldsymbol{v}_3 \\ \boldsymbol{e}_3^T\boldsymbol{v}_1 & \boldsymbol{e}_3^T\boldsymbol{v}_2 & \boldsymbol{e}_3^T\boldsymbol{v}_3 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \\ &= \begin{bmatrix} s_x\boldsymbol{e}_1^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_1^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_1^T\boldsymbol{v}_3 \\ s_x\boldsymbol{e}_2^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_2^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_2^T\boldsymbol{v}_3 \\ s_x\boldsymbol{e}_3^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_3^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_3^T\boldsymbol{v}_3 \end{bmatrix} \end{aligned} \tag{9}$$

Combing (4), (5) and (9) we can derive the affine transformation matrix A_θ as:

$$A_\theta = \begin{bmatrix} s_x\boldsymbol{e}_1^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_1^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_1^T\boldsymbol{v}_3 & t_x \\ s_x\boldsymbol{e}_2^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_2^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_2^T\boldsymbol{v}_3 & t_y \\ s_x\boldsymbol{e}_3^T\boldsymbol{v}_1 & s_y\boldsymbol{e}_3^T\boldsymbol{v}_2 & s_z\boldsymbol{e}_3^T\boldsymbol{v}_3 & t_z \end{bmatrix} \tag{10}$$

After mapping a target point to a source point in the original CTPA volume, we then apply a trilinear interpolation to get the computed point's intensity. By doing so, we successfully get a vessel-aligned 3D cube.

REFERENCES

[1] S. V. Konstantinides, S. Barco, M. Lankeit, and G. Meyer, "Management of pulmonary embolism: An update," *J. Amer. College Cardiol.*, vol. 67, no. 8, pp. 976–990, 2016.
[2] S. B. Smith, J. B. Geske, J. M. Maguire, N. A. Zane, R. E. Carter, and T. I. Morgenthaler, "Early anticoagulation is associated with reduced mortality for acute pulmonary embolism," *Chest*, vol. 137, no. 6, pp. 1382–1390, 2010.
[3] H. Özkan, O. Osman, S. Şahin, and A. F. Boz, "A novel method for pulmonary embolism detection in CTA images," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 757–766, 2014.
[4] S. C. Park, B. E. Chapman, and B. Zheng, "A multistage approach to improve performance of computer-aided detection of pulmonary embolisms depicted on CT images: Preliminary investigation," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1519–1527, Jun. 2011.
[5] Y. Masutani, H. MacMahon, and K. Doi, "Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1517–1523, Dec. 2002.
[6] H. Bouma, J. J. Sonnemans, A. Vilanova, and F. A. Gerritsen, "Automatic detection of pulmonary embolism in CTA images," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1223–1230, Aug. 2009.
[7] N. Tajbakhsh, M. B. Gotway, and J. Liang, "Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Nov. 2015, pp. 62–69.
[8] J. Liang and J. Bi, "Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.* Munich, Germany: Springer, 2007, pp. 630–641.
[9] *CAD-PE Challenge*. [Online]. Available: http://www.cad-pe.org
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[11] A. F. Members, A. Torbicki, A. Perrier, S. Konstantinides, G. Agnelli, N. Galié, P. Pruszczyk, F. Bengel, A. J. B. Brady, and D. Ferreira, "Guidelines on the diagnosis and management of acute pulmonary embolism: The task force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC)," *Eur. Heart J.*, vol. 29, no. 18, pp. 2276–2315, 2008.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, May 2015, pp. 234–241.

[13] 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.

[14] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5648–5656.

[15] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[16] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[17] M. Masoudi, H.-R. Pourreza, M. Saadatmand-Tarzjan, N. Eftekhari, F. S. Zargar, and M. P. Rad, "A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism," *Nature Sci. Data*, vol. 5, 2018, Art. no. 180180.

[18] K. Suzuki, I. Horiba, and N. Sugie, "Linear-time connected-component labeling based on sequential local operations," *Comput. Vis. Image Understand.*, vol. 89, no. 1, pp. 1–23, Jan. 2003.

[19] (2016). *Lung Nodule Analysis*. [Online]. Available: https://luna16.grand-challenge.org