





VINet: A Visually Interpretable Image Diagnosis Network

Donghao Gu, Yaowei Li , Feng Jiang , *Member, IEEE*, Zhaojing Wen, Shaohui Liu , *Member, IEEE*, Wuzhen Shi , Guangming Lu, *Member, IEEE*, and Changsheng Zhou, *Member, IEEE*

Abstract—Recently, due to the black box characteristics of deep learning techniques, the deep network-based computer-aided diagnosis (CADx) systems have encountered many difficulties in practical applications. The crux of the problem is that these models should be explainable the model should give doctors rationales that can explain the diagnosis. In this paper, we propose a visually interpretable network (VINet) which can generate diagnostic visual interpretations while making accurate diagnoses. VINet is an end-to-end model consisting of an importance estimation network and a classification network. The former produces a diagnostic visual interpretation for each case, and the classifier diagnoses the case. In the classifier, by exploring the information in the diagnostic visual interpretation, the irrelevant information in the feature maps is eliminated by our proposed feature destruction process. This allows the classification network to concentrate on the important features and use them as the primary references for classification. Through a joint optimization of higher classification accuracy and eliminating as many irrelevant features as possible, a precise, fine-grained diagnostic visual interpretation, along with an accurate diagnosis, can be produced by our proposed network simultaneously. Based on a computed tomography image dataset (LUNA16) on pulmonary nodule, extensive experiments have been conducted, demonstrating that the proposed VINet can produce state-of-the-art diagnostic visual interpretations compared with all baseline methods.

Index Terms—Machine learning, neural network, image classification, medical diagnostic imaging.

I. INTRODUCTION

IN RECENT years, deep learning technologies have made tremendous progress, and many algorithms have achieved

Manuscript received March 24, 2019; revised September 15, 2019 and November 28, 2019; accepted January 17, 2020. Date of publication February 3, 2020; date of current version June 23, 2020. This work was supported by the National Key Research and Development Program of China under Grants 2018YFC0806802 and 2018YFC0832105. The code of this work is released at <https://github.com/plantabrick/VINet>. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu. (Donghao Gu and Yaowei Li contributed equally to this work.) (Corresponding author: Feng Jiang.)

Donghao Gu, Yaowei Li, and Zhaojing Wen are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: 18s103170@stu.hit.edu.cn; ywli@hit.edu.cn; 18s103172@stu.hit.edu.cn).

Feng Jiang, Shaohui Liu, and Wuzhen Shi are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: fjiang@hit.edu.cn; shliu@hit.edu.cn; wzshi@hit.edu.cn).

Guangming Lu and Changsheng Zhou are with the Department of Medical Imaging, Nanjing Jinling Hospital, Nanjing 210002, China (e-mail: cjr.luguangming@vip.163.com; 66368823@qq.com).

Color versions of one or more of the figures in this article are available online at <https://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2020.2971170

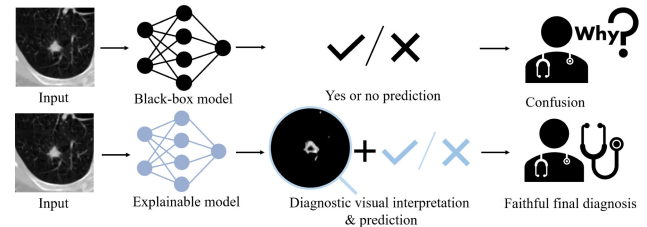


Fig. 1. Explaining model prediction. A black-box model only predicts whether a patient has lung cancer. With the provided rationales, doctors can make an accurate final diagnosis when they disagree with the models prediction.

higher accuracy than human experts in some computer vision tasks [1], [2]. The field of biomedical imaging has also been significantly influenced. With the capabilities of classification, detection, and segmentation tasks in the field of biomedical imaging, deep learning algorithms have helped doctors with manual diagnosis and decision-making [3]. Computer-aided diagnosis (CADx) systems using convolutional neural networks (CNNs) have been developed for the diagnosis of various diseases, such as lung and breast cancer, and the average accuracy of CADx systems for classifying pulmonary nodules has exceeded that of humans [4]. However, these techniques have encountered considerable difficulties when they are attempted to be applied in hospitals, and the most important problem behind this dilemma is the trust problem. Science should be show me, not trust me. As a tool, a CADx system should provide its rationales for making such decisions in order to be helpful in clinical diagnoses. For example, early diagnosis is critical for the treatment of patients with lung cancer [5]. However, the features that doctors use to diagnose lung cancer are very subtle in early small nodules, which poses a great challenge for radiologists [5]–[7]. At this point, a visually interpretable model can help doctors locate these subtle features and improve accuracy. Besides, when the diagnoses of a doctor and a CADx system are different, the system should be capable to display the basis of its diagnosis and the focus of the algorithm so that doctors can make a final diagnosis based on these diagnostic visual interpretations, like shown in Fig. 1. Furthermore, the results of an interpretable algorithm can also be used as reference information for insurance companies, law enforcement, etc., thus giving impetus for better applications of the technique.

Many studies have attempted to improve the interpretability of the decisions made by CNNs in medical imaging tasks [3], [8]–[10]. Some studies attempt to interpret the network through

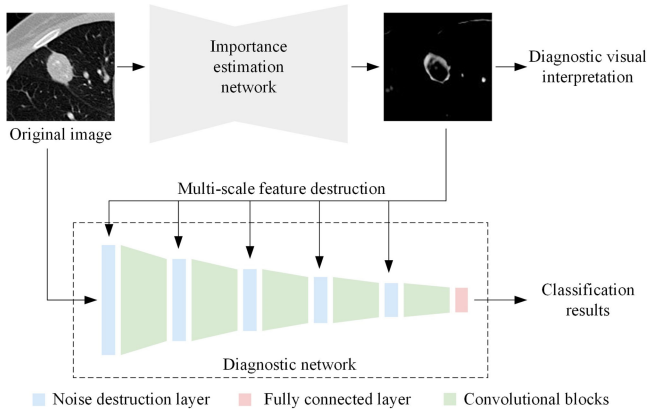


Fig. 2. An overview of VINet. As shown, the input to the importance estimation network and classification network are the same. The output of the importance estimation network is the diagnostic visual interpretation of the same size as the input, which serves as the guidance in the feature destruction layer to remove the unimportant and irrelevant parts from the feature maps of each pooling layer in the classifier. Trained with an end-to-end manner and with appropriate regularization parameters setting, VINet can locate the most important parts of the original input image and make accurate diagnoses.

semantic segmentation, and some other studies use the attention mechanism to visualize the regions of interest. However, the explanations generated by most algorithms are ambiguous and unclear, making it difficult to help doctors. In order to effectively assist doctors to diagnose, the algorithm-generated explanation which can only indicate the location of a lesion is not enough, and it is necessary to elaborate on which parts and features play an important role in the decision-making process.

In this paper, we propose a visually interpretable network (VINet), which can generate a visually interpretable result to help doctors make a faithful final diagnosis. The overall framework of VINet is illustrated in Fig. 2. VINet consists of an importance estimation network and a deep classification network. Based on the diagnostic visual interpretation generated by the former component, the importance value corresponding to each pixel in the input data has been learned. Since the classifier uses the same input as the importance estimation network, the importance of the corresponding pixels of the feature maps in the classifier is obtained. We then apply our proposed irrelevant feature destruction process to the feature maps of each pooling layer in the classifier to remove the unimportant and irrelevant features. When the network is trained with an end-to-end manner, its ability to make accurate diagnoses and to determine the importance of the various parts of the input image grows. Therefore, VINet can generate accurate diagnosis and diagnostic visual interpretation which precisely indicate important features in the input image.

In order to evaluate and validate VINet, we have tested it with a computed tomography (CT) scans dataset on pulmonary nodule - the Lung Nodule Analysis 2016 (LUNA16) challenge dataset [11]. The experimental results show that the diagnostic visual interpretation generated by VINet is superior to all baseline methods.

In summary, the main contributions of this work are as follows:

- 1) We propose a Visually Interpretable Image Diagnosis Network, namely VINet. The proposed structure, which is compatible with various popular deep learning-based classifiers and image generation network, can produce both classification and diagnostic visual interpretation.
- 2) The visual interpretation accurately captures the diagnosis basis of lung nodules. These features, which are highly in line with medical pathology, can help solve the trust problem and assist doctors to make a faithful final diagnosis.
- 3) With the visual interpretation of VINet, the classification network can produce satisfactory classification results by very few pixels (averaged 2.37% on the LUNA16 dataset). This inspiring discovery gives a new direction and possibility for biomedical image analysis.

The rest of this paper is organized as follows. Section II summarizes the related work on visualization research, interpretable computer-aided systems. Section III explains the proposed VINet in detail. Section IV presents the related experimental results, and Section V concludes our work.

II. RELATED WORK

A. Interpretability and Visualization Research

The interpretability of a machine learning model is usually defined as the degree to which the model can be understood by human users [12]. The main purpose of interpretability research is to allow more people to generate more trust in the algorithm, or more practically, to make computer-assisted systems more effective in helping people. To this end, researchers have proposed a variety of algorithms, which can be roughly divided into two categories according to the objects they interpret: one is to interpret the model, the other is to explain the decision. The difference between the two is how the responses are projected to the input space. The former is to solve the optimization problems of reconstructing the image input [13]–[15], and the latter is to reconstruct the decision of the classifier [16], [17].

As a milestone method for interpreting models, deconvolution network was originally designed for unsupervised learning tasks [15], then Zeiler *et al.* [14] use a deconvolution network attached to a convolutional network to project the feature activation of the intermediate layer to the input space. More specifically, the method repeatedly performs unpooling, rectification, and convolution on the feature map of a certain intermediate layer until the feature map and the input image have the same resolution, thereby identifying the pattern that activates a certain neuron the most. Guided propagation [13] replaces maxpooling with strided convolution. Compared to [14], the method additionally introduces the guidance signal from higher layers in the network, thus achieves a better performance. Another way to interpret the model is to find the specific input image in the dataset that can maximize the activation of neurons in higher layers in the deep network, as described in [18], [19].

Some algorithms explain the model decisions to improve interpretability. Authors in [20], [21] draw on the method in [14], which visualize the correct classification probability of partially occluded images. The method quantifies the effect of each input component on the output. Through this prediction difference

analysis, it visualizes the importance of each input component as a heatmap. Other sensitivity-based methods [22]–[24] use partial derivatives to calculate the contribution of input components to classifier decisions. Class activation mapping (CAM) [25] and gradient-weighted class activation mapping (Grad-CAM) [26] construct the weighted sum of the feature maps for visualization. The difference between the two methods is the way the weights are obtained; the former uses global average pooling and the latter uses the gradient. However, neither method can produce high-resolution, fine-grained results. The layer-wise relevance propagation (LRP) method [17] back-propagates the predictions of the deep network without using gradients. The method calculates the relevance between the prediction and each neuron from top to bottom and obtains the contribution of each pixel in the input image to the classifier prediction. Samek *et al.* [27] compare the deconvolution method, the sensitivity-based method, and LRP, and LRP outperforms other considered methods in interpreting deep network decisions. Visual back propagation (VBP) [16] was originally developed as a debugging tool for CNN-based systems for steering self-driving cars. The method first averages the multi-channel feature map into a single-channel feature maps for each layer and obtains the visualization by repeating the resizing and point-wise multiplication from the top to the bottom.

The above visualization methods all use the trained classifier as the analysis object, and no training is involved in the visualization process. As a special case, Dosovitskiy *et al.* [28] trained an up-convolutional network to reconstruct input images from feature representations. The method can reconstruct the original image from shallow representations such as histogram of oriented gradient (HOG) and scale invariant feature transform (SIFT) with excellent performance. For deep networks, the colors and the rough contours of an image can be reconstructed from activations in higher network layers and even from the predicted class probabilities.

B. Interpretable CAD System for Biomedical Images

When doctors use traditional methods to solve the classification tasks of medical images, they often use certain features weights in a linear classifier [29], [30] or their P-values [31] to explain the importance of each feature for the final diagnosis. However, these interpretations of the weights may be misleading [32]. Besides, these methods cannot directly interpret the input image.

When deep learning algorithms are used in the field of medical imaging, it is the most common means to interpret individual classification decisions by generating visual heatmaps through different processes. Cireş *et al.* [3] used a deep network to accurately locate mitotic cells in breast cancer histology images. Kim *et al.* [8] adopted a model that can diagnose glaucoma and locate the suspicious areas. CLEAR-DR, a diagnostic model for diabetic retinopathy proposed by Kumar *et al.* [9], can accurately locate lesions in the retina, thus enabling the clinician to visualize the factors taken by the system in predicting disease grades. MDNet proposed by Zhang *et al.* [10] is an interpretable medical imaging diagnostic algorithm that combines image models

and language models, achieving impressive results in pathology bladder cancer images. However, these explainable models face two important deficiencies. First, their visual interpretations are vague and can only be used to indicate the approximate location of the lesion. Second, the correctness and accuracy of their explanations can not be verified.

III. VISUALLY INTERPRETABLE NETWORK (VINET)

The overall structure of our model is shown in Fig. 2. Our proposed model consists of two sub-modules: importance estimation network and deep classification network. The importance estimation network predicts and quantifies each pixels importance for the classification. In the classification process, the information of the unimportant pixels is eliminated, and only the useful information will be preserved. Two constraints are applied to train VINet: (a) achieving the highest classification accuracy; (b) using the least information in the diagnostic process. Therefore, the preserved information must be crucial for the diagnosis.

A. Importance Estimation Network

The single-channel output of the importance estimation network (i.e., the diagnostic visual interpretation) has the same size as the input of the network. Each pixel on the generated diagnostic visual interpretation has a value between 0 and 1, indicating the pixels importance. Besides, the network structure is very flexible and can be replaced by any popular network if its input and output have the same size. Multiple networks are compared in the experiment in Section IV-E

The activation function applying on the last layer of the importance estimation network has a great impact on the convergence of the network. The most straightforward idea to set the output range to [0,1] is to use the sigmoid activation function. However, in experiments, we find that using sigmoid function may lead to training failure. With a sigmoid activation function, the importance estimation network tends to generate all-zero diagnostic visual interpretations, causing that the noise destruction layer in the classifier to remove all features, which further makes the classifier fail to work. Since our loss function consists of two parts, the network easily falls into local optimum with sigmoid function.

To address this problem, we propose a normalized softmax activation function applying to the final layer of the importance estimation network:

$$M = \frac{\text{softplus}(M_0)}{\max(\text{softplus}(M_0))} \quad (1)$$

where M_0 is a single-channel feature map generated by the last convolutional layer in the importance estimation network, M is the final diagnostic visual interpretation, and $\text{softplus}(\cdot)$ [33] is a function with output in a scale of $(0, +\infty)$:

$$\text{softplus}(x) = \ln(1 + e^x) \quad (2)$$

In Eq. (1), the output of the importance estimation network is normalized. With this normalization, it is impossible for the network to generate all-zero diagnostic visual interpretations and

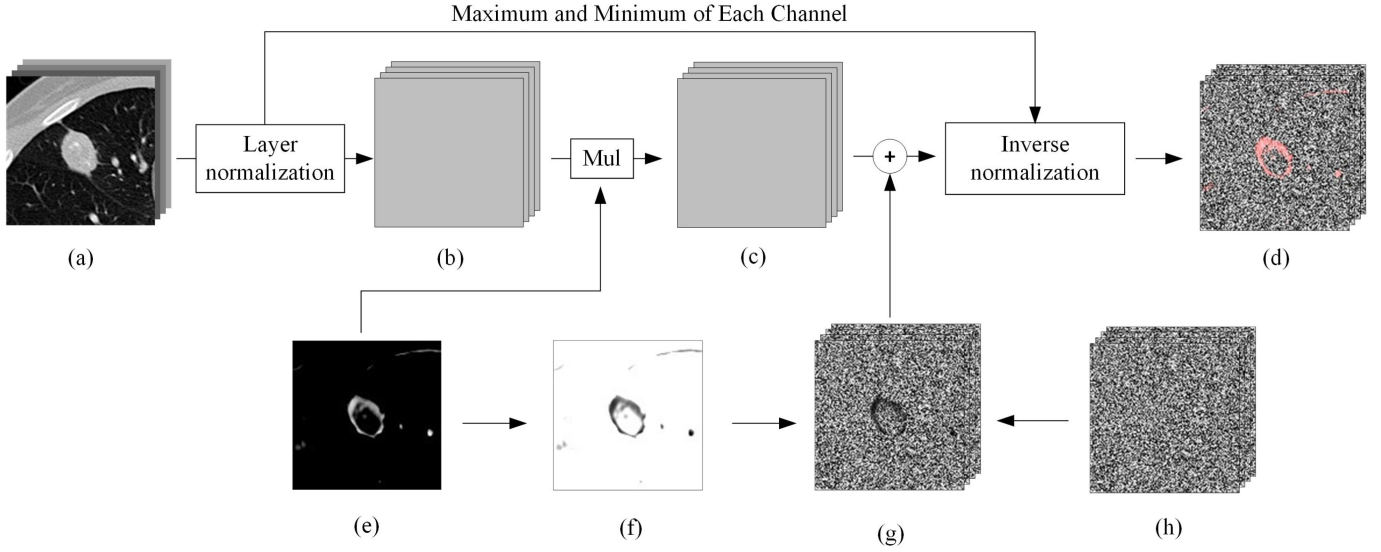


Fig. 3. The irrelevant feature destruction process. (a) is the original feature maps/input. (b) is the normalized feature maps. (c) is the images obtained by multiplying (a) and (e), which is the diagnostic visual interpretation M . (d) is the final output images of the feature destruction process. (f) is the inversion of the diagnostic visual interpretation M (i.e. $1-M$). (g) contains the images obtained by multiplying (f) and (h), which is the uniform noise.

some pixels must be preserved with values close to one, ensuring the training stability of the importance estimation network.

B. Classification Network

Same as the importance estimation network, the structure of the classification network is very flexible. Most of the popular network structures proposed in recent years such as DenseNet [34] and ResNet [35] can all be used in our framework to perform the classification task. The input of the classification network is the same as the importance estimation network. The output of the classification network is a prediction of whether the input image contains malignant pulmonary nodules. Inside the classifier, the feature maps obtained after each pooling layer are destroyed by the irrelevant feature destruction process.

C. Irrelevant Features Destruction

The irrelevant feature destruction process (noise destruction layer in Fig. 2), which inhibits the classifier from learning the unimportant or irrelevant information determined by the importance estimation network, can directly act on feature maps of any layer in the classification network. The whole process is illustrated in Fig. 3.

1) *Noise-Based Feature Destruction*: When using the noise to corrupt the feature maps inside a convolution network, a key problem is to effectively eliminate the unwanted information. It must be ensured that the feature maps inside the classification network are irreversibly removed and cannot be recovered by the learning process of the classification network. For example, a straightforward idea would be multiplying the original feature map and the diagnostic visual interpretation to make the values of those unimportant pixels smaller. However, this is obviously a bad idea, because if all the pixels are multiplied by the same small

coefficient, no information is virtually removed. The relative distribution of the data remains the same, and the original distribution can be restored by simply multiplying it by a large number. To ensure that unimportant features are irreversibly removed, we propose a novel noise-based feature destruction process.

First, we normalize each feature map P to set the values of pixels in P between 0 and 1, which works as:

$$P' = \frac{P - \min(P)}{\max(P) - \min(P) + \delta} \quad (3)$$

where δ is a small number set to 1×10^{-8} , P is the normalized output of the feature map P . After the normalization, the uniform noise is introduced to eliminate to irrelevant information, which is defined as:

$$P'' = P' \circ M + N \circ (1 - M) \quad (4)$$

where N is the uniform noise, M is the diagnostic visual interpretation, P'' is the output image of the noise-adding process. When M and P have different sizes, M is adjusted to have the same size as P by bilinear interpolation.

The noise-added feature map P'' is then scaled to the same scale as the original feature map P in order to ensure that the uniform noise can destroy the unimportant features completely. Therefore, we perform the inverse operation of Eq. (3) on P'' to obtain the final noise-damaged feature map P^* . The inverse operation is defined as:

$$P^* = \min(P) + (\max(P) - \min(P) + \delta) \cdot P'' \quad (5)$$

where P^* is the final output of the irrelevant feature destruction process.

It should be noted that changing the scales of the feature maps before and after adding uniform noise is necessary. When the noise is absolute rather than relative, the convolutional layers can amplify the original inputs by adjusting the weights during the learning process to counteract the damage. Our proposed

process perfectly solves this problem by keeping the features and noise always on the same scale.

2) *Multi-Scale Feature Destruction*: Although using uniform noise of the same scale as the input can cause irreversible damage to the original image, it does not mean that all irrelevant information in the image is removed. To address this problem, we implement a multi-scale feature destruction in the classification network as shown in Fig. 2. Not only the input image of the classification network is applied with the feature destruction process, but the irrelevant features of every downsampling layers are correspondingly removed as well. This ensures that the network cannot extract any useful information from the spatial distribution of noise to help classification. For a clearer illustration of the importance of the multi-scale feature destruction process, we compare the performances of using multi-scale feature destruction and single-scale feature destruction in Section IV-E.

D. Network Optimization

The design of VINet allows for an end-to-end training. In order to combine the importance estimation network and the classification network into one model, we design the loss function of VINet as the sum of two parts. The process of generating the diagnostic visual interpretation M and the process of classification are supervised by two losses L_M and L_C . The importance loss L_M aims to lead VINet to eliminate as many features as possible, and the classification loss L_C is to obtain more accurate classification results. The overall loss of the VINet is defined as:

$$L = L_C(v_c, l_{gt}) + \alpha \cdot L_M(M) \quad (6)$$

where l_{gt} is the class label. v_c is the output of classification network. M is the output of importance estimation network. α is set to 0.5. The classification loss L_C in (6) is the cross-entropy loss:

$$L_C(v_c, l_{gt}) = -\log \left(\frac{\exp(v_c[l_{gt}])}{\sum_j \exp(v_c[j])} \right) \quad (7)$$

The importance loss L_M in (6) is defined as:

$$L_M(M) = \frac{1}{H \cdot W} \cdot \sum_{(x,y)} M_{x,y} \quad (8)$$

where H is the height of M and W is the width of M .

Since L_C and L_M have mutually restrictive effects, coefficient α is used to balance the interaction between the two losses. When the value of L_M is small, the noise damage to the features in the classifier can be severe, resulting in a decrease in classification accuracy and an increase in L_C .

IV. EXPERIMENTS

A. Dataset

The LUNA16 dataset [11], which includes 1,186 nodule labels in 888 patients annotated by radiologists, is used to train the model. The LUNA16 dataset is a subset of the LIDC dataset [36], [37], and the malignancy score labels used in training come from the LIDC dataset. The nodules are classified into five

categories according to the given malignancy score. The reason why we use LUNA16 instead of LIDC directly is that LUNA16 ignores the nodules which were annotated by less than 3 doctors, which enhances the reliability of the experimental results.

In the following experiments, we use slices of CT images with a size of 128×128 as the input images. Each slice contains a lung nodule and has a corresponding image-level label indicating the malignancy score of the nodule.

B. Implementation Details

The hyper-parameter α introduced in Section III.D is set to 0.25 in the following experiments. The model is optimized via the Adam method [38] with batch size 8. For other hyper-parameters of Adam, the initial learning rate is set to 1×10^{-4} , and we set the exponential decay rates for the first and second moment estimates to 0.9 and 0.999, respectively. We use the method described in [1] to initialize the weight of the convolutional filters. Our model is trained with the Python toolbox PyTorch [39] on an NVidia Titan X GPU.

C. Diagnostic Visual Interpretation on LUNA16

The most common causes of lung nodules overall include granulomas (clumps of inflamed tissue due to an infection or inflammation) and hamartomas. In the clinic, the features of the edge contour of a nodule are used by doctors as a reference to diagnose lung cancer, such as spiculation mass, which is defined as a lump of tissue with spikes or points on the surface [40]. This is in line with the pathology of malignant tumors. A malignant pulmonary nodule can pull the blood vessels in the lungs, causing the blood vessels to concentrate, which in turn appear as non-smooth features of the edges of the lung nodules in the CT scans. In addition, the tumor cells stimulate the surrounding connective tissues to form reactive fiber bands and infiltrate adjacent bronchial vessels or local lymphatic, all of which cause lung nodules to exhibit certain edge features [5], [6], [16]. Besides, ground-glass nodule (GGN), which is defined as a nodular shadow with ground-glass opacity, is generally associated with the early-stage lung adenocarcinoma [41]. GGNs noted at thin-section CT scans have been shown to have a histopathologic relationship with atypical adenomatous hyperplasia, bronchioloalveolar carcinoma (BAC), and adenocarcinoma with a predominant BAC component [42].

In VINet, the diagnostic visual interpretation of each individual classification is the output of the importance estimation network. The diagnostic visual interpretation contains only the most important parts of the original input for classification, such as some cases shown in Fig. 4. Throughout the experiments, we observe that: (a) In the condition of lung nodules having clear edge contours, as shown in the three sets of images in the left column, the pixels that VINet highlights in the original input image are mainly the edge contour of the lung nodules. This indicates that the edge contour of the lung nodules plays an important role in the decision-making process. (b) In the three sets of images in the middle column, there are spiculation signs around

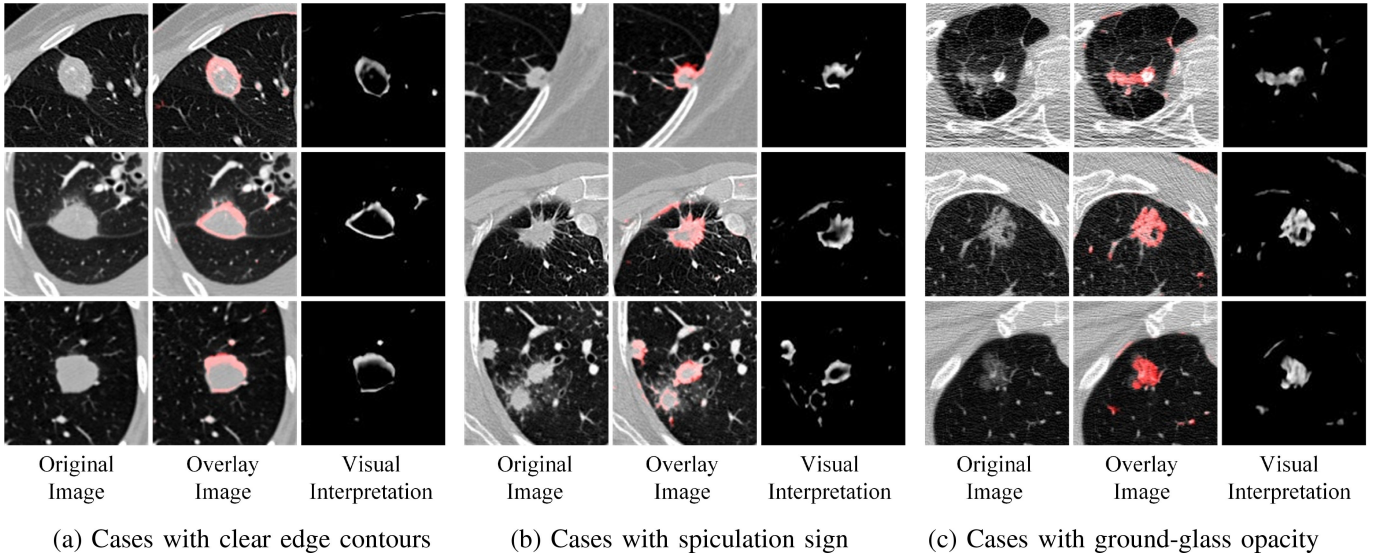


Fig. 4. Diagnostic visual interpretation of the diagnosis generated by VINet. Each case consists of three pictures. **Left:** the original CT scans. **Middle:** images obtained by superimposing the original images and the diagnostic visual interpretations, the red areas in the images correspond to the diagnostic visual interpretations. **Right:** diagnostic visual interpretation generated by VINet, the whiter and brighter areas in the images are considered more critical for lung cancer diagnosis.

the nodules. These signs are perfectly captured by VINet in diagnostic visual interpretations. (c) The three sets of images in the right column are GGNs. Because the main feature of GGN is the shadow with unclear edges, so the diagnostic visual interpretation of GGNs accurately covers the area where GGN is located.

D. Evaluating Correctness

In this section, we evaluate the correctness and accuracy of the visualization result. To show that our proposed VINet can generate state-of-the-art visualization results, we choose three popular visualization methods for CNNs as baseline for comparison. These three methods are CAM [25], VBP [16], and LRP [17], which are very popular visualization methods for CNNs in recent years.

1) *Classification Performance of Important Features:* A reasonable visual interpretation can accurately capture the important features for diagnosis. In this part, we reserve the important features of the original image and set the rest pixels to zero. A classification network is trained and tested with the processed images. If the classification network still has satisfactory classification accuracy, the visual interpretation can be considered as accurate and correct.

Specifically, the visual interpretation generated by different methods is binarized as a mask. With a constant as the threshold, the pixels whose weight is greater than the threshold value are reserved, and the pixels whose weight is less than the threshold value are set to zero. The ResNet-18 [35] network is trained and tested with the processed datasets and the five categories malignancy labels. The accuracy of these diagnostic visual interpretations is evaluated by comparing two indicators. One is the classification accuracy of the ResNet-18 network, and the other is the proportion of pixels reserved in the original image. The experimental results are shown in Table I. The reserved

TABLE I
CLASSIFICATION PERFORMANCE OF IMPORTANT FEATURES. THE RESERVED PIXELS HAVE LOW WEIGHT IN THE IMPORTANCE MAP

| Method | Reserved pixels proportion (%) | Classification accuracy (%) |
|--------------|--------------------------------|-----------------------------|
| Original | 100.00 | 82.57 |
| VINet | 2.37 | 82.15 |
| VBP [16] | 17.71 | 78.58 |
| CAM [25] | 36.22 | 78.35 |
| LRP [17] | 7.68 | 77.57 |

pixels proportion is defined as:

$$S = \frac{1}{H \cdot W} \cdot \sum_{(x,y)} \text{floor}(I_{x,y} + T) \quad (9)$$

where I is the visual interpretation, H and W are the width and height of I , and the constant T is set to 0.9.

The LRP algorithm removes the second most pixels, and it severely damages classification performance (drops by 5.00%), which means that it incorrectly removes some useful features. The VBP algorithm achieves the second-highest classification accuracy. However, it reserves too many pixels, including many irrelevant features. In summary, the overall performance of VINet on both indicators exceeds all baseline methods. The average number of pixels in the diagnostic visual interpretation produced by VINet is only 2.37% of the number of pixels in the original input image. With such a limited number of pixels, a network trained from scratch must perform poorly if the diagnostic visual interpretation is not correct and accurate. Compared with the network that uses the raw image as input, the classification accuracy of the network trained and tested on the diagnostic visual interpretation produced by VINet only drops by 0.42%, which strongly supports the accuracy and correctness of the diagnostic visual interpretation generated by VINet.

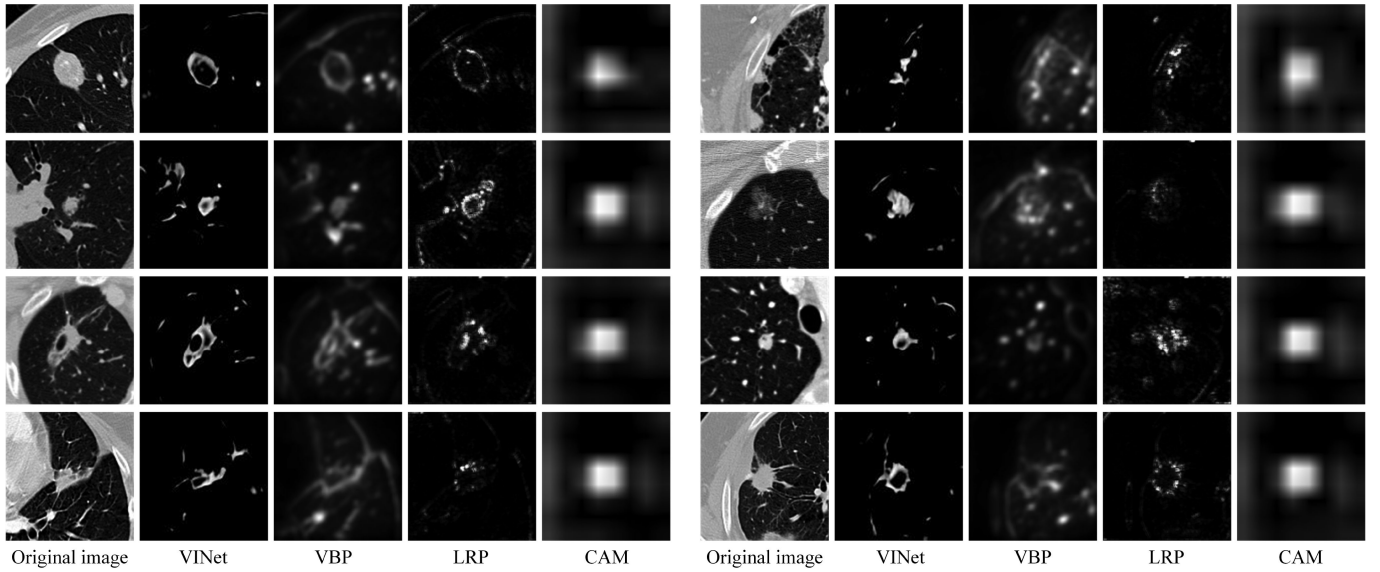


Fig. 5. Diagnostic visual interpretation of the classification of some pulmonary nodules by VINet and other baseline methods. The white position in the diagnostic visual interpretation indicates that the position is important in the decision-making process.

TABLE II
COMPARISON OF OCCLUSION SENSITIVITY. THE RESERVED PIXELS HAVE LOW WEIGHT IN THE IMPORTANCE MAP

| Method | Reserved pixels proportion (%) | Classification accuracy (%) |
|--------------|--------------------------------|-----------------------------|
| Original | 100.00 | 82.57 |
| VINet | 97.63 | 8.11 |
| VBP [16] | 82.29 | 17.08 |
| CAM [25] | 63.78 | 10.60 |
| LRP [17] | 92.32 | 40.03 |

2) *Occlusion Sensitivity*: In this part, we verify the correctness of visual interpretation by image occlusion [14]. Specifically, we test the difference in classification accuracy when important pixels of input images are masked.

A classification network ResNet-18 [35] is trained separately on the original dataset for testing. Different visualization methods generate different visual interpretations on the test set. With a constant as the threshold, pixels whose weight greater than the threshold are set to zero. Table II shows the classification accuracy differences of the same classification network on different masked test sets. The reserved pixels proportion is defined as:

$$S = \frac{1}{H \cdot W} \cdot \sum_{(x,y)} 1 - \text{floor}(I_{x,y} + T) \quad (10)$$

where I is the visual interpretation, H and W are the height and width of I , and the constant T is set to 0.9. With T as the threshold, the important pixels in visual interpretation are set to zero in the original image. The test results show that the visual interpretation of VINet has the highest occlusion sensitivity. With masking the least pixels (2.37%), the effect on classification accuracy is the most significant (from 82.57% to 8.11%).

E. Visual Comparison

We use the same LUNA16 dataset to train a VGG-16 [43] network and the above three visualization methods are used on the trained network. The comparison between these three visualization methods and VINet are shown in Fig. 5.

By comparing the results obtained by each visualization method, we observe that: (a) the CAM algorithm can only indicate the approximate position of the nodule, and the other three visualization methods can remove most of the redundant information and accurately locate the nodule positions. (b) the result of the LRP algorithm removes a significant portion of the redundant information and only focuses on a very small number of pixels, which however causes it to ignore many features that should be considered important. (c) the result of the VBP algorithm is better than those of the CAM and the LRP algorithms, but it often contains edges that are not related to lung nodules. (d) VINet can effectively capture features of nodules and remove irrelative features. The diagnostic visual interpretation produced by VINet is significantly clearer and more accurate than the other three baseline methods.

Compared with other methods, the visualization results of VINet are much sharper. Most of the pixels of the visualization results are high-weight (close to 1) or low-weight (close to 0). A reasonable explanation is that the medium-weight pixels in the importance map can not effectively reduce the loss function and therefore can not survive in the training process. Since the feature maps in the classification network are added with high-intensity noise on multi-scale, only the locations contain the key information for diagnosis will be high valued and the corresponding image information will be retained. For the other locations, in the training process, even if the pixels are medium-valued, the corresponding image information will be seriously destroyed by the proposed multi-scale feature destruction. In this condition, to minimize the loss function, the importance loss will depress

TABLE III
PERFORMANCE COMPARISON OF VINET WITH DIFFERENT IMPORTANCE ESTIMATION NETWORKS

| Importance estimation network | classification network | Classification accuracy (%) | Importance loss ($\times 10^{-2}$) |
|-------------------------------|------------------------|-----------------------------|--------------------------------------|
| - | SEResNet-50 | 83.43 | - |
| DeepLabv3+ [44] | SEResNet-50 | 81.37 | 2.95 |
| GridNet [45] | SEResNet-50 | 80.75 | 3.46 |
| U-Net [46] | SEResNet-50 | 80.63 | 3.85 |
| SegNet [47] | SEResNet-50 | 76.78 | 7.57 |

TABLE IV
PERFORMANCE COMPARISON OF VINET WITH DIFFERENT CLASSIFICATION NETWORKS

| Importance estimation network | Classification network | Classification accuracy (%) | Importance loss ($\times 10^{-2}$) |
|-------------------------------|------------------------|-----------------------------|--------------------------------------|
| DeepLabv3+ | SEResNet-50 [48] | 81.37 | 2.95 |
| DeepLabv3+ | DenseNet-121 [34] | 80.58 | 3.24 |
| DeepLabv3+ | ResNet-50 [35] | 79.90 | 3.19 |
| DeepLabv3+ | VGG-19 [43] | 75.54 | 8.25 |

the pixels value of the importance map significantly. As a result, when the importance estimation network converges, the pixels will be high or low valued, and the visualization results look sharp.

F. Ablation Study

1) *VINet with Different Network Structures*: In this ablation study, we aim to explore the effects of the different network structures of the importance estimation network and the classification network on experimental results under the VINet framework. The structure of VINet is extremely flexible. The importance estimation network and the classification network inside VINet can use a variety of different network structures. For the importance estimation network, VINet only requires its input and output to have the same size, therefore almost all networks designed for semantic segmentation and keypoint detection can be used as the importance estimation network.

In the ablation experiment for the importance estimation network, network structures such as U-Net, SegNet, GridNet and DeepLabv3+ are selected as the importance estimation network. The experimental results are shown in Table III. The importance loss is defined as the Equation 8. To demonstrate the impact of noise on classification performance, we also tested the performance of the classification network without noise corruption.

In the ablation experiment for classification network, VGG-19, ResNet-50, DenseNet-121, and SEResNet-50 are selected as the classifier. The experimental results are shown in Table IV. After considering the experimental results in Tables III and IV, in the standard version of VINet, we use DeepLabv3+ as the importance estimation network and SEResNet-50 as the classification network.

2) *VINet with Different Feature Destruction Processes*: To verify the necessity of multi-scale feature destruction process, we compare the diagnostic visual interpretations generated by VINet with different feature destruction strategies. As a comparison of multi-scale feature destruction process, the single-scale feature destruction process only corrupts the original input image of the classifier. The experimental results of some random

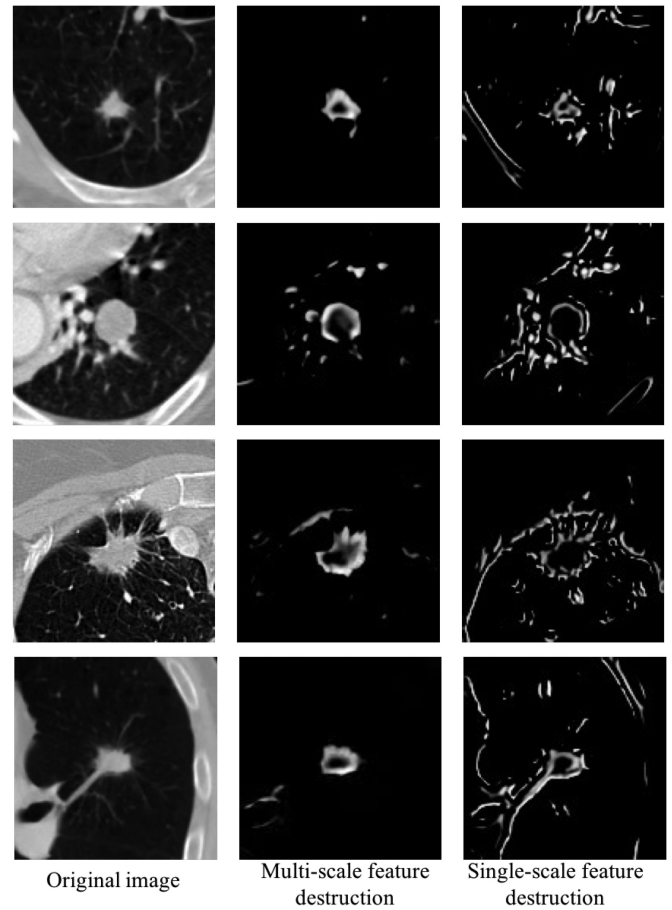


Fig. 6. Diagnostic visual interpretations generated by VINet with multi-scale and single-scale feature destruction processes. VINet with multi-scale feature destruction accurately captures pathological features whereas VINet with single-scale feature destruction targets too many irrelevant parts.

cases are shown in Fig. 6. As shown by the results of single-scale feature destruction, the classification network tends to focus on the edge-like features in the input image. It is easy to notice that these edge-like features produced by a single-scale feature destruction process contain a lot of misinterpretation, targeting areas which is not related to the pathological signs. However, the multi-scale feature destruction process corrects the outliers produced by a single-scale process, generating a more accurate and clearer pathology-consistent diagnostic visual interpretation.

G. Limitations of VINet

The network structure has a great influence on the visualization and classification results. However, the classification network and importance estimation network of VINet are the common network structures used in the field of natural images, which do not take into account the characteristics of medical images. One of our important tasks in the future is to design the new network structures which make use of medical image characteristics to improve the accuracy of classification and visualization. In addition, the visualization results of the proposed VINet need to be further verified in clinical practice. We will

further improve our method according to the feedback from the doctors.

V. CONCLUSION

In this study, we propose VINet, a visually interpretable image diagnosis network. Combining an importance estimation network and a classification network in a coupled way, VINet can generate diagnostic visual interpretations and classification decisions at the same time. The accuracy of our models diagnostic visual interpretation is validated by the experiments, and its correctness is also supported by the pathology of malignant tumors. Besides, experimental results show that VINet can explain the decision-making basis of the deep network more accurately than all the baseline visualization methods, reaching the state-of-the-art level.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.
- [2] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.
- [3] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2013, pp. 411–418.
- [4] 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.*, vol. 30, no. 11, pp. 3484–3495, Nov. 2019.
- [5] S. S. Siegelman *et al.*, "Solitary pulmonary nodules: CT assessment," *Radiology*, vol. 160, no. 2, pp. 307–312, 1986.
- [6] C. Zwirerich, S. Vedal, R. Müller, and N. Müller, "Solitary pulmonary nodule: High-resolution CT and radiologic-pathologic correlation," *Radiology*, vol. 179, no. 2, pp. 469–476, 1991.
- [7] K. Kuriyama *et al.*, "CT-pathologic correlation in small peripheral lung cancers," *Amer. J. Roentgenology*, vol. 149, no. 6, pp. 1139–1143, 1987.
- [8] M. Kim *et al.*, "Web applicable computer-aided diagnosis of glaucoma using deep learning," *Mach. Learn. Health Workshop at 2018 Neural Inf. Process. Syst.*, 2018.
- [9] D. Kumar, G. W. Taylor, and A. Wong, "Discovery radiomics with CLEAR-DR: Interpretable computer aided diagnosis of diabetic retinopathy," *IEEE Access*, vol. 7, pp. 25 891–25 896, 2019.
- [10] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6428–6436.
- [11] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, 2017.
- [12] O. Lahav, N. Mastronarde, and M. Van Der Schaar, "What is interpretable? using machine learning to design interpretable decision-support systems," *Mach. Learn. Health Workshop at 2018 Neural Inf. Process. Syst.*, 2018.
- [13] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Workshop at Int. Conf. on Learn. Representations*, ICLR, 2015.
- [14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 818–833.
- [15] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 2528–2535.
- [16] M. Bojarski *et al.*, "Visualbackprop: Efficient visualization of cnns for autonomous driving," *IEEE Int. Conf. Robot. Automat.*, pp. 1–8, 2018.
- [17] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.
- [18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 580–587.
- [19] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [20] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Tran. Knowl. Data Eng.*, vol. 20, no. 5, pp. 589–600, May 2008.
- [21] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *Int. Conf. on Learn. Representations*, ICLR, 2017.
- [22] D. Baehrens *et al.*, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.
- [23] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop at Int. Conf. on Learn. Representations*, ICLR, 2014.
- [24] P. M. Rasmussen *et al.*, "Visualization of nonlinear classification models in neuroimaging: Signed sensitivity maps," *BIOSIGNALS 2012 - Proc. of the Int. Conf. on Bio-Inspired Syst. and Signal Process.*, pp. 254–263, Jan. 2012.
- [25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2921–2929.
- [26] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 618–626.
- [27] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.
- [28] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4829–4837.
- [29] S. Klöppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.
- [30] C. Ecker *et al.*, "Describing the brain in autism in five dimensions: magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," *J. Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.
- [31] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, "Support vector machine learning-based FMRI data group analysis," *Neuroimage*, vol. 36, no. 4, pp. 1139–1151, 2007.
- [32] S. Haufe *et al.*, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.
- [33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML-10)*, 2010, pp. 807–814.
- [34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
- [36] 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.
- [37] S. G. Armato III *et al.*, "Data from LIDC-IDRI: the cancer imaging archive," vol. 9, p. 7, 2015, <http://doi.org/10.7937/K>.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learn. Representations*, ICLR, 2014.
- [39] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Inf. Process. Syst.*, pp. 8024–8035, 2019.
- [40] S. J. Swensen *et al.*, "CT screening for lung cancer: Five-year prospective experience," *Radiology*, vol. 235, no. 1, pp. 259–265, 2005.
- [41] D. Chen *et al.*, "New horizons in surgical treatment of ground-glass nodules of the lung: Experience and controversies," *Therapeutics clin. Risk Manage.*, vol. 14, p. 203, 2018.
- [42] H. Y. Lee and K. S. Lee, "Ground-glass opacity nodules: Histopathology, imaging evaluation, and clinical implications," *J. Thoracic Imag.*, vol. 26, no. 2, pp. 106–118, 2011.
- [43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Representations*, ICLR, 2015.

- [44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2016.
- [45] D. Fourure *et al.*, "Residual conv-deconv grid network for semantic segmentation," in *British Mach. Vision Conf.*, BMVC, 2017.
- [46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
- [47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7132–7141.



Donghao Gu received the bachelor's degree in automation from Harbin Institute of Technology (HIT), China, in 2018, where he is currently working toward the master's degree with the School of Computer Science and Technology. His research interests include image processing and computer vision.



Yaowei Li received the bachelor's degree of automation from the Harbin Institute of Technology (HIT), Harbin, China, in 2018, where he is currently working toward the master's degree with the School of Computer Science and Technology. His research interests include image processing, computer vision, and biomedical image analysis.



Feng Jiang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2001, 2003, and 2008, respectively. He is currently a Professor with the Department of Computer Science, HIT and a Visiting Scholar with the School of Electrical Engineering, Princeton University, Princeton, NJ, USA. His research interests include computer vision, pattern recognition, and image and video processing.



Zhaojing Wen received the bachelor's degree in measurement and control technology and instrument from Harbin Institute of Technology (HIT), China, in 2018, where he is currently working toward the master's degree with the School of Computer Science and Technology. His research interests include image processing and computer vision.



Shaohui Liu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computation mathematics and its application software, computation mathematics, and computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1999, 2001, and 2007, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, HIT. His research interests include data compression, pattern recognition, image and video processing, and multimedia security.



Wuzhen Shi received the bachelor's degree from Shenyang Agricultural University, Shenyang, China, in 2012, and the master's degree from Northwest A & F University, Yangling, China, in 2014. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Since 2018, he has been with the Peng Cheng Laboratory. His research interests include image processing, computer vision, and image/video coding and transmission.

Guangming Lu (Member, IEEE), photograph and biography not available at the time of publication.

Changsheng Zhou (Member, IEEE), photograph and biography not available at the time of publication.