This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

Diagnosis of benign and malignant thyroid nodules using combined conventional ultrasound and ultrasound elasticity imaging

Pinle Qin, Kuan Wu, Yishan Hu, Jianchao Zeng, Xiangfei Chai

***Abstract*—Ultrasonography is one of the main imaging methods for diagnosing thyroid nodules. Automatic differentiation between benign and malignant nodules in ultrasound images can great assist inexperienced clinicians in their diagnosis. The core of problem is the effective utilization of the features of ultrasound images. In this study, we propose a method that is based on the combination of conventional ultrasound and ultrasound elasticity images based on a convolutional neural network and introduces richer feature information for the classification of benign and malignant thyroid nodules. First, the conventional network model performs pretraining on ImageNet and transfers the feature parameters to the ultrasound image domain by transfer learning so that depth features may be extracted and small samples may be processed. Then, we combine the depth features of conventional ultrasound and ultrasound elasticity images to form a hybrid feature space. Finally, the classification is completed on the hybrid feature space, and an end-to-end CNN model is implemented. The experimental results demonstrate that the accuracy of the proposed method is 0.9470, which is better than that of other single data-source methods under the same conditions.**

***Index Terms*—Image classification, Transfer learning, Deep learning, Ultrasound image, Elastic ultrasound**

I. Introduction

THYROID nodules are one of the most common endocrine carcinomata. According to global epidemiological data, the incidence of thyroid malignancy is increasing every year. Ultrasonography has become the most widely used modality for detecting and diagnosing thyroid cancer. Ultrasonography is a safe, convenient, non-invasive, and repeatable diagnostic technique that can accurately locate thyroid masses, identify echogenic characteristics within the thyroid nodules, and find annular, point-like blood flow signals inside the masses so that small lesions may be detected and the blood flow inside them may be evaluated. Compared with computed tomography (CT) and magnetic resonance imaging (MRI), ultrasonography has a better ability to distinguish benign and malignant nodules in pathological features. This greatly facilitates early clinical diagnosis and the choice of treatment options [1]. With the rapid development of medical imaging technology, computer aided diagnosis (CAD) can replace subjective diagnoses, which are largely dependent on personal experience. CAD, which functions as a well-trained "expert", has wide application prospects in several cases that rely on experience. A fully-automatic CAD process includes image preprocessing (i.e. denoising, data augmentation, and image reconstruction), region-of-interest (ROI) extraction, segmentation, and classification. Recently, the first two phases have attracted considerable attention, whereas studies on classification by using ultrasound images are still rare, particularly in the case of thyroid nodule classification. The challenge of such a classification problem primarily lies in the selection of distinguishable features. Thus, most studies have focused on feature design for various types of conventional ultrasound imaging, such as morphometric and texture features. In [2], Owjimehr et al. chose completed LBP texture to classify liver images, and Zakeri et al. proposed certain effective texture features to distinguish breast nodules [3]. Luo et al. proposed classifying thyroid nodules by linear discriminant analysis [4]. Ding et al. combined B-mode images and elastograms to obtain both local texture and global elastic features, achieving an accuracy of 0.9680 on 40×40-size thyroid nodules [5]. Raghavendra et al. combined spatial gray-scale dependence and fractal texture features to discriminate thyroid nodules in conventional ultrasound images, achieving an area under the curve (AUC) index of 0.9445 [6]. However, experimental results demonstrate that these features perform quite unsatisfactorily on the dataset of the present study. This may be due to the internal simplicity and locality of low-level features. Moreover, these methods are not suitable for practical clinical applications primarily because they require hand-crafted feature information to label the contour of nodules.

Deep-learning methods, particularly convolutional neural networks (CNNs), are widely used and perform highly satisfactorily in various visual recognition tasks, such as object detection and image classification [7]–[9]. CNNs extract features that can be regarded as complex hierarchical representations of inputs; moreover, they can well capture recessive image features [10]. Owing to their successful application to non-medical images, CNNs have been applied to several medical-image classification problems. For example, Spanhol et al. [11] proposed using a CNN to classify breast cancer histopathological images. Wu et al. compared radiological and Bayesian techniques, support vector machines, and neural networks in the classification of thyroid nodules. Neural networks can achieve an accuracy of 0.8474 and an AUC of 0.9103, which are close to those of radiological methods [12]. Li et al. directly trained ResNet-50 [13] and DarkNet-19 [14] on a large conventional ultrasound dataset. Then, they combined these deep-learning models by weighting their performance

P. Qin and K. Wu contributed equally, and they are both the first author.
J. Zeng is the corresponding author, email: zjc@nuc.edu.cn
P. Qin, K. Wu, Y. Hu and J. Zeng with Shanxi Medical Imaging and Data Analysis Engineering Research Center, School of Big Data, North University of China
X. Chai with Huiying Medical Technology (Beijing) Co, Ltd

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

and assessed the ensemble DCNN model by using internal and external validation sets. This method achieves an accuracy of 0.8980 and an AUC of 0.9470, which exceed radiological indicators [15]. They used large-scale ultrasound datasets for direct training, thus overcoming the problem of small-data training. Wang et al. proposed a semi-supervised learning method, designed an effective EM algorithm to train a CNN, and classified weakly labeled conventional ultrasound data with an accuracy of 0.8825 and an AUC of 0.9286 [16]. However, these CNNs consist of a large number of nodes and configurations, this implies that only large datasets can support the training process. Unfortunately, this hampers medical applications because such large datasets are usually unavailable. The lack of sufficient image data will lead to overfitting problems. Two possible solutions are transfer learning [17] and data augmentation. Initial studies have demonstrated that transfer learning may be effective by applying well-trained deep neural network models to other image datasets for feature extraction because the aim of CNNs is not handling a specific problem but learning to capture the inherent features of visual objects [18], as shown in Figure 1. Liu et al. proposed combining low-dimensional texture features, such as HOG and LBP, with high-dimensional semantic features after transfer. This can effectively compensate for the lack of features caused by insufficient data [19]. However, extracting low-dimensional features requires fine contour annotation. The acquisition of this annotation information is quite difficult and is not available in the present dataset. Regarding data augmentation, traditional image enhancement methods, such as cropping, rotation, flipping, and scaling, are often used. Similarly, GAN is also a data augmentation method. Zhu et al. proposed a data augmentation method based on CNNs for improved thyroid nodule classification performance in conventional ultrasound images [20]. However, data augmentation methods for natural images are not suitable for medical images, because the annotation of medical images requires an experienced clinician.

All the above methods perform feature extraction or data augmentation on conventional ultrasound images without using ultrasound elasticity data. From the perspective of clinical diagnosis, ultrasound elasticity imaging is an important supplement to conventional ultrasound imaging [21]. Ultrasound elasticity imaging has different principles from those of conventional ultrasound imaging, as shown in Figure 2. Elasticity images reflect biomechanical characteristics. The underlying imaging principle is that an external force is applied to the lesion by means of the probe, and the tissue hardness is indirectly reflected by detecting the degree of deformation of the lesion under the external force. Tissues with a larger elastic coefficient exhibit larger hardness and less deformation, whereas tissues with a smaller elastic modulus exhibit the opposite behavior. Owing to the multi-source structure of tumors, tumor images are complicated. There is some overlap between the features of benign and malignant masses, and diagnostic specificity is low. Conventional ultrasound imaging has certain limitations in identifying benign and malignant masses. The development of ultrasound elastography further expanded the scope of conventional ultrasound diagnosis regarding the identification of benign and malignant thyroid nodules, as information on tissue elastic hardness can be directly obtained, thus compensating for the deficiency of traditional medical imaging [22]–[24]. Moreover, ultrasound elastography objectively quantifies the lesion hardness of thyroid nodules by computer technology, and greatly reduces the subjective error of the operator, thereby improving the specificity, sensitivity, and reliability of the diagnosis. Therefore, ultrasound elasticity imaging is of considerable value in relation to the thyroid nodular disease and should be considered for clinical application. According to the diagnostic and management guidelines for thyroid nodules published jointly by the American Society of Clinical Endocrinology (AACE), the American Endocrine Institute (ACE), and the Italian Clinical Endocrine Association (AME), elasticity imaging falls under the category of thyroid ultrasound imaging. When ultrasound imaging and cytology are inconclusive, elasticity imaging can be used as a supplementary examination but cannot completely replace a B-ultrasound scan [25]. Therefore, ultrasound elasticity imaging may significantly aid in the diagnosis of thyroid cancer. Combined with the pathological features and tissue characteristics of thyroid cancer obtained by conventional ultrasound imaging, it can more effectively identify benign and malignant thyroid lesions.

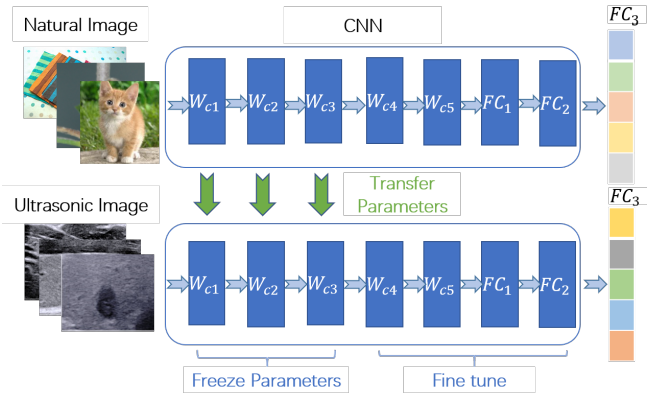

Fig. 1. Flowchart of deep models for transfer learning fine-tuning

In this study, we transfer the VGG16 model parameters pretrained on ImageNet [26] to the present ultrasound image dataset and verify that selecting the first six convolutional layers as a fixed feature extractor is the best choice for transfer learning on ultrasound data. Accordingly, we also extract the features of conventional ultrasound and ultrasound elasticity images to form a hybrid feature space and achieve an end-to-end classification model. Furthermore, it is confirmed that the fully connected layer plays an indispensable role in cross-domain knowledge transfer with large differences. To the best of our knowledge, conventional ultrasound and ultrasound elasticity imaging have not been combined to distinguish benign and malignant of thyroid nodules using a CNN. The experimental results demonstrate that the classification performance of the proposed method is obviously better than that of other methods.

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

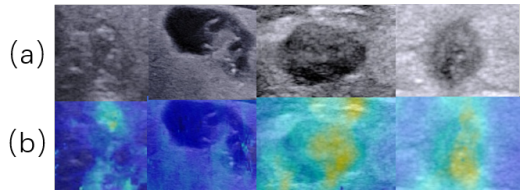

Fig. 2. Conventional ultrasound versus ultrasound elasticity imaging: (a) Conventional ultrasound nodule images (b) Ultrasound elasticity nodule images. Both correspond to the extract same position.

II. Method

CNNs are a class of deep-learning models and can extract high-level features. A CNN consists of an input and an output layer, as well as multiple hidden layers. A CNN adopts a feed-forward operating mode, and the features generated by a layer are input to an intermediate layer and are then passed on to the next layer [27]. In classification tasks, a CNN can be used as a feature extractor, and its learned parameters can be transferred to distinct but related problems. Owing to the large differences between two domains, it is an open problem to determine the extract process of knowledge transfer from the natural image domain so that the classification performance for medical applications may be improved. Our conclusion is that although the high-level features of ultrasound and natural images are distinct, low-level features, such as oriented edges, corners and junctions, do share similar statistical patterns. Thus, low-level representations learnt from natural images can be transferred to the medical image domain. The proposed transfer learning is based on ImageNet pretraining. He and Grishick proved that ImageNet pre-training can speed up convergence, particularly at early training stages, because ImageNet pre-training provides the low-/mid-level features (e.g., edges and textures). ImageNet pretraining is a historical workaround when the target data or computational resources are insufficient for training on the target task [28]. In [29], Tajbakhsh et al. transferred intima-media boundary segmentation in ultrasonographic images for training and fine-tuning. The results were better than those by conventional learning. This demonstrates that the deep fine-tuning of a CNN is subitable for medical image analysis, and a well-trained CNN is useful even if limited training data are available. Moreover, knowledge transfer from natural to medical images is feasible although there are relatively large differences between source and target data, and for specific applications, neither "shallow" nor "deep" is the best choice. Hierarchical fine-tuning allows the understanding of the effective depth of fine-tuning, as it depends on the specific application and the amount of tagged data that can be used. Based on the amount of available data, layer-by-layer fine-tuning provides a practical way to achieve the best performance for a specific application. Similarly, Chen et al. use dictionaries learnt from natural ultrasound and images corrupted with random Gaussian noise to calculate the residual of the reconstructed target ultrasound image. It is demonstrated that dictionary bases learnt from the natural image domain could reconstruct ultrasound data well, and hence the knowledge learnt from natural images

Method	Accuracy%	Sensitivity%	Specificity%	AUC%
VGG16	72.99	73.68	72.15	75.71
ResNet18	70.11	79.76	61.11	73.74
GoogleNet	68.96	70.00	67.57	73.15
Inception-V3	66.09	71.74	64.63	70.17
AlexNet	67.81	75.51	57.89	69.20

TABLE I
COMPARISON OF CONVENTIONAL TRAINING RESULTS OF DIFFERENT METHODS ON CONVENTIONAL ULTRASOUND IMAGES

Method	Accuracy%	Sensitivity%	Specificity%	AUC%
VGG16	72.41	66.29	78.82	79.56
ResNet18	70.69	68.42	73.42	78.44
GoogleNet	69.54	64.21	75.95	73.94
Inception-V3	68.29	65.06	71.43	69.73
AlexNet	67.24	64.58	70.51	73.69

TABLE II
COMPARISON OF CONVENTIONAL TRAINING RESULTS OF DIFFERENT METHODS ON ULTRASOUND ELASTICITY IMAGES

may effectively be transferred to the medical image domain. It is observed that the statistical patterns in low-level features extracted from natural and ultrasound images are quite similar [30]. Hence, the knowledge transferred from the natural image domain to the ultrasound domain has the potential to enhance the learning performance with limited ultrasound data.

A. Transfer Learning and Pre-trained Model

We first trained five networks on the present training set, VGG16 [31], ResNet18, GoogleNet [32], inception-V3 [33], and AlexNet. The training results for conventional ultrasound and ultrasound elasticity images are shown in Tables I and II, respectively. It can be seen that if the data are seriously insufficient, then for both deep models with a large number of parameters (such as VGG16, with a total of 13 convolutional layers), and for shallower networks (such as AlexNet, with only five convolutional layers), conventional training is not satisfactory. Therefore, transfer learning should be used to alleviate the problem of insufficient data. Moreover, we chose the VGG model (which achieved the best results in the table above) as the basic network structure for subsequent transfer learning and feature or data fusion.

The VGG16 model, designed by the Oxford Visual Geometry Group, consists of 13 convolutional layers, five pooling layers, and three FC layers. The convolution part for feature extraction and the fully connected (FC) layers have 4096-dimentional outputs that can be directly used as feature descriptors for classification. During this process, the output of each layer can be regarded as a certain feature, and the features of different layers have different meanings. The convolution in the front layer extracts local image features, and in subsequent layers, the receptive field is expanded by downsampling to extract more abstract semantic features.As shown in Figure 3, we use a deconvolutional network to project visual features onto the pixel space [34]. The output of the lower layers, corresponds to low-level features. For example, layer 2 responds to corners and other edge/color conjunctions. By contrast, in the output from the last several layers, various composite features emerge. Layer 7 exhibits entire objects with

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

Freeze layers	Accuracy% (Conventional ultrasound images)	Accuracy% (ultrasound elasticity images)
0	75.05 ± 1.65	76.89 ± 1.54
1	76.21 ± 0.65	77.47 ± 1.96
2	78.85 ± 1.59	78.16 ± 0.41
3	78.74 ± 0.57	79.34 ± 1.26
4	80.57 ± 1.31	80.57 ± 1.59
5	81.95 ± 1.19	82.29 ± 0.94
6	**82.76 ± 0.74**	**85.06 ± 0.66**
7	81.72 ± 1.36	83.10 ± 1.32
8	81.26 ± 1.97	83.56 ± 1.18
9	80.23 ± 0.31	82.52 ± 1.75
10	79.88 ± 1.21	81.03 ± 1.77
11	79.54 ± 0.51	81.15 ± 1.59
12	78.05 ± 1.10	80.23 ± 1.55
13	76.44 ± 1.40	79.77 ± 1.24

TABLE III
COMPARISON OF TRANSFER-LEARNING FINE-TUNING CROSS-VALIDATION RESULTS ON DIFFERENT IMAGES.

significant pose variation, and these features are related to specific classification tasks. Therefore, in CNNs, the features of the first few layers can usually be trained on a certain dataset and applied to another related dataset.

We conducted a comparative experiment using the VGG16 model pretrained on the present training set to determine the number of convolutional layers required for a general feature extractor. As shown in Figure 1, we freeze the weights and bias parameters of the previous n-layer convolution from the beginning to the end, so that the gradient of the back-propagation will not be updated when the pre-training network fine-tunes on the thyroid ultrasound dataset. Thus, the previous n-layer convolution is equivalent to a fixed-parameter feature extractor, that is, the gradient of the back-propagation will update only the parameters of the subsequent convolutional and FC layers to adapt to the new classification task. In the present dataset, two groups of different data, namely, conventional ultrasound and ultrasound elasticity images, were compared experimentally. The experimental results are shown in Table III. Under conventional ultrasound and ultrasound elasticity data, the convolution parts of the first six layers are used as general feature extractors in transfer learning, and fine-tuning of the subsequent convolutional and FC layers on ultrasound images results in the best performance. Furthermore, as the number of fixed prefix convolution layers increases, the data transfer performance imporves, until a certain critical point is reached; thereafter, it gradually worsens, resembling a quadratic curve. The transfer performance at this critical point is what we hope to see and is also the basis for the subsequent experiments.

B. Feature Fusion and Classification

The dataset contains both conventional ultrasound and ultrasound elasticity images; thus, we should combine the data from these two modalities to achieve better classification performance. Ultrasound data based on different imaging principles have different feature distributions and different effects on the classification of thyroid nodules. This combination of features will produce a more comprehensive feature space for representing the pathological characteristics of the nodules.

Accordingly, we use the following three methods, as shown in Figure 4:

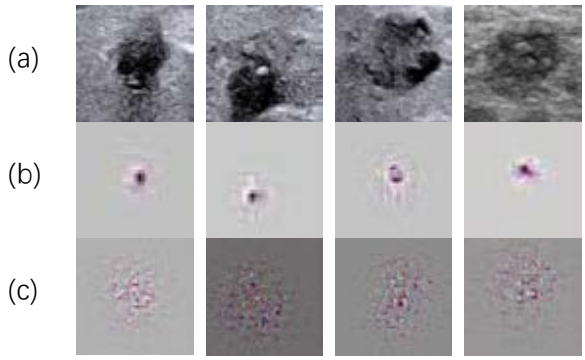

Fig. 3. Feature visualization. (a) Input image. (b) Layer 2 feature map project. (c) Layer 7 feature map project. For layer 2 and 7, the top 1 activations in a validation dataset are shonw, projected down to the pixel space using a deconvolutional network.

1) Mixed training fine-tuning

The first is the most intuitive idea. As images by different imaging principles have different feature distributions, we hope that the network can learn the common features of two different feature distributions. Accordingly, we directly mix conventional ultrasound and ultrasound elasticity images into one dataset for training. As shown in Figure 4.a, we should first preprocess the original ultrasound data, then extract the conventional ultrasound and ultrasound elasticity images, and finally perform the related data augmentation to ensure better generalization ability. Gaussian random sampling is used for data loading for each batch as the input of the pre-trained VGG16 model. The model outputs 4096 dimensions and classifies them in the output vector.

2) Fusion followed by feature re-extraction

For mixed training, the feature representations learned by the network may be wavering between two different distributions. From a clinical point of view, for the same original ultrasound image, its conventional and elastography parts should be treated as a sample. Their feature information is complementary. Therefore, we should fuse two different sets of data and integrate them into the network. As shown in Figure 4.c.2, we combine the 3-channel conventional ultrasound and

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

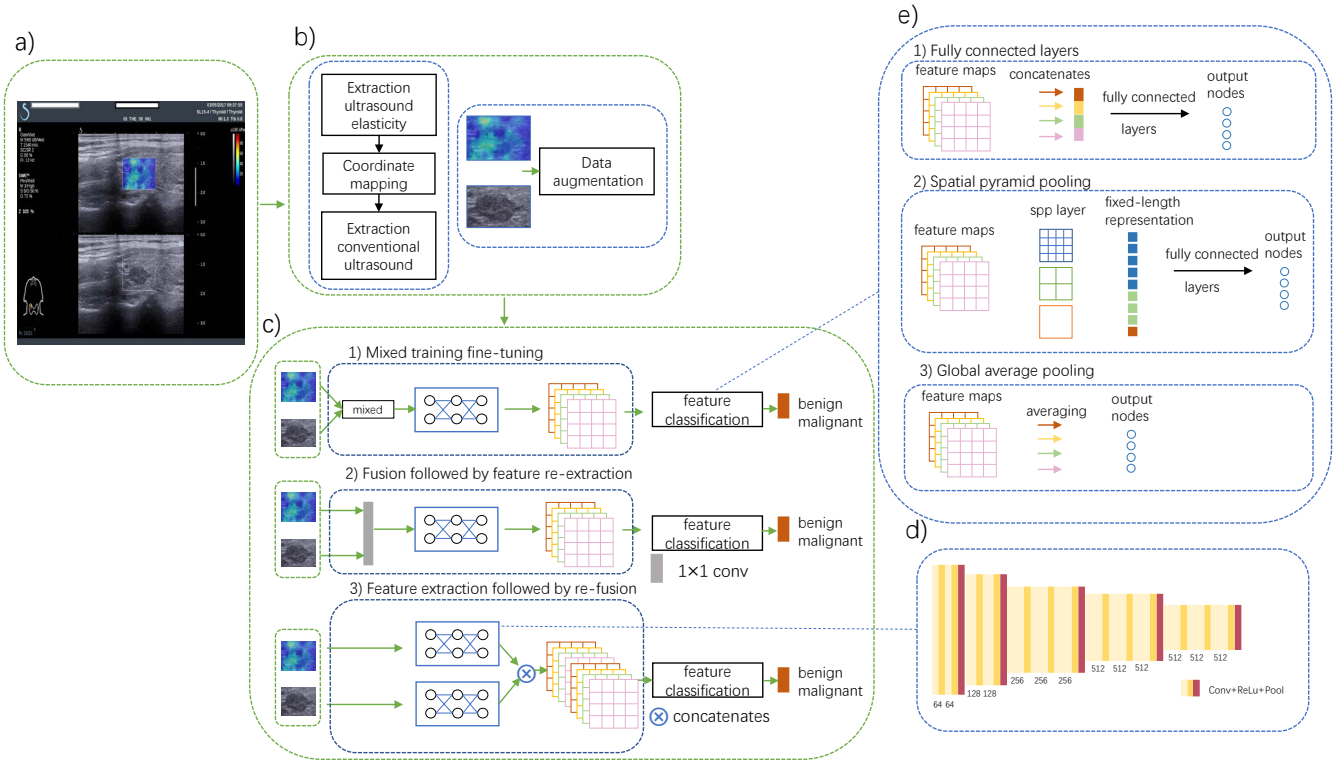

Fig. 4. Overview of the proposed method. (a) Raw input data. (b) Data preprocessing and data augmentation. (c) Fusion methods for two types of data. Orange boxes represent the nodule classification output. (d) Basic network architecture (VGG16). (e) Feature classification methods.

the 3-channel ultrasound elasticity images into a 6-channel tensor. The first three channels of the tensor are conventional ultrasound data information and the last three channels are ultrasound elasticity data, thus forming a 224×224×6 tensor. Then, we use 1×1 convolution for channel dimension reduction and convert the tensor to a 224×224×3 tensor, which is the input to the pre-trained VGG16 model. The output is a 4096-dimensional vector, which is to be classified in the final step. This is similar to the early fusion method in [35], where the entire frame information of a complete sliding window is input, whereas in the proposed method (fusion followed by feature re-extraction), two different data sources are fused to obtain better and more complete features before feature re-extraction.

However, the input is not a single conventional ultrasound or ultrasound elasticity image. Thus, fine-tuning in the six layers is no longer applicable. Therefore, we re-fine-tuning layer-by-layer in the fusion followed by feature re-extraction structure to select the optimal level. The experimental results are shown in Table V. The result is the same as that by choosing the six layers.

3) Feature extraction followed by re-fusion

Considering that different ultrasound images have different feature distributions, the prerequisites for information complementarity are independent of one another. Therefore, directly mixing two different data sources may result in an inconspicuous feature representation. Moreover, as shown in Figure 4.c.3, we independently sample the conventional ultrasound and the ultrasound elasticity image, and input two different sets of data

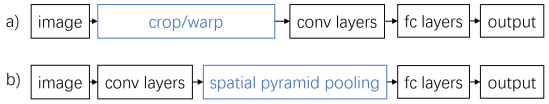

Fig. 5. (a) Hierarchical CNN structure of CNN using the crop or warp method. (b) The spatial pyramid pooling layer is added before the convolutional layer and the first FC layer.

into two independent pre-trained VGG16 models for feature extraction. After the fifth pooling layer, the features extracted by the two independent feature extractors are cascaded, so that the output feature dimension of this model is twice that of the ordinary VGG16 model. The feature dimension is reduced to 4096 by using the FC layer, and classification is performed in the last step. This can ensure the independence of different feature extraction processes for different data sources. According to the chain rule, the cascade operation of the final feature blending also makes different CNN parameter updates independent of different data sources when the gradient is transferred by back propagation.

This method is similar to the late fusion method in [35] and [36]. The former uses different front and back frames for feature extraction and final fusion, which is based on information fusion in the time dimension, whereas the latter extracts features from different 2D planes of the same 3D data, which is based on the spatial dimension. The proposed approach is based on the fusion of different modal dimensions.

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

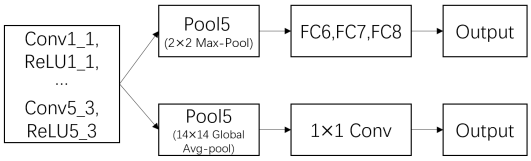

Fig. 6. Network architectures of VGG-16 (above) and VGG-16-GAP (below).

	Total Cases	Samples	Positive Cases	Samples	Negative Cases	Samples
Total	233	1156	126	617	107	539
Training	183	908	99	484	84	424
Testing	40	248	27	133	23	115

TABLE IV
DATA DISTRIBUTION FOR TRAINING AND TEST SETS

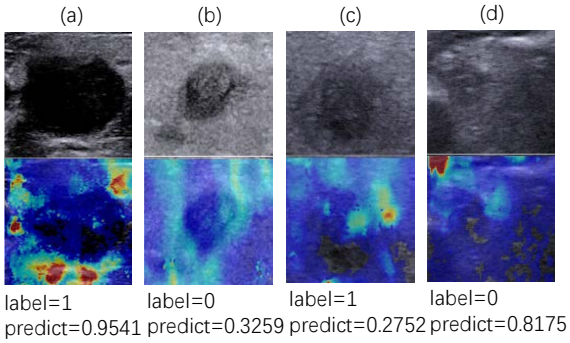

Fig. 7. Visualization of inference results (a) malignant; (b) benign; (c) false negative; (d) false positive;

C. Input Image Shape

In the above method, the nodules of different samples have different sizes, but we are forced to reshape the input image to 224×224 for processing. This distorts the spatial resolution, size, and shape of the structures, and will affect performance. In traditional CNNs, the size of the input image is severely limited owing to the FC layer. Normally, we should perform a crop or warp operation on the original image to resize it so that it may fit in the CNN. However, the cropped image may not contain all the required information, and the deformation operation that changes the aspect ratio may also cause undesired deformation of key parts. The loss or distortion of the image content may greatly affect the accuracy of the model. In fact, the FC layer is the key factor that restricts the input size. As the convolution and pooling layers ignore input size, they only take the feature map of the previous layer as input and then output the convolution or pooling result to the next layer. However, as the weight dimension is not fixed, only the FC layer can not be changed, so that the layers look back. This will lead to the condition that all the dimensions should be fixed. We used the following two structures to resolve this problem separately.

1) Spatial pyramid pooling

The convolutional layers of a CNN (and related layers, such as pooling and local response normalization) are able to process variable-size input. Therefore, the problem of variable-size input propagates down to the first FC/inner product layer, which requires a vector of fixed size. He et al. proposed adding a spatial pyramid pooling layer immediately before the first FC layer. This layer operates by hierarchically partitioning the feature maps of the last convolutional layer (or a subsequent pooling or response normalization layer) into a fixed number of bins. Within these bins, responses are pooled in the usual manner, generating a fixed-size output (where the size depends on the hierarchy and number of bins) [37].

Therefore, as shown in Figure 5, we add a spatial pyramid pooling layer before the first FC layer. When the input is an image of any size, we can arbitrarily perform convolution and pooling operations before the FC layer, and abstract the image into fixed-size features through the SPP layer (fixed feature vector extraction under multi-scale features).

2) Global average pooling

The FC layer is one of the most fundamental modules in CNN. It is widely used in traditional CNN models. However, it is known that FC may cause overfitting and requires millions of parameters. As the FC layer is the key factor that limits the size of the input, it is natural to ask whether its removal may be beneficial. The global average pooling strategy has been proposed to replace the FC layers [38]. The global average pooling layer has no parameters, summarizes the spatial information using an average, and can be seen as a regularizer. GoogLeNet and ResNet are two typical examples of CNN without FC layers. Both GoogLeNet and ResNet use the global average pooling layer to replace the FC layers, and have achieved the best results in the ImageNet competition in 2014 and 2015, respectively.

Experiments are conducted on VGG models with and without FC based on two architectures. Thus we have two CNN models from the source domain, i.e., the ImageNet data. In addition, to obtain a VGG model without FC from the source domain, we replace the pool5 layer with a global average pooling layer and remove all subsequent FC layers in the original VGG-16. Then, a 1×1×1000 convolutional layer is added to output the predicted results. The modified model is called as VGG-16-GAP. Then, the ImageNet datasets are used to fine-tuning this model until convergence is achieved. The architectures of VGG-16 and VGG-16-GAP are shown in Figure 6.

III. EXPERIMENT

A. Data and Evalution Indexes

The thyroid ultrasound data used in the experiments were provided by Huiying Medical Technology (Beijing) Co, Ltd, and were verified by clinical pathology. The experimental data were obtained from Aixplorer's ultrasonic machine, and the frequency of the detector was 10–14MHz. 1156 thyroid nodule ultrasound images in 233 patients were used in the experiments, including 578 cross-sectional images and 578 longitudinal images, that is, 539 benign images and 617 malignant images in total. As there are multiple images in the same patient collection, the division of the training validation set is at the patient level. Each image consist of conventional

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

Freeze layers	Accuracy%	Sensitivity%	Specificity%	AUC%
0	75.86 ± 1.77	75.00 ± 1.45	76.83 ± 1.60	85.17 ± 1.54
1	77.01 ± 0.63	80.41 ± 0.26	72.73 ± 0.94	85.15 ± 0.62
2	78.74 ± 1.35	82.35 ± 1.57	73.61 ± 1.27	86.61 ± 1.24
3	79.31 ± 1.64	82.76 ± 1.18	75.86 ± 1.22	87.79 ± 1.13
4	81.03 ± 0.70	75.82 ± 0.98	86.75 ± 0.82	91.07 ± 0.94
5	83.91 ± 0.60	82.98 ± 0.85	85.00 ± 0.46	90.59 ± 0.89
6	**86.21 ± 0.84**	**90.10 ± 0.91**	**80.82 ± 0.62**	**92.50 ± 0.94**
7	82.76 ± 0.74	81.11 ± 0.77	84.52 ± 0.82	88.89 ± 1.03
8	84.48 ± 0.61	81.25 ± 0.43	87.23 ± 0.77	89.19 ± 0.26
9	81.61 ± 0.64	82.98 ± 0.54	80.00 ± 0.33	87.11 ± 0.73
10	77.59 ± 0.33	73.53 ± 0.76	83.33 ± 0.20	85.36 ± 0.94
11	77.01 ± 0.91	77.32 ± 0.69	76.62 ± 0.76	84.34 ± 0.63
12	75.86 ± 1.14	76.24 ± 1.35	75.34 ± 0.98	83.66 ± 0.66
13	73.56 ± 1.30	74.23 ± 1.62	72.73 ± 0.94	79.49 ± 1.38

TABLE V
COMPARISON OF FINE-TUNING BY FUSION RE-EXTRACTION FEATURE METHOD

Method	Accuracy%	Sensitivity%	Specificity%	AUC%
Conventional ultrasound fine-tuning	82.76 ± 0.74	86.14 ± 0.62	78.08 ± 0.94	89.61 ± 0.58
Ultrasonic elastography fine-tuning	85.06 ± 0.66	82.14 ± 0.38	87.78 ± 0.51	91.75 ± 0.56
Mixed training fine-tuning	85.30 ± 0.97	83.05 ± 0.71	87.65 ± 0.46	90.95 ± 0.91
Fusion followed by feature re-extraction	86.21 ± 0.84	90.10 ± 0.91	80.82 ± 0.62	92.50 ± 0.94
Feature extraction followed by re-fusion	**92.42 ± 0.63**	**92.31 ± 0.81**	**92.54 ± 0.81**	**98.07 ± 0.29**
Conventional ultrasound fine-tuning with SPP	87.93 ± 0.88	91.30 ± 0.55	84.15 ± 0.56	92.84 ± 0.51
Ultrasonic elastography fine-tuning with SPP	87.36 ± 0.80	86.17 ± 0.93	88.75 ± 0.75	94.28 ± 0.44
Mixed training fine-tuning with SPP	87.90 ± 0.42	85.08 ± 1.17	90.96 ± 0.91	93.55 ± 1.03
Fusion followed by feature re-extraction with SPP	89.08 ± 0.76	91.21 ± 1.13	86.75 ± 1.39	93.13 ± 1.12
Feature extraction followed by re-fusion with SPP	**94.70 ± 0.53**	**92.77 ± 1.04**	**97.96 ± 1.13**	**98.77 ± 1.05**
Conventional ultrasound fine-tuning with GAP	86.20 ± 0.76	86.52 ± 0.57	85.88 ± 1.09	92.02 ± 2.01
Ultrasonic elastography fine-tuning with GAP	84.48 ± 0.89	83.48 ± 1.09	86.44 ± 0.47	92.50 ± 1.25
Mixed training fine-tuning with GAP	86.74 ± 1.39	87.36 ± 1.40	86.13 ± 0.86	92.93 ± 1.98
Fusion followed by feature re-extraction with GAP	87.93 ± 0.91	88.76 ± 0.98	87.06 ± 0.90	94.06 ± 1.17
Feature extraction followed by re-fusion with GAP	**93.94 ± 0.97**	**93.24 ± 0.88**	**93.10 ± 0.94**	**98.60 ± 1.11**

TABLE VI
COMPARISON OF DIFFERENT METHODS USING FOUR EVALUATION INDEXES

ultrasound and ultrasound elasticity data. The nodule type is marked by a radiologist, and there is no annotation of the contour information. The detailed distribution of the experimental data is shown in Table IV, where the "Cases" column in the Table IV indicates the number of patients, and the "Samples" column indicates the number of images.

We pre-extracted the ROI area of each nodule image by a color channel transformation. Moreover, we separately extracted the conventional ultrasound and ultrasound elasticity data parts of each image at the preprocessing stage according to the radiologist's annotation information.

The evaluation indexes are as follows:

(1) Accuracy:=$(TP + TN)/(TP + TN + FP + FN)$;

(2) Sensitivity:=$TP/(TP + FN)$;

(3) Specificity:=$TN/(TN + FP)$;

(4) Receiver operating characteristic curve (ROC) and AUC;

TP (true positive) and TN (true negative) denote the number of positive and negative samples, respectively, that were correctly classified. FP (false positive) and FN (false negative) denote the number of negative and positive samples, respectively, that were misclassified. In the classification of thyroid nodule samples, negative samples correspond to benign nodules.

Therefore, sensitivity and specificity determine the possibility of predicting malignant and benign nodules, respectively.

B. Results and Discussion

We first performed fine-tuning of the pre-training model under conventional ultrasound and ultrasound elasticity data. For conventional ultrasound and ultrasound elasticity datasets, it is best to freeze the parameters of the first six conventional layers convolution for a feature extraction to subsequent layers.

Accordingly, we performed experimental comparisons on three different structures: original, SPP, and GAP structure. For each structure, we compared five methods: conventional ultrasound transfer, ultrasonic elastography transfer, mixed training transfer, fusion followed by feature re-extraction and feature extraction followed by re-fusion. For each method, we performed 10-fold cross verification involving sample partitioning into training and verification sets. The details of all evaluation indexes are shown in Table VI.

First, we compare original structure. In the comparison of single data sources, it can be found that the accuracy on ultrasound elasticity data is higher than that on conventional ultrasound data. The sensitivity index for conventional ultrasound data is better, and the discriminative performance on malignant nodules is higher; the specificity index for ultrasound elasticity data is better, and the discriminative performance on benign nodules is higher.

Compared with three different data fusion methods, mixed training transfer and fusion followed by feature re-extraction exhibit a slight improvement in terms of accuracy and AUC,

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

but the performance improvement is not evident. Essentially, after different data mixing and fusion, the extractive performance for key features is not obvious for different feature distributions.

In the data mixing method, the data mixing training of different imaging principles will affect the feature extractor's judgment on the feature distribution of the target data; in the fusion followed by feature re-extraction method, oweing to the limitation in the input parameters of the basic network, channel reduction of the data should be performed. The use of dimension reduction for the channel of the 1×1 convolution is essentially equivalent to pixel-level superposition of the data from two different sources with different weights. This also affects the independence of the distribution of different data features, and does not essentially change the feature space representation to improve model performance. In the feature extraction followed by re-fusion method, compared with the previous methods, the performance indexes are obviously improved. This demonstrates that combining feature extraction feature cascading is more appropriate in this problem, as it dose not affect the feature distribution of each data source.

In the SPP structure, we resolve the problem that the network can receive only fixed-size input; thus, we alleviate the performance degradation caused by the necessity for image reshaping. Therefore, the indicators of each method are significantly improved.

The redundancy of network parameters has been extensively discussed, and the parameters of the FC layer account for 80% of the network parameters. FC layers have serious disadvantages: They are prone to overfitting, do not easily attain convergence during training, and hamper the generalization ability. Some recent high-performance networks, such as ResNet and GoogleNet use global average pooling instead of FC to fuse the deep features learned by convolutional layers, and finally use softmax and other loss functions as the objective function to guide the learning process. When we further remove the FC layer and want to improve the classification performance by reducing the number of model parameters, we find that the performance indicators of each method under the GAP structure are lower than those under the SPP structure. Despite the obtained results, we believe that in the case of large differences between the source and target domains, the FC layer is irreplaceable in fine-tuning. Global average pooling reduces the number of network parameters and improves prediction efficiency, but it is effective in transfer-learning fine-tuning. Zhang et al. [39] concluded that when the target domain is not far from the source domain, FC layers can be replaced by global average pooling for better efficiency and accuracy. If there is a large difference in either the image properties or task objective, FC layers are essential in visual representation transfer. Therefore, FC can be regarded as a "firewall" of model representation capability. Particularly, if the source and target domain are different, FC can maintain a large model capacity to ensure the transfer of model representation capabilities. (Redundant parameters are not worthless.)

In the case of unbalanced datasets, AUC can convincingly reflect the performance of the classifier, and higher AUC value for the extraction feature re-fusion demonstrates that this method has better classification performance.

Figure 7 is a visualization of the different inference results of the proposed method. It can be seen that the difference in the quality of ultrasound images will cause discriminant model error, as in, Figures 7.c and 7.d.

Table VII shows a comparison of several state-of-the-art methods with the proposed method in classifying benign and malignant thyroid nodule images.

IV. Conclusion

In this study, we proposed a method for feature extraction and fusion of conventional ultrasound and ultrasound elasticity images for the differentiation of benign and malignant thyroid nodules. Considering the clinical practicality of ultrasound imaging, we used two different data sources, that is, conventional ultrasound and ultrasound elasticity images, to distinguish benign and malignant thyroid nodules. The feature extraction method also ensures the independence of different feature extractors and the data distribution classification of the target data domain, higher sensitivity for conventional ultrasound images, higher specificity for ultrasound elasticity images, and better performance improvement. Furthermore, only end-to-end implementation of high-level features is used; this results in higher computational efficiency in the training and inference stages, does not require fine labeling information, and is more widely used in clinical practical.

The proposed approach achieved a 94.70% accuracy, and the comparison with other single data-source methods also indicated an more obvious advantage. In future work, we plan to optimize the basic network, attempt to use other more efficient network structures, and introduce low-level features for further improvement of the classification performance. In addition, given the large number of medical imaging data sources, we will attempt to extend this approach to other medical tasks for clinical application.

References

[1] M. Schlumberger, M. Tahara, L. J. Wirth, B. Robinson, M. S. Brose, R. Elisei, M. A. Habra, K. Newbold, M. H. Shah, A. O. Hoff, *et al.*, "Lenvatinib versus placebo in radioiodine-refractory thyroid cancer," *New England Journal of Medicine*, vol. 372, no. 7, pp. 621–630, 2015.

[2] M. Owjimehr, H. Danyali, and M. S. Helfroush, "Fully automatic segmentation and classification of liver ultrasound images using completed lbp texture features," in *22nd Iranian Conference on Electrical Engineering*, pp. 1956–1960, 2014.

[3] F. S. Zakeri, H. Behnam, and N. Ahmadinejad, "Classification of benign and malignant breast masses based on shape and texture features in sonography images," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1621–1627, 2012.

[4] S. Luo, E.-H. Kim, M. Dighe, and Y. Kim, "Thyroid nodule classification using ultrasound elastography via linear discriminant analysis," *Ultrasonics*, vol. 51, no. 4, pp. 425–431, 2011.

[5] J. Ding, H.-D. Cheng, J. Huang, and Y. Zhang, "Multiple-instance learning with global and local features for thyroid ultrasound image classification," in *7th International Conference on Biomedical Engineering and Informatics*, pp. 66–70, 2014.

[6] U. Raghavendra, U. R. Acharya, A. Gudigar, J. H. Tan, H. Fujita, Y. Hagiwara, F. Molinari, P. Kongmebhol, and K. H. Ng, "Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions," *Ultrasonics*, vol. 77, pp. 110–120, 2017.

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2950994, IEEE Journal of Biomedical and Health Informatics

Method	Testing Sample			Accuracy%	Sensitivity%	Sensitivity%	AUC%
	Total	Positive	Negative				
[12]	-	-	-	84.74	92.31	76.00	91.03
[15]	3734	-	-	89.80	93.40	86.10	94.70
[16]	400	200	200	88.25	90.00	86.50	92.86
[19]	103	-	-	93.10	90.80	94.50	97.70
[20]	240	200	40	93.75	93.96	92.68	-
Proposed	248	133	115	**94.70 $\pm$ 0.53**	**92.77 $\pm$ 1.04**	**97.96 $\pm$ 1.13**	**98.77 $\pm$ 1.05**

TABLE VII
COMPARISON OF DIFFERENT METHODS IN CLASSIFYING BENIGN AND MALIGNANT THYROID NODULE IMAGES

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[10] J. Hu, R. Ji, S. Zhang, X. Sun, Q. Ye, C.-W. Lin, and Q. Tian, "Information competing process for learning diversified representations," *arXiv preprint arXiv:1906.01288*, 2019.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *International joint conference on neural networks*, pp. 2560–2567, 2016.

[12] H. Wu, Z. Deng, B. Zhang, Q. Liu, and J. Chen, "Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography," *American Journal of Roentgenology*, vol. 207, no. 4, pp. 859–864, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[15] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, *et al.*, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study," *The Lancet Oncology*, vol. 20, no. 2, pp. 193–201, 2019.

[16] J. Wang, S. Li, W. Song, H. Qin, B. Zhang, and A. Hao, "Learning from weakly-labeled clinical data for automatic thyroid nodule classification in ultrasound images," in *25th IEEE International Conference on Image Processing*, pp. 3114–3118, 2018.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

[19] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 919–923, 2017.

[20] Y. Zhu, Z. Fu, and J. Fei, "An image augmentation method using convolutional network for thyroid nodule classification by transfer learning," in *3rd IEEE International Conference on Computer and Communications*, pp. 1819–1823, 2017.

[21] J. A. Sipos, "Advances in ultrasound for the diagnosis and management of thyroid cancer," *Thyroid*, vol. 19, no. 12, pp. 1363–1372, 2009.

[22] F. Sebag, J. Vaillant-Lombard, J. Berbis, V. Griset, J. Henry, P. Petit, and C. Oliver, "Shear wave elastography: a new ultrasound imaging mode for the differential diagnosis of benign and malignant thyroid nodules," *The Journal of Clinical Endocrinology & Metabolism*, vol. 95, no. 12, pp. 5281–5288, 2010.

[23] J.-B. Veyrieres, F. Albarel, J. V. Lombard, J. Berbis, F. Sebag, C. Oliver, and P. Petit, "A threshold value in shear wave elastography to rule out malignant thyroid nodules: a reality?," *European journal of radiology*, vol. 81, no. 12, pp. 3965–3972, 2012.

[24] A. Dizeux, T. Payen, G. Barrois, C. Baldini, D. L. G. Buffelo, E. Comperat, J.-L. Gennisson, M. Tanter, and S. L. Bridal, "Complementarity of shear wave elastography and dynamic contrast-enhanced ultrasound to discriminate tumor modifications during antiangiogenic and cytotoxic therapy," in *IEEE International Ultrasonics Symposium*, pp. 1144–1147, 2014.

[25] H. Gharib, E. Papini, J. R. Garber, D. S. Duick, R. M. Harrell, L. Hegedüs, R. Paschke, R. Valcavi, and P. Vitti, "American association of clinical endocrinologists, american college of endocrinology, and associazione medici endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules–2016 update," *Endocrine practice*, vol. 22, no. s1, pp. 1–60, 2016.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[27] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[28] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," *arXiv preprint arXiv:1811.08883*, 2018.

[29] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[30] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE journal of biomedical and health informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[34] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, 2014.

[35] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[36] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[38] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[39] C.-L. Zhang, J.-H. Luo, X.-S. Wei, and J. Wu, "In defense of fully connected layers in visual representation transfer," in *Pacific Rim Conference on Multimedia*, pp. 807–817, 2017.

2168-2194 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.