WSAN: An Effective Model of Weakly Supervised Similarity Analysis Network for the Lung CT Images

Yi Zhuang 1, Member, IEEE, Nan Jiang 2, Shuai Chen 1
1 School of Computer & Information Engineering, Zhejiang Gongshang University, Hangzhou, P.R.China
2 Affiliated Hangzhou First People’s Hospital, Zhejiang University School of Medicine, Hangzhou, P.R.China

Corresponding author: Yi Zhuang (e-mail: zy158cn@gmail.com)

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant No.Y22F021788; the Zhejiang Public Welfare Technology Application Research Project under grant No.GF22H185665; the Zhejiang Medical and Health Research Project under grant No. 2019RC070.

ABSTRACT With the rapid development of medical imaging technologies, the high-resolution CT image data is of great value for medical research as well as a clinical diagnosis. The paper takes lung CT image as an example. Retrieving images similar to the input one is helpful to assist physicians in clinical diagnosis. Compared with the traditional content-based image retrieval, the similarity retrieval of lung CT images requires higher retrieval accuracy, with similar requirements in external shape as well as internal vascular and lesion location similarity. In the state-of-the-art supervised deep learning networks, the learning of the network relies on labeling. The labeling of medical images, however, requires time and effort for the professionals to label each image, which is too costly. In this paper, we propose a weakly supervised deep learning network model for similarity analysis of lung CT images that is called a Weakly Supervised similarity Analysis Network (WSAN). Extensive experiments show that the WSAN model achieves satisfactory results in measuring the similarity between lung CT images and can be used for similarity retrieval tasks.

INDEX TERMS content-based retrieval; CT image; deep learning; similarity retrieval,

I. INTRODUCTION

With the explosive growth of the number of medical images, as one of the important types of medical images, high-resolution lung CT images have exploded in number. Retrieving similar lung CT images to an input one from the large-scale lung CT image databases enables physicians to efficiently refer to previous similar cases, so as to achieve accurate computer-aided diagnosis and treatment.

Generally, a lung CT image sequence is composed of hundreds of lung images that represent different sections of a patient's lungs with temporal information. The lung CT images not only contain information about the images themselves, but also indicate patient-specific treatment measures and treatment outcomes to greatly assist physicians in treating the current clinical cases. Through similarity matching, physicians will find similar images from different patients who have a high chance of causing the same disease because they contain similar pathological features.

Different from the content-based similarity retrieval of ordinary images, the medical image (e.g., CT image) similarity retrieval requires higher retrieval accuracy. As a whole, CT images of lungs are similar from one person to another. Basically, all images contain blood vessels, bones and soft tissues (e.g. thorax, trachea, bronchi, etc.) inside the left and right lung lobes. What differs from patient to patient is the shape of the lung lobes and the information about the bronchi, blood vessels, and nodules inside the lungs. All of which differ at the pixel level. Therefore, high precision similarity retrieval of the lung CT images is generally required. In addition, the objects inside the lung lobes have complex characteristics such as location and shape. So they are difficult to describe and quantify. Meanwhile, deep learning-based similarity assessment usually requires a large amount of data and labels. Such labels require medical experts to manually label the images, which is a time-consuming, laborious and expensive task, especially in the medical field. Content-based medical image retrieval usually combines feature descriptors extracted by deep learning networks with traditional mathematical methods, then calculates their
similarity using distance formula. This retrieval method, however, requires a large number of labels. The enormous number of image descriptors consumes a large amount of storage space.

To tackle the above challenges, the paper proposes a deep learning network model architecture called the Weakly Supervised similarity Analysis Network (WSAN). The WSAN model consists of two layers of deep learning networks: 1) the first layer is used to compute the similarity of contour information and add a spatial transformation layer before training; 2) the second layer is responsible for calculating the similarity of the details inside the lung lobe. In terms of the network training, the addition of the spatial transformation layer allows us to generate a dataset for training the contour similarity calculator. The labeling method in this dataset, however, is not perfect since it only tells the deep learning network which images are similar to each other in terms of the contour shape, which belongs to the inexact supervised learning method [1]. To solve the problem, after training a contour similarity calculator with this dataset, we further generate a training set with higher labeling accuracy using contour similarity calculator, i.e., labels with accuracy categories focusing on the internal structure of the lung lobes. With this high-precision dataset, we can train a set of deep learning networks (i.e., WSAN) for discriminating the similarity of lung CT images with high accuracies. The WSAN model is testified to be effective for similarity analysis of lung CT images and can be used for lung CT image similarity retrieval.

The main contributions of this paper include the following three aspects.
1. We propose a new weakly supervised deep learning network model called the WSAN, for lung CT image similarity measurement.
2. We present a new automatic labeling method for similarity labeling between medical images.
3. We perform extensive experiments to demonstrate the effectiveness and efficiency of our proposed WSAN model in lung CT image similarity measurement.

The remainder of the paper is organized as follows: Section II provides the background. In Section III, a WSAN-based lung CT image similarity analysis method is proposed. Section IV discusses the training details. Extensive experiments are performed in Section V. Section VI concludes the paper.

II. BACKGROUND

The content-based medical image retrieval has been extensively studied which involves the feature extraction, similarity measurement and image ranking. The key lies in the feature extraction and the use of feature vectors to represent each image. The study of feature extraction has gone through two stages.

In the first stage, the traditional feature extraction algorithm initially focused on global feature extraction, such as color, texture, shape, and other features of images. Since global features are easily affected by external noise, the research focus began to shift to local features. The Idiap research team [3] coupled LBP and modSIFT [4]. The two descriptors obtained an error of 178.93 on the IRMA dataset. Mizotin et al. [1] proposed a brain magnetic resonance image (MRI) retrieval method based on SIFT features of visual bag-of-words (BoVWs) for the diagnosis of Alzheimer’s disease [3]. Recently, Pan et al. [6] used an Uncertain Location Graph (ULG) structure to model the brain CT images, then the similarity computation of the brain CT images was transformed to the similarity measurement between the ULG structures. Using this structure, they obtained an accuracy of up to 80%.

In the second stage, semantic feature extraction algorithms, i.e., deep learning methods are used to bridge the gap between traditional feature extraction algorithms and advanced semantic features of human vision. Calculating the similarity between such advanced features can lead to more accurate retrieval results. Shin et al. [7] applied the transfer learning ideas to fine-tune CNN models pre-trained on ImageNet datasets for the interstitial lung disease dataset (ILD) and thoracoabdominal lymph node (LN) dataset. Demonstrating the value of converting non-medical images to medical images in computer-aided detection (CADe) through CNN activation visualization, Sundararajan et al. [8] proposed a method to retrieve avascular necrosis-free (AN) images using deep bilinear convolutional network (DB-CNN) feature representation. After preprocessing the input dataset using median filter (MF) to eliminate image noise, DB-CNN was then used to obtain image features in binary code, preferably using modified Hamming distance for similarity measure. Khatami et al. [9] first predicted the most probable class of the retrieved image by CNN network, and then applied it in the search space consisting of this class Radon transform for further retrieval. After that, Khatami et al. [10] proposed two more retrieval schemes. The first one is to expand the original dataset by two different data expansion methods to obtain two augmented datasets, one of which is used for CNN pre-training and the other for fine-tuning. This network will output a search space with only one category, and then use Radon orthogonal projection; the second one uses 3 different networks for parallel retrieval, and then refines the retrieval results of multiple networks using LBP, HOG, and Radon transformation features. The space is further compressed by the construction, and finally, the LBP features are used for refined retrieval. Karthik et al. [11] proposed a hybrid
feature model to represent a hybrid feature vector of shape and texture properties, in which the four different feature extraction methods (e.g., GLSM, HCSD, HOG, and HBF) are adopted to optimal represent of image information. Liu et al. [17] designed a deep supervised hashing approach for fast image retrieval. Ma et al. [18] proposed a new method of content based medical image retrieval and its applications to CT imaging sign retrieval. Cai et al. [19] presented a medical image retrieval method based on convolutional neural network and supervised hashing. Sampathila et al. [20] proposed a computational approach for content-based image retrieval of K-similar images from brain MR image database.

III. THE WSAN MODEL

In this section, we present an overall framework of the WSAN model for lung CT image similarity measurement.

A. MOTIVATIONS & OVERALL ARCHITECTURE

The motivations of the WSAN model are based on the following analysis: 1) a considerable part of lung diseases can be found from the contour shape of the lung CT images. For example, in the CT image of lobar pneumonia, there will be large areas of high-density shadows in the lung lobes, and it looks as if the lung lobe is divided into two parts; in CT images of atelectasis, the lobes of the lungs will atrophy and collapse; 2) Some lung diseases are discovered by simultaneously observing the contour shape and internal details of the lungs. For example, the basic CT manifestations of pneumothorax are extremely low-density gas shadows in the pleural cavity, accompanied by varying degrees of compression and shrinkage of the lung tissue. The rest of the lung diseases are only discovered by observing the details of the interior of the lung lobes: for example, the basic CT appearance of pulmonary nodules is that there will be small nodules inside the lung lobes. Based on the above analysis, both contour similarity and detail similarity are very important for the similarity retrieval of the lung CT images.

Fig. 1 illustrates the architecture of the WSAN model which includes two main modules: 1) the dataset generator and 2) the similarity calculator.

The dataset generator contains the spatial transformation layer, which is mainly used to generate the initial training set for network training. Based on that the subsequent network training tends to be inexact supervision, i.e., the training samples have only coarse-grained labels.

The similarity calculator contains not only the contour similarity calculation but the detail similarity calculation, which is the core part of the WSAN used for network training. The WSAN model is trained and tested on the LUNA16 dataset [21].

B. DATA GENERATOR

- Spatial Transformation Generator

As the prototype of Spatial Transformation Layer (STL), the Spatial Transformation Network (STN) is originally proposed by the DeepMind [12]. The aim of the STN is to analyze the objects in different scenes in computer vision, where the model can be invariant to changes in object pose or position. In this paper, the STN is simplified by removing the localization net module for learning, and the STL is mainly used to construct images similar to the lung CT images in the original dataset for subsequent network training. So the STL is the key to reflect the weak supervisory nature of this network structure. In this system, the STL is mainly responsible for providing a Thin Plate Spline (TPS) based spatial transformation for each input [13]. The internal architecture of the STL is shown in Fig. 2.

In Fig. 2, $U$ is an input image matrix (512*512*1), $V$ is an output image matrix (512*512*1), and each spatial transformation layer consists of a grid generator and sampler, $\theta$ is a randomly generated 25*2 tensor, which enters grid generator after normalization. The $G$ in Grid Generator is a duplication of $U$, $\tau_\theta(G)$ represents a mapping relationship. The goal of the grid generator is to get the corresponding value of each pixel point of the output feature map. Then enter the sampler to insert the corresponding point into the new matrix $U$ using the thin-slab sample interpolation transform. Finally get the output matrix $V$ after spatial transformation.

C. SIMILARITY CALCULATOR

In this subsection, we study the similarity calculator which consists of: 1) contour similarity calculator, and 2) detail similarity calculator.

1) CONTOUR SIMILARITY CALCULATOR

- Problem Formulation
The contour similarity calculator is used to calculate the similarity of the contour shapes of the lungs in the two input lung CT images without taking into account the details of the interior of the lungs.

**Definition 1.** Given two lung CT images: $I_1$ and $I_2$, the similarity between them can be represented by:

$$\text{sim}_p = S(I_1, I_2)$$

where $S(I_1, I_2)$ means the contour similarity function.

Based on Definition 1, the function $S(I_1, I_2)$ is implemented by a deep learning network to be introduced in this subsection. The scale size of both $I_1$ and $I_2$ is $512 \times 512 \times 1$, $\text{sim}_p$ denotes the contour similarity between $I_1$ and $I_2$.

Due to the continuity and diversity of lung contour shapes of lung CT images, there is a contextual linkage between different patches of the two images. Given an input image $I_1$, the candidate image $I_2$, Partition $I_1$ and $I_2$ into several image patches, respectively. Then we have $I_1 = \{i_11, i_{12}, ..., i_{1k}\}$, $I_2 = \{i_{21}, i_{22}, ..., i_{2k}\}$, where $k$ is the number of image patches. The image patch $i_{1j}$ is tiled and expanded to form a vector $x_{1j}$ that is synthesized into an image patch vector $x_j = (x_{1j}, x_{2j})$ with the vector $x_{2j}$, which formed from the image patch $i_{2j}$.

**Definition 2.** There is a contextual linkage $(x_1, x_2, ..., x_k)$ between image patch vectors, which is represented in Eq.(2):

$$C = \alpha_1^T x_1 + \alpha_2^T x_2 + ... + \alpha_k^T x_k$$

where $C$ is a constant, $\alpha_i$ is the weight coefficient of the vector, and $x_j$ is a $k$-dimensional column vector.

**Network Design**

According to the contextual linkage of different patches of the lung CT images in Definition 2, and some image patches are important and others are secondary. The weight coefficients of some vectors may be significantly larger than those of other vectors. Based on these facts, a self-attentive mechanism [14] is required to filter out a small amount of important information from a large amount of information in the contour similarity calculation. Based on the self-attentive mechanism [15], as illustrated in Fig. 3, the Vision Transformer (ViT) is used as a network for the contour similarity calculator. In the self-attentive mechanism, the lung lobe contour part which changes significantly is given more weight. In other words, the ViT pays more attention to the lung lobe contour part.

2) DETAIL SIMILARITY CALCULATOR

**Problem Formulation**

After the contour similarity calculation, a set of contour similar image set $\mathcal{C}$ is obtained, where $\mathcal{C}=\{I_1, I_2, ..., I_n\}$. The images with similar information on bronchi, blood vessels, nodules, etc. inside the lung lobes of the input image need to be further identified in the image set $\mathcal{C}$. In Fig. 4, since the detail similarity calculation focuses entirely on the parenchymal part of the lung, so it can be defined as a medical useful area (MUA) in the images $I_i$. The definition of detail similarity is described below.

**Definition 3.** Given a lung CT image, its corresponding medically useful area (MUA) can be defined below:

$$\text{MUA} = \{G_1, G_2, ..., G_m\}$$

where $G_i$ represents the $i$-th white stripe in the lung and $m$ denotes the number of white stripes in the lung.

**FIGURE 3.** The network architecture for the vision transformer

**FIGURE 4.** Two MUAs in a lung CT image

The white stripe in Definition 3 refers to the objects (e.g., bronchi, blood vessels, and nodules, etc.) that are shown in the lung lobes of the lung CT images. Since the white stripes have different shapes and sizes with their unique patterns in the images, it is hard for a computer to effectively discriminate them. So the graph model is adopted to describe the white stripe.

**Definition 4.** Given a white stripe $G_i$, it can be modeled by: $G_i = \{V, E\}$, where $V$ denotes the set of vertices of the graph, $V = \{v_1, v_2, ..., v_k\}$, $v_i$ denotes the $i$-th pixel point that do not equal to zero after binarization; $E$ denotes the set of edges, $E = \{e_1, e_2, ..., e_n\}$, in which $e_k = v_i \sim v_j$, ...
means that \( v_i \) is connected to \( v_j \).

In Definition 4, if the Euclidean distance between two-pixel points at their positions in the matrix does not exceed \( \sqrt{2} \) (i.e., the distance between two neighboring pixels is 1), these two pixels are considered adjacent. A vertex \( v_i \) is taken as the center and a breadth search is performed around it to connect all adjacent vertices to generate the edge set \( E \). Finally, a connected graph \( G \) is formed, which is the formation process of the stripe.

Suppose that the lung parenchyma to be retrieved is \( MUA_1 \) and the candidate lung parenchyma is \( MUA_2 \). Then, the similarity of lung internal details in the two lung CT images can be described by Definition 5.

**Definition 5.** Given two MUAs (i.e., \( MUA_1 \) and \( MUA_2 \)), their similarity can be defined by:

\[
\text{Similarity}(MUA_1, MUA_2) =
\begin{cases} 
1, & \text{if } \sum_{i=1}^{m} \sum_{j=1}^{n} G_{ij} \sim G_{2j} \geq \alpha \cdot \text{Len}(MUA_1) \\
0, & \text{otherwise}
\end{cases}
\]

(4)

where \( G_{ij} \) denotes the \( i \)-th stripe in the \( MUA_1 \); \( G_{2j} \) means the \( j \)-th stripe in the \( MUA_2 \); \( m \) is the number of stripes in \( MUA_1 \); \( n \) is the number of stripes in \( MUA_2 \); \( \text{Len}(MUA) \) refers to the number of stripes in the \( MUA \), and \( \alpha \) is a similarity threshold. \( G_{ij} \sim G_{2j} \) indicates that two stripes are similar.

Note that, in Definition 5, the similarity between two stripes is determined by the number of vertices and the shape constructed by all the vertices.

### Network Design

The internal structure of the detail similarity calculator is illustrated in Fig. 5. Based on the similarity definition, the similarity of details inside the lung is highly dependent on the number, shape and position of the stripes inside the lobe. Since the fully connected layer has some translation invariance, the fully connected layer at the resnet18 [16] is replaced with a more position-sensitive convolutional layer as a variant of resnet18. The resnet18 variant is used to construct the detail similarity calculator.

The round rectangle in the figure indicates a convolution operation which includes the convolution kernel size and the number of convolution kernels. ‘SD’ means the number of convolution kernel move steps. ‘PD’ refers to the blank fill size around the image. ‘Bn’ indicates the batch normalization layer. ‘Relu’ is an activation function and ‘maxpool’ refers to the maximum pooling layer. The entire detail similarity calculator has five blocks (i.e., one conv1 and four Bblocks). In each Bblock, an add operation on the matrix of the input Bblock and the matrix of the output from it is performed, which is the core residual idea of resnet. It can eliminate the gradient disappearance problem in network training.

![Network Architecture](image)

**FIGURE 5.** The network architecture for the details similarity calculator

Given two MUAs (e.g., \( MUA_1 \) and \( MUA_2 \)) as the inputs of the detail similarity calculator, after passing through the conv1 and the four Bblocks, the ‘SimScore’ is the output after the last convolution layer, which denotes the similarity between \( MUA_1 \) and \( MUA_2 \).

### IV. TRAINING

#### A. DATA PRE-PROCESSING

In this paper, we use the LUNA16 (*Lung Nodule Analysis 16*) public dataset [21] as the training dataset with a total of 1018 cases, which is derived from a larger dataset LIDC-IDRI.

A complete lung CT image contains not only both lung lobes but also other details (e.g., human muscle or organ soft tissue, bone tissue). For the WSAN model, to focus on learning the features of the MUAs, we need to extract the lung parenchyma information from the lung CT image.

Based on the characteristics of the CT images, CT images are used to reflect the degrees of X-ray absorption by organs or tissues with grayscale values. They are used to quantitatively assess the density levels of organ tissues in the image, called CT value in HU (Hounsfield unit). The
CT value of lung tissue in CT images is -500 but is usually greater than this value. The image is binarized by different CT values corresponding to different gray values in the image. Then the lung mask is obtained by dividing the external air and the internal torso using a seed filling algorithm. There are many fibers inside the lung, so it seems hollow (relative to the lung). To fill these hollows, the closing operation in morphology is used (expansion first, then erosion). First, expand the lung to fill the small cavities; then corrode to restore the original size; finally keep the largest connected domain. At this time the largest connected domain is the lung. The before and after image preprocessing is shown in Fig.6.

In order to extract the lung parenchyma information, we need to get a lung mask that masks the area other than the lung parenchyma. Before introducing the method of extracting lung parenchyma, it is necessary to introduce the characteristics of CT images. CT images use gray values to reflect the degree of X-ray absorption of organs or tissues, and use this to quantitatively evaluate the density of organs and tissues in the image. The level is called CT value, and the unit is HU (hounsfield unit). Lung tissue has a CT value of -500 in CT images, but is usually greater than this value.

Given an initial lung CT image $I_i$, where $I_i$ is a 512*512 HU matrix, as shown in Fig. 6(a), we extract lung parenchyma as follows:
1) In Fig. 6(b), take \( HU = -600 \) as the critical value, binarize \( I_i \), if the pixel value in \( I_i \) is less than -600, set it to 1; otherwise, set it to 0, and get the binarization matrix \( B_1 \).

2) As shown in Fig. 6(c), clear the boundary information of \( B_1 \). That is, set the pixels with the pixel value of 1 on the \( B_1 \) boundary and the pixels with the pixel value of 1 connected to it as 0.

3) In Fig. 6(d), to classify the pixel values in \( B_1 \), pixels with the same pixel value and adjacent pixels are grouped into one category and assigned a label value (i.e., \( label \)), where \( label \in \{1, 2, ..., n\} \) and \( n \) represents the number of categories. A class matrix \( L \) is obtained.

4) In Fig. 6(e), each category in \( L \) has an area, and the two areas with the largest area are reserved and set to 0 and 1 respectively to obtain a new binary matrix \( B_2 \).

5) In Fig. 6(f), morphological erosion was performed on \( B_2 \), that is, the white part was expanded. The significance of this operation was to separate the lung nodules connected to the blood vessels.

6) In Fig. 6(g), morphological closure operation was performed on \( B_2 \), and the significance of this operation was to attach the nodule to the lung wall.

7) In Fig. 6(h), the roberts operator is used to filter \( B_2 \) to fill the small holes in the lungs to obtain the lung mask.

8) In Fig. 6(i), use the lung mask (i.e., Mask) to mask the input image \( I_i \) to obtain the final lung parenchyma image.

**B. GENERATION OF TRAINING SET AND TEST SET**

This subsection studies how to automatically generate the datasets required for training the contour similarity calculator as well as the detail similarity calculator.

1) **CONTOUR SIMILARITY CALCULATOR DATASET**

First, the STL is used to generate the dataset and labels needed for the contour similarity calculator. Input a CT image of lung \( I_1 (512*512*1) \), which is slightly deformed by the STL to generate image \( I_2 (512*512*1) \), then it is considered that \( I_1 \) and \( I_2 \) are similar, and they are synthesized into a 2*512*512 tensor and given label 1 (similar).

However, a dataset with a label of 1 does not allow the contour similarity calculator to learn which images are contours dissimilar to each other. So the next is to find the set of images with dissimilar contours. Since a CT scan case consists of hundreds of slices, the contour shape of different sections of the lung will change dramatically as the levels change. Images with contour dissimilarity to the input image can be found in the CT scan images of the same case with different levels of sections. Synthesize them into a 2*512*512 tensor with the label 0.

The groups of the lung CT images with similar and dissimilar contours identified by the above method are shown in Fig. 7 and Fig. 8.

2) **DETAIL SIMILARITY CALCULATOR DATASET**

After training the contour similarity calculator, use the contour similarity calculator to randomly retrieve the database for image \( I_2 (512*512*1) \) with contour similarity higher than a certain threshold to the input image \( I_1 (512*512*1) \). Assume that these two images are detail dissimilar, then assign the label 0 (dissimilar) to this set of images and synthesize it into a tensor of 2*512*512. Next, it is necessary to find the images with similar details to be given to the detail similarity calculator to learn. Given input image \( I_1 (512*512*1) \), the image \( I_2 \) in the same group of CT scan lung images with adjacent layers must be similar to \( I_1 \) in details, i.e., in a group of CT scan lung images, if the input image \( I_1 \) is the \( r \)-th layer lung section, then the image \( I_2 \) with similar details to it is the \( (r-1) \)-th layer section or the \( (r+1) \)-th layer section. In this way, a set of lung CT images with similar details can be found, given the label 1 (similar), and synthesized as a 2*512*512 tensor.

It is uncertain to find a set of images with dissimilar details by the above method because it is likely that an image similar to \( I_1 \) in details can be retrieved by the
contour similarity calculator. Let the probability of retrieving similar details be \( p \). Then \( p = \text{num}_e / \text{num}_s \), where \( \text{num}_e \) is the number of images in the database with similar details to \( I_1 \) and \( \text{num}_s \) is the number of all images in the database with similar contour to \( I_1 \). To attenuate the uncertainty caused by \( p \), the ratio of similar to dissimilar labels in the training set is adjusted to 3:1, allowing the detail similarity calculator to learn more similar details.

The groups of similar and dissimilar lung CT images identified by the above method can be shown in Fig. 9 and Fig. 10, respectively.

C. LOSS FUNCTION DESIGN

For the similarity calculation of the lung CT images, the goal of the network is to score the similarity between two images. In essence, it is a classification task into one category, i.e., similar or not. We adopt the sigmoid function as the activation function and the cross-entropy function as the loss function for the training of the similarity calculators (e.g., the contour similarity calculator and the detail similarity calculator). The loss function is designed to measure the difference between the predicted and the ground truth, which is shown in Eq. (5).

\[
\text{Loss}(h(x), y) = \frac{1}{n} \sum_{x \in \{\text{batch}\}} -y \log[g[h(x)]] - (1 - y) \log[1 - g[h(x)]]
\]

where

- \(|\text{batch}|\) denotes the batch size in the network training;
- \(h(x)\) refers to the similarity score of the output in the network and \( h(x) \in [0, 1] \);
- \( y \in \{0, 1\} \) means the corresponding ground truth;
- \( g(x) \) is the sigmoid function and \( g(x) = \frac{1}{1 + e^{-x}} \).

V. EXPERIMENTAL EVALUATION

In this section, we perform extensive experiments to evaluate the effectiveness of the WSAN model.

A. EXPERIMENTAL SETUP

Our proposed WSAN method is implemented with the PyTorch open-source deep-learning library and uses an NVIDIA 1080Ti GPU for accelerated training. The system runs on the following platform configuration information: Intel i5-11400F CPU, 16GB running memory, and 4T mechanical hard disk. The Adam optimizer is used to train the network with an initial learning rate of 0.001. To verify the effectiveness of the WSAN model, the metrics are compared with those of the SIFT-BoVWs.

The database we used is the LUNA16 dataset [21] which includes 239,232 lung CT images. In this database, there are 712 sets of lung CT images. Each set containing lung section images of different levels ranging from 200-600, with an average of 336 lung CT images per set.

As shown in Table I, we divide the images in the database into training set, test set and cross-validation set according to the ratio of 6:2:2, in which the images in the training set and test set are used to implement the operations in Section IV.B.

| Dataset                | Data Size |
|------------------------|-----------|
| Training Set           | 26713     |
| Test Set               | 8902      |
| Cross-validation Set   | 8896      |

Tables II and III list the experiments’ settings of the WSAN, respectively. Note that, \( \text{Epochnum} \) is the number of iterations when training the model; \( \text{Dropout} \) means that during the training process, for the neural network unit, it is temporarily discarded from the network based on a certain probability to prevent overfitting. \( \text{Heads} \) refers to the number of the heads of the Multi-Head Attention module in the Transformer Encoder in the contour similarity calculator. The more heads, the more features that have received attention.
TABLE III. EXPERIMENTS’ SETTING OF DETAIL SIMILARITY CALCULATOR

| Parameter       | Value       |
|-----------------|-------------|
| Epoch num       | 50          |
| Learning rate   | 0.0001      |

B. PROTOTYPE SYSTEM

The prototype system is illustrated in Fig. 11. Fig. 11(a) is the input window, which can be used to set the \( k \) value in Top-\( k \) retrieval by the Setting button. Fig. 11(b) shows the Top-\( k \) similar images based on the input image. After submitting an image as an input, the WSAN-based similarity computation is conducted with all candidate images in the database and the similarities are sorted in the sequel. Finally, the Top-\( k \) similar images can be obtained.

To shorten the response time of the algorithm, the size of the candidate image set needs to be minimized, and the system uses the comparison of grayscale histograms to filter the dissimilar images. That is, before inputting two images into the WSAN model for similarity evaluation, the gray value statistics arrays of the two images are obtained separately. The length of the gray value statistics arrays of the images is 26 in a set of 10 gray values. Compare the maximum value in the gray value statistics array of the two images. If the difference between the two values is higher than 3000, it is no need to input the WSAN model and filter directly since they are not similar.

C. ACCURACY RATE

For the Top-\( k \) similarity retrieval of the lung CT images, the accuracy of the retrieval results is needed. So in this experiment, we use precision to represent the retrieval accuracy rate in Eq.(6):

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

where \( TP \) means the number of images output correctly and \( FP \) refers to the number of images output incorrectly. For the ground truth of the retrieval result, in fact, we invited several professional physicians to judge our retrieval results. The \( TP \) and \( FP \) are determined by the physicians.

(a). WSAN-based (b). SIFT-BoVWs-based

FIGURE 12. Example of Top-5 retrieval results (a red tick indicates that the output image is similar to the input one both in detail and the contour; a red circle indicates that the output image is similar to the input one only in contour)

First, the \( k \) in the Top-\( k \) retrieval is set to 5, i.e., the 5 most similar lung CT images are returned and the experimental results are compared. Fig. 12 shows an example of the Top-\( k \) retrieval in which each column represents the input and output. The first row represents the input lung CT images and the following 5 rows are the 5 most similar images identified by the retrieval system in the database. Fig. 12(a) shows the retrieval results based on the WSAN model. The third column has only four outputs representing only 4 similar images retrieved by the system in the database. Fig. 12(b) represents the retrieval results based on the SIFT-BoVWs scheme. In the lower right corner of the output image, a red tick is used to represent that it is similar to the input image detail while the contour is similar, and a red circle indicates that it is similar to the input image contour only. If we consider only the accuracy of contour similarity in the examples,
the accuracy of the WSAN-based retrieval system can reach 96%, while the accuracy of the SIFT-BoVWs retrieval system is only 32%. If not only contour similarity but also detail similarity is considered, the accuracy of the WSAN-based retrieval system is 72%, while the accuracy of the SIFT-BoVWs retrieval system is only 4%. Therefore, the retrieval accuracy of the proposed WSAN model outperforms that of the SIFT-BoVWs one in the Top-k retrieval of the lung CT images.

Next, the k of the Top-k retrieval is set to 10. The comparisons of the retrieval accuracies between the two methods are conducted for 50 times, which are shown in Fig.13 and Fig. 14, respectively. In Fig. 13, there are four methods in the horizontal-axis, where the WSAN model indicates that only the retrieval accuracy of contour similarity is considered using the WSAN-based retrieval system, while the WSAN+ considers both the detail similarity accuracy and SIFT-BoVWs and SIFT-BoVWs+ do the same. Three metrics (i.e., max, min and avg) of the retrieval accuracy are provided. It can be seen that the accuracy rate of the WSAN-based retrieval system is superior to that of the SIFT-BoVWs retrieval system, regardless of which metric is used for comparison. The accuracy rates of the 50 retrievals are shown in Fig. 14.

D. MEAN AVERAGE PRECISION(mAP)

Mean average precision(mAP) is also an important metric for evaluating a retrieval system. mAP is derived based on the average precision(AP) which is defined in Eq.(7):

\[ AP = \frac{\sum_{RK} Precision_{RK}}{Num_{RK}} \]  

(7)

where RK means the index of the image among all the images retrieved, Precision_{RK} the retrieval accuracy rate until the output image with the index of RK, and Num_{RK} denotes the total number of the RKs. Then the formula of the mAP is derived below.

\[ mAP = \frac{\sum_{i=0}^{n} AP_i}{n} \]  

(8)

where n is the number of retrieval samples. The AP and mAP of our algorithm and SIFT-BoVWs algorithm in terms of contour similarity and detail similarity are likewise calculated for the retrievals in our experiments, as illustrated in Fig. 15 and Fig. 16. If only contour similarity is considered, the mAP@10 of our algorithm can reach 94.91%, while the SIFT-BoVWs retrieval system is only 26.84%. If detail similarity is further considered, the
mAP@10 of our algorithm can reach 66.40%, while the latter is only 2.68%. Both considering the contour and detail similarities may make the retrieval system not able to find such similar lung CT images because they do not exist in the database, thus resulting in a lower AP. So it is reasonable that further consideration of the detail similarity causes the mAPs of both methods to decrease.

VI. CONCLUSIONS

The paper proposes a new deep learning network called the WSAN model for similarity analysis of the lung CT images, using which content-based lung CT image retrieval can be effectively performed. Compared with the state-of-the-art deep learning-based medical image similarity retrieval systems, the biggest advantage of this system is its weak supervision, i.e., there is no need to hire a professional physician to label the images to provide the training task of the network. Meanwhile, the WSAN-based retrieval system achieves satisfactory retrieval performance under weakly supervised training. The extensive experiments demonstrated that our proposed scheme can obtain 94.91% contour similarity mAP@10 and 66.40% contour as well as detail similarity mAP@10 on the LUNA16 dataset, which basically meets the retrieval requirements.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their helpful comments.

REFERENCES

[1] ZH. Zhou. "A brief introduction to weakly supervised learning." National Science Review 5.1 (2018): 44-53.
[2] DG Lowe "Distinctive image features from scale-invariant keypoints." Int. J. of Computer Vision 60.2 (2004): 91-110.
[3] H Müller, J Kalpathy-Cramer, et al. "Overview of the CLEF 2009 medical image retrieval task." in Proc. of Workshop of the Cross-Language Evaluation Forum for European, Heidelberg, 2009.
[4] A Krizhevsky, I Sutskever, and GE. Hinton. "Imagenet classification with deep convolutional neural networks." in Proc. of NIPS 25, pp.1097-1105. 2012.
[5] M Mizotin, J Benois-Pineau, M Allard, G Catheline. "Feature-based brain MRI retrieval for Alzheimer disease diagnosis." in Proc. of IEEE ICIP, 2013.
[6] H. Pan, P. Li, Q. Li, Q. Han, X. Feng, L. Gao. "Brain CT image similarity retrieval method based on uncertain location graph." IEEE J. of Biomedical and Health Informatics vol.18, no.2, pp. 574-584. 2013.
[7] HC. Shin, HR. Roth, M. Gao, L. Le, Z. Xu, I. Nogues, J. Yao, D. Mallura, RM. Summers. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Trans. on Medical Imaging, vol.35, no.5, pp.1285-1298. 2016.
[8] SK. Sundararajan, B Sankaragomathi, DS Priya. "Deep belief CNN feature representation based content based image retrieval for medical images." J. of Medical Systems. vol.43, no.6, pp. 1-9. 2019.
[9] A Khatami, et al. "A deep-structural medical image classification for a radon-based image retrieval." in Proc. of 2017 IEEE 30th Canadian Conf. on Electrical and Computer Engineering, 2017.
[10] A. Khatami, M. Babaie, HR. Tizhoosh, A. Khoosravi, T. Nguyen, S. Nahavandi. "A sequential search-space shrinking using CNN transfer learning and a Radon projection pool for medical image retrieval." Expert Systems with Applications, vol.100, pp. 224-233. 2018.
[11] K. Karthik, and S.S.Kamath. "A hybrid feature modeling approach for content-based medical image retrieval." in Proc. of IEEE 13th Int’l Conf. on Industrial and Information Systems, 2018.
[12] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu. "Spatial transformer networks." in Proc. of NIPS 28 (2015): 2017-2025.
[13] FL. Bookstein. "Principal warps: thin-plate splines and the decomposition of transformations." IEEE Trans. on Pattern Analysis and Machine Intelligence. vol.11, no.6 pp.567-585. 1989.
[14] A Vaswani, et al. "Attention is all you need." in Proc. of NIPS. 2017.
[15] A Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
[16] K He, et al. "Deep residual learning for image recognition." in Proc. of the IEEE CVPR. 2016.
[17] H Liu, et al. "Deep supervised hashing for fast image retrieval." in Proc. of the IEEE CVPR. 2016.
[18] L Ma, et al. "A new method of content based medical image retrieval and its applications to CT imaging sign retrieval." J. of biomedical informatics. vol.66, pp.148-158. 2017.
[19] YH. Cai, et al. "Medical image retrieval based on convolutional neural network and supervised hashing." IEEE access, 7, pp. 51877-51885. 2019.
[20] N Sampathila, and R J Martis. "Computational approach for content-based image retrieval of K-similar images from brain MR image database." Expert Systems, e12652. 2020.
[21] AAA Setio, et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge." Medical Image Analysis 42 (2017): 1-13.

Yi Zhuang received Ph.D degree in computer Science from Zhejiang University in May 2008. He is currently a full professor at the School of Computer & Information Engineering in Zhejiang Gongshang University where he joined as faculty member since May 2008. Dr. Zhuang is a recipient of the CCF Doctoral Dissertation Award conferred by Chinese Computer Federation in 2008 and IBM Ph.D. Fellowship 2007–2008. His research interests mainly focus on multimedia database and cloud computing. He has published 40+ papers in the leading journals and conferences, i.e., ACM TOIT, TALIP, Information Sciences, KAIS, and ESWA, etc.

Nan Jiang received the bachelor degree of medical science and the master degree of medical science both from the Zhejiang University in 2004 and 2007, respectively. Ms. Jiang is with Affiliated Hangzhou First People’s Hospital, Zhejiang University School of Medicine. Her research interests mainly focus on medical image processing. She has published 10+ papers in international conferences and journals.
**Shuai Chen** is now a master student in Computer Science at Zhejiang Gongshang University. Before that, he received the bachelor degree of Computer Science from Hangzhou Dianzi University. His research interests mainly focus on medical image processing, machine learning, etc.