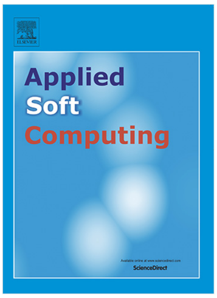


Segmentation and classification of knee joint ultrasonic image via deep learning

Long Zhili, Zhang Xiaobing, Li Cong, Niu Jin, Wu Xiaojun, Li Zuohua

PII: S1568-4946(20)30703-1
DOI: https://doi.org/10.1016/j.asoc.2020.106765
Reference: ASOC 106765

To appear in: *Applied Soft Computing Journal*

Received date : 29 June 2019
Revised date : 12 September 2020
Accepted date : 27 September 2020

Please cite this article as: Z. Long, X. Zhang, C. Li et al., Segmentation and classification of knee joint ultrasonic image via deep learning, *Applied Soft Computing Journal* (2020), doi: https://doi.org/10.1016/j.asoc.2020.106765.

Segmentation and Classification of Knee Joint Ultrasonic Image via Deep Learning

Zhili Long[1] Xiaobing Zhang[1] Cong Li[1] Jin Niu[1] Xiaojun Wu[1] Zuohua Li[1]

1. Harbin Institute of Technology, Shenzhen;

Shenzhen, Guangdong, China

Correspondence information: Li Zuohua, Harbin Institute of Technology, Shenzhen, lizuohua@hit.edu.cn

Segmentation and Classification of Knee Joint Ultrasonic Image via Deep Learning

Long Zhili[a], Zhang xiaobing[a], Li cong[a], Niu Jin[a], Xiaojun Wu[a], Li Zuohua[a,*]

[a] Harbin Institute of Technology, Shenzhen, 518055 Shenzhen, China

Abstract

The knee is one of the most complicated joints in the human body, but it could be easily injured. Ultrasound imaging is an important technology for the diagnosis of the knee disease. To assist doctors in the treatment and reduce errors of judgment, we investigate the segmentation of disease regions and the automated identification of the typical knee joint diseases. First, we use deep learning to segment the Region of Interest (ROI). To solve the mis-segmentation and poor edge segmentation that occur when the ultrasound image is directly fed into the deep neural network, an image segmentation framework is proposed that integrates snake preprocessing, dilated convolution to expand the receptive fields, and multi-channel learning. Second, due to the small difference in features among various categories of ultrasound images, a hybrid algorithm is proposed based on the Resnet rough classification and quadratic training with graph embedding. Finally, the experiments show that the proposed image segmentation framework achieves 10% greater accuracy than a common segmentation network. By visualizing the feature vectors extracted from the classification network,

[*] Li Zuohua. E-mail address: lizuohua@hit.edu.cn

we verify that the feature vectors are closer on similar images after quadratic training by graph embedding. Employing the optimization with quadratic training, we increase the classification accuracy by 11% compared to the Resnet approach.

Keywords: Deep learning, image segmentation, image classification, graph embedding, snake algorithm, knee ultrasound image.

1. Introduction

The knee is one of the most complicated joints in the human body, but it is susceptible to infections and injuries when people engage heavily in physical exercise or experience physical trauma in collisions or accidents. Common knee diseases include synovitis, synovial thickening, and cysts. Medical imaging technology is important for diagnosing knee joint conditions [1-3]. Generally, the liquid area of a lesion shows darker in a medical image providing doctors with key information. The corresponding accuracy to locate the lesion areas directly affects related diagnosis. At present, diagnosis of knee diseases depends on doctors' observations and personal experience, which requires a great deal of manpower and time, and unfortunately has limited accuracy. On the other hand, the use of artificial intelligence to assist or even replace doctors' diagnosis can not only improve diagnostic accuracy, but also decrease the burden on doctors [4-5].

Currently, medical imaging of knee joints is performed using computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US). CT images are acquired by utilizing a professional CT scanner and allow the bone image of the knee joint to be observed. Li et al. proposed a method for three-dimensional (3-D) reconstruction based on CT images of knees [6]. This method involves the selection of an appropriate threshold for segmentation according to the histogram of the CT image sequence and then the processing of the images with region growing segmentation. Finally, the femur, tibia, patella and etc. are reconstructed. To enhance the understanding of the 3-D anatomical structure of the knee, Maruyama et al. reconstructed the lateral image of the knee joint using CT images, and established a planar image generation method corresponding to the position of the curved knee joint [7]. The knee joint is divided into upper and lower parts, and the planar images of the two parts are combined to obtain a planar image corresponding to the joint bending at different angles. MRI can produce high-precision stereoscopic images of the internal structure of knee joints using a nuclear magnetic resonance instrument. It is also an important means to obtain images of the knee joints. Liu et al. combined a deep convolutional neural network (CNN), three-dimensional fully connected conditional random field (CRF), and three-dimensional simplex model to segment tissue in MRI images of the knee joint [8]. High resolution pixel level multi-class tissue classification was performed on 12 different joint structures. The results show that four types of tissue existed, with an average Dice coefficient above 0.9. This method is suitable to perform rapid and accurate comprehensive tissue segmentation of the knee joint, however, the convolutional encoder-decoder (CED) network training is a complicated computation and requires a large amount of pixel-wise annotated training data sets for each new evaluated tissue in contrast. Norman et al. established a deep learning model based on a U-net convolutional network to achieve automatic segmentation of knee articular cartilage [9]. The model has a high Dice coefficient, especially for the three-dimensional double-echo steady-state (3D-DESS) image, which has a dice coefficient being 0.770~0.878 for dilated cartilage and 0.809~0.753 for the lateral meniscus and medial meniscus. However, the model was not verified on any 2D-US image. US imaging technology is another technology used to obtain knee images by sending and receiving ultrasound echo signals allowing a 2D image of the knee to be reconstructed. Desai et al. applied a new framework to automatically segment cartilage tissue in the enhanced ultrasound knee joint images, which improves US bone surfaces by calculating local phase image features, dynamically programming for the bone segmentation, and segmenting bone surface as the initial seed of a random walk algorithm [10]. In a case study, a hundred scan images were validated on eight healthy volunteers, and the corresponding average Dice coefficient was 0.8758. Lai et al. utilized a snake model to segment US images of the knee joint meniscus and provided a multi-active contour frame called multiple LREK to handle multiple objects [11]. Although the model is clear and concise, it can only implement rough segmentation of images. Conventional intelligent machine learning such as Support Vector Machine (SVM) is also applied to segment and classify the US images, since it not only has a rigorous theoretical background, but also can find the global optimal solutions. Lei et al. applied the SVM method to classify the normal and cirrhotic liver with the accuracy being 0.87 [12]. However, they need more distinctive features and enhanced methods to improve the accuracy. Mehdi et al. proposed a novel ultrasound RF time series analysis and an extended version of SVM. They reported an area under receiver operating characteristic curve being 0.95 in tenfold cross validation and 0.82 for the detection of prostate cancer [13]. However, the method is limited to the routine clinical examination process. It has been proved that the conventional SVM method has a good classification, especially in the task of small sample set. However, it cannot meet the requirement of ultrasonic medical image classification due to the challenge of high similarities and irregularities in the US knee image.

Recently, a variety of advanced algorithms have been proposed for medical image processing, ranging from traditional algorithms to machine learning and deep learning, such as the level set [12], snake model, and U-net. With the development of Alexnet [13], deep learning has been widely applied to the segmentation and classification of medical images. Common classification networks include VGG network [14], GoogLeNet [15], and Resnet [16]. Conventional segmentation frameworks include DeconvNet [17], Segnet [18], FCN [19], PSP network [20] and Deeplabv3 network [21-24]. BenTaieb et al. proposed a topology-aware FCN to achieve gland segmentation [25]. Samundeeswari et al. adopted the K-Means algorithm to segment US images of breasts [26]. Ronneberger et al. constructed a U-net for US image segmentation in a symmetric contracting path and an expansive path [27]. To improve the output resolution, the maximum pooling layer was replaced by the up-sampling layer. The expansive path was more or less symmetric to the contracted path, and produced a U architecture that reduced the loss of image information during down-sampling. Milletari et al. presented a volumetric convolutional neural network that performed segmentation of MRI prostate volumes in a fast and accurate manner [28]. Bi et al. improved the FCN and proposed a new stack structure with multi-channel learning [29]. Wu et al. constructed a cascaded FCN to automatically segment US images of the head and abdomen of prenatal fetuses [30]. These classification networks can solve some problems such as severe boundary incompleteness and achieve the acceptable segmentation accuracy on fetal head and abdomen US image without the introduction of the validation to the knee joint image.

Among medical imaging technologies, MRI and CT are most useful for diagnosing cartilage, bone, and other parts of the knee joint, and the corresponding accuracy is high. However, the shortcomings of high cost and the emission of radiation are inevitable. Most importantly, the liquid area of knee lesions cannot be clearly observed by these two technologies. Currently, US imaging technology is applied to obtain knee joint images that can be used to diagnose knee diseases, such as effusion, synovitis, synovial thickening, and cysts. US imaging technology is low-cost, simple to operate, radiationfree, non-intrusive, and allows for repeated scans. However, current research on the detection of knee diseases with US imaging is limited. Moreover, most processing methods for US knee joint images focus on traditional algorithms such as histogram equalization, median filtering, and the neighborhood de-mean method. Therfore, the corresponding processing accuracy is low and the processing time is long. As a result, it is important to investigate how the knee disease can be diagnosed by the deep learning approach based on US images.

In this paper, we present a deep learning method to segment and classify lesion areas in the US knee joint image. Figure 1 presents an overview of our investigation of a knee disease diagnosis system, which includes a training phase, in which a segmentation and classification network is proposed. The main contributions of our work are threefold: 1) a multi-channel learning model is proposed to improve the segmentation result of lesion areas; 2) graph embedding method in natural language processing is adopted for the disease recognition; 3) a novel optimization network based on quadratic training of feature vector is proposed. The experiments show that the proposed networks can improve the classification performance compared to conventional methods using only VGG or Resnet. Thus, it can be extended to other medical image for segmentation and classification with retraining parameters.

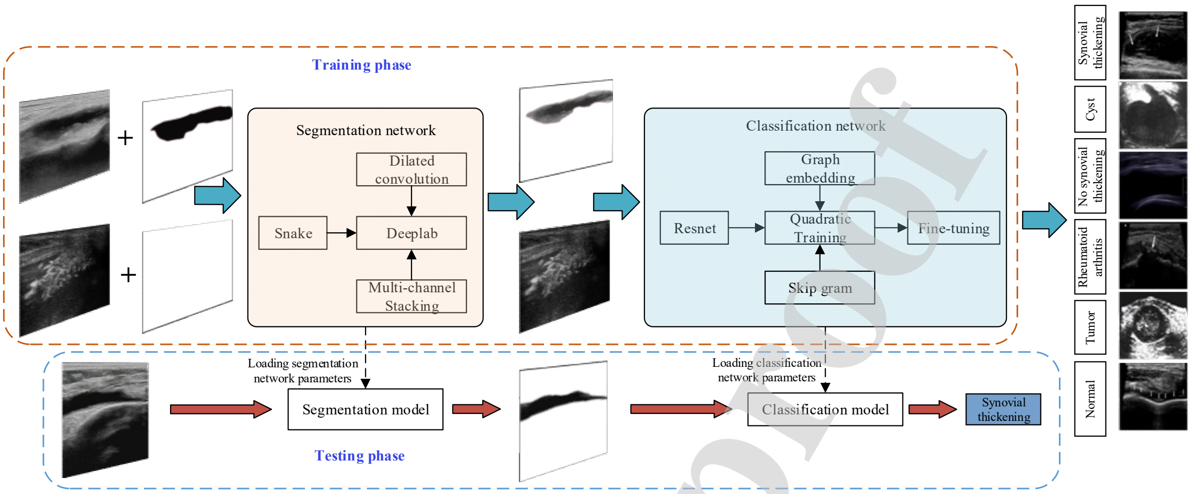


Fig. 1. Overview of the proposed method.

This paper is organized as follows. Section 2 adopts deep learning to achieve semantic segmentation of the effusion region in US knee joint images. Section 3 develops an automatic recognition system for common knee joint diseases from the dataset. Section 4 presents experimental verifications of the segmentation network and the proposed classification algorithm. Section 5 draws the conclusion.

2. Effusion area segmentation

2.1. Snake preprocessing

In the US knee joint image dataset, it is difficult to attain accurate segmentation, as the effusion and background are very similar in color. In addition, the traditional cropping method can only solve the fixed area because the effusion shape is variable. In this study, we propose to use a snake algorithm to preprocess the image and eliminate background areas similar to the effusion.

In the snake model, it is necessary to randomly or manually define a controllable and deformable initial contour curve. The region within the contour curve is considered as the segmentation region. The contour curve is deformed by minimizing the energy function, and the closed curve with the minimum energy value is the final contour. The energy function of the control point on the contour curve is expressed as

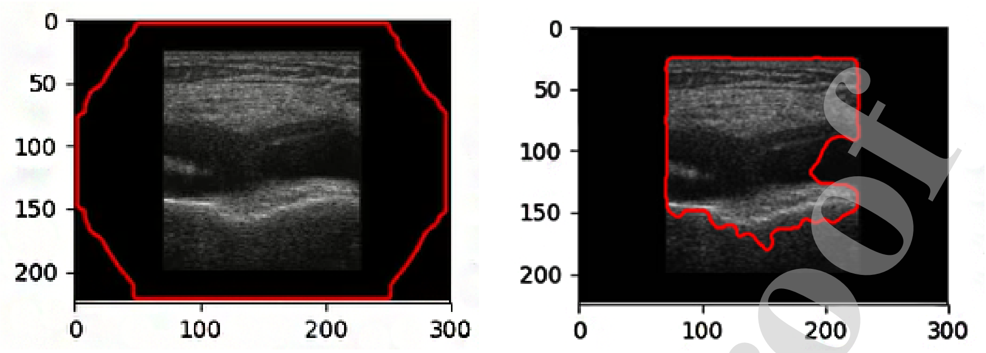
$$\boldsymbol{E}_{total}=\int(\alpha\left|\frac{\partial}{\partial s}\boldsymbol{v}\right|+\beta\left|\frac{\partial^2}{\partial s^2}\boldsymbol{v}\right|^2+\boldsymbol{E}_{ext}(\boldsymbol{v}(s)))\mathrm{d}s \tag{1}$$

$$\boldsymbol{v}(s)=[x(s),y(s)]\ \ s\in[0,1] \tag{2}$$

$$\boldsymbol{E}_{ext}\big(\boldsymbol{v}(s)\big)=-|\nabla I(\boldsymbol{v})|^2 \tag{3}$$

where $\boldsymbol{E}_{ext}$ represents external energy, $\boldsymbol{v}(s)$ represents the coordinates of the control point, and $\boldsymbol{E}_{total}$ denotes the total energy of the control point. α and β are the coefficients that control the proportion of the elastic and rigid energy, respectively. This expression reflects the deformation of the contour curve by the combination of different energies and the relationship between the energy and the control points on the contour curve.

In this study, the outline curve is used as the initialization curve and the target is to minimize the energy function. The curve in the image is deformed so that it gradually approaches the edge of the target area. The snake process is demonstrated in Figure. 2(a). Because the contour is limited by the energy function, the control point stops shrinking when the gradient of the gray image is maximized and the velocity of the control point is reduced to zero. Figure. 2(b) shows the preprocessing results of the snake algorithm. The red contour curves gradually shrink from Figure. 2 (a) to Figure. 2 (b) by minimizing the energy function, and the surrounding black is the easily mis-segmented area. The peripheral contour approximates the target area, and the area with mis-segmentation is removed successfully. Lastly, the peripheral rectangle of the contour is chosen as the preprocessed image.



(a) Process image

(b) Contour shrinking area

Fig. 2. Snake processing.

2.2. Image segmentation

To improve the accuracy of the disease recognition, we propose the semantic segmentation of the effusion area in the US knee image. In our study, the semantic segmentation is a process used to classify each pixel in the US image into foreground and background. A large amount of textural and spatial information exist between each pixel. By combining the convolution layer and the pooling layer with dilated convolution and multi-channel stacking, the information between each pixel is integrated and an optimized segmentation result is obtained. Figure 3 is the encoder-decoder structure in our study, which consists of multiple convolutional layers and pooling layers. The input and output sizes of each layer are shown in the figure, where the digital combination (h, w, c) represents the height, width and channels of image, respectively. To accelerate training, the image size is rescaled to (400, 400) as the input. Feature extraction is performed through convolution and pooling. Finally, the scale is resized to (400, 400) by the quadratic linear interpolation.

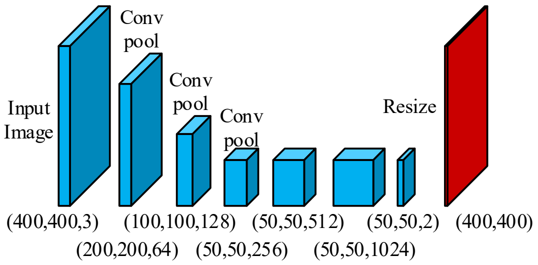


Fig. 3. Encoder-decoder structure.

2.2.1. Dilated Convolution

The dilated convolution to increase the receptive fields and improve the result of feature extraction is introduced. The main parameter in the dilated convolution is the expansion rate *r*, meaning that the zero values of *r*-1 numbers are added to the elements of the non-dilated convolution kernel. With dilated convolution, the receptive field can be expanded while the parameter number remains fixed, and more image information can be attained without an increase in the processing time [23]. Therefore, an effective mechanism to achieve an optimal balance between small and large fields is proposed.

Figure 4 is the Deeplabv3 network, which contains the Atrous Spatial Pyramid Pooling (ASPP) structure. Where, conv and ReLu are the convolution layer and activation function, respectively, and Pooling represents the down-sampling layer. ASPP structure captures objects and context at multiple scales by multiple filters with different rates. k denotes the size of the convolution kernel and s is stride. In this study, to expand the receptive field and balance small and large fields, segmentation without dilated convolution, with double-layer dilated convolution, and with multi-layer dilated convolution are analyzed. Moreover, dilated convolution layers from back to front are added to the Deeplabv3 network [22].

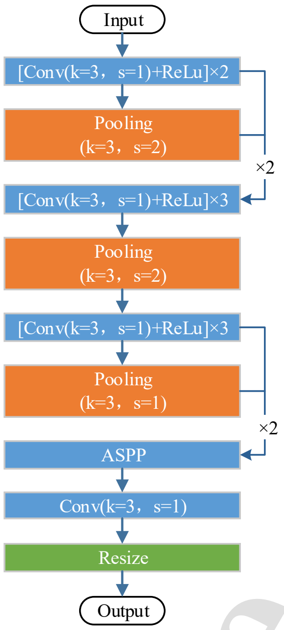


Fig. 4. Block diagram of the network structure.

2.2.2. Multi-channel Stacking

Because the edge of the effusion area is blurry in US knee images, the corresponding segmentation accuracy is low when images are directly fed into the Deeplabv3 network with dilated convolution. In this study, the foreground and background are used as training targets. The final segmentation is then determined by the weighting result of the two channels.

The foreground and background of the image label are complementary; if the foreground is predicted, the background is automatically generated. However, if the foreground and background are predicted separately, a cross-section in the results will exist, as they are not absolutely complementary. The category of each pixel in the intersection area is determined by the prediction probability. Figure 5 demonstrates the flow chart for combining the predictive probabilities of the two models. In the figure, the red curve in the stacking result represents the result of segmentation combining the predictive probabilities of the two models, and the Single means learning through a single channel. The two-channel probability maps are defined as

$$pred=\gamma\cdot Q+(1-\gamma)\cdot(1-R) \tag{4}$$

where *pred* is the prediction probability map after multi-channel stacking, and the index of maximum *pred* is considered as the classification result of each pixel. Q and R represent predicted results obtained by training the effusion area and the background as ROI, respectively. γ is the weighting factor that is a crucial parameter representing the preference for foreground or background. γ is initialized to 0.5 for the equation, while it becomes an optimized value during learning. The classification result is 0 or 1, which represents the background or the effusion area. The squared sum of different pixel values from the label is used as a loss function as

$$loss=\sum_{j}^{batch\_size}\sum_{i}^{num}(\boldsymbol{pred}_i\text{-}\boldsymbol{label}_i)^2 \tag{5}$$

where $\boldsymbol{pred}_i$ is the predicted value of the *i*-th pixel stacked by two channels, $\boldsymbol{label}_i$ denotes the label of the *i*-th pixel, and *loss* is the output of the loss function. *batch_size* represents the minimum batch size in the training, and *num* is the total number of pixels.

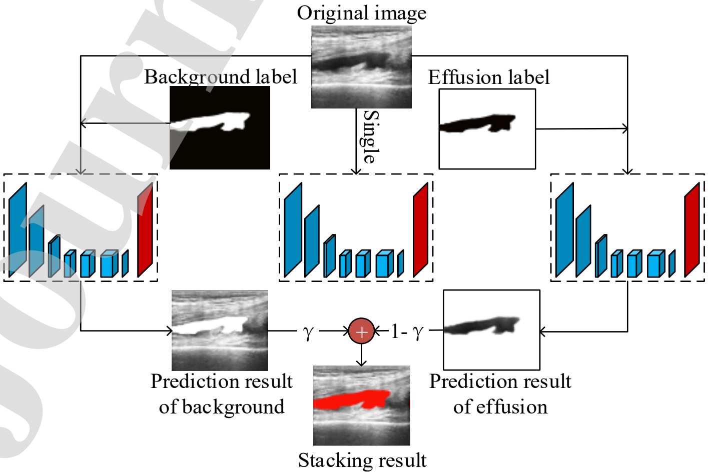


Fig. 5. Multi-channel stacking.

3 Disease recognition

3.1. Disease description

The structure of the knee joint includes the joint cavity, suprapatellar capsule, and patella. Figure 6 presents six typical manifestations of knee joints in ultrasound diagnosis. Figure 6 (a) is a normal knee joint, and the area identified by white arrows is the cartilage. Knee diseases often appear in conjunction with effusion and synovial thickening. In general, knee diseases are divided into five categories. Figure 6 (b) and (c) show the non-synovial thickening and synovial thickening, respectively. Figure 6 (d) shows a cyst, which is usually in elliptical echoless area with clear boundary and appearing hyperechoic. Figure 6 (e) shows a tumor as it is a circular or elliptical hypoechoic area with clear boundary, hyperechoic capsule and a constriction for surrounding tissues. Figure 6 (f) shows typical signs of rheumatoid arthritis. More hypoechoic areas in the lesions of articular cavity will exist, and uneven isoechoic tissues in the hypoechoic areas may appear.

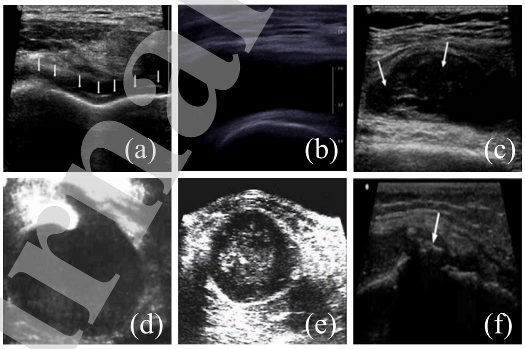


Fig. 6. Manifestations of knee joint. (a) A normal knee joint, (b) Non-synovial thickening, (c) Synovial thickening, (d) Cyst, (e) Tumor, (f) Rheumatoid arthritis.

3.2. Graph embedding method

As the segmentation result exhibits a binary classification value for all the pixels, we map it to the original image and attain the corresponding pixel. Then, the feature vectors from the mapped original images are extracted by Resnet to achieve a rough disease classification. With its residual structure and multiple convolution layers, Resnet can learn deeper features. The overall structure of Resnet is shown in Figure 7. Conv and FC are the convolution layer and fully connected layer, respectively. We adopt ReLu as an activation function and average pooling is used in Pooling. BN indicates batch normalization. The top half of the picture is ID Block, which is implemented by shortcut connections.

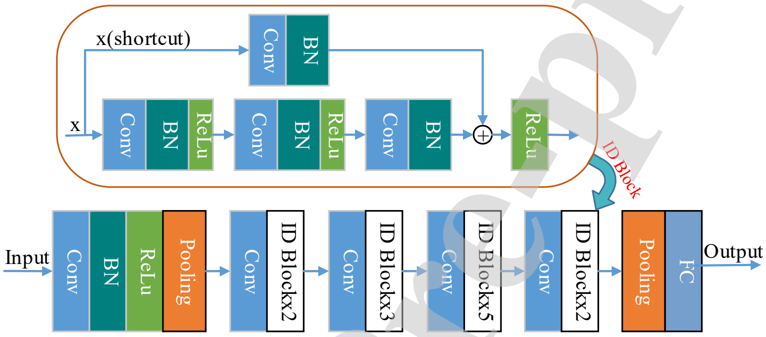


Fig. 7. Rough classification network.

As the feature vectors extracted by Resnet are the high-dimensional non-sparse matrix, to optimize the feature vectors with more separability in different US images, we use the graph embedding method in natural language processing to map the high-dimensional non-sparse matrix into low-dimensional non-sparse vectors [31]. In this study, the random walk method is adopted and the main steps for processing US images are as follows.

3.2.1. Constructing Isomorphic Graphs

First, an ID is given to each image in the dataset, then an isomorphic network of all US images is constructed according to their categories and IDs. Each image corresponds to a node in the network. The edges connecting the two nodes represent the relationship between the two images. Figure 8 shows an isomorphic graph constructed by randomly selecting some samples from six US image categories. Nodes of the same color represent one image category. Clearly, there is only connection between the same image categories while no any connection between different image categories exists.

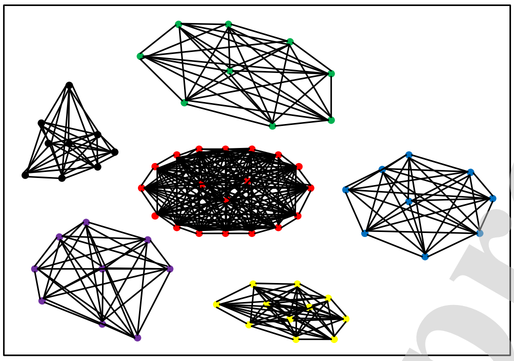


Fig. 8. Isomorphic graph constructed from six US image categories.

3.2.2. Random Walk Sampling

To extract multiple relationships between nodes, a sequence of other nodes connected to a given node is randomly sampled from the constructed isomorphic graph. From each node, a sequence with the specified length can be obtained. Figure 9 shows a sequence of four nodes randomly generated from the isomorphic graphs of two US image categories. In the figure, 1, 2, 3 … represent image IDs, and the two colors of nodes correspond to two US image categories. It is clear that the sequence generated belongs to an US image category. The length of sequence is variable.

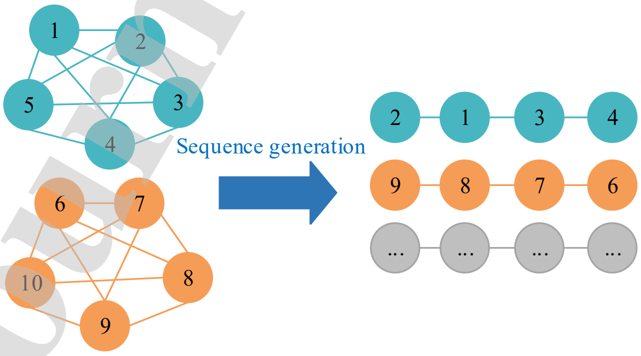


Fig. 9. Sequence generation by random walk.

In the random sampling, to avoid repeatedly selecting the same node, we adopt the probability model in the following expression (6). Moreover, the traversal times can be adjusted according to the sampling numbers in each US category.

$$\boldsymbol{P}_{ij}=\begin{cases} 1/(n_i\text{-}n_{Seq}) & (j \in Seq_k) \\ 0 & (\text{other}) \end{cases} \tag{6}$$

where $\boldsymbol{P}_{ij}$ represents the probability of selection from node i to j, n_i is the node number adjacent to i, Seq_k is sequences generated, and n_{seq} denotes the node number adjacent to i in the generated sequence.

3.2.3. Training

In our study, skip gram is used to train random sampled image sequences [32-34]. First, a node in the sampling sequence is selected as the input node, then a source sequence is obtained by a sliding window. The length of source sequence depends on the size of the sliding window, which is represented as *window_size*. If *window_size*=k, the sequence length is 2k+1, where k nodes are on both sides of the input node. Lastly, we select a node in the source sequence and the input node as a training sample, as shown in Figure 10. Where, n denotes the length of the sampling sequence, the k+1 node is the input node, and red box is the range of the sliding window. The training sample consists of the input node and prediction node.

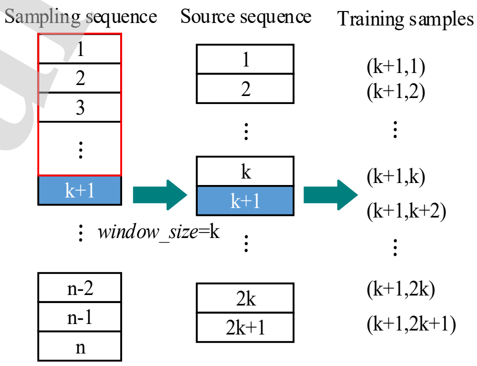


Fig. 10. Generating training samples.

The embedding vectors for each image can be attained by skip gram. Thus, we can calculate the distance between vectors, which represents the degree of similarity of US images. For an input vector, we encode the input node as 1 and the rest as 0's; output vectors are encoded in the same way. The implementation program for the skip gram in US images is demonstrated in Figure 11. There is no activation function and the number of neurons is equal to the dimension of the embedding vector in the hidden layer. The Softmax function is used in the output layer with a probability distribution for the whole sequence. The dimensions of the input and output vector are determined by the total number of input images.

Finally, the output of the hidden layer is considered as an embedding vector after training the neural network with one hidden layer. The embedding vectors are calculated as

$$\begin{aligned} \boldsymbol{E}_v &= \boldsymbol{I}_v \cdot \boldsymbol{E}_m \\ &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times m} \cdot \begin{bmatrix} a_{00} & \cdots & a_{0n} \\ \vdots & \ddots & \vdots \\ a_{m0} & \cdots & a_{mn} \end{bmatrix}_{m\times n} \\ &= \begin{bmatrix} a_{10} & a_{11} & \cdots & a_{1n} \end{bmatrix} \end{aligned} \tag{7}$$

where m denotes the number of images in the dataset and n is the dimension of the embedding vectors, $\boldsymbol{E}_m$ is the predicted parameter matrix, $\boldsymbol{E}_v$ is the embedding vector extracted from the parameter matrix, and $\boldsymbol{I}_v$ represents the one-hot coding vector.

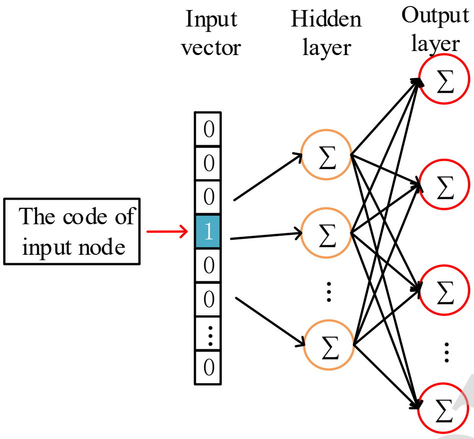


Fig. 11. Skip gram model.

3.3. Optimized-classification-based Resnet

The classification of knee diseases is a task of fine-grained recognition, as no obvious difference in US image characteristics exists. Thus, detailed characteristics should be classified in a larger category. We use two-stage training and one-stage testing in fine-grained classification, as shown in Figure 12. A rough classification result is obtained using only Resnet, while an optimized classification result is attained after combining Resnet with quadratic optimization.

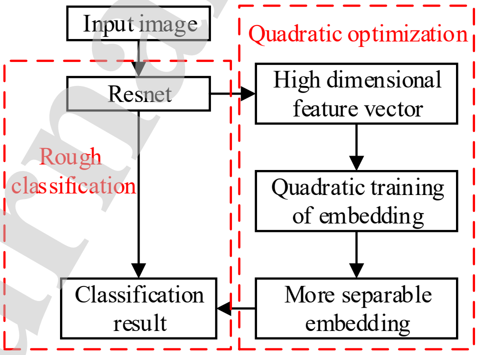


Fig. 12. Flow chart of optimized classification based on Resnet.

3.3.1. Quadratic Training

Embedding vectors with more separability can be attained by graph embedding. However, graph embedding is limited to comparing known images. It is necessary to classify unknown images as well. Thus, quadratic training combining graph embedding and Resnet is proposed, as shown in Figure 13. In the figure, the feature vector is extracted from Resnet, and FC is a fully connected layer with six output units. A weighted undirected graph is constructed based on image ID and category. Each node in the undirected graph is traversed, and a random walk is carried out to extract sequences, in which the size of the extraction window is set to 10. A neural network with a single hidden layer is trained to obtain embedding vectors with more separability; it is initialized using feature vectors from Resnet. The embedding vectors are used to fine-tune the fully connected network, and the final classification results are obtained.

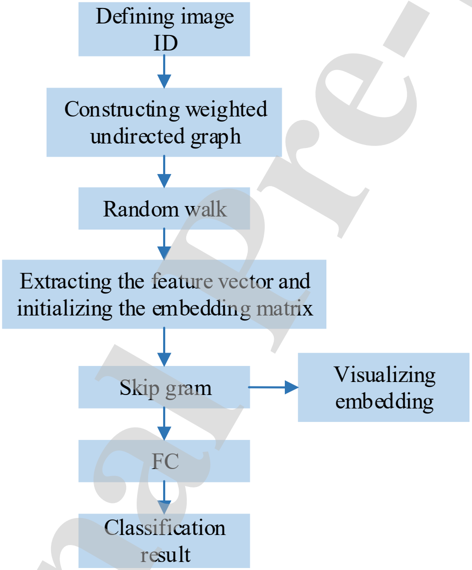

Defining image ID
Constructing weighted undirected graph
Random walk
Extracting the feature vector and initializing the embedding matrix
Skip gram
Visualizing embedding
FC
Classification result

Fig. 13. Flow chart of quadratic training.

3.3.2. Testing

The trained model based on quadratic training and Resnet is then tested. The parameters of Resnet and graph embedding are loaded to the testing model, and the classification results are obtained by inputting an image.

4. Experiments and results

4.1. Dataset

The US knee image dataset was provided by the affiliated hospital after secret data masking. It includes 600 US images that feature six types of manifestation: normal joints, effusion with synovial thickening, effusion without synovial thickening, cyst, tumor, and rheumatoid arthritis. At the same time, the labels of segmentation and classification are determined by the doctors. As deep learning needs a large amount of data, it is necessary to augment the data. Because the biological features of US knee joint images influence diagnosis, some image augmentation approaches that influence biological features, such as flipping or mirroring, are excluded. Three types of transformation, i.e., enlargement, reduction, and image distortion, are applied to augment the US dataset to 6000 images, and each category in the dataset is divided according to the ratio 7:3 between the training set and testing set.

4.2. Comparison of segmentation results

4.2.1. Segmentation Index

Because the effusion show individual and its category appears distinctive in segmentation result, we use mean intersection over union (MIoU) and pixel accuracy (PA) as the segmentation indexes instead of dice coefficient as follows:

$$\mathrm{MIoU}=\frac{1}{k}*\frac{1}{m}\sum_{i=1}^{k}\frac{\boldsymbol{P}_{ii}}{\sum_{j=0}^{k}\boldsymbol{P}_{ij}+\sum_{j=0}^{k}\boldsymbol{P}_{ji}\text{-}\boldsymbol{P}_{ii}} \tag{8}$$

$$\mathrm{PA}=\frac{\sum_{i=0}^{k}\boldsymbol{P}_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}\boldsymbol{P}_{ij}} \tag{9}$$

where MIoU is the average intersection-union ratio, k and m are the numbers of categories without background and image of category k, respectively, and k=1 in our study. $\boldsymbol{P}_{ij}$ means that pixels of category i are predicted to belong to category j. Clearly, $\boldsymbol{P}_{ii}$ denotes the pixels that are predicted correctly, and PA denotes all the pixel accuracies including target and background.

4.2.2. Comparison with Different Networks

An experiment is carried out that uses a backbone VGG network and transfer learning. The parameters of a classification network trained on ImageNet are taken as the initial parameters of the convolution layer for VGG network, which reduces the training time and improves segmentation accuracy. To decrease the sensitivity of the training process to the learning rate, the Adam optimizer is adopted in our study.

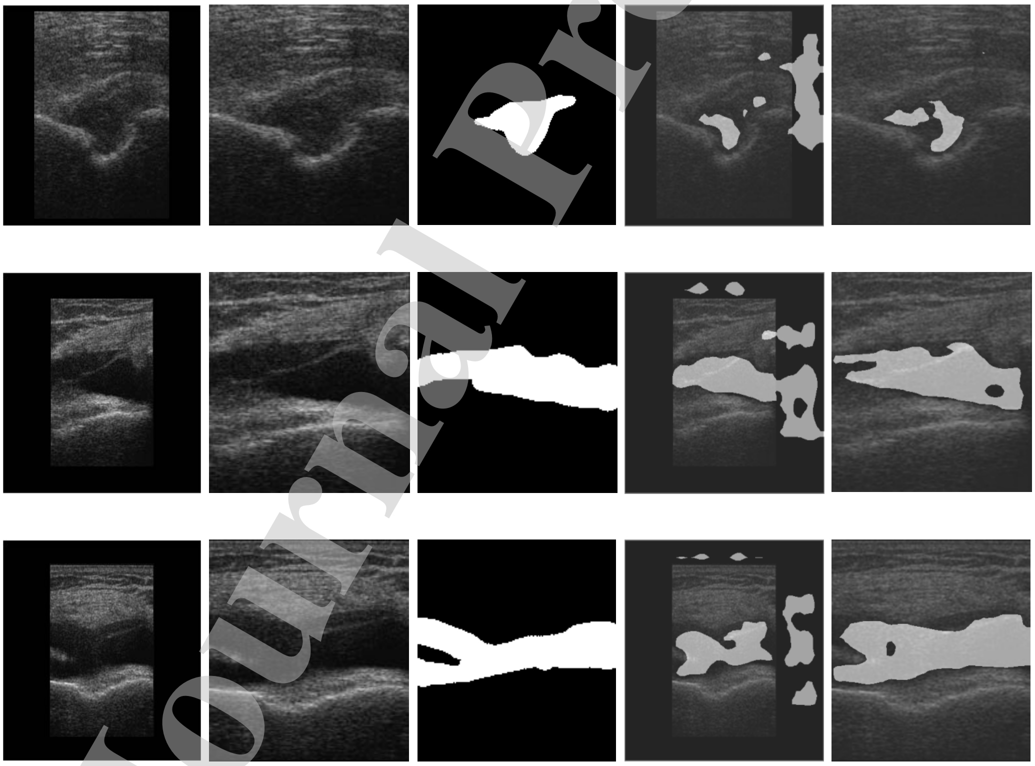
Based on the Deeplabv3 network, a comparison among non-dilated convolution, double-layer dilated convolution and multi-layer dilated convolution is carried out. The multi-layer dilated convolution may have three layers, four layers, or five layers. At the same time, the segmentation results with FCN and U-net models are also compared. The results are shown in Table 1, where the double-layer dilated convolution has the optimal result. For the multi-layer dilated convolution, the eigenvectors extracted are poor resulting in a poor MIoU. Double-layer dilated convolution can achieve a balance between large and small fields, which enlarges the receptive fields and ensures that the feature extraction is not over-grid. Therefore, double-layer dilated convolution is chosen to segment the US images.

Tab. 1 Segmentation index comparison

Network	Validation		Testing	
	MIoU	PA	MIoU	PA
U-net	0.801	0.9823	0.690	0.9528
FCN	0.771	0.9754	0.663	0.9387
No dilated convolution	0.919	0.9948	0.758	0.9437
Double-layer dilated convolution	0.930	0.9935	0.801	0.9574
Multi-layer dilated convolution	0.930	0.9936	0.775	0.9576

4.2.3. Comparison with/without Snake Algorithm

To confirm the necessity of using a snake algorithm for preprocessing, we use images with and without snake processing as the input of the Deeplabv3 network with double-layer dilated convolution; the segmentation results are shown in Figure 14. Figure 14 (d) is the segmentation result from inputting the original image, and figure 14 (e) is the segmentation result from inputting the snake preprocessed image. Clearly, snake preprocessing greatly reduces the mis-segmentation areas. The segmentation indexes with and without snake preprocessing are compared in Table 2. It can be seen that the indexes in the training set are not significantly improved, while the MIoU of the testing set is increased by over 10%.



(a) Original image (b)Image after snake (c) Label (d) Original result (e) Snake result

Fig. 14. Results of segmentation with and without snake algorithm.

Tab. 2 Comparison of segmentation indicators with and without snake algorithms

Network	Training		Testing	
	MIoU	PA	MIoU	PA
Without snake	0.820	0.9054	0.693	0.9242
With snake	0.930	0.9935	0.801	0.9574

4.2.4. Comparison to Multi-channel Stacking

To verify the effect of multi-channel stacking, the weight γ is tuned to find the optimal MIoU; the result is given in Figure 15, which shows an optimal weight γ of 0.51.

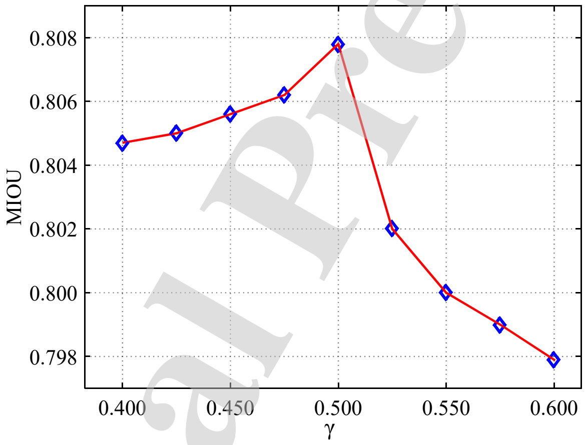


Fig. 15. Relationship between γ and MIoU. MIoU rises first and then decreases as γ increases. MIoU is highest when γ is 0.51.

Figure 16 presents the training result of the segmentation network with foreground, background, and their stacking. It can be observed that the image details are more accurate after merging the foreground and background. The segmentation indexes for different channels are listed in Table 3, which shows the MIoU increased by 1% on the training set, but only by 0.7% on the testing set. The enlarged single- and multi-channel images are shown in Figure 17. Comparing (a) and (b), we can see that the edges are more delicate and closer after combining multi-channel information. The segmentation index improves slightly, but the edge regions greatly improve with multiple channels.

Tab. 3 Comparison of segmentation results with different channels

Network	Training		Testing	
	MIoU	PA	MIoU	PA
Single channel	0.900	0.9535	0.801	0.9574
Multi-channel	0.940	0.9805	0.828	0.9600

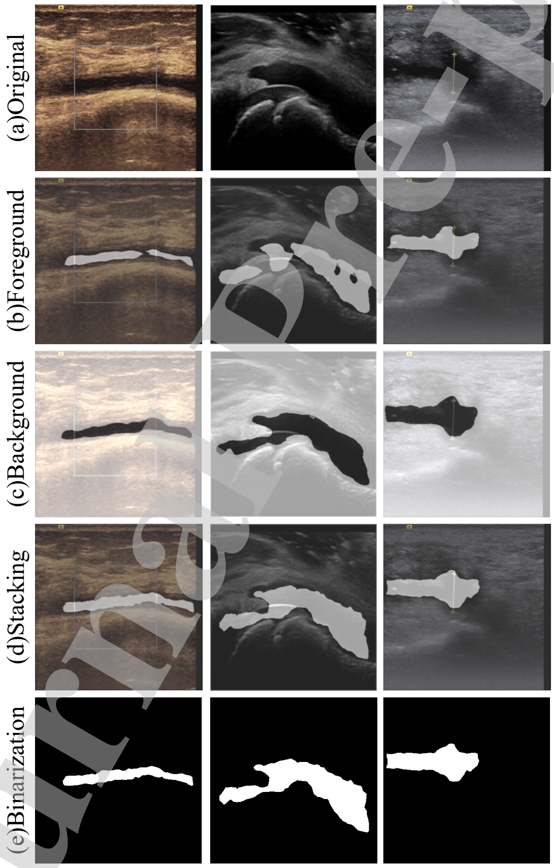
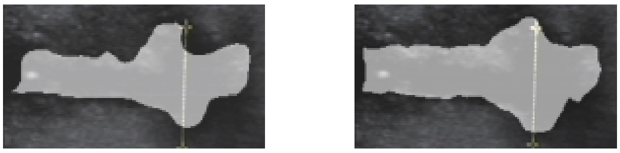


Fig. 16. Multi-channel stacking processes and segmentation results. (a) Original image. (b) Prediction of ROI-trained model based on effusion area, with the predicted results integrated into the original map (the gray part is the predicted results). (c) Prediction of ROI-trained model based on background. (d) Weighted summation of (b) and (c) using two-channel stacking. (e) Results after binarization of (d).



(a) Single channel (b) Multi-channel

Fig. 17. Enlarged image of single channel and multi-channel learning.

4.2.5. Comprehensive Results

Table 4 is a comprehensive comparison of image segmentation. We compared the deep learning network U-net, FCN, and the segmentation results after adding Snake algorithm and dilated convolution in different layers. Finally, we found that the algorithm we proposed (Snake + double-layer dilated convolution + multi-channel learning) has higher MIoU and PA than the results of other related works such as U-net, FCN, and etc. In Table 4, the accuracy of conventional deep learning U-net, FCN, or Double-layer dilated convolution is less than 0.70 in MIoU and less than 0.96 in PA. The accuracy is improved when combining the Snake and non-dilated convolution, or combining the Snake and Double-layer dilated convolution, or combining the Snake and Multi-layer dilated convolution. However, the segmentation can achieve the highest accuracy, as the PA is 0.828 and the PA is 0.96 when combining the Snake, double-layer dilated convolution, and multi-channel learning.

Tab. 4 Comparison of the network effects on the testing set

Snake processing	U-net	FCN	No dilated convolution	Double-layer dilated convolution	Multi-layer dilated convolution	Multi-channel stacking	Testing	
							MIoU	PA
	√						0.690	0.9528
		√					0.663	0.9387
√			√				0.758	0.9437
√				√			0.801	0.9574
				√			0.693	0.9542
√					√		0.775	0.9576
√				√		√	0.828	0.9600

4.3. Disease Classification

To simplify symptom classification in our study, the six types of ultrasound manifestation are transformed into two three-part classifications, as shown in Figure 18, where the "target area" refers to the region corresponding to 1 in the binary result of segmentation, that is, the liquid region of the lesion. On the contrary, the "targetless area" refers to the region corresponding to 0 in the binary result of segmentation, i.e., the background. Two results are obtained from segmentation, and each consists of three symptoms. The final classification network is constructed based on Resnet and graph embedding.

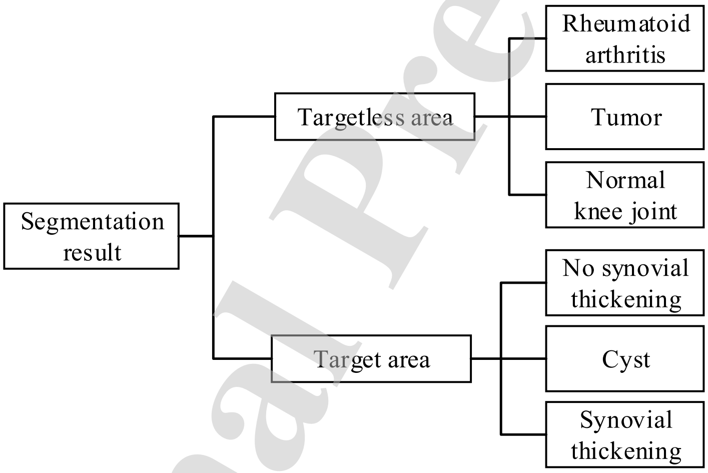


Fig. 18. Relationship between symptom identification and segmentation results.

4.3.1. Resnet Classification

Resnet is used as the basic framework in the experiment. The image features are extracted from the deep network to generate the feature vectors. Next, the feature vectors are fed into the fully connected layer to complete the classification. The ultimate precision attained is 0.701 on the training set and 0.618 on the testing set. To analyze the separability of feature vectors, which directly affects the classification results, we use T-SNE to reduce the dimension of feature vectors as shown in Figure. 19. The six colors represent different symptoms. Many feature vectors extracted by Resnet are inseparable after dimensionality reduction.

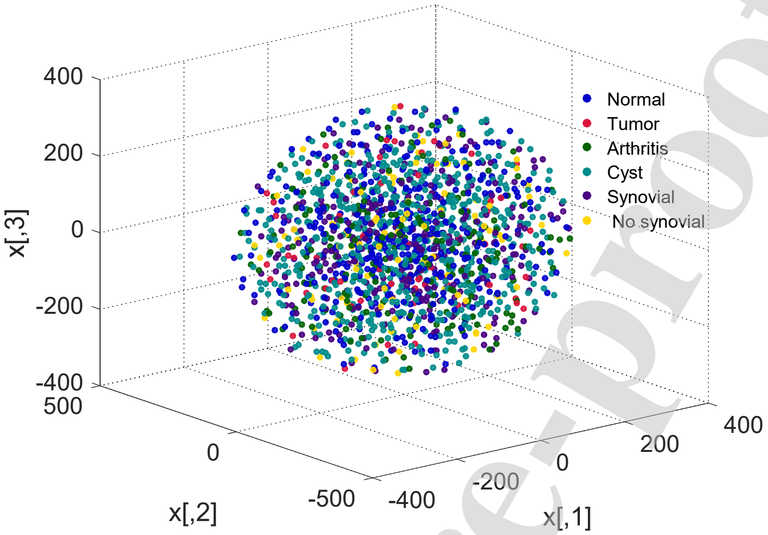


Fig. 19. Visual feature vectors.

4.3.2. Quadratic Training Classification

To obtain more separable embedding vectors of uncategorized images, we use a skip gram model with graph embedding for the additional training. First, the feature vectors of stage 5 in Resnet are extracted, and the dimension of feature vectors is (8, 8, 2048). The feature vectors are pooled into (1, 1, 2048) dimensions by an average pool layer. Next, the feature is used as the initial value of the embedding vector. The 2,048 dimension features are calculated as

$$\boldsymbol{O}=\boldsymbol{V}\cdot\boldsymbol{W}+\boldsymbol{b} \tag{10}$$

$$\boldsymbol{output}=\sigma(\boldsymbol{O}) \tag{11}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weighting matrix and bias matrix, respectively. $\boldsymbol{V}$ represents the 2,048 dimension constant feature vectors, and ***Output*** is the embedding vector trained by skip gram; the result is shown in Figure 20. It can be seen that the embedding vectors after quadratic training are more separable compared with the result shown in Figure 19. It is evident that the embedding vectors trained by graph embedding are more separable, and the whole training set is divided into six categories.

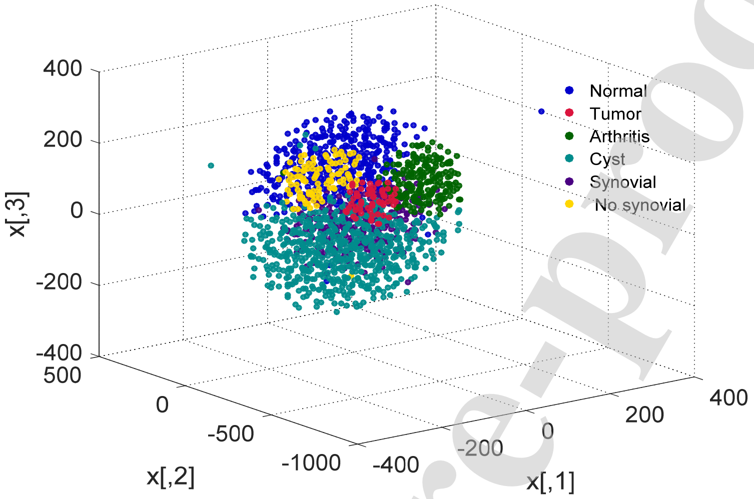
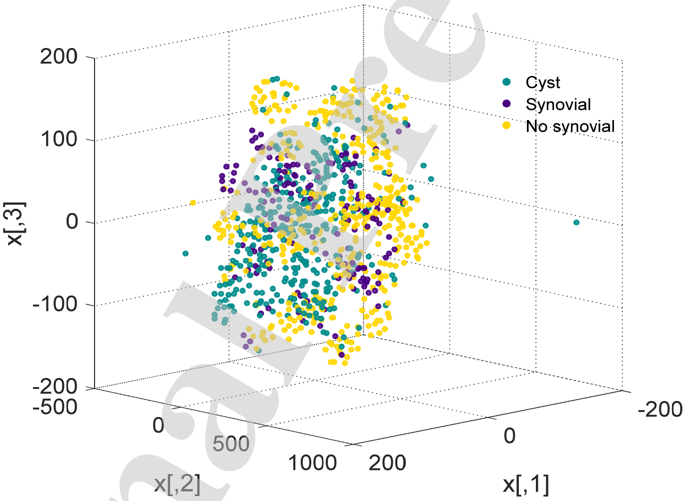
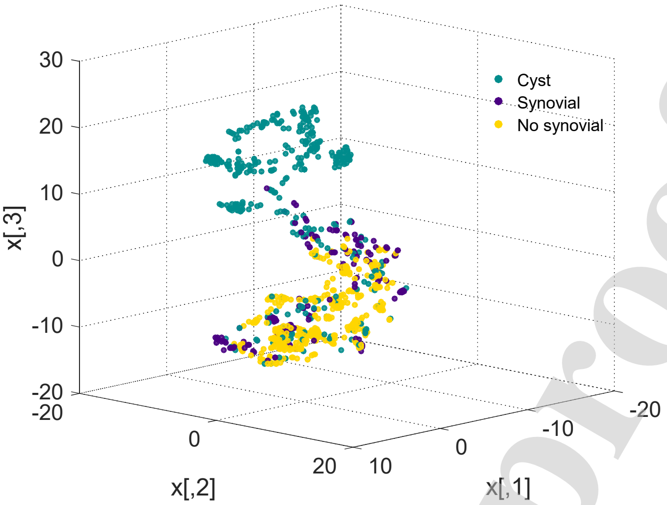


Fig. 20. Visual embedding vectors.

In our study, we perfectly divide original images into two targets by segmentation network, which are with and without the liquid lesion area. The T-SNE visualizations of images with and without segmented regions are shown in Figures 21 and 22. In figure 21, embedding vectors are dimensionally reduced by T-SNE, and it is clear that there is more separability with quadratic training. US images with target areas are classified into three categories. In figure 22, embedding vectors by quadratic training still shows more separability, and US images with targetless areas are classified into three categories. To obtain an optimized classification result for a small dataset, the Resnet is initialized by model parameters trained on ImageNet by transfer learning. Only the parameters of the fully connected layer are trained, and the other convolution layers for feature extraction are fixed. This method has higher accuracy, compared to the method of training the whole network. Table 5 presents a comparison of the classification algorithms. By comparing our proposed algorithms and the deep learning Resnet or VGG, we found that the classification accuracy is less than 0.7 in testing for the individual Resent, or individual VGG, or the combining Resnet + Graph embedding, or combining Segmentation + Resnet + Fine-tuning, while our cascaded algorithm of Segmentation + Resnet + Graph embedding + Fine-tuning can achieve 0.757 in testing. Clearly, Resnet combined with image segmentation and graph embedding has the highest classification accuracy for the testing set. The experiments prove that the conventional network is low in accuracy to classify the ultrasonic images while it is improved if we firstly segment the disease region and then classify the disease types by using our cascaded algorithm.

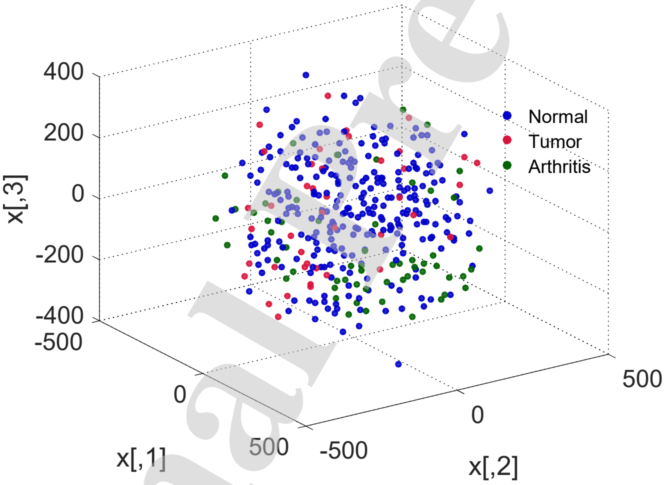


(a) Embedding vectors without quadratic training

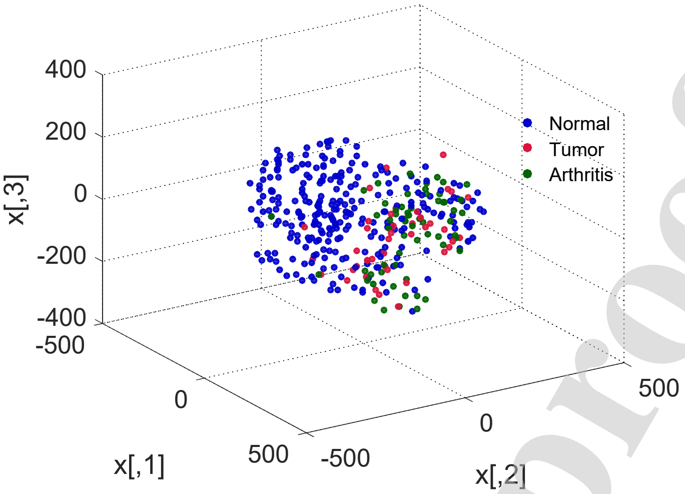


(b) Embedding vectors with quadratic training

Fig. 21. Visualization with target area images.



(a) Embedding vectors without quadratic training



(b) Embedding vectors with quadratic training

Fig. 22. Visualization with targetless area images.

Tab. 5 Comparison of classification algorithms.

Segmentation in table indicates that segmentation results are used, and **Fine-tuning** means that only the last FC layer is trained, while the parameters of other layers are fixed in the network.

Algorithms	Training	Testing
Resnet	0.685	0.533
Resnet + Graph embedding	0.733	0.589
Resnet + Fine-tuning	0.701	0.618
Segmentation + Resnet + Fine-tuning	0.812	0.669
Segmentation + Resnet + Graph embedding + Fine-tuning	**0.803**	**0.757**
VGG	0.633	0.495

5. Discussion and conclusion

Automatic recognition of medical images is an important way to assist doctors in diagnosing diseases. In this study, a deep neural network with dilated convolution, multi-channel stacking, and graph embedding was used to segment and recognize the lesion areas of US knee images. To reduce negative information such as similar color and vague edges in the images, a snake model was introduced to preprocess the images. The result makes the peripheral contour shrink to the rough target area and eliminates non-target areas that are easily mis-segmented. Experimental results on the same network show that the MIoU of the preprocessed image is 0.93 on the training set that is 11% higher than the treatment without snake processing, and the MIoU is 0.801 on the test set that is 10.8% higher than the treatment without snake processing. Based on a comparison with the segmentation results of FCN and U-net, Deeplabv3 has a higher accuracy and was selected as the basic network for segmentation. To reduce loss of information in the feature extraction of the segmentation network, dilated convolution was introduced to expand the receptive fields. At the same time, to balance the large and small fields, double-layer dilated convolution was selected through multiple dilated convolution layer comparisons. Finally, multi-channel stacking was adopted to reduce the hollows and edge discontinuity in segmentation results. Although the MIoU of multi-channel learning is less than 1% than the MIoU of single-channel learning, multi-channel is optimal for image segmentation.

To perform disease recognition, the US image dataset was first classified using Resnet, but this did not lead to satisfactory results. To improve classification accuracy, graph embedding was introduced to optimize the feature vectors, and skip gram was adopted to train the feature vectors. The results show that the feature vectors with quadratic training have more separable characteristics than those without quadratic training, after TSNE is used to reduce the dimension. The classification accuracy of the testing set is 0.757, which is 10.8% higher than that of the Resnet. The presented approaches can be beneficial to the other biomedical image tasks, such as tissue classification. However, when the proposed algorithmes are applied to the other biomedical image, the proposed network must be trained and the model parameters must be appropriately tuned according to the image features.

Snake preprocessing significantly improves the segmentation accuracy, but unfortunately its execution speed is slow. In the future research, a target detection algorithm should be investigated to extract the effusion area. In addition, the loss of information in the up-sampling needs to be reduced in the segmentation network.

ACKNOWLEDGMENT

This work was supported by the Key Research and Development Program of Guangdong Province (2020B090926001), National Natural Science Foundation of China (U1913215, U1713206), and Basic Research Plan of Shenzhen (JCYJ20170413112645981, GJHZ20180928154402130). The authors would like to thank the questions and suggestions of the anonymous reviewers who helped to improve this document. Knee image data used in this article were obtained from the Department of Orthopedics at Jinzhou Hospital in China.

REFERENCES

[1] C. Huang, L. Shan, H. C. Charles, W. Wirth, M. Niethammer and H. Zhu, "Diseased Region Detection of Longitudinal Knee Magnetic Resonance Imaging Data," IEEE Transactions on Medical Imaging, vol. 34, no. 9, pp. 1914-1927, Sept. 2015.

[2] X. Zhang, Y. Guo and P. Du, "The Contour Detection and Extraction for Medical Images of Knee Joint," 2011 5th International Conference on Bioinformatics and Biomedical Engineering, Wuhan, 2011, pp. 1-4.

[3] A. Gandhamal, S. Talbar, S. Gajre, A. F. M. Hani and D. Kumar, "A Generalized Contrast Enhancement Approach for Knee MR Images," *2016 International Conference on Signal and Information Processing (IConSIP)*, Vishnupuri, 2016, pp. 1-6.

[4] J. Antony, K. McGuinness, N. E. O'Connor and K. Moran, "Quantifying Radiographic Knee Osteoarthritis Severity Using Deep Convolutional Neural Networks," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, 2016, pp. 1195-1200.

[5] J. Fripp, S. Crozier, S. K. Warfield and S. Ourselin, "Automatic Segmentation and Quantitative Analysis of the Articular Cartilages From Magnetic Resonance Images of the Knee," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 55-64, Jan. 2010.

[6] J. Li, H. He, H. Huang and J. Lei, "Research on the Three Dimensional Reconstruction of Knee from CT Images," *2012 International Conference on Computer Science and Service System*, Nanjing, 2012, pp. 1911-1914.

[7] T. Maruyama and H. Yamamoto, "Reconstruction of Knee Joint Image from CT Data Using Positioning Doll," *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, Macau, 2015, pp. 1-6.

[8] Z. Zhou, G. Zhao, R. Kijowski, *et al.* "Deep Convolutional Neural Network for Segmentation of Knee Joint Anatomy," *Magnetic Resonance in Medicine*, 2018(Pt 2).

[9] B. N. Bs, V. Pedoia and S. Majumdar. "Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry," *Radiology*, 2018, 288(1): 172322.

[10] P. R. Desai and I. Hacihaliloglu, "Enhancement and Automated Segmentation of Ultrasound Knee Cartilage for Early Diagnosis of Knee Osteoarthritis," *2018 IEEE 15th International Symposium on Biomedical Imaging* (*ISBI 2018*), Washington, DC, 2018, pp. 1471-1474.

[11] A. Faisal, S. C. Ng, S. L. Goh, *et al.* "Multiple LREK Active Contours for Knee Meniscus Ultrasound Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 2162-2171, 2015.

[12] Y. Zhang, B. J. Matuszewski, L. K. Shark, *et al.* "Medical Image Segmentation Using New Hybrid Level-Set Method," *Fifth International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics*, 2008, pp. 71-76.

[12] Y. Lei, X. Zhao , *et al.* " Cirrhosis recognition of liver ultrasound images based on SVM and uniform LBP feature," *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018.

[13] M. Moradi, P. Abolmaesumi, et al. “Augmenting Detection of Prostate Cancer in Transrectal Ultrasound Images Using SVM and RF Time Series,” *IEEE Transactions on Biomedical Engineering, 2008*, pp. 2214-2224

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks,” *International Conference on Neural Information Processing Systems*, 2012, pp. 1097-1105.

[14] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ArXiv preprint arXiv, 2014, pp. 1549-1556.

[15] C. Szegedy, W. Liu, Y. Jia, *et al*. “Going Deeper with Convolutions,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[16] K. He, X. Zhang, S. Ren, *et al*. “Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[17] H. Noh, S. Hong and B. Han. “Learning Deconvolution Network for Semantic Segmentation,” *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520-1528.

[18] V. Badrinarayanan, A. Kendall and R. Cipolla. “Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation,” ArXiv preprint arXiv: 1511.00561, 2015.

[19] J. Long, E. Shelhamer, T. Darrell. “Fully Convolutional Networks for Semantic Segmentation,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[20] H. Zhao, J. Shi, X. Qi, *et al*. “Pyramid Scene Parsing Network,” *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881-2890.

[21] L. C. Chen, G. Papandreou, I. Kokkinos, *et al*. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” ArXiv preprint arXiv: 1412.7062, 2014.

[22] L. C. Chen, G. Papandreou, I. Kokkinos, *et al*. “Deeplabv3: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2016, pp. 834-848.

[23] L. C. Chen, G. Papandreou, F. Schroff, *et al*. “Rethinking Atrous Convolution for Semantic Image Segmentation,” ArXiv preprint arXiv: 1706.05587, 2017.

[24] L. C. Chen, Y. Zhu, G. Papandreou, *et al*. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," ArXiv preprint arXiv: 1802.02611, 2018.

[25] A. BenTaieb and G. Hamarneh. "Topology Aware Fully Convolutional Networks for Histology Gland Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2016, pp. 460-468.

[26] E. S. Samundeeswari, P. K. Saranya and R. Manavalan. "Segmentation of Breast Ultrasound Image Using Regularized K-Means (ReKM) Clustering," *International Conference on Wireless Communications*, 2016.

[27] O. Ronneberger, P. Fischer and T. Brox. "U-net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2015, pp. 34-241.

[28] F. Milletari, N. Navab and S. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, 2016, pp. 565-571.

[29] L. Bi, J. Kim, A. Kumar, *et al*. "Stacked Fully Convolutional Networks with Multi-channel Learning: Application to Medical Image Segmentation," *The Visual Computer*, vol. 33, no. 6, pp. 1061-1071, 2017.

[30] L. Wu, Y. Xin, S. Li, *et al*. "Cascaded Fully Convolutional Networks for Automatic Prenatal Ultrasound Image Segmentation," *IEEE International Symposium on Biomedical Imaging*, 2017.

[31] P. Goyal, E. Ferrara. "Graph Embedding Techniques, Applications, and Performance: A Survey," *Knowledge-Based Systems*, 2018, pp. 78-94.

[32] B. Perozzi, R. Al-Rfou and S. Skiena. "Deepwalk: Online Learning of Social Representations," *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701-710.

[33] A. Grover and J. Leskovec. "node2vec: Scalable Feature Learning for Networks," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855-864.

[34] J. Tang, M. Qu, M. Wang, et al. "Line: Large-scale Information Network Embedding," Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015, pp. 1067-1077.

Segmentation and Classification of Ultrasound Knee Joint Images via Deep Learning

Highlights

- A multi-channel learning model is proposed to improve segmentation result of lesion areas.

- Graph embedding method in natural language processing is applied to disease recognition.

- A novel optimization network based on quadratic training of feature vector is proposed.

Credit author statement

Zhili Long: Conceptualization, Methodology
Xiaobing Zhang: Data processing, Writing- Original draft preparation
Cong Li: Visualization, Investigation
Jin Niu: Writing- Reviewing and Editing
Xiaojun Wu: Validation
Zuohua Li: Supervision

Conflict of interest

We declare that we have no conflicts of interest to this work.

Zuohua Li