IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 February 2023, accepted 1 March 2023, date of publication 6 March 2023, date of current version 10 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3252893*

RESEARCH ARTICLE

# Chest X-Ray Quality Assessment Method With Medical Domain Knowledge Fusion

**SHANSHAN DU[1], YINGWEN WANG[2], XINYU HUANG[3], RUI-WEI ZHAO[4], XIAOBO ZHANG[5], RUI FENG[3], QUANLI SHEN[6], AND JIAN QIU ZHANG[1], (Senior Member, IEEE)**

[1]School of Information Science and Technology, Fudan University, Shanghai 200433, China
[2]Nursing Department, Children's Hospital of Fudan University, Shanghai 201102, China
[3]School of Computer Science, Fudan University, Shanghai 200433, China
[4]Academy of Engineering and Technology, Fudan University, Shanghai 200433, China
[5]Respiratory Department, Children's Hospital of Fudan University, Shanghai 201102, China
[6]Department of Radiology, Children's Hospital of Fudan University, Shanghai 201102, China

Corresponding author: Quanli Shen (qlshen8@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62172101, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 22511106003 and Grant 20511100800.

**ABSTRACT** Digital X-ray radiography is widely used in clinical diagnosis. High quality chest X-ray is conducive to the accurate diagnosis of diseases by clinicians. However, the quality assessment of the chest X-ray images mainly depends on the subjective evaluation of doctors, and the results are influenced by the skill level and experience of the evaluators, involving many issues such as heavy workload and various uncertain factors in such subjective judgment. In this paper, we propose a chest X-ray quality assessment method that combines image-text contrastive learning with medical domain knowledge fusion. Based on pretraining the model from contrastive text-image pairs, large-scale real clinical chest X-ray and diagnostic report text information are fused, and the model is fine-tuned to achieve cross domain transfer learning. While improving the prediction accuracy of the algorithm, the cost of massive sample data annotation is avoided. The local visual patch features of the X-ray images are aligned with multiple text features to ensure that the visual features contain more fine-grained image information. Theoretical analysis and experimental results show that the contrastive learning algorithm based on the fusion of triplet information in medical knowledge graph and chest X-ray multi-modal data has achieved good performance in terms of accuracy. In addition, the method proposed in this paper can be easily extended to complete other tasks such as medical image multi-lesion segmentation and disease progression prediction.

**INDEX TERMS** Chest X-ray, image quality assessment, no reference assessment, image-text contrastive learning, knowledge graph.

## I. INTRODUCTION

Digital X-ray radiography, boasting the advantages of low radiation dose, fast imaging speed and high image definition, is a common imaging method widely used in clinical diagnosis [1]. During the chest X-ray imaging process, there are many factors, e.g., patient's position, breathing status, exposure, irradiation field, center line, projection angle, etc., that will affect the imaging quality and directly influence the doctor's diagnosis of the disease. High quality chest X-ray images are helpful for the clinicians to accurately diagnose the diseases, while low quality X-ray images are likely to cause misdiagnosis or missed diagnosis [2]. At present, the quality judgment of the chest X-ray images mainly depends on the subjective evaluation from the radiologists. The results depend on the skill level and experience of the evaluators. The evaluation process requires high concentration and heavy workload. The image quality assessment method based on machine recognition can achieve objective evaluation of chest X-ray image quality, avoid subjective uncertainties, and effectively improve the efficiency.

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

Image Quality Assessment (IQA) is one of the basic technologies in image processing [3]. It is widely used in algorithm design and analysis, system performance evaluation, etc., and can help complete image super-resolution, image restoration, evaluation and debugging of imaging product parameters and other tasks.

IQA can be divided into two methods: subjective quality assessment and objective quality assessment according to the evaluation method. The subjective IQA method judges the image quality through the observer's subjective scoring. The advantage is that it can truly reflect people's subjective visual experience. The results are direct, accurate and reliable. Mean Opinion Score (MOS) and Differential Mean Opinion Score (DMOS) are the main forms of obtaining subjective scores [4]. However, this method is usually affected by objective factors such as the observation environment and the number of experimenters, thus having many limitations. The objective IQA method establishes a mathematical model based on the visual system of the human eye and evaluates or scores the images to be tested. It has the advantages of batch processing and reproducible results, which is easier to be applied in various scenarios.

Objective image quality assessment methods can be divided into Full-Reference Image Quality Assessment (FR-IQA), Reduced-Reference Image Quality Assessment (RR-IQA) and No-Reference Image Quality Assessment (NR-IQA) according to their processing methods. Among them, FR-IQA algorithms rely entirely on the reference image and evaluate the image by detecting the difference between the distorted image and its corresponding original undistorted image. There are already many FR-IQA methods, such as SSIM [5], MS-SSIM [6], VIF [7], IFC [8], FSIM [9], etc. Due to the existence of reference images, FR-IQA can usually accurately measure the quality of distorted images. However, their usefulness is limited due to the difficulty in obtaining reference images in practical applications. RR-IQA algorithms utilize the representative information of the reference image and its distorted image (e.g., some statistical features of the reference image, such as power spectrum). Image quality evaluation is conducted by measuring the similarity of these features in the reference image and the distorted image [10], [11], [12], [13]. NR-IQA algorithms evaluate the image quality without using the original reference image. In terms of practicality, NR-IQA methods are widely used because no information about the reference image is required [14], [15], [16], [17]. So far, designing an effective NR-IQA method remains a challenging research topic.

Since there is often no gold standard image in medical images, the quality assessment is mainly based on NR-IQA. Early NR-IQA methods mainly used handcrafted features, including lots of local features and global features. However, its limited feature expression ability resulted in low performance. With the rapid development of deep learning, deep features are increasingly used in NR-IQA. The performance of deep features far exceeds that of handcrafted features.

Although NR-IQA has made a lot of progress, however, due to the complexity of real clinical image data and many factors causing low quality, there are still many problems in directly applying it to the quality assessment of chest X-rays. Firstly, the most used NR-IQA data sets such as LIVE, TID2008, TID2013 are either natural distortion data or artificially simulated distortion data. Their data distribution and characteristics are far from real medical imaging data. Secondly, the accuracy of algorithm prediction is mainly measured based on the correlation with the subjective scoring of the human eye, which cannot meet the relevant requirements of the medical domain for medical image quality [18]. Finally, most of the existing methods use quality assessment methods based on the chest X-ray image features [3], [19], [20], and use image segmentation algorithms to semantically segment the diagnostic regions, and then use classification algorithms to judge image quality. However, such methods require large amount of annotated chest X-ray data, which brings huge data annotation workload to doctors and challenges the current data-driven deep learning methods.

At present, chest X-ray quality assessment methods that only rely on medical image feature learning still rely heavily on expensive or expert-knowledge supported datasets. The collation of these data requires heavy work in data collection, sampling, and manual labeling, which it is difficult to scale up. This expensive data collation process limits the size of the data set, which in turn hinders the effective improvement of the performance of the algorithm model. In recent years, with the continuous development of transfer learning technology, a number of unsupervised multimodal pre-training models using image-text pairs have emerged, and have achieved excellent performance in downstream tasks such as image classification and image segmentation [21], [22]. Considering that each patient's chest X-ray image will have a corresponding diagnosis report and quality control report, it is possible to construct the X-ray image-text pairs by performing natural language processing on the diagnosis report and quality control report and obtain the knowledge descriptions related to the quality of chest X-ray images. Then they can be fine-tuned based on the CLIP pre-trained model to obtain better performance. In addition, sicne the text feature has a certain generalization ability, it can effectively improve the generalization performance of the model.

Each single chest X-ray images may have many quality problems. For example, a chest X-ray image may have two quality problems of uneven clavicle and abnormal external objects at the same time. Thus, there will be multiple quality judgment rules, and every quality judgment rule is related to the local area of the image, leading to a typical multi-label image feature learning problem. Using the fine-grained comparative learning method of local image features in DualCoOp [23] can further improve the performance.

Therefore, inspired by the multimodal pre-training big models CLIP [21] and ALIGN [22], we propose a chest X-ray quality assessment method that organically combines image-text contrastive learning with medical domain knowledge fusion. First, an algorithm framework for chest X-ray quality assessment is presented. It achieves cross-domain transfer learning by fusing large scale real clinical chest X-ray images and diagnosis report text information based on Contrastive Language-Image Pre-training (CLIP) and model fine-tuning. While improving the prediction accuracy of the algorithm, the cost of massive data labeling is avoided, and the local visual patch features of the X-ray image are aligned with multiple text features to ensure that the visual features contain more fine-grained image information. The contribution of this paper can be summarized as:

1) A medical image quality assessment method integrating medical domain knowledge is proposed. The text data from chest X-ray image diagnosis reports and quality control reports are converted into triplet information through knowledge extraction and knowledge fusion, which are used as the guidance information from medical domain knowledge to medical images in the process of model training.
2) Exploiting the text annotation data of large-scale medical images together with their corresponding diagnostic reports and quality control reports, cross domain transfer learning is accomplished by fine-tuning the pre-trained model based on the contrastive text-image pairs, which effectively help overcome the problem of insufficient training data and avoid the heavy manual data labeling work.
3) By aligning local visual patch features of the X-ray image to multiple text features, the visual features are able to contain more fine-grained image information, so that the problem that global visual information cannot reflect the local image quality for disease diagnosis is addressed.

## II. RELATED WORK

### A. FACTORS AFFECTING THE DISQUALIFICATION OF THE CHEST X-RAY IMAGESS

According to the working experience and technical guidelines of radiologists [18], high-quality chest X-ray images should meet the following conditions, as illustrated in Figure 1: 1) The image should have no artifacts or severe image noise. Bad cases include abnormal external objects not belonging to the body (except those that cannot be removed), motion blur, post-processing, and equipment artifacts, etc. No granular noise should be visible to the naked eye in soft tissue density areas. 2) The position of the patient should be correct and appropriate, and 80% of the scapula in the case should be moved outside the lung field. (3) Both sides of the thorax should be symmetrical; the double clavicles should be flatp; and the sternoclavicular joints should be symmetrical. (4) The irradiation field should be properly selected, and there should

**FIGURE 1.** Example of qualified chest X-ray image.

be no case where chest tissue cannot be displayed due to occluded cuts (including both lung apexes, the lateral edge of the ribs above the diaphragm, and the double costophrenic angle). (5) The chest projection should be in the center of the image. The upper part of the chest X-ray should include the lung apices on both sides, and a 3-5 cm empty exposure area can be seen above the soft tissue shadow of the shoulder; the lower part should include the bilateral costophrenic angle and about 1-3 cm below; the outer sides of the chest should include the outer edge of the ribs and the soft tissue of the chest wall. (6) The bilateral lung fields, trachea and adjacent bronchi, heart and aortic border, diaphragm and bilateral costophrenic angles, lung fields and mediastinum behind the heart shadow should be clearly displayed. (7) The level of lung field and mediastinum, lung field and chest wall, lung field and shoulder soft tissue should be distinguishable. The structure of lung field texture should be clear and sharp, mediastinal soft tissue can be vaguely distinguished, and lung texture overlapping with heart shadow can be clearly displayed. (8) There should be no image blur caused by patient movement, and there should be no visible breathing, heartbeat motion blur, or diaphragm ghosting in the diagnostic area.

According to the above requirements, we identified 13 common problems in chest X-ray image quality assessment. As shown in Figure 2, we provide 13 common unqualified chest X-ray image examples with comparisons, including (a) overlapping of scapula and lung field, (b) misalignment of clavicles on both sides, (c) inconsistent clavicle height on both sides, (d) asymmetrical sternoclavicular joints, (e) unclear lung apex, (f) lung apex not included, (g) too little empty exposure above the shoulders, (h) diaphragm not included, (i) costophrenic angle not included, (j) asymmetrical ribcage on both sides, (k) asymmetrical lung fields on both sides, (l) existence of abnormal objects, (m) motion blur. The left part of each image is the original image, and the right part is the unqualified image. Figure 2-(n)

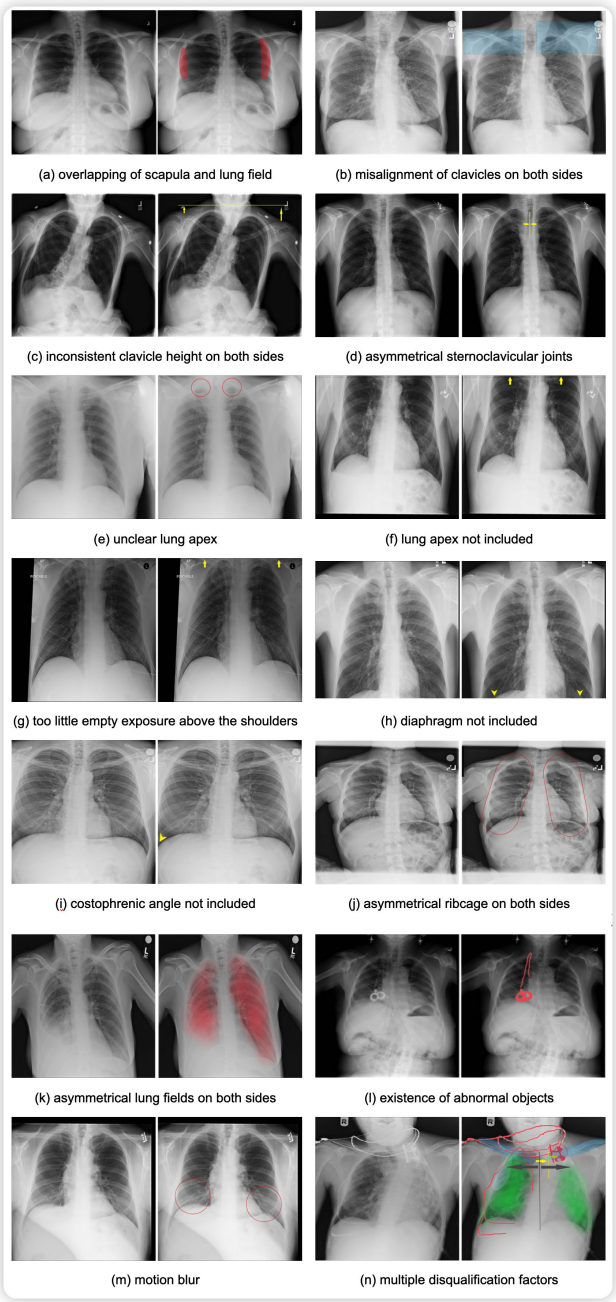



**FIGURE 2.** Examples of disqualified chest X-ray images: (a) overlapping of scapula and lung field, (b) misalignment of clavicles on both sides, (c) inconsistent clavicle height on both sides, (d) asymmetrical sternoclavicular joints, (e) unclear lung apex, (f) lung apex not included, (g) too little empty exposure above the shoulders, (h) diaphragm not included, (i) costophrenic angle not included, (j) asymmetrical ribcage on both sides, (k) asymmetrical lung fields on both sides, (l) existence of abnormal objects, (m) motion blur; (n) multiple disqualification factors. Marks on the images: lake blue indicates that the scapula overlaps with the lung field; sky blue indicates that the clavicles on both sides are not in the same position and height; yellow indicates asymmetry of the sternoclavicular joints on both sides; dark red indicates that the apex of the lung is not clear; gray indicates thoracic asymmetry; green indicates lung field asymmetry; pink indicates abnormal objects.

contain 7 disqualification factors; they are overlapping of scapula and lung field, misalignment of clavicles on both sides, inconsistent clavicle height on both sides, asymmetrical sternoclavicular joints, unclear lung apex, asymmetrical ribcage on both sides, asymmetrical lung fields on both sides, existence of abnormal objects, which are marked in different colors.

### B. NO-REFERENCE IMAGE QUALITY ASSESSMENT METHODS

No-Reference Image Quality Assessment (NR-IQA) is an important type of objective quality evaluation methods. Since no original reference image is needed, it has a broad application prospect. Feichtenhofer et al. proposed a sharpness measurement method based on the statistical analysis of local edge gradients for the no-reference assessment of distorted images with blurs and noises [24]. Mittal et al. did not use subjective opinion scores for training, but performed the NR-IQA by calculating locally normalized brightness coefficients in the spatial domain, which is more competitive than the two methods of peak signal to noise ratio and structural similarity [16]. Min et al. proposed a NR-IQA framework based on pseudo reference images, which generated pseudo reference images as new reference images for the distorted images [25].

In recent years, with the continuous development of deep learning technology, a variety of NR-IQA methods based on depth learning have been proposed. Kang et al. first proposed to use CNN to effectively measure the distortion degree of local image regions, and the experimental results showed that the performance was better than the traditional methods [26]. Gu et al. proposed a vector regression framework for NR-IQA, which estimated the trust score vector by CNNs and improved the evaluation performance by using object-oriented pooling [27]. Gao et al. used two objective image quality assessment indicators of peak signal to noise ratio and structural similarity, instead of subjective scoring, to train the network for the lack of CT annotation data [28]. The end-to-end NR-IQA methods based on deep learning integrate feature extraction and fitting/regression into a unified framework and optimize them simultaneously, which are the current mainstream NR-IQA scheme. The main ideas include using GAN to restore the distorted image; using the generated restored image and distorted image to perform loss calculation and distance measurement; generating a quality score; and using the idea of rank learning to solve the problem of lack of large-scale data sets in IQA and to improve the accuracy of the model. In addition, the attention mechanism is used to improve the weight of the region of interest so as to assess the image quality. And the prior knowledge is learned through the meta-learning method to solve the problem that the scale of the no-reference image dataset is too small.

High quality medical images are the prerequisite for radiologists to accurately diagnose and treat diseases. Considering that deep learning has achieved remarkable results in object detection, organ segmentation and image classification, the quality assessment of medical images without reference

by deep learning has gradually attract more attention and become a research hotspot. For example, Eck et al. proposed a CT quality assessment algorithm to realize the NR-IQA under the premise of ensuring the detection rate of lesions [29]. Mortamet et al. proposed a method for fully automatic assessment of 3D MRI quality by analyzing the air background pattern in MRI [30]. Esses et al. used neural networks to assess the image quality of the whole image of T2-weighted MRI of the liver [31]. Wang et al. used the two-step convolution neural network to evaluate the image quality of the region of interest on MRI of liver sites [32]. Xiao-Qian et al. used neural networks to achieve the evaluation of DR chest radiographs into four grades of excellent, good, medium and poor [33]. Wu et al. realized the automatic quality assessment of prenatal fetal ultrasound images based on two-step convolution neural network [34]. Li et al. used the improved AlexNet to score the quality of CT images, proving the potential of CNN in no-reference quality assessment of medical images [35]. However, it was still unable to effectively solve the problem of lack of high-quality scoring data sets. Due to the particularity of pathological images, the assessment of medical image quality cannot be determined solely by the whole image alone, while the quality of local images for disease diagnosis is often more important. At present, most research work is based on the overall structure of medical images. Their used scale of quality assessment is relatively large, leaving no analysis of local focus on the images for disease diagnosis.

Currently, there are only a few research work on no-reference medical images quality assessment based on deep learning. One of the main reasons is that the network needs a large amount of manually scored data for training. However, it is time and labor consuming to perform such manual quality scoring and annotation on massive medical images. At the same time, the current mainstream data driven deep learning methods have achieved satisfactory results in many cases. In order to solve these problems, many scholars have carried out adding prior knowledge in the machine learning process to improve the performance of the algorithm model. For instance, logical rules [36], [37] or algebraic equations [38], [39] have been added as constraints of loss functions; feature representation of the neural network was enhanced by using the association information between instances in the form of a knowledge graph [40]. These methods have achieved better performance in image classification tasks [41], [42]. Increasing amount of research results show that the data and knowledge driven methods are playing an important role in more and more fields.

At the same time, the two-stage training paradigm of ''pretraining and fine-tuning'' has gradually become one of the mainstream learning schemes in deep learning. Pretraining on large-scale datasets can significantly improve the model performance and generalization ability. In fact, large-scale data not only help define the approximation of the target problem, but is also necessary to ensure asymptotic convergence [43]. In the field of visual-language pre-training (VLP), CLIP [21] and ALIGN [22] collected millions of image-text pairs for learning visual representation from natural language supervision, which has been proved to be transferable to various downstream tasks, such as vision and language [44], image [45] and video tasks [46]. They directly aligned the visual and linguistic features through image-text contrastive (ITC) loss, which can also extended to large-scale data sets with high training efficiency.

The multi-label image recognition task obtains a more comprehensive understanding of an image by identifying multiple semantic labels present in the image [47], [48], [49], [50]. Prompt learning provides an effective way to transfer pre-trained visual-language models to other tasks and has achieved success in many tasks [51], [52], [53], [54], [55], [56]. However, these methods mainly focus on matching a single label per image and cannot handle the case where there are multiple multi-labels for an image. DualCoOp [23] proposes a fine-grained contrastive learning method, which effectively improves the ability to recognize multiple objects in multi-label recognition tasks. Considering that each chest X-ray image may have multiple image quality factors as shown in Figure 2 (n), we can learn from its designing idea to realize the identification of multiple chest X-ray image quality factors.

To summarize, NR-IQA method is an important kind of objective quality assessment method, which has important application potential in the field of medical image quality assessment. And currently, the deep learning based NR-IQA has become a research hotspot. However, the existing research works suffer from the lack of high quality labeled data; the global feature evaluation method of medical images cannot reflect the data quality of local focusing information used by the doctors for disease diagnosis; and the end-to-end data driven learning methods lacks medical domain knowledge fusion, which cause the bottleneck in quality assessment. In this paper, we combine image-text contrastive learning with medical domain knowledge. By using large-scale chest X-ray images, diagnostic reports and quality control reports information, the pretrained model based on contrastive text-image pairs is fine-tuned to achieve cross domain transfer learning. The local visual patch features of X-ray images are aligned with multiple text features to ensure that the visual features contain more fine-grained image information.

## III. CHEST X-RAY QUALITY ASSESSMENT WITH MEDICAL KNOWLEDGE

Our proposed framework contains the chest X-ray quality control knowledge triplets, text encoder, and image encoder, as illustrated in Figure 1. First, in the training phase, the knowledge rules of chest X-ray standardized quality control [18] are converted to triplets of text description corresponding to X-ray images, which are denoted as $G_{pos} = (h, r, t)$ and treated as positive samples. And then the corresponding

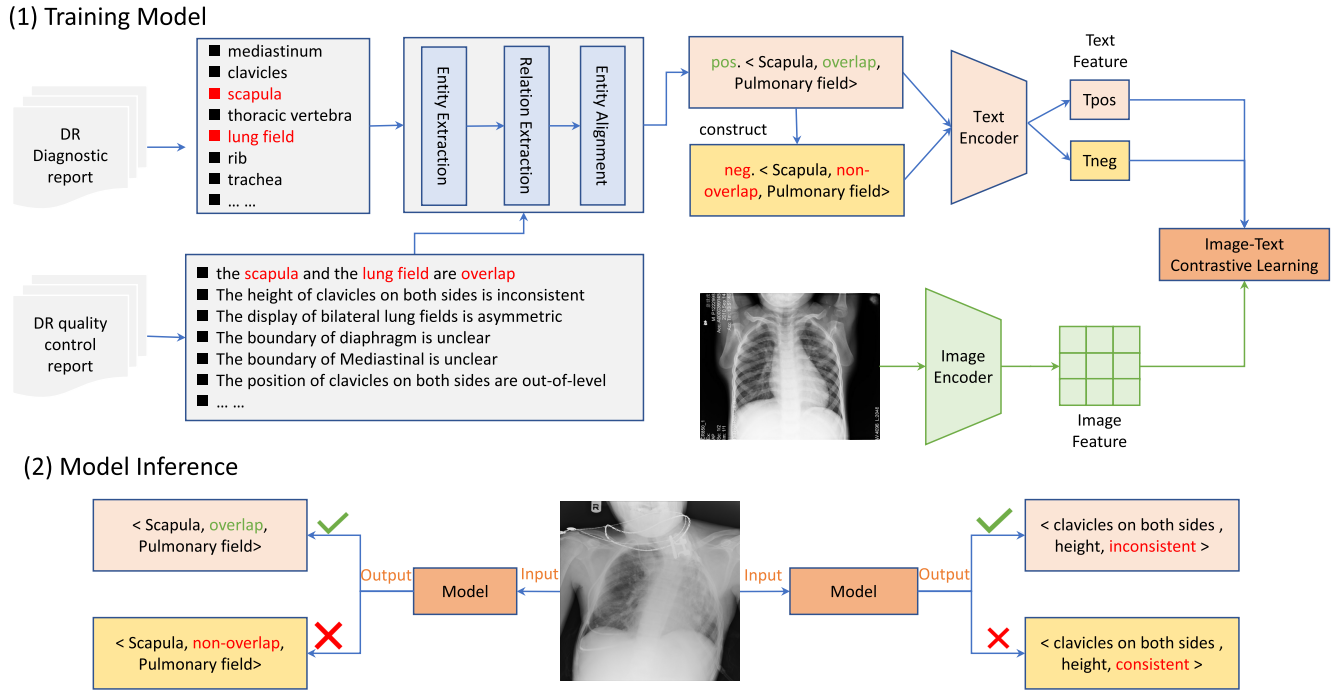



**FIGURE 3.** The overall framework of our proposed method. In the training phase, through entity extraction, entity relationship extraction and entity alignment, the text data from chest X-ray diagnosis report and quality control report are converted to the knowledge graph triplets for chest X-ray image quality assessment. The obtained triples (e.g., <scapula, overlap, lung field>) are treated as positive samples, whereas the corresponding negative samples (e.g., <shoulder blades, no overlap, lung fields>) are manually constructed. Image-text contrastive learning (ITC) is applied to the text encoder and image encoder to obtain text and image features respectively (Tpos, Tneg, I). The image-text contrastive loss is adopted to maximize the feature similarity of positive pairs. In the model inference phase, only input images are needed to feed into the model to obtain quality assessments for different aspects of the images. Such combination of knowledge graph triplets and ITC can effectively improve the accuracy of chest X-ray image quality assessment.

negative samples$G_{neg} = (h, r, t)$ are manually constructed. With the diagnostic report and quality control report paired to the chest X-ray image, we conduct entity extraction, entity relationship extraction and entity alignment process, so as to map the organ entity and its attribute/relationship in each report to the knowledge triplet information, and build the quality assessment triplets of each patient's chest X-ray image.

Image-text contrastive learning (ITC) is applied to text encoder and image encoder to respectively obtain text and image features. The standard vision Transformer (ViT) model [29] is used as the image encoder. The chest X-ray image $I \in R^{H\times W\times 1}$ is divided into $N$ pieces of $p \times p$ sized patches $X_p \in R^{N\times(P^2\times 1)}$, $N = \frac{H\times W}{p\times p}$. Then they are encoded into $N+1$ feature vector sequences $\{I_{gal}, I_1, I_2, \cdots, I_N\}$. Similarly, the positive and negative text samples are encoded into $\{T_{pos_{gal}}, T_{pos_1}, T_{pos_2}, \cdots, T_{pos_N}\}$, $\{T_{neg_{gal}}, T_{neg_1}, T_{neg_2}, \cdots, T_{neg_N}\}$ using the Transformer structure. Traditional image-text contrastive learning (ITC) uses global visual feature vectors to align visual language features with text features. Since each chest X-ray image contains multiple triples of text corresponding to different organ regions, we propose to use the patch level visual feature vectors of for more fine-grained feature alignment. Feature vector embeddings $f_{pos(i)} = g_i(I_i) \cdot g_t\left(T_{pos_{gal}}\right), f_{neg(i)} = g_i(I_i) \cdot g_t(T_{neg_{gal}})$ are constructed by inner product operation. In the training process, the feature vectors provided by the image encoder are expected to be as close as possible to the positive sample $T_pos$ and far away from the negative sample $T_neg$.

In the model inference phase, only input images are needed to feed into the model to obtain quality assessments for different aspects of the images. Combining knowledge graph triplets with ITC has the following two significant advantages: (1) redundant information in X-ray diagnosis report and X-ray quality control report can be removed, (2) medical domain knowledge can be naturally integrated with image-text contrastive learning model; thus effectively improving the model performance.

## IV. CHEST X-RAY QUALITY ASSESSMENT TRIPLET

### A. JOINT ENTITY RELATION EXTRACTION

Entities (e.g., mediastinum, scapula, lung field, trachea, etc.) and attributes (e.g., clear, unclear, overlapping, aligned, etc.) are respectively identified from the chest X-ray diagnosis report and quality control report. In this paper, the joint entity relationship extraction method [57] based on parameter sharing is applied to the chest X-ray diagnosis report to obtain the triplets $G_d(h_d, r_d, t_d)$, and to the quality control report to obtain the triplets $G_q(h_q, r_q, t_q)$.

Given a sentence $X_j$ in the training set $D$ and all possible triplets $T_i = \{(h, r, t)\}$ in the sentence, the optimization

objective of obtaining triplet entities and relations turns into the maximization of the likelihood function in the training set [57]:

$$\prod_{j=1}^{|D|}\left[\prod_{(h,r,t)\in T_j} p((h,r,t)|x_j)\right] \tag{1}$$

$$=\prod_{j=1}^{|D|}\left[\prod_{h\in T_j} p(h|x_j)\prod_{(r,t)\in T_j|s} p((r,t)|h,x_j)\right] \tag{2}$$

$$=\prod_{j=1}^{|D|}\left[\prod_{h\in T_j} p(h|x_j)\prod_{r\in T_j|h} p(t|h,x_j)\prod_{r\in R\setminus T_j|h} p_r(t_\varnothing|h,x_j)\right] \tag{3}$$

where $h \in T_j$ it an entity in the triplet $T_i$; $T_i|h$ represents the related triplets of the entity $h$ in $T_i$; $(r,t) \in T_j|h$ stands for a $(r,t)$ pair related to entity $h$ in the triplet $T_i$; $R$ is the set of all possible entity relationships; $R \setminus T_j|h$ represents other relationships associated with the entity $h$. $t_\varnothing$ denotes the empty entity.

The entity extraction model is used to find all possible entities in the sentence. Then for each found entity, the relationship extraction model is posed to find all its correlated relations and their corresponding entities. By formula (1), (2) and (3), multiple entity relation triples can be extracted in the sentence.

In this paper, the pretraining language representation model (BERT) [58] based on multi-layer bidirectional Transformer is used to conduct the feature extraction of sentence $X_j$. The sentence representation learning is completed by jointly mediating the context of each word. The model includes $N$ identical Transformer blocks $Trans(x)$, where $x$ represents the input vector. The specific calculation is:

$$h_0 = SW_h + W_p \tag{4}$$

$$h_\alpha = Trans(h_{\alpha-1}), \alpha \in [1,N] \tag{5}$$

where $S$ represents a word vector matrix in the input sentence; $W_h$ represents the word embedding matrix; $W_p$ denotes a position embedding matrix. $p$ is the position of a word in a sentence; $h_\alpha$ is an implicit state vector; and $N$ is the number of Transformer blocks. Since the input is a single sentence, piece-wise embedding is not used.

Head entity identification adopts the binary classification method to directly decode the above encoding results, so as to identify all possible head entities. We use a linear layer and sigmoid activation function to determine whether each token is a beginning token or end token of the head entity, as shown in the following formula:

$$P_i^{start\_h} = \delta(W_{start}X_i + b_{start}) \tag{6}$$

$$P_i^{end\_h} = \delta(W_{end}X_i + b_{end}) \tag{7}$$

where $P_i^{start\_h}$ and $p_i^{end\_h}$ denote the probability that the $i_{th}$ token serves as the head or tail entity in the input sentence, respectively. If it exceeds a certain threshold, the corresponding token represents a head entity. Based on the nearest matching principle, the identified start and end tokens are paired to obtain the candidate head entity set.

After identifying the head entity, joint recognition of the relationship and the tail entity is carried out. Here, a group of relationship-based tail entity recognition is implemented. The structure of the tail entity recognition layer is similar to that of the head entity recognition layer. The main difference lies in the input data. The input of the head entity recognition layer is the output of encoding layers, while the input of the tail entity recognition layer takes joint considerations on the head entity feature $v_{sub}^k$:

$$P_i^{start\_t} = \delta(W_{start}^r(X_i + v_{sub}^k) + b_{start}^r) \tag{8}$$

$$P_i^{end\_t} = \delta(W_{end}^r(X_i + v_{sub}^k) + b_{end}^r) \tag{9}$$

where $v_{sub}^k$ represents the vector average of all tokens contained in the $k_{th}$ candidate head entity.

### B. ENTITY ALIGNMENT

After the triplets $G_d(h_d, r_d, t_d)$ and $G_q(h_q, r_q, t_q)$ are obtained from the chest X-ray diagnostic report and quality control report respectively, they need to be aligned with the standard triplets $G_{pos}(h,r,t)$ constructed by the knowledge rules for the chest X-ray standardization quality control standards [4].

Without loss of generality, we use the interaction model based on BERT [59] to achieve alignment of the two triplets $G_d(h_d, r_d, t_d)$ and $G_{pos}(h,r,t)$. The alignment of $G_q(h_q, r_q, t_q)$ and $G_{pos}(h,r,t)$ adopts the same method.

The pretrained BERT model is used to calculate the name/description of the entity. For the given entity input, the corresponding value of CLS downstream classification task tag is taken and mapped by multi-layer perceptron (MLP). Given $e \in E$, the vector representation of the entity can be calculated by the following formula:

$$C(e) = MLP(CLS(e)) \tag{10}$$

For the given two entities in $G_q(h_q, r_q, t_q)$ and $G_{pos}(h,r,t)$, i.e., $h_{qi} \in h_q$ and $h_i \in h$, the vector representation $C(h_{qi})$ and $C(h_i)$ of the name/description of the two entities can be calculated using Equation (10). By calculating the cosine distance of the two vector representations, the similarity of the two entities can be obtained.

$C(h_{qi})$ and $C(h_i)$ are calculated respectively for the names/descriptions of similar nodes of entities in the two graphs to obtain two vector sets, i.e., $\{C(h_{qi})\}_{i=1}^{|N(h_q)|}$ and $\{C(h_i)\}_{i=1}^{|N(h)|}$. The similarity matrix of the two vector sets can be obtained by cosine similarity calculation. The calculation formula is as follows:

$$S_{ij} = \frac{C(h_{qi})\cdot C(h_i)}{\|C(H_{qi})\|\cdot\|C(h_i)\|} \tag{11}$$

Then the row direction and column direction of the matrix are aggregated respectively. In the process of the row direction aggregation of the matrix, the maximum pooling operation is carried out for each row. For vector

$S_i = \{S_{i0}, S_{i1}, \cdots, S_{in}\}$ in the $i_{th}$ row, considering the heterogeneity of the triplet, the neighbor entities of the two aligned entities are not identical. Therefore, the similarity between one of the neighbor entities in $h_{qi}$ and its most similar entity among the neighbor entities of $h_i$ is considered. Their the maximum value $S_i^{max}$ is taken, that is:

$$S_i^{max} = \max_{j=0}^{n}\{S_{i0}, S_{i1}, \cdots, S_{in}\} \tag{12}$$

where $n$ indicates the maximum number of neighbors.

The Gaussian kernel function is used for one-to-many mapping for $S_i^{max}$, and multiple mapping values are obtained to form a vector $K^r(S_i)$. We average the matrix $K^r(S)$ logarithmically in the column direction and obtain a vector of length $L$. The calculation process is as follows:

$$K_l(S_i^{max}) = exp\left[-\frac{(S_i^{max} - \mu_l)^2}{2\delta_l^2}\right] \tag{13}$$

$$K^r(S_i) = \left[K_1(S_i^{max}), \cdots, K_l(S_i^{max}), \cdots, K_L(S_i^{max})\right] \tag{13.1}$$

$$\phi^r(N(h_{qi}), N(h_i)) = \frac{1}{|N(h_{qi})|}\sum_{i=1}^{|N(h_{qi})|} logK^r(S_i) \tag{13.2}$$

where $L$ represents the number of Gaussian kernels, and $r$ represents row aggregation.

Column aggregation is completed in the same steps as above. Then, the two aggregation result vectors are added according to Equation (14) to obtain the neighbor view interactive similarity vector $\phi(N(h_{qi}), N(h_i))$:

$$\phi(N(h_{qi}), N(h_i)) = \phi^r(N(h_{qi}), N(h_i)) \oplus \phi^c(N(h_{qi}), N(h_i)) \tag{14}$$

where $\oplus$ represents the concatenation calculation.

For the two aligned entities $h_q$ and $h$, their corresponding neighbor triplets are $G_q(h_q, r_q, t_q)$ and $G_{pos}(h, r, t)$, respectively. If the tail entity $t_q$ and $t$ are similar, the relationship $r_q$ and $r$ are also similar. The mask matrix can be calculated to calibrate the neighbor entity similarity matrix.

The vector averages of $C(e)$ in the head entity set and the tail entity set of the entity relation are calculated respectively, and the vector representation of the entity relation is obtained by concatenating the these vectors. The similarity matrix $M$ is obtained according to the neighbor relation vector of entity $h_q$ and $h$, i.e., $M_{ij} = sim(C(r_q i), C(r_j))$, where $M_{ij}$ represents the cosine similarity between the entity $h_q$'s $i_{th}$ neighbor relation $r_{qi}$ and the entity $h$'s $j_t h$ neighbor relation $r_j$. The matrix $S$ is then modified by $M$, i.e., $S = S \otimes M$, where $\otimes$ stands for elementwise multiplication.

Similar to the construction of entity similarity matrix, the entity attribute similarity matrix can be obtained. And finally the entity attribute similarity vector is obtained. Based on this, the entity similarity vector, the neighbor entity similarity vector and the attribute similarity vector are concatenated to obtain the similarity vector of two triplet pairs. Then, the similarity score between the entities can be calculated using the multi-layer perceptron.

In the process of entity triplet alignment, the candidate aligned entities with the highest cosine similarity are first calculated according to the entity vector $C(e)$. Then the similarity scores of the candidate entities and entity $h$ are calculated respectively, and the results are sorted in descending order.

## V. IMAGE-TEXT CROSS-MODAL CONTRASTIVE LEARNING

After the triplets are obtained from the chest X-ray diagnostic report and quality control report and entity alignment is completed, a positive sample triplet description is obtained for the corresponding chest X-ray image. Then negative sample triplet description is constructed by changing entity relationships/attributes. The image and text encoding features are obtained by using feature extractors of different modalities, and then the improved fine-grained image-text contrastive function is used for cross-modal feature alignment.

### A. MULTI-MODAL FEATURE EXTRACTION

For a given chest X-ray image and its corresponding positive/negative sample triplets, image features and text features are extracted by the image encoder and text encoder respectively. Image encoder utilizes standard vision Transformer (ViT) model [60]. The chest X-ray image $I \in R^{H\times W\times 1}$ is divided into $N$ patches $X_p \in R^{N\times(P^2\times 1)}$, $N = \frac{H\times W}{p\times p}$ with resolution of $p\times p$, and then encoded into a sequence of visual feature vectors $\{I_{gal}, T_1, T_2, \cdots, T_N\}$, with sequence length of N+1, where $\{I_{gal}\}$ is used to represent the global features of the image. The text encoder uses the standard Transformer model to encode the triplets of text representing positive and negative samples as $\{T_{pos_{gal}}, T_{pos_1}, T_{pos_2}, \cdots, T_{pos_N}\}$, $\{T_{neg_{gal}}, T_{neg_1}, T_{neg_2}, \cdots, T_{neg_N}\}$ respectively. Both image encoder and text encoder can use the feature encoder in pretrained CLIP model. Since it has been pretrained on large-scale natural image-text pairs, the model boasts strong image-text matching ability. Based on this, parameters of the model are fine-tuned, which can not only retain the original strong cross-modal matching ability, but also adapt the model to the data domain of chest medical image.

### B. CROSS-MODAL FEATURE ALIGNMENT

For the traditional cross-modal contrastive learning of image-text, such as CLIP shown in Figure 4 (a), they generally align the global feature $\{I_{gal}\}$ of the image with the global feature $\{T_{pos_{gal}}\}$ of the positive sample text, while the global feature $\{I_{gal}\}$ of image and the global feature $\{T_{neg_{gal}}\}$ of negative sample text are pulled apart. Specifically, the global feature vector of the image and the feature vector of the text are mapped to the same dimension by the corresponding projection layers $g_i$ and $g_t$ respectively, and the embedded vector features are constructed by the inner product operation: $f_{pos} = g_i\left(I_{gal}\right)\cdot g_t\left(T_{pos_{gal}}\right), f_{neg} = g_i(I_{gal})\cdot g_t(T_{neg_{gal}})$. Then, image-text contrastive loss function (ITC) is used to optimize

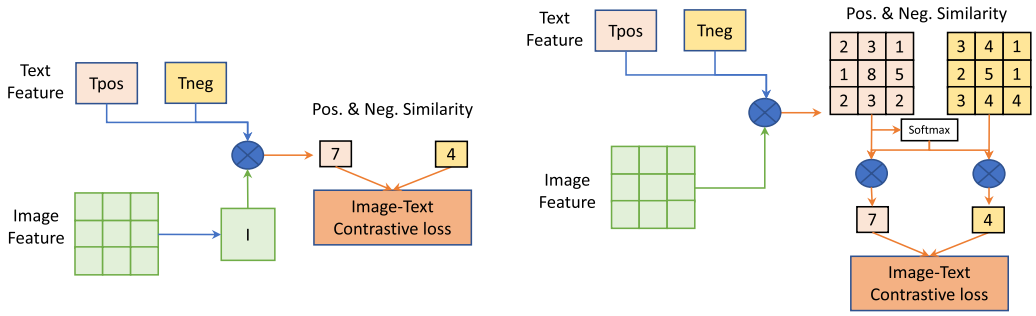



**FIGURE 4.** Illustration of the different image-text cross-modal contrastive learning methods. (a) Image-text cross-modal contrastive learning. (b) Fine-grained image-text cross-modal contrastive learning, which is more suitable for the situation in this paper where each image has multiple quality failure factors corresponding to multiple different regions of the image, as shown in Figure 2 (n).

**TABLE 1.** Data distribution with factors influencing chest X-ray image quality.

| Influencing Factors | #Pos. | #Neg. |
|---|---|---|
| Overlapping of scapula and lung field | 1025 | 1000 |
| Misalignment of clavicles on both sides | 1012 | 1000 |
| Inconsistent clavicle height on both sides | 1009 | 1000 |
| Asymmetrical sternoclavicular joints | 995 | 1000 |
| Unclear lung apex | 978 | 1000 |
| Lung apex not included | 869 | 1000 |
| Too little empty exposure above the shoulders | 895 | 1000 |
| Diaphragm not included | 1056 | 1000 |
| Costophrenic angle not included | 975 | 1000 |
| Asymmetrical ribcage on both sides | 996 | 1000 |
| Asymmetrical lung fields on both sides | 879 | 1000 |
| Existence of abnormal objects | 1098 | 1000 |
| Motion blur | 985 | 1000 |

the model parameters:

$$Loss_{itc} = \frac{exp(f_{pos}/\pi)}{exp(f_{pos}/\pi + exp(f_{neg}/\pi))} \tag{15}$$

where $\pi$ is the learnable temperature coefficient. As each chest X-ray image corresponds to multiple positive/negative sample triplets, and different triplets contain different organ regions of the image, fine-grained image information will be lost when the global features of an image are aligned with the corresponding text features. Therefore, inspired by [23], we improve the traditional cross-modal contrastive learning of image-text and propose a fine-grained image-text alignment method.

Specifically, instead of aligning the global features of an image with multiple text features, we align all visual features of an image with multiple text features, since these visual features contain more fine-grained image information. As shown in Figure 4 (b), the visual feature vector sequence $\{I_1, I_2, \cdots, I_N\}$ apart from the global image feature and the feature vector of the text are mapped to the same dimension by the corresponding projection-layer $g_i$ and $g_t$ respectively, and the embedded vector features are constructed by inner product operation: $f_{pos(i)} = g_i(I_i) \cdot g_t\left(T_{pos_{gal}}\right), f_{neg(i)} = g_i(I_i) \cdot g_t(T_{neg_{gal}})$. The obtained embedded vector feature is also a vector sequence with length N. In order to obtain the feature vector of length 1 for each image-triplet, we embed positive sample into the feature vector sequence to aggregate the features of positive/negative sample embedding vector respectively:

$$f_{pos} = \sum_{i}^{N} Softmax(f_{pos(i)}) \cdot f_{pos(i)} \tag{16}$$

$$f_{neg} = \sum_{i}^{N} Softmax(f_{pos(i)}) \cdot f_{neg(i)} \tag{17}$$

$$Softmax(f_{pos(i)}) = \frac{exp(f_{pos(i)})}{\sum_j exp(f_{pos(j)})} \tag{18}$$

Finally, the image-text contrastive loss function (ITC) is also used to optimize the model parameters:

$$Loss_{itc} = \frac{exp(f_{pos}/\pi)}{exp(f_{pos})/\pi + exp(f_{neg}/\pi)} \tag{19}$$

## VI. EXPERIMENT AND ANALYSIS

### A. DATA PREPARATION

The experimental data used in this study is the large-scale chest X-ray image dataset ChestX-ray8 [61] released by NIH. The dataset contains 112,120 frontal view X-ray images from 30,805 patients. A total number of 1000 qualified image are selected from the dataset as negative set; while 4846 disqualified images are selected as positive set. The 13 disqualification factors are: overlapping of scapula and lung field, misalignment of clavicles on both sides, inconsistent clavicle height on both sides, asymmetrical sternoclavicular joints, unclear lung apex, lung apex not included, too little empty exposure above the shoulders, diaphragm not included, costophrenic angle not included, asymmetrical ribcage on both sides, asymmetrical lung fields on both sides, existence of abnormal objects, and motion blur. Finally, a total number of 5846 chest X-ray images are included in the experimental

**TABLE 2.** Chest X-ray image quality assessment results.

| Positive Sample Class | Precision | Recall | F1-Score | FPR |
|---|---|---|---|---|
| Overlapping of scapula and lung field | 0.954 | 0.974 | 0.964 | 0.0460 |
| Misalignment of clavicles on both sides | 0.957 | 0.958 | 0.957 | 0.0439 |
| Inconsistent clavicle height on both sides | 0.966 | 0.944 | 0.955 | 0.0348 |
| Asymmetrical sternoclavicular joints | 0.950 | 0.967 | 0.958 | 0.0491 |
| Unclear lung apex | 0.959 | 0.945 | 0.952 | 0.0406 |
| Lung apex not included | 0.972 | 0.951 | 0.961 | 0.0245 |
| Too little empty exposure above the shoulders | 0.922 | 0.971 | 0.946 | 0.0670 |
| Diaphragm not included | 0.926 | 0.959 | 0.942 | 0.0753 |
| Costophrenic angle not included | 0.970 | 0.941 | 0.956 | 0.0299 |
| Asymmetrical ribcage on both sides | 0.967 | 0.938 | 0.952 | 0.0341 |
| Asymmetrical lung fields on both sides | 0.958 | 0.975 | 0.966 | 0.0365 |
| Existence of abnormal objects | 0.897 | 0.954 | 0.924 | 0.1061 |
| Motion blur | 0.960 | 0.946 | 0.953 | 0.3960 |

**TABLE 3.** Comparative results on different features.

| Positive Sample Class | Feature Setting | Precision | Recall | F1-Score | FPR |
|---|---|---|---|---|---|
| Overlapping of scapula and lung field | Image Feature | 0.878 | 0.888 | 0.883 | 0.1238 |
| | Image Feature + Text Feature | 0.897 | 0.935 | 0.921 | 0.0720 |
| | Image Feature + Text Feature (Fine-grained) | 0.954 | 0.974 | 0.964 | 0.0460 |
| Misalignment of clavicles on both sides | Image Feature | 0.861 | 0.872 | 0.866 | 0.1391 |
| | Image Feature + Text Feature | 0.905 | 0.926 | 0.927 | 0.0745 |
| | Image Feature + Text Feature (Fine-grained) | 0.957 | 0.958 | 0.957 | 0.0439 |
| Inconsistent clavicle height on both sides | Image Feature | 0.889 | 0.855 | 0.872 | 0.1167 |
| | Image Feature + Text Feature | 0.914 | 0.905 | 0.912 | 0.0616 |
| | Image Feature + Text Feature (Fine-grained) | 0.966 | 0.944 | 0.955 | 0.0348 |
| Asymmetrical sternoclavicular joints | Image Feature | 0.902 | 0.883 | 0.892 | 0.0994 |
| | Image Feature + Text Feature | 0.926 | 0.947 | 0.934 | 0.0663 |
| | Image Feature + Text Feature (Fine-grained) | 0.950 | 0.967 | 0.958 | 0.0491 |
| Unclear lung apex | Image Feature | 0.873 | 0.859 | 0.866 | 0.1264 |
| | Image Feature + Text Feature | 0.921 | 0.913 | 0.922 | 0.0634 |
| | Image Feature + Text Feature (Fine-grained) | 0.959 | 0.945 | 0.952 | 0.0406 |
| Lung apex not included | Image Feature | 0.914 | 0.851 | 0.881 | 0.0799 |
| | Image Feature + Text Feature | 0.935 | 0.918 | 0.941 | 0.0425 |
| | Image Feature + Text Feature (Fine-grained) | 0.972 | 0.951 | 0.961 | 0.0245 |
| Too little empty exposure above the shoulders | Image Feature | 0.848 | 0.891 | 0.869 | 0.1104 |
| | Image Feature + Text Feature | 0.897 | 0.944 | 0.927 | 0.0832 |
| | Image Feature + Text Feature (Fine-grained) | 0.922 | 0.971 | 0.946 | 0.0670 |
| Diaphragm not included | Image Feature | 0.871 | 0.870 | 0.870 | 0.1368 |
| | Image Feature + Text Feature | 0.909 | 0.939 | 0.918 | 0.0926 |
| | Image Feature + Text Feature (Fine-grained) | 0.926 | 0.959 | 0.942 | 0.0753 |
| Costophrenic angle not included | Image Feature | 0.922 | 0.894 | 0.908 | 0.0786 |
| | Image Feature + Text Feature | 0.951 | 0.932 | 0.941 | 0.0486 |
| | Image Feature + Text Feature (Fine-grained) | 0.970 | 0.941 | 0.956 | 0.0299 |
| Asymmetrical ribcage on both sides | Image Feature | 0.919 | 0.853 | 0.885 | 0.0879 |
| | Image Feature + Text Feature | 0.936 | 0.917 | 0.923 | 0.0696 |
| | Image Feature + Text Feature (Fine-grained) | 0.967 | 0.938 | 0.952 | 0.0341 |
| Asymmetrical lung fields on both sides | Image Feature | 0.900 | 0.907 | 0.904 | 0.0869 |
| | Image Feature + Text Feature | 0.930 | 0.949 | 0.949 | 0.0577 |
| | Image Feature + Text Feature (Fine-grained) | 0.958 | 0.975 | 0.966 | 0.0365 |
| Existence of abnormal objects | Image Feature | 0.843 | 0.898 | 0.870 | 0.1613 |
| | Image Feature + Text Feature | 0.875 | 0.936 | 0.909 | 0.1265 |
| | Image Feature + Text Feature (Fine-grained) | 0.897 | 0.954 | 0.924 | 0.1061 |
| Motion blur | Image Feature | 0.912 | 0.887 | 0.893 | 0.0902 |
| | Image Feature + Text Feature | 0.945 | 0.926 | 0.924 | 0.0560 |
| | Image Feature + Text Feature (Fine-grained) | 0.960 | 0.946 | 0.953 | 0.0396 |

dataset. Sample distribution of the image quality factors is shown in Table 1.

In the experiment, randomly 70% of the data (4092 chest films) are selected as the training set, 20% of the data (1169 chest films) as the validation set, and 25% of the data (1461 chest films) as the test set.

In data preprocessing of the training and inference phases, the visual encoder adopts the same image resolution as in [19], i.e., 512×512 pixels, and all intensities are normalized to 0-1. The text encoder uses the same Transformer as in [54].

### B. MODEL TRAINING DETAILS

For each label in the chest X-ray image, we use an SGD optimizer with an initial learning rate of 0.002, decayed by the cosine annealing rule. We obtain the best performance over

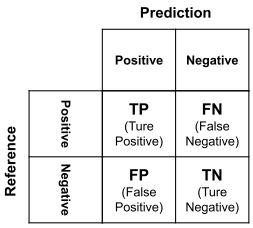



**FIGURE 5.** Confusion matrix.

the past 200 epochs by training the model for 1000 epochs. The entire model training is done on a server with 8 NVIDIA RTX 3090 GPU cards.

### C. EVALUATION METRIC

The FR-IQA methods mostly use SSIM, MS-SSIM, VIF, IFC, FSIM and other quality assessment metrics. The NR-IQA cannot use the above-mentioned quality assessment due to the lack of reference images.

In our experiment, precision, recall, F1-Score, and false positive rate (FPR) are used to evaluate the performance of our proposed algorithm. The confusion matrix is defined as in Figure 5.

The definitions of the evaluation metrics are as follows.

$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$
$$F_1 = (2 \times Precision \times Recall)/(Precision + Recall)$$
$$FPR = FP/(FP + TN)$$

### D. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental results shown in Table 2, it can be observed that among the evaluation results on the 13 influencing factors, 10 F1-Scores are higher than 0.95, 3 F1-Scores are between 0.92 and 0.95. The result on the existence of abnormal objects class is comparatively lower. Elaborately designed methods can be adopted in the future to further improve the model. The overall performance of our proposed model is satisfactory and it can be used for chest X-ray image quality assessment.

In order to further verify the effectiveness of our method proposed in this paper, we further carried out the following ablation study. On the same training data, we experimented with (a) only image features, (b) global image features + text features, and (c) fine-grained image features + text features. As shown in Table 3, the experimental results show that using the visual text model can achieve better performance than only using image features, and the fine-grained visual text model can achieve the best performance.

## VII. CONCLUSION

In this paper, we propose a chest X-ray quality assessment method combining both image-text contrastive learning and medical domain knowledge. Specifically, it integrates large-scale real clinical chest X-rays and diagnostic report text information, and fine-tunes the pretrained model based on contrastive text-image pairs. It achieves cross-domain transfer learning and can save the huge workload caused by doctors in labeling multi-modal medical data. The integration of the triplet information from knowledge graph into the deep learning model proposed in this paper provides a new solution for knowledge and data-driven machine learning methods. The proposed method can be extended to complete other tasks such as multi-lesion segmentation on medical images and prediction of disease progression. The experimental results and analysis show that the proposed method boasts good performance.

## REFERENCES

[1] J. N. Itri, "Patient-centered radiology," *RadioGraphics*, vol. 35, no. 6, pp. 1835–1846, Oct. 2015.
[2] H.-C. Zhang, Z.-P. Chen, Y.-M. Huang, and Y.-H. Guo, "Establishment of prenatal ultrasound diagnosis quality control system and its application value study," *Modern Hospital*, vol. 17, no. 2, p. 3, 2017.
[3] T. Venkat and N. Rao, "Assessment of diverse quality metrics for medical images including mammography," *Int. J. Comput. Sci. Netw. Secur.*, vol. 14, no. 11, pp. 1–6, 2014.
[4] M. Guan, Y. Lyu, W. Cao, X. Wu, J. Lu, and S. K. Zhou, "Perceptual quality assessment of chest radiograph," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Strasbourg, France: Springer, Oct. 2021, pp. 315–324.
[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[7] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[8] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[9] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[10] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.
[11] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.
[12] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.
[13] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1700–1705, Nov. 2013.
[14] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[15] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010.
[16] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[17] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.
[18] K. Chen, X. Gong, C. Li, Y. Yu, and S. Zhan, *Quality Control Standards for Clinical Radiology in Yangtze River Delta Region*, 1st ed. Shanghai, China: Shanghai Science and Technology Press, 2022.
[19] J. Hu, C. Zhang, K. Zhou, and S. Gao, "Chest X-ray diagnostic quality assessment: How much is pixel-wise supervision needed?" *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1711–1723, Jul. 2022.

[20] J. von Berg, S. Krönke, A. Gooßen, D. Bystrov, M. Brück, T. Harder, N. Wieberneit, and S. Young, "Robust chest X-ray quality assessment using convolutional neural networks and atlas regularization," in *Proc. Med. Imag.*, vol. 11313, Mar. 2020, pp. 391–398.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[22] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[23] X. Sun, P. Hu, and K. Saenko, "DualCoOp: Fast adaptation to multi-label recognition with limited annotations," 2022, *arXiv:2206.09541*.

[24] C. Feichtenhofer, H. Fassold, and P. Schallauer, "A perceptual image sharpness metric based on local edge gradient analysis," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 379–382, Apr. 2013.

[25] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[26] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.

[27] J. Gu, G. Meng, J. A. Redi, S. Xiang, and C. Pan, "Blind image quality assessment via vector regression and object oriented pooling," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1140–1153, May 2018.

[28] Q. Gao, S. Li, M. Zhu, D. Li, Z. Bian, Q. Lyu, D. Zeng, and J. Ma, "Blind CT image quality assessment via deep learning framework," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, Oct. 2019, pp. 1–4.

[29] B. L. Eck, R. Fahmi, K. M. Brown, S. Zabic, N. Raihani, J. Miao, and D. L. Wilson, "Computational and human observer image quality evaluation of low dose, knowledge-based CT iterative reconstruction," *Med. Phys.*, vol. 42, no. 10, pp. 6098–6111, 2015.

[30] B. Mortamet, M. A. Bernstein, C. R. Jack Jr., J. L. Gunter, C. Ward, P. J. Britson, R. Meuli, J.-P. Thiran, and G. Krueger, "Automatic quality assessment in structural brain magnetic resonance imaging," *Magn. Reson. Med., Off. J. Int. Soc. Magn. Reson. Med.*, vol. 62, no. 2, pp. 365–372, 2009.

[31] S. J. Esses, X. Lu, T. Zhao, K. Shanbhogue, B. Dane, M. Bruno, and H. Chandarana, "Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture," *J. Magn. Reson. Imag.*, vol. 47, no. 3, pp. 723–728, Mar. 2018.

[32] Y. Wang, Y. Song, F. Wang, J. Sun, X. Gao, Z. Han, L. Shi, G. Shao, M. Fan, and G. Yang, "A two-step automated quality assessment for liver MR images based on convolutional neural network," *Eur. J. Radiol.*, vol. 124, Mar. 2020, Art. no. 108822.

[33] J. Xiao-Qian, Z. Xiang-Li, L. Zhe, Z. Qiang, Z. Zhi-Fu, L. Yan-Shou, H. Huang, D. A-Mei, Y. Jian, and G. Jian-Xin, "Application value of convolutional neural network in quality control of direct digital chest X-ray images," *J. Xi'an Jiaotong Univ. Med. Sci.*, vol. 40, no. 5, p. 784, 2019.

[34] L. Wu, J. Cheng, S. Li, B. Lei, T. Wang, and D. Ni, "FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1336–1349, May 2017.

[35] S. Li, J. He, Y. Wang, Y. Liao, D. Zeng, Z. Bian, and J. Ma, "Blind CT image quality assessment via deep learning strategy: Initial study," *Proc. SPIE*, vol. 10577, pp. 293–297, Mar. 2018.

[36] M. Diligenti, S. Roychowdhury, and M. Gori, "Integrating prior knowledge into deep learning," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 920–923.

[37] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Broeck, "A semantic loss function for deep learning with symbolic knowledge," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5502–5511.

[38] A. Daw, A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," 2017, *arXiv:1710.11431*.

[39] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.

[40] P. Battaglia, R. Pascanu, and M. Lai, "Interaction networks for learning about objects, relations and physics," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[41] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," 2016, *arXiv:1612.04844*.

[42] C. Jiang, H. Xu, X. Liang, and L. Lin, "Hybrid knowledge routed modules for large-scale object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[43] F. Li, H. Zhang, Y.-F. Zhang, S. Liu, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, "Vision-language intelligence: Tasks, representation learning, and large models," 2022, *arXiv:2203.01922*.

[44] S. Shen, L. Harold Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How much can CLIP benefit vision-and-language tasks?" 2021, *arXiv:2107.06383*.

[45] X. Gu, T. Lin, W. Kuo, and Y. Cui, "Zero-shot detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.

[46] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, Oct. 2022.

[47] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5177–5186.

[48] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 700–708.

[49] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, and S. Wen, "Multi-label classification with label graph superimposing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12265–12272.

[50] V. O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, and J. van de Weijer, "Orderless recurrent models for multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13440–13449.

[51] T. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," 2022, *arXiv:2204.03649*.

[52] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *Computer Vision–ECCV*. Cham, Switzerland: Springer, Oct. 2022, pp. 105–124.

[53] T. Lüddecke and A. S. Ecker, "Image segmentation using text and image prompts," 2021, *arXiv:2112.10003*.

[54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[55] C. Zhou, C. Change Loy, and B. Dai, "DenseCLIP: Extract free dense labels from CLIP," 2021, *arXiv:2112.01071*.

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.

[57] Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang, "A novel cascade binary tagging framework for relational triple extraction," 2019, *arXiv:1909.03227*.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

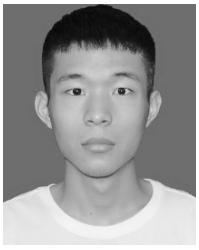
[59] X. Tang, J. Zhang, B. Chen, Y. Yang, H. Chen, and C. Li, "BERT-INT: A BERT-based interaction model for knowledge graph alignment," *Interactions*, vol. 100, p. e1, Jan. 2020.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[61] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chest-X-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

**SHANSHAN DU** received the B.S. degree in computer science and technology from Yangtze University, Hubei, China, in 2006, and the M.S. degree in computer science from Fudan University, Shanghai, China, in 2015, where she is currently pursuing the Ph.D. degree with the School of Information Science and Technology. Her research interests include computer version, multimedia information analysis and processing, machine learning, and medical image analysis.

**YINGWEN WANG** received the bachelor's degree in nursing from China Medical University, Shenyang, Liaoning, China, in 2005, and the M.Sc. degree in nursing from Fudan University, Shanghai, China, in 2015. She was with the Hematology and Oncology Department, Children's Hospital of Fudan University, as a Bedside Nurse, from 2005 to 2011, and a Head Nurse, from 2011 to 2021. Then, she was transferred to the Nursing Department, Children's Hospital of Fudan University. She is a member of the Nursing Group of China Anti-Cancer Association and the Shanghai Anti-Cancer Association. She is also a Senior Member of China Pediatric Nursing Alliance Oncology Group.



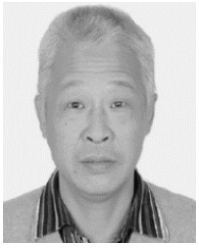
**XINYU HUANG** received the B.S. degree from the School of Computer Science, China University of Geosciences, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, China, under the supervision of Rui Feng. His research interests include machine learning and computer vision, especially on multi-modal learning and multi-label learning.

**RUI-WEI ZHAO** received the B.S. degree in automation and the M.S. degree in control theory from Tongji University, Shanghai, China, in 2009 and 2013, respectively, and the Ph.D. degree in applied computer science from Fudan University, Shanghai, in 2018. He was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. He is currently an Associate Research Fellow with the Academy of Engineering and Technology, Fudan University. His research interests include artificial intelligence, cross-modal computing, and medical image analysis.

**XIAOBO ZHANG** received the bachelor's degree in medicine from Shanghai Jiaotong University, Shanghai, China, in 1992, and the M.M. and M.D. degrees from Fudan University, Shanghai, in 2008 and 2013, respectively. She is currently a Chief Physician, a Ph.D. Supervisor, and the Vice President of the Children's Hospital of Fudan University. Her major research interests include standardized diagnosis, treatment, etiology and intervention strategies of congenital lung disease, chronic lung disease, and allergic disease in children. She also focusing on the construction of big-data base for pediatric diseases, pediatric evidence-based clinical guidelines library, large knowledge graph for pediatrics diseases, and pediatric disease standard resource cloud platform.

**RUI FENG** received the B.S. degree in industrial automatic from Harbin Engineering University, Harbin, China, in 1994, the M.S. degree in industrial automatic from Northeastern University, Shenyang, China, in 1997, and the Ph.D. degree in control theory and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2003. In 2003, he joined the Department of Computer Science and Engineering (now School of Computer Science), Fudan University, as an Assistant Professor, and then he became an Associate Professor and a Full Professor. His research interests include medical image analysis, intelligent video analysis, and machine learning.

**QUANLI SHEN** received the B.S. and M.M. degrees in medicine and the Ph.D. degree in imaging medicine and nuclear medicine from Fudan University, Shanghai, China, in 2000, 2004, and 2010, respectively. He majors in pediatric thorax imaging. Since 2000, he has been with the Department of Radiology, Children's Hospital of Fudan University, to write and check the reports of X-ray, CT, and MRI. He is currently studying the application of artificial intelligence in pediatric chest imaging.



**JIAN QIU ZHANG** (Senior Member, IEEE) received the B.Sc. degree from East China Institute of Engineering, Nanjing, in 1982, and the M.S. and Ph.D. degrees from Harbin Institute of Technology (HIT), Harbin, China, in 1992 and 1996, respectively. From 1982 to 1987, he was an Assistant Electronic Engineer with 544th Factory, Hunan, China. From 1989 to 1994, he was a Lecturer with the Department of Electrical Engineering, HIT, where he was an Associate Professor, from 1995 to 1997. In 1998, he was a Visiting Research Scientist with the Institute of Intelligent Power Electronics, Helsinki University of Technology, Espoo, Finland. From 1999 to 2002, he was a Senior Research Fellow with the School of Engineering, University of Greenwich, Gillingham, U.K. He is currently a Professor with the Department of Electronic Engineering, Fudan University, Shanghai, China. His main research interest includes signal processing and its application.