Displays 84 (2024) 102785

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Chest CT-IQA: A multi-task model for chest CT image quality assessment and classification ☆

Siyi Xun [a], Mingfeng Jiang [b], Pu Huang [c], Yue Sun [a], Dengwang Li [c], Yan Luo [d], Huifen Zhang [d], Zhicheng Zhang [e], Xiaohong Liu [f], Mingxiang Wu [d,*], Tao Tan [a]

[a] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao Special Administrative Region of China
[b] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, 310018, China
[c] Shandong Key Laboratory of Medical Physics and Image Processing, Shandong Institute of Industrial Technology for Health Sciences and Precision Medicine, School of Physics and Electronics, Shandong Normal University, Jinan, Shandong, 250358, China
[d] Department of Radiology, Shenzhen People's Hospital, Luohu, Shenzhen, 518020, China
[e] JancsiLab, JancsiTech, Hongkong, China
[f] Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

MSC:
68U10
92C55

Keywords:
Computerized tomography
Image quality assessment
Classification
Multi-task model

ABSTRACT

In recent years, especially during the COVID-19 pandemic, a large number of Computerized Tomography (CT) images are produced every day for the purpose of inspecting lung diseases. However, the diagnosis accuracy depends on the quality of CT imaging and low quality images may greatly affect clinical diagnosis, resulting in misdiagnosis. It is difficult to effectively rate the quality of massive CT images. To solve the above problems, we first constructed a dataset of 800 CT volumes for chest CT image quality assessment. Then we propose a multi-task model for chest CT image quality assessment and classification. This model can automatically classify CT image sequences of different visual inspection windows, and automatically estimate CT image quality score, to match the visual score from clinicians. The experimental results show that the window classification accuracy and the dose exposure classification accuracy of our model can reach 0.8375 and 0.8813 respectively. The Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) between the model prediction results and the two radiologist's annotation average result reached 0.3288 and 1.9264. It shows that our model has a potential to mimic quality evaluation of experts.

1. Introduction

Computerized Tomography (CT) image has become an important mode in medical image because of its high resolution, multi-plane reconstruction, three-dimensional imaging and fast scanning speed [1]. CT image has also been widely used in many medical fields, and has played an important role in clinical practice, and has important clinical significance for the diagnosis and treatment of diseases.

In recent years, especially during the COVID-19 pandemic, a large number of CT images are produced every day. At the same time, CT images can be adjusted to different scanning parameters according to different clinical needs, so the same patient may also have multiple CT scan images with different parameters. In clinical practice, the most common images are lung window and soft tissue window, as shown in Fig. 1.

However, it is difficult to clean and classify massive CT images effectively, which brings great difficulties to the use, storage and management of medical data. Low quality images will greatly affect clinical diagnosis, resulting in misdiagnosis [2–5]. At the same time, the quality assessment of CT images must also be carried out under the same visual inspection window. Accurate window classification can effectively help CT data cleaning and horizontal comparison of image quality. In the process of CT scanning, because of the settings those information might be missing in the dicom tags. At the same time, in existing public datasets, such information is often missing because of the process of data anonymization. In addition, many other medical image formats, such as Nifti, do not have provide the relevant information. In certain cases, this prediction is still be useful.

At present, the daily Image Quality Assessment (IQA) [6] of the imaging department mainly relies on the subjective assessment of radiologists. However, manually labeling the quality of each image

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author.
E-mail addresses: xiaohongliu@sjtu.edu.cn (X. Liu), wu.mingxiang@szhospital.com (M. Wu), taotan@mpu.edu.mo (T. Tan).

https://doi.org/10.1016/j.displa.2024.102785
Received 7 February 2024; Received in revised form 3 June 2024; Accepted 19 June 2024
Available online 24 June 2024

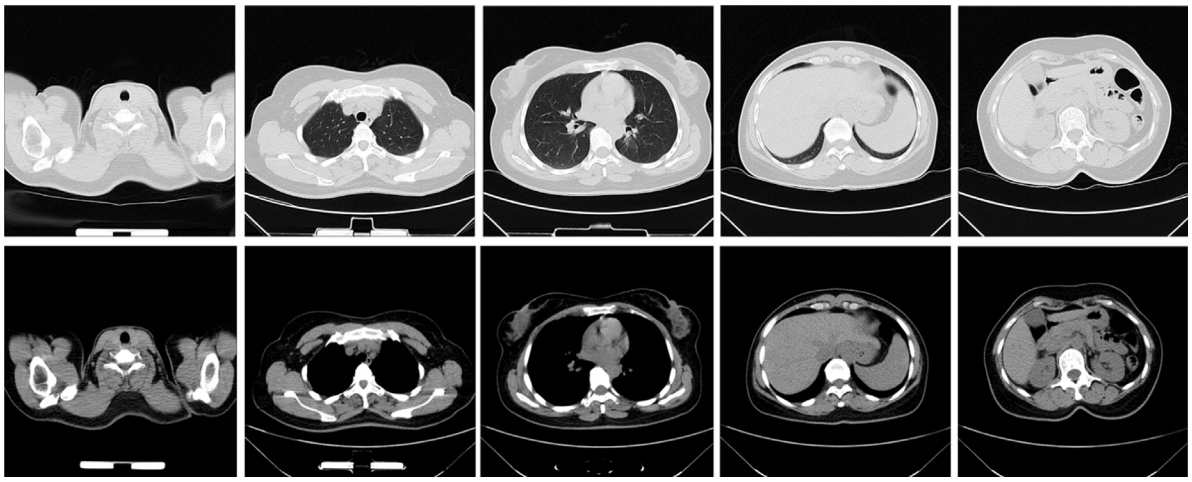


Fig. 1. Chest CT images of lung and soft tissue visual inspection windows at different levels. The first row is lung window image, the second row is soft tissue window image.

is a time-consuming and laborious task, and the assessment results will also be affected by the experience and subjective factors of the evaluator. Therefore, accurate and objective automated medical IQA is of great significance. With the continuous development of Artificial Intelligence (AI), how to use machine learning and deep learning technology to learn image quality characteristics from large-scale data and automatically predict image quality has become an urgent problem to be solved. However, due to the subjectivity and complexity of image quality, completely accurate assessment of image quality remains a challenging task.

To solve the above problems, we first constructed a dataset for chest CT image IQA. Then we propose a multi-task model for chest CT image IQA and classification. This model can automatically classify CT image sequences of different visual inspection windows, and automatically rate CT image quality score, to help clinicians to assessment images. Specific contributions are as follows:

- First, by screening Chest images, a chest CT image IQA dataset was constructed by using 800 CT volumes with different dose exposure parameters and different visual inspection windows, named Chest-CT-QA. To our knowledge, this is the first dataset to assessment the quality of chest CT images without using synthetic data.
- Second, based on the constructed dataset, a multi-task model integrating CT image sequence window classification, dose exposure, Signal-to-Noise Ratio (SNR) and Blind/Reference less Image Spatial Quality Evaluator (BRISQUE) metric calculation and automatic regression is proposed. Using deep learning method, the image quality is automatically assessment and the corresponding image quality score is given.
- Finally, we fit the prediction results of the model with the scores of different clinicians to verify the effectiveness of the model. We envision that this model can be used as a tool to assess the quality of chest CT images.

2. Related work

2.1. CT image window classification

In recent years, CT imaging technology has been widely used in clinical diagnosis and treatment. In CT image data, accurate window sequence classification can effectively help data cleaning and horizontal comparison of image quality. CT image as a digital reconstruction image, window technology is an important means to analyze CT image. The window technology in the field of medical imaging refers to the selection of the CT value range of interest by adjusting the Window Width (WW) and Window Level (WL), to obtain clear images with different contrasts according to clinical needs, and further help diagnosis. The CT value is often called the Hounsfield Unit (HU) and reflects the degree of X-ray absorption by the tissue. WW refers to the range of CT values displayed on a CT image, within which tissues and lesions are displayed in different simulated gray scales. WL is the central position of the window [7]. Since various tissue structures and lesions have different CT values, it is necessary to classify image sequences of different windows.

2.2. Medical image quality assessment

IQA is an important research direction in the fields of digital image processing, computer vision and image transmission by analyzing the features of images, and then evaluating the degree of image distortion.

IQA can be divided into subjective assessment and objective assessment according to whether it is judged by human subjects [8]. The subjective assessment method is to assess the quality of images through human visual perception. Objective assessment method is to assess image quality through computer algorithm or mathematical model, common metrics include SNR, Structural Similarity Index (SSIM), Visual Information Fidelity (VIF) and so on. IQA can be divided into Full Reference-IQA (FR-IQA), Reduced Reference-IQA (RR-IQA) and No Reference-IQA (NR-IQA) according to whether reference images are used as assessment criteria [9].

With the continuous development of AI technology, how to use machine learning and deep learning methods to automatically evaluate image quality has become a current research hotspot. Sim et al.. proposed a two-dimensional FR-IQA method based on depth feature maps and local similarity [10]. In the face of the absence of reference images, Zhu et al.. used the prior knowledge of meta learning to design an NR-IQA method [11]. Messai et al.. designed NR-IQA method for stereo images by using the idea of Ada Boost in machine learning [12].

The current AI-based IQA methods, metrics and datasets are mainly oriented towards natural images, videos [13–15] and images in specific modalities [16–20]. However, there are few IQA methods and open source datasets designed for medical images. As a result, many researchers have migrated natural image IQA methods to medical image IQA. Mudeng et al.. analyzed the application prospect of SSIM, an important metric, in medical IQA [21]. Outtas et al.. evaluated the usability of two advanced metrics for natural image quality evaluation (NIQE) [22] and the blind image quality evaluator based on scales (BIQES) [23] in medical images and further improved them [24].

Due to its particularity, there is no perfect image as the gold standard. Therefore, NR-IQA is the best assessment method. Chow et al..

improved the classic BRISQUE [25] for the quality assessment of Magnetic Resonance Imaging (MRI) images [26]. Oszust et al.. developed a new NR-IQA method for automatic quality assessment of MRI images [27]. Guo et al. presented a NR-IQA model for pathological microscopic images [28].

For the IQA of CT images, Li et al.. used NR-IQA method based on deep learning to evaluate the quality of CT images [29]. On this basis, Gao et al.. combined the global and local information of CT images to conduct a more detailed quality assessment [30]. However, Greffier et al.. used deep learning methods to evaluate the quality of CT images of different parts [31–33]. In these studies, deep learning methods were used for medical image IQA. These methods have achieved certain research objectives, but their effectiveness needs to be further verified.

3. Method

3.1. Dataset

3.1.1. Data collection and screening

A high quality medical image dataset is the premise of quality assessment. To explore a more usable medical IQA model, we selected the most common chest CT in clinical diagnosis and treatment and daily physical examination as our data. All data we use is desensitized and ethically compliant.

To screen out the appropriate data, we mainly use the imaging modalities, dose exposure, and visual inspection windows as our screening basis. First, to avoid the difference in brightness and contrast of images produced by different devices, we uniformly selected the images taken by the same CT machine produced by Philips.

Secondly, the CT image quality is related to the strength of the exposure dose during scanning (equivalent to the energy of the photon). The magnitude of the exposure dose is proportional to the exposure dose, the greater the exposure dose, the stronger the X-ray intensity and penetration, the higher the image contrast and clarity, and the better the image quality. However, higher exposure doses may also cause an increase in noise in the image, affecting the quality of the image. Therefore, we chose two kinds of exposure doses commonly used in clinical practice, 30 mAs and 50 mAs, as our initial images of two different quality levels. Fig. 2 shows CT images of different exposure doses.

Then, we selected the lung window and soft tissue window image sequences commonly used in clinical practice. The lower WL and larger WW of the lung window can improve the contrast and resolution of the lung tissue. The soft tissue window is set with a higher WL and a narrower WW, which can increase the contrast and resolution of the soft tissue structure.

Finally, we screened 800 available chest CT image sequences. There were 400 cases of lung window sequence and 400 soft tissue window sequences, including 200 cases of 30 mAs and 50 mAs current sequences.

3.1.2. Subjective assessment

To explore whether the results of the automatic assessment of the model are consistent with the results of the subjective assessment, we invited two experienced clinicians to assess 100 new sequences with the same conditions as the above constructed dataset but not included in the dataset 50 cases of lung window sequences and 50 cases of soft tissue window sequences.

Assessment criteria are set on a 3 level, 10 point scale. According to the quality of the image is divided into bad (1–3 points, level 1), general (4–7 points, level 2), good (8–10 points, level 3). In the process of assessment, the quality of medical image is mainly defined by sharpness, supplemented by the amount of noise and contrast (such as the contrast between different anatomical structures). The final score should be an integer ranging from 1 to 10. The higher the score, the better the image quality.

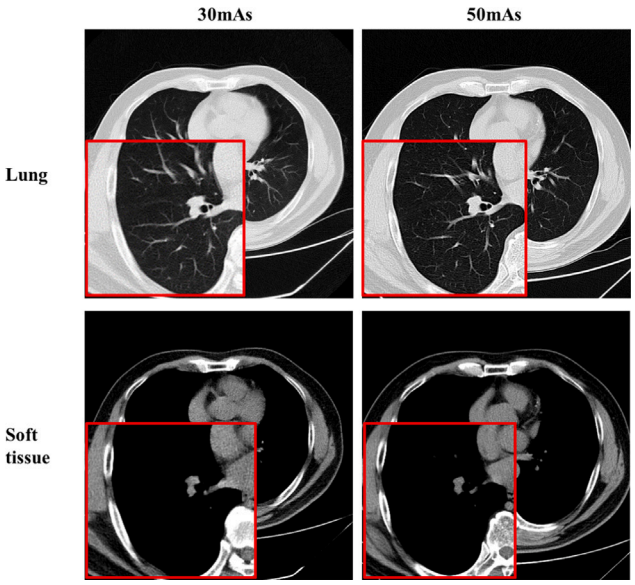


Fig. 2. CT images of different exposure doses. From the three dimensions of sharpness, noise and contrast, the visual perception quality of 50 mAs dose image is higher.

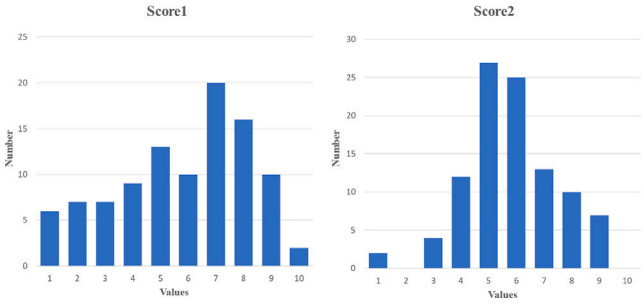


Fig. 3. The score distribution of subjective assessment. Score 1 and Score 2 are the subjective evaluation results of two clinicians respectively.

Through subjective experiments, we obtained the quality level and score of 100 chest CT sequences, and verified our model based on this standard. The score distribution of subjective assessment is shown in Fig. 3.

3.2. Model

3.2.1. Network structure

Based on the dataset we constructed, we proposed a multi-task model Chest CT-IQA for chest CT image quality assessment and classification. The model is composed of backbone network modified by 3D VGG [34], visual inspection windows and dose exposure classification module, assessment metric regression module and quality assessment module, as shown in Fig. 4.

Specifically, we first modified the 3D VGG network for image deep learning semantic features, window contrast features, and quality features extraction. Inspired by VGG, the backbone network can be divided into 5 parts for feature extraction. Each part mainly includes 4 layers, which are Convolution layer, BatchNorm layer, Nonlinear layer, Maximum pooling layer. Unlike the original VGG, in addition to the image size and the number of channel parameters such as modification, we in each feature extraction module adds a BatchNorm layer to prevent gradient explosion and gradient disappeared, speed up the convergence speed of the network. In addition, compared with the

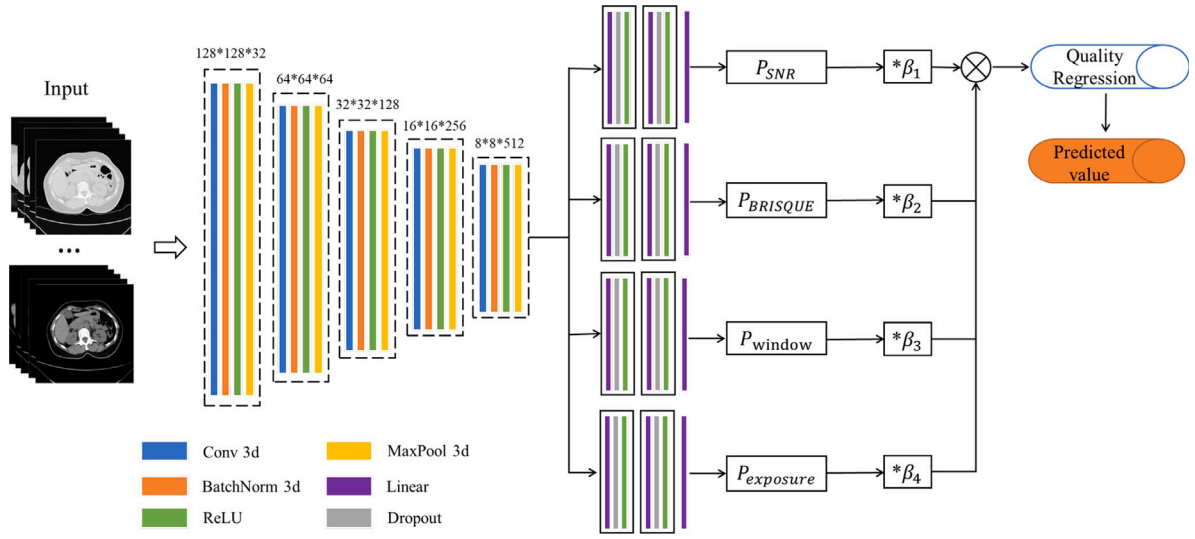


Fig. 4. Schematic diagram of our proposed model structure.

original VGG, because we proposed a 3D model, in order to speed up training and inference and reduce running parameters, we did not use too many convolutional layers in the process of feature extraction.

After feature extraction, we designed four fully connected modules to output the results of window classification, dose exposure classification, SNR metric regression and BRISQUE metric regression respectively. Each fully connected module includes three fully connected layers, and in each of the two fully connected layers, a dropout layer with parameter 0.5 is added, as well as a Relu activation layer.

3.2.2. Classification

The window and dose exposure classification modules are used to give the classification results of image visual inspection windows and the prediction results of image quality. At the same time, window and dose information can provide labels of image types, which as sub-task can help the network learn image features and effectively improve the performance of regression tasks. The learning of this module adopts the Cross Entropy loss function [35]. For each category our prediction yields probabilities p and $1-p$, in which case the expression is:

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot log(p_i) + (1-y_i)\cdot log(1-p_i)] \quad (1)$$

where y_i represents the label of sample i, where class 1 is treated as 1 and class 2 is treated as 0. p_i represents the probability that sample i predicts a positive class.

3.2.3. Regression

The regression module is used to automatically regression the values of SNR and BRISQUE. Firstly, the SNR and BRISQUE values of all images in the dataset are calculated using the traditional algorithm.

SNR refers to the size of the relationship between the effective signal and the noise signal. For medical images, there is no standard image that we can refer to. Therefore, we simplify the calculation formula of SNR to the ratio of the mean and standard deviation of the signal (gray value). For the 3D sequential image, we split the whole image into several small volumes and calculate the SNR value of each small volume respectively, and finally average the final SNR value. BRISQUE metric is a non-reference spatial IQA algorithm. Here we use the "brisque" toolkit in python to do the calculations.

Then, we use the metric value calculated by the traditional algorithm as label for automatic regression. According to the calculation, it takes 2–3 min to calculate these two indicators using the traditional method, while it only takes 10 s to perform automatic regression for the trained neural network, which greatly improves the speed of quality assessment. The learning of this module adopts Mean Square Error (MSE) loss function [36]. MSE measures the average squared error between the model's predicted value and the real value. The smaller the value of MSE, the smaller the difference between the predicted value and the true value of the model, and the better the performance of the model. The calculation formula is as follows:

$$MSE = \frac{1}{n}\sum(x_i - y_i)^2 \quad (2)$$

where n is the number of samples, x_i and y_i is the true value and the predicted value of the model.

3.2.4. Model optimization

For the optimization of the model, the weight summation of the loss functions of the above four tasks is carried out, and the optimal weight parameters are adjusted experimentally. Therefore, the loss function of our overall model is:

$$Loss = \lambda_1 L_{window} + \lambda_2 L_{exposure} + \lambda_3 L_{SNR} + \lambda_4 L_{BRISQUE} \quad (3)$$

where L_{window} is the loss of the window classification, $L_{exposure}$ is the loss of the dose exposure classification, L_{SNR} is the loss of the SNR regression, and $L_{BRISQUE}$ is the loss of the BRISQUE regression.

To avoid the inaccuracy of single prediction results and metric values on quality assessment results, weighted fitting was carried out for window classification results, dose exposure classification results, regression results of SNR and BRISQUE metrics in the quality assessment module. Specifically, we took the four values predicted by the model as independent variables X_1, X_2, X_3, X_4 and the clinician's manual score as the dependent variable Y to carry out multivariate linear fitting, the fitting form is as follows:

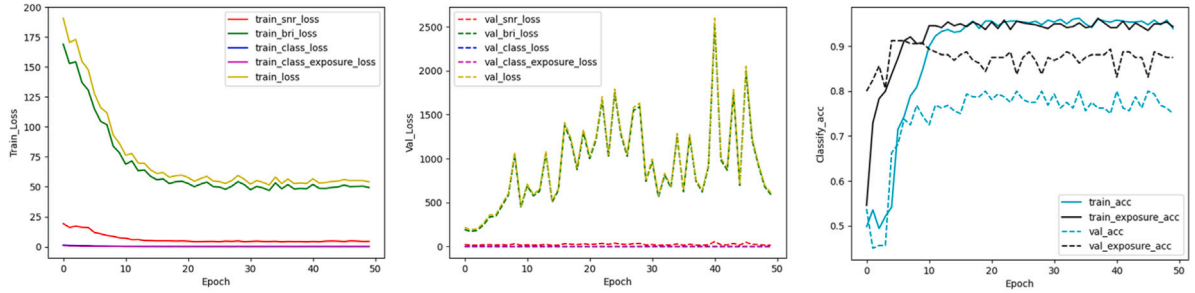
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e \quad (4)$$

where β_0 is intercept of regression model, β_1, β_2, β_3, β_4 called regression coefficient, e is random error. X_1, X_2, X_3, and X_4 correspond to P_{window}, $P_{exposure}$, P_{SNR}, $P_{BRISQUE}$ in Fig. 4. We use human label to fit the regression coefficient and the four independent variables as inputs, performs the final assessment of CT image quality, and generate the final prediction score.

4. Experiments

4.1. Experimental environment and parameters

This research experiment is based on the constructed dataset, and the model is built based on Python3.9 environment and PyTorch framework. The optimizer uses the Adam optimizer with an initial learning



(a) Train loss (b) Validation loss (c) Train and Validation accuracy

Fig. 5. Diagram of experimental results of the proposed model.

Table 1
Experimental results of our proposed model on train, validation and test dataset.

Model	Results						
	Accuracy	Precision	Recall	F1-score	SRCC	PLCC	RMSE
Train	W: 0.8312 E: 0.8604	W: 0.9563 E: 0.7352	W: 0.7749 E: 0.9776	W: 0.8561 E: 0.8393	S: 0.5079 B: 0.3308	S: 0.4471 B: 0.3133	S: 4.0228 B: 21.7341
Validation	W: 0.7500 E: 0.8750	W: 0.8630 E: 0.7500	W: 0.6774 E: 1.0000	W: 0.7590 E: 0.8571	S: 0.4918 B: 0.2379	S: 0.4211 B: 0.2083	S: 4.0692 B: 24.2904
Test	W: 0.8375 E: 0.8813	W: 0.9600 E: 0.7805	W: 0.7578 E: 0.9846	W: 0.8470 E: 0.8707	S: 0.5998 B: 0.3015	S: 0.5591 B: 0.1943	S: 3.4063 B: 19.3409

W: Window; E: Exposure; S: SNR; B: BRISQUE.

rate of 0.0001. Regularization step size is set to 1 and gamma is set to 0.75. Set the number of training iterations to 50 and the batch size to 1. We trained and tested on an Intel(R) Xeon(R) W2245 CPU @ 3.90 GHz, NVIDIA RTX A6000 computer.

4.2. Experimental metric

In the experiment, we used Accuracy (Acc), Precision, Recall, F1-score to evaluate the performance of our classification tasks. These indicators assess the accuracy of the classification task, and the closer the value is to 1, the better the classification task performance. The performance of regression task was evaluated by Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) and Root mean square error (RMSE). SRCC and PLCC measure the monotonicity and linear correlation of the model, while RMSE assesses the consistency of the model's predictions.

4.3. Experimental results

Based on the above experimental environment and experimental metrics, the proposed model is tested. The experimental results of model are shown in Fig. 5 and Table 1. Among them, (a) shows the change of loss during training. During the training process, the loss of the four tasks continued to decrease, and all converged and gradually leveled off after 20 epochs. The training loss of the model is greatly affected by the BRISQUE regression task. (b) shows the change of loss during validation. It can be seen that the effect of model verification is still greatly affected by BRISQUE regression. (c) shows the window and dose exposure classification of accuracy. Through the learning of the model, accuracy continues to rise and gradually becomes stable after 20 rounds. The training Acc of the final window classification task is stable at about 0.94 and the validation is stable at about 0.75. The training Acc of the dose exposure classification task is stable at about 0.94 and the validation is stable at about 0.88. The final window and dose exposure classification test results are 0.8375 and 0.8813 respectively.

4.3.1. Branch experiments

To explore the influence of the proposed modules on the overall model performance, we conducted ablation experiments on the model. Firstly, we explored the impact of the timing of bifurcation of the fully connected layer and the weight of the loss function on the performance of the multi-task model, and the experimental results are shown in Table 2. Three fully connected layers are designed in our model, and we carry out experiments on three different situations: (1) Four tasks share the first two fully connected layers, and finally separate four different fully connected layers to output the results of four tasks (Model 1). (2) Two classification tasks and two regression tasks share the first two connection layers respectively, and then separate two different full connection layers to output two classification results and regression results (Model 2). (3) Each task enters the fully connected module separately after extracting features, and does not share the fully connected layer (Model 3). At the same time, we adjust the total loss function weight of the model according to the loss function of each task. Through experiments, the model achieves the best performance when the weights (λ_1, λ_2, λ_3, λ_4) are adjusted to 50,50,5,1 (Model 4). The experimental results show that our model achieves the best performance in two classification tasks. And in the other two regression tasks, also achieved a good performance. From the perspective of model performance equilibrium, we finally choose the case Model 4 as our model.

4.3.2. Multi-task experiments

Secondly, we investigate the effect of multitasking on model performance and the experimental results are shown in Table 3. For two classification tasks and two regression tasks, we conducted experiments on the following three situations: (1) Two classification/regression tasks were trained together. (2) The four tasks are trained separately. In case (1), only the regression task of the SNR metric achieved the best results, and the other three tasks did not show particular potential. In case (2), all the four tasks showed poor results. Therefore, multi-task models, especially association models (the same classification and regression tasks), share information through the bottom sharing layer, complement each other and help to learn features, and can improve each other's performance to a certain extent. However, when the correlation between tasks is small, it is difficult for tasks to use each

Table 2
Experimental results of the effect of fully connected layer branch and loss function weight on the performance of the multi-task model.

Model	Branch			Results						
	N/A	2	4	Accuracy	Precision	Recall	F1-score	SRCC	PLCC	RMSE
1	✔			W: 0.7125 E: 0.6562	W: 0.4400 E: 0.3902	W: 0.8919 E: 0.8648	W: 0.5892 E: 0.5378	**S: 0.6126** B: 0.0758	S: 0.6560 B: 0.0406	S: 3.2307 B: 22.3644
2		✔		W: 0.6437 E: 0.5125	W: 0.2533 E: 0.4268	**W: 0.9500** E: 0.5303	W: 0.3999 E: 0.4729	S: 0.3775 **B: 0.3536**	S: 0.4416 **B: 0.2626**	S: 5.2416 **B: 14.3117**
3			✔	W: 0.7750 E: 0.6375	W: 0.6000 E: 0.3292	W: 0.8823 E: 0.9000	W: 0.7142 E: 0.4821	S: 0.6022 B: 0.1211	**S: 0.6663** B: 0.0386	**S: 3.1422** B: 17.8507
4 (ours)			✔	**W: 0.8375** **E: 0.8813**	**W: 0.9600** **E: 0.7805**	W: 0.7578 **E: 0.9846**	**W: 0.8470** **E: 0.8707**	S: 0.5998 B: 0.3015	S: 0.5591 B: 0.1943	S: 3.4063 B: 19.3409

W: Window; E: Exposure; S: SNR; B: BRISQUE.

Table 3
Experimental results of the effect of multitasking on model performance.

Model	Classification		Regression		Results						
	C1	C2	R1	R2	Accuracy	Precision	Recall	F1-score	SRCC	PLCC	RMSE
1	✔	✔			W: 0.8188 E: 0.8563	**W: 0.9733** E: 0.7683	W: 0.7300 E: 0.9403	W: 0.8342 E: 0.8456			
2			✔	✔					**S: 0.6275** B: 0.0498	**S: 0.6642** B: 0.0250	**S: 3.0062** B: 23.9157
3	✔				W: 0.6375 E: 0.8500	W: 0.4533 E: 0.7195	W: 0.6666 E: 0.9833	W: 0.5397 E: 0.8309			
4		✔									
5			✔						S: 0.3685 B: 0.0584	S: 0.3606 B: 0.0300	S: 3.8386 B: 22.1978
6				✔							
7 (ours)	✔	✔	✔	✔	**W: 0.8375** **E: 0.8813**	W: 0.9600 **E: 0.7805**	**W: 0.7578** **E: 0.9846**	**W: 0.8470** **E: 0.8707**	S: 0.5998 **B: 0.3015**	S: 0.5591 **B: 0.1943**	S: 3.4063 **B: 19.3409**

C1–Window C2–Dose R1–SNR R2–BRISQUE.

Table 4
Comparison of model prediction results and assessment results of MOS scores.

	SCORE1			SCORE2			SCORE-MEAN		
	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
SNR	−0.1085	−0.1282	4.8172	−0.0974	−0.1273	4.4075	−0.1311	−0.1425	4.5136
BRISQUE	0.0504	0.0346	31.6993	−0.0799	−0.1355	31.8221	0.0059	−0.0387	31.7458
Window	−0.1241	−0.1191	5.7723	−0.0368	−0.0321	5.4927	−0.1201	−0.0932	5.5500
Exposure	**0.2866**	0.2921	5.6789	**0.1558**	0.1561	5.4552	**0.2805**	0.2638	5.4829
Fit	0.1889	**0.3128**	**2.4310**	−0.0479	**0.1972**	**1.7703**	0.0597	**0.3288**	**1.9264**

other's training part as a priori, and it is easy to overfit, resulting in the improvement effect of the multi-task model is not significant.

4.3.3. Fitting experiments

Then, we conduct a fitting experiment of MOS score. We use the trained model to predict the artificially scored data in the dataset and get the corresponding metric values of the four tasks. After multivariate linear fitting of the manually labeled scores and the predicted values of the model, a regression model with weight −0.0536, −0.0256, −0.1192, 1.3802 and intercept 6.1930 was obtained, and the final prediction quality score was further given.

At the same time, three metrics of SRCC, PLCC and RMSE were used for evaluation, and the experimental results were shown in Tables 4 and 5. We calculated the correlation between the model predictions and the scores of the two radiologists, and calculated the correlation between the scores of the two radiologists as a comparison. The results show that our model has a good effect on RMSE metric. At the same time, we found that there was no high correlation between the scores of the two radiologists. This further indicates that the subjective IQA for medical images is greatly influenced by the subjective factors of the raters. Therefore, how to improve the scoring method of subjective assessment, reduce the influence of subjective factors, and provide a more accurate ground truth for objective assessment is also a problem worth thinking about.

Table 5
Assessment results between the predictions of our model and the results of two radiologists' scores and between two radiologists' scores.

Compare	SRCC	PLCC	RMSE
1	0.1889	0.3127	2.4309
2	−0.0479	0.1971	**1.7703**
MEAN	0.0597	0.3288	1.9246
MOS	**0.5692**	**0.5952**	1.9416

4.3.4. Comparison experiment

In the process of conducting comparative experiments, we reviewed the existing studies on CT-IQA using deep learning methods [37–39]. However, these models were trained with artificially synthesized 2D images with small amount of data and were trained with clinicians' scores as labels. As mentioned above, it costs a lot of time and labor to completely use clinicians for annotation, and according to our research when constructing the data set, the subjective error between different clinicians' scores is very large (see MOS score comparison in Table 5), so it is difficult to effectively verify the effect of the model.

Our data came from patients undergoing chest physical examination, all of which were close to normal images, with small individual differences. After clinician interpretation, image quality did not affect the detection and diagnosis of lesions. Under this condition, it can be regarded as a special natural image, which is suitable for the evaluation of natural images. The assessment of natural images has

Table 6
Comparison of correlation between different methods and human scores.

Model	Clip-IQA + [40]	DBCNN [41]	NIMA [42]	NIQE [43]	PaQ2PiQ [44]	CNNIQA [45]	ILNIQE [46]	Ours
SRCC	0.2612	0.2239	0.1164	0.0111	**0.2775**	0.0675	−0.1994	0.0597
PLCC	0.1972	0.1596	0.1309	−0.0510	0.2283	0.0317	−0.2415	**0.3288**
RMSE	5.4818	38.7027	2.1114	2.0289	65.1108	7.9608	44.1136	**1.9246**

been extensively verified, and the performance is stable and reliable. Therefore, we used the IQA method of natural images to assess our data and calculated the correlation with human scores, and the experimental results are shown in Table 6. The results show that our model is more beneficial to CT image quality assessment.

5. Limitation

However, our model still has some limitations. Firstly, in the construction of datasets, we need to increase the number of images, and use multi-center data to avoid possible limitations. At the same time, as mentioned above, we also need to further improve the subjective assessment method of data. Secondly, BRISQUE, as a no-reference metric of natural images, has certain computational errors when dealing with such special images as medical images. How to improve this situation, find more appropriate medical image quality assessment indicators and develop more suitable medical image characteristics assessment indicators are also an important direction that we need to continue to explore in the future. Finally, the performance of our model still needs to be improved. In the future, we can improve the learning ability of neural network by adding modules such as attention mechanism and transformer to obtain a quality assessment model that is more in line with human perception.

6. Conclusion

In this paper, the window classification and quality assessment of common CT images in medical images are studied. Specifically, we first constructed a dataset for chest CT image quality assessment. Based on this dataset, we propose a model for automatic window classification and quality assessment of CT images. Based on the dataset and the model, we explore the bifurcation of the fully connected layer and the influence of the multi-task module on the model performance. At the same time, the model prediction results and human perception scores were fitted and verified. The experimental results show that our model can achieve automatic window classification and quality assessment of CT image to a certain extent. And There is no significant difference between the prediction results and the mean MOS scores and the correlation between the two different MOS scores. In the future, increasing the number of datasets and using multi-center data, finding and developing metrics more suitable for medical IQA, adding innovative modules and improving model performance will become our important research directions.

CRediT authorship contribution statement

Siyi Xun: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Mingfeng Jiang:** Writing – review & editing. **Pu Huang:** Writing – review & editing. **Yue Sun:** Writing – review & editing. **Dengwang Li:** Writing – review & editing. **Yan Luo:** Data curation. **Huifen Zhang:** Data curation. **Zhicheng Zhang:** Writing – review & editing. **Xiaohong Liu:** Conceptualization, Methodology. **Mingxiang Wu:** Conceptualization, Methodology. **Tao Tan:** Conceptualization, Methodology, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is partly supported by Macao Polytechnic University Grant (RP/FCSD-01/2022, RP/FCA-05/2022), and Science and Technology Development Fund of Macao (0105/2022/A, 0021/2022/AGJ).

References

[1] H. Kasban, M. El-Bendary, D. Salama, A comparative study of medical imaging techniques, Int. J. Inf. Sci. Intell. Syst. 4 (2) (2015) 37–58.
[2] J. Carmichael, European Guidelines on Quality Criteria for Diagnostic Radiographic Images, Office for Official Publications of the European Communities, 1996.
[3] C. Maccia, B. Moores, B. Wall, et al., The 1991 CEC trial on quality criteria for diagnostic radiographic images: detailed results and findings, (No Title) (1996).
[4] J. Boita, R.E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, A. Bolejko, S. Zackrisson, M.G. Wallis, D.M. Ikeda, C. Van Ongeval, et al., How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography? Eur. Radiol. 31 (2021) 5335–5343.
[5] Z. Al-Ameen, G. Sulong, Prevalent degradations and processing challenges of computed tomography medical images: A compendious analysis, Int. J. Grid Distrib. Comput. 9 (10) (2016) 107–118.
[6] U. Pilania, A. Dagar, S. Aggarwal, A. Pathak, A study of issues and challenges with digital image processing, Comput. Intell. Anal. Inf. Syst. (2023) 3–14.
[7] L.W. Goldman, Principles of CT and CT technology, J. Nucl. Med. Technol. 35 (3) (2007) 115–128.
[8] G. Zhai, Recent advances in image quality assessment, Vis. Signal Qual. Assess.: Qual. Exp. (QoE) (2014) 73–97.
[9] L.S. Chow, R. Paramesran, Review of medical image quality assessment, Biomed. Signal Process. Control 27 (2016) 145–154.
[10] K. Sim, J. Yang, W. Lu, X. Gao, Mad-DLS: mean and deviation of deep and local similarity for image quality assessment, IEEE Trans. Multimed. 23 (2020) 4037–4048.
[11] H. Zhu, L. Li, J. Wu, W. Dong, G. Shi, MetaIQA: Deep meta-learning for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14143–14152.
[12] O. Messai, F. Hachouf, Z.A. Seghir, Adaboost neural network and cyclopean view for no-reference stereoscopic image quality assessment, Signal Process., Image Commun. 82 (2020) 115772.
[13] Y. Gao, Y. Cao, T. Kou, W. Sun, Y. Dong, X. Liu, X. Min, G. Zhai, VDPVE: VQA dataset for perceptual video enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1474–1483.
[14] T. Kou, X. Liu, W. Sun, J. Jia, X. Min, G. Zhai, N. Liu, Stablevqa: A deep no-reference quality assessment model for video stability, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1066–1076.
[15] Y. Dong, X. Liu, Y. Gao, X. Zhou, T. Tan, G. Zhai, Light-VQA: A multi-dimensional quality assessment model for low-light video enhancement, 2023, arXiv preprint arXiv:2305.09512.
[16] M. Hu, G. Zhai, R. Xie, X. Min, Q. Li, X. Yang, W. Zhang, A wavelet-predominant algorithm can evaluate quality of THz security image and identify its usability, IEEE Trans. Broadcast. 66 (1) (2019) 140–152.
[17] Z. Zhang, C. Li, W. Sun, X. Liu, X. Min, G. Zhai, A perceptual quality assessment exploration for AIGC images, 2023, arXiv preprint arXiv:2303.12618.
[18] Z. Zhang, W. Sun, Y. Zhou, H. Wu, C. Li, X. Min, X. Liu, G. Zhai, W. Lin, Advancing zero-shot digital human quality assessment through text-prompted evaluation, 2023, arXiv preprint arXiv:2307.02808.
[19] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, W. Lin, AGIQA-3K: An open database for AI-generated image quality assessment, 2023, arXiv preprint arXiv:2306.04717.
[20] C. Li, H. Wu, Z. Zhang, H. Hao, K. Zhang, L. Bai, X. Liu, X. Min, W. Lin, G. Zhai, Q-Refine: A perceptual quality refiner for AI-generated image, 2024, arXiv preprint arXiv:2401.01117.
[21] V. Mudeng, M. Kim, S.-w. Choe, Prospects of structural similarity index for medical image analysis, Appl. Sci. 12 (8) (2022) 3754.

[22] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2012) 209–212.
[23] A. Saha, Q.M.J. Wu, Utilizing image scales towards totally training free blind image quality assessment, IEEE Trans. Image Process. 24 (6) (2015) 1879–1892.
[24] M. Outtas, L. Zhang, O. Deforges, W. Hammidouche, A. Serir, C. Cavaro-Ménard, A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images, in: 2016 International Symposium on Signal, Image, Video and Communications, ISIVC, IEEE, 2016, pp. 308–313.
[25] A. Mittal, A.K. Moorthy, A.C. Bovik, Blind/referenceless image spatial quality evaluator, in: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers, ASILOMAR, IEEE, 2011, pp. 723–727.
[26] L.S. Chow, H. Rajagopal, Modified-BRISQUE as no reference image quality assessment for structural MR images, Magn. Reson. Imaging 43 (2017) 74–87.
[27] M. Oszust, A. Piórkowski, R. Obuchowicz, No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features, Magn. Reson. Med. 84 (3) (2020) 1648–1660.
[28] Y. Guo, M. Hu, X. Min, Y. Wang, M. Dai, G. Zhai, X.-P. Zhang, X. Yang, Blind image quality assessment for pathological microscopic image under screen and immersion scenarios, IEEE Trans. Med. Imaging (2023).
[29] S. Li, J. He, Y. Wang, Y. Liao, D. Zeng, Z. Bian, J. Ma, Blind CT image quality assessment via deep learning strategy: initial study, in: Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, volume 10577, SPIE, 2018, pp. 293–297.
[30] Q. Gao, S. Li, M. Zhu, D. Li, Z. Bian, Q. Lv, D. Zeng, J. Ma Sr, Combined global and local information for blind CT image quality assessment via deep learning, in: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, volume 11316, SPIE, 2020, pp. 242–247.
[31] J. Greffier, J. Frandon, S. Si-Mohamed, D. Dabli, A. Hamard, A. Belaouni, P. Akessoul, F. Besse, B. Guiu, J.-P. Beregi, Comparison of two deep learning image reconstruction algorithms in chest CT images: a task-based image quality assessment on phantom data, Diagn. Interv. Imaging 103 (1) (2022) 21–30.
[32] J.H. Lee, B.R. Grant, J.H. Chung, I. Reiser, M. Giger, Assessment of diagnostic image quality of computed tomography (CT) images of the lung using deep learning, in: Medical Imaging 2018: Physics of Medical Imaging, volume 10573, SPIE, 2018, pp. 399–405.
[33] C.T. Jensen, X. Liu, E.P. Tamm, A.G. Chandler, J. Sun, A.C. Morani, S. Javadi, N.A. Wagner-Bartak, Image quality assessment of abdominal CT by use of new deep learning image reconstruction: initial experience, Am. J. Roentgenol. 215 (1) (2020) 50–57.
[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[35] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, Adv. Neural Inf. Process. Syst. 31 (2018).
[36] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623.
[37] S. Li, J. He, Y. Wang, Y. Liao, D. Zeng, Z. Bian, J. Ma, Blind CT image quality assessment via deep learning strategy: initial study, in: Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, volume 10577, SPIE, 2018, pp. 293–297.
[38] Q. Gao, S. Li, M. Zhu, D. Li, Z. Bian, Q. Lyu, D. Zeng, J. Ma, Blind CT image quality assessment via deep learning framework, in: 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), IEEE, 2019, pp. 1–4.
[39] Q. Gao, M. Zhu, D. Li, Z. Bian, J. Ma, CT image quality assessment based on prior information of pre-restored images, Nan Fang yi ke da xue xue bao J. South. Med. Univ. 41 (2) (2021) 230–237.
[40] J. Wang, K.C. Chan, C.C. Loy, Exploring clip for assessing the look and feel of images, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 2555–2563.
[41] A. Khmag, N. Kamarudin, Natural image deblurring using recursive deep convolutional neural network (R-DbCNN) and second-generation wavelets, in: 2019 IEEE International Conference on Signal and Image Processing Applications, ICSIPA, IEEE, 2019, pp. 285–290.
[42] H. Talebi, P. Milanfar, NIMA: Neural image assessment, IEEE Trans. Image Process. 27 (8) (2018) 3998–4011.
[43] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2012) 209–212.
[44] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, A. Bovik, From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3575–3585.
[45] R. Li, H. Yang, T. Yu, Z. Pan, CNN model for screen content image quality assessment based on region difference, in: 2019 IEEE 4th International Conference on Signal and Image Processing, ICSIP, IEEE, 2019, pp. 1010–1014.
[46] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.