Computers and Electrical Engineering 81 (2020) 106529

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A bone age assessment system for real-world X-ray images based on convolutional neural networks☆

Jiajia Guo[a], Jianyue Zhu[b], Hongwei Du[a,*], Bensheng Qiu[a]

[a] University of Science and Technology of China, Hefei, Anhui 230026, China
[b] National Mobile Communications Research Lab, School of Information Science and Engineering, Southeast University, Nanjing, Jiangsu 211189, China

ARTICLE INFO

Article history:
Received 12 October 2018
Revised 8 October 2019
Accepted 26 November 2019
Available online 26 December 2019

Keywords:
Bone age
Poor quality X-ray image
Automated assessment system
Convolutional neural networks
Quality improvement
Deep learning

ABSTRACT

It is of vast significance to assess the bone age of hand radiographs automatically in pediatric radiology and legal medicine. In the literature, many papers focus on improving the assessment accuracy but neglecting the existence of poor-quality X-ray images. However, in real medical scenarios, the existence of poor-quality X-ray images is unavoidable. To tackle this problem, we propose a bone age assessment system for real-world X-ray images. Specifically, we first establish a regression model 'BoNet+' based on densely connected convolutional networks. Then, to handle poor-quality X-ray images, we introduce three model architectures that are different in the way of improving image quality. Experiment results show that the proposed models can estimate the bone age of poor-quality images accurately. We also tentatively put forward that if the expressivity of CNN model is enough high, multiple tasks can be handled together just by a single model.

1. Introduction

Bone age assessment (BAA) is one of the most commonly used clinical diagnosis technologies in the evaluation of children's skeletal maturation in pediatric radiology and legal medicine [1,2]. In medical scenarios, doctors usually compare bone age with chronological age when diagnosing children growth, metabolic disorders, genetic disorders, etc. Roughly estimated, about 76% of radiologists assess bone age via comparing X-ray images of left hand with a reference atlas for its high speed and simplicity. However, this method is intuitive, experience-based, and prone to inter- and intra-observer variablility, i.e., the accuracy of the BAA highly depends on the medical practitioners' ability.

For the above-mentioned reasons, about 30 years ago, a BAA system called HANDX [3] was proposed to replace manual assessment by a semi-automated computer-aided diagnosis system. Since then, many other semi-automated and automated systems [4–9] have been continuously established. However, most of them just focused on the assessment of X-ray images of good quality while there are still many X-ray images of poor quality needing to be assessed in a real-world scenario [4,6,10–14].

Specifically, in [10], the authors found corrupted images (e.g., a rectangle label across the target bone) were rejected by their proposed system; Reference [13] also showed the most commonly provided file format in teleconsultations was

☆ This paper is for CAEE special section SI-mip. Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Li He.

* Corresponding author.
E-mail addresses: jjguo@mail.ustc.edu.cn (J. Guo), zhujy@seu.edu.cn (J. Zhu), duhw@ustc.edu.cn (H. Du), bqiu@ustc.edu.cn (B. Qiu).

https://doi.org/10.1016/j.compeleceng.2019.106529
0045-7906/

lossy JPEG image files with low resolution, and it solved this problem by creating new DICOM images dependent on the existence of the caliper scale. In addition, the authors in [9,15] found their system rejected poor-quality X-ray images with much noise. The noise in X-ray image can be regarded as Poisson noise, because the quantum fluctuations in the number of X-rays generated follows a Poisson statistic [15].

If the system rejects these poor-quality X-ray images, patients will need another X-ray examination [10–14]. Unfortunately, repeated X-ray examinations not only bring patients inconvenience and financial implications but also increase radiation exposure. A much more serious situation is that some patients have difficulty in taking an additional examination under special conditions (e.g., remote medical treatment) [13].

To tackle this problem, we build a fully automated BAA system based on convolutional neural networks (CNN) to assess good-quality and poor-quality X-ray images together. Specifically, we first design a baseline CNN model called "BoNet+" based on DenseNet [16] to assess good-quality X-ray images and investigate the choice of loss function. Then, we evaluate some factors of building a real-world BAA system, including the strategies of processing poor-quality X-ray images. Generally, we propose a more accurate real-world BAA system based on these existing works. The key idea of this paper is to tackle the problem of poor-quality X-ray images, and the contributions in this paper are summarized as follows:

1. Among the existing works, we are the first to find mean absolute error (MAE) loss is much more suitable in BAA problem than mean square error (MSE) loss, which is helpful to build a more accurate BAA system.
2. We propose a new BAA system BoNet+ and achieve a MAE of 0.76 years, which outperforms the state-of-the-art performance [8] using the same dataset [5].
3. To the best of our knowledge, it is the first study on the assessment of poor-quality images. We propose a new architecture for the real-world assessment, which overcomes the limitations of the existing BAA system in the assessment of poor-quality X-ray images.
4. We find it possible for a single model to handle multiple kinds of poor-quality X-ray images when quality improvement model is of high expressivity.

The remainder of this paper is organized as follows: Section 2 briefly summarizes several representative BAA systems. Section 3 describes the proposed baseline model 'BoNet+', real-world BAA systems, etc. Section 4 gives the simulation results and analyzes these results in detail, while the conclusion is given in Section 5.

2. Related work

In this section,we chiefly introduce the most prominent two BAA systems, BoneXpert [6] and BoNet [8].

BoneXpert is the first commercial BAA system and has been applied in over 100 European hospitals. It automatically segmented hand bones by active appearance model and then assessed bone age by the intensity, shape, and texture scores of the 13 particular bones derived from principal component analysis [6]. Many similar works [4,17,18] have been done over the past 20 years, all of which mainly focused on designing a more accurate segmentation algorithm, a better feature extractor, and a better learning algorithm. Inspired by this kind BAA method, [19] proposed an automated neural network-based pipeline including segmenting, standardizing and preprocessing input radiographs, and performing BAA.

Essentially different from traditional machine learning-based BoneXpert, BoNet is the first to apply deep learning to BAA problem. This model was composed of multiple neural network layers to learn representations of input X-ray images with multiple levels of abstractions end-to-end. In other words, good BAA features were learned automatically without domain expertise and engineering skill while classical machine learning methods highly depend on researchers' experience. Experiment results showed BoNet outperformed all classical machine learning solutions by a margin.

Recently, [1] found a deep-learning BAA system could assess bone age with an accuracy much higher than classical models and similar to expert radiologists. This result represents the great potential of deep learning and encourages other researchers to study further in the CNN-based BAA system (e.g., [2,20]).

3. Real-world BAA system

In this section, we first propose a baseline BAA model based on densely connected convolutional networks for high-quality X-ray images and investigate the choice of loss function. Then, we introduce three different architectures for the assessment of poor-quality X-ray images.

To the best of our knowledge, there are no works focusing on the choice of loss function in BAA problem and the assessment of poor-quality X-ray images. However, on one hand, we investigate the performance of different loss functions,e.g., MAE loss and MSE loss. On the other hand, we propose three different representative system architectures for real-world X-ray images especially for poor-quality images.

3.1. BoNet+ model

3.1.1. A baseline CNN architecture

BAA problem can be seen as a regression problem [8], or a classification problem [1,19]. In [21], the authors compared the performance of regression model and classification model in age estimation problem. Experiment results showed regression

Table 1
BoNet+ architecture.

Layers	Output Size	BoNet+
Conv(1)	96 256 × 256	7 × 7 conv, stride 1
Conv(2)	96 256 × 256	5 × 5 conv, stride 1
Pooling(1)	96 128 × 128	3 × 3 max pool, stride 2
DB(1)	160 128 × 128	3 × 3 conv, stride 1, *gr* 16
TL(1)	128 128 × 128	1 × 1 conv
	128 64 × 64	3 × 3 max pool, stride 2
DB(2)	256 64 × 64	3 × 3 conv, stride 1, *gr* 32
TL(2)	128 64 × 64	1 × 1 conv
	128 32 × 32	3 × 3 max pool, stride 2
DB(3)	256 32 × 32	3 × 3 conv, stride 1, *gr* 32
TL(3)	128 32 × 32	1 × 1 conv
	128 16 × 16	3 × 3 max pool, stride 2
DB(4)	25616 × 16	3 × 3 conv, stride 1, *gr* 32
TL(4)	128 16 × 16	1 × 1 conv
	128 8 × 8	3 × 3 max pool, stride 2
DB(5)	384 8 × 8	3 × 3 conv, stride 1, *gr* 64
TL(5)	256 8 × 8	1 × 1 conv
	256 4 × 4	3 × 3 max pool, stride 2
DB(6)	512 4 × 4	3 × 3 conv, stride 1, *gr* 64
TL(6)	256 4 × 4	1 × 1 conv
	256 2 × 2	3 × 3 max pool, stride 2
TL(7)	1024 2 × 2	1 × 1 conv
	1024 1 × 1	2 × 2 max pool, stride 2
Conv(3)	1024 1 × 1	1 × 1 conv, stride 1
Conv(4)	512 1 × 1	1 × 1 conv, stride 1
OutLayer	1 × 1	1 × 1 conv, stride 1

Note that in this table all 'conv' layers correspond the block BN_RELU_CONV in Fig. 2, all 'DB' layers correspond the dense block, all 'TL' layers correspond the transition layer and parameter *gr* corresponds the growth rate for dense block [16].

model performed much better than classification model. Thus, we directly regard BAA problem as a regression problem in this work.

Owing to the great success achieved by DenseNet in computer vision problem, our BAA model is mainly composed of several dense blocks. As can be seen in Fig. 3, dense block is an iterative concatenation of previous feature maps. Therefore, each layer is able to access all previous layers easily and the information of preceding feature maps is fully reused. Furthermore, from the parameter usage perspective, this block is much more efficient than any other model.

As illustrated in Fig. 4, the input of BoNet+ is a 256 × 256 normalized gray X-ray image. The first two BN_RELU_CONV layers respectively have 96 7 × 7 filters and 96 5 × 5 filters with a stride 1, followed by a 3 × 3 max pooling layer with a stride 2. The following parts are six dense blocks, respectively followed by a transition layer. Then, there is a BN_RELU_CONV layer with 1024 1 × 1 filters and a stride 1, followed by a 2 × 2 max pooling layer with a stride 2. The last three layers are three BN_RELU_CONV layers respectively with 1024, 512, and 1 1 × 1 filters with a stride 1. The details are shown in Table 1.

3.1.2. BAA dataset

Although many researches have been done on BAA problem, there are nearly no public bone age datasets, leading to the difficulty in the comparison of different algorithms. The Digital Hand Atlas Database [5], which contains about 1400 X-ray left-hand images, is one of the few public datasets.

Different from other BAA datasets, ages in this dataset range from 0 to 19 years and the race includes Asian, Black, Caucasian, and Hispanic. The large age range and multiple kinds of the race make it a much more challenging work to estimate the bone age of this dataset. The bone age of each X-ray image was provided by two expert radiologists and in our experiment, we simply use the mean value of the two prediction ages as the ground truth. Some example X-ray images in this dataset are shown in Fig. 5. We can see that these X-ray images vary greatly in brightness, contrast, and hand placement, dramatically increasing the difficulty of age prediction.

To simulate the real-world X-ray images of poor-quality, we artificially add Poisson noise or black labels to the images and decrease the resolution by 4x and the input of the network is the blurred low resolution image, as illustrated in Fig. 1. To prevent the labels placed at black background, we only add them on the central part (60% of the full image). To prevent overfitting and improve the model's performance, we conduct our experiment on this dataset via using 5-fold cross validation when compared with the state-of-the-art BAA method and data augmentation including random rotation, shift, zoom, etc, as illustrated in Fig. 6.

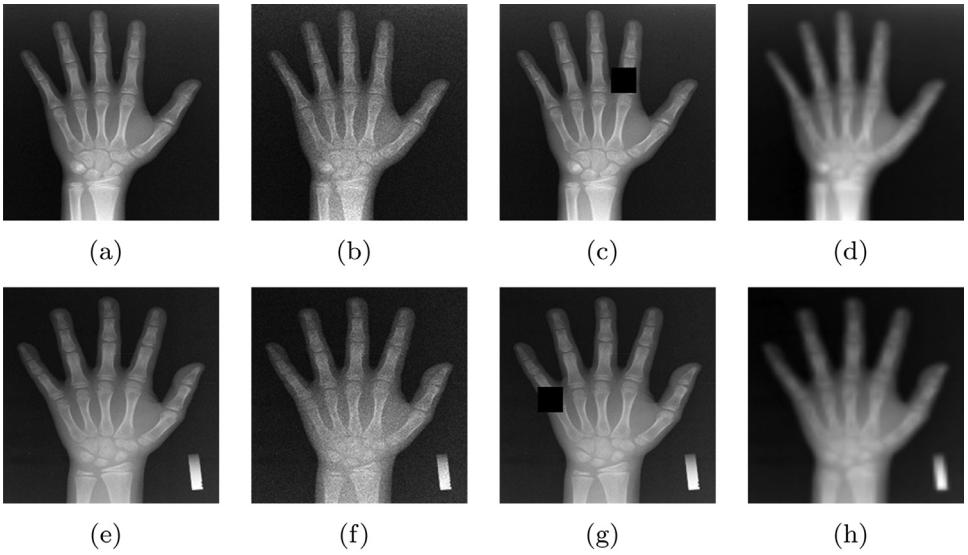


Fig. 1. Simulation examples of X-ray images of good or poor quality. a e: origin images. b f: images with Poisson noise. c g: corrupted images with label. d h: low-resolution images.

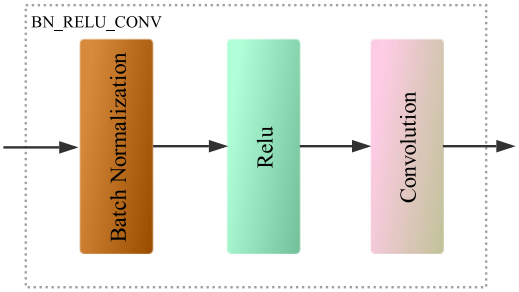


Fig. 2. Convolutional block.

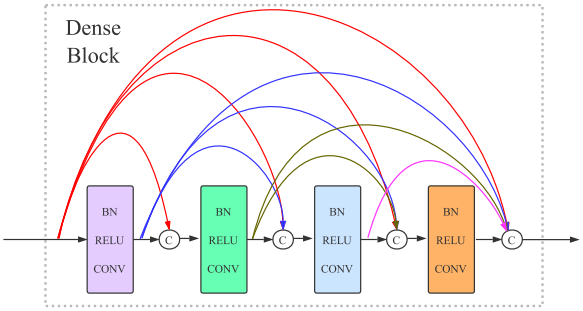


Fig. 3. Dense block.

3.1.3. Evaluation metrics

The most commonly used evaluation metrics in regression problem are MAE and median absolute error (MdAE), which are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\overline{y_i} - y_i|, \tag{1}$$

$$MdAE = median|\overline{y} - y|, \tag{2}$$

where N is the number of training samples, $\overline{y_i}$ is the prediction bone age and y_i is the ground truth of the ith sample. As described in [8], MAE has been employed in nearly all previous works using the above-mentioned dataset. Hence, in order to directly compare BoNet+ with previous techniques, we use MAE as our basic metric.

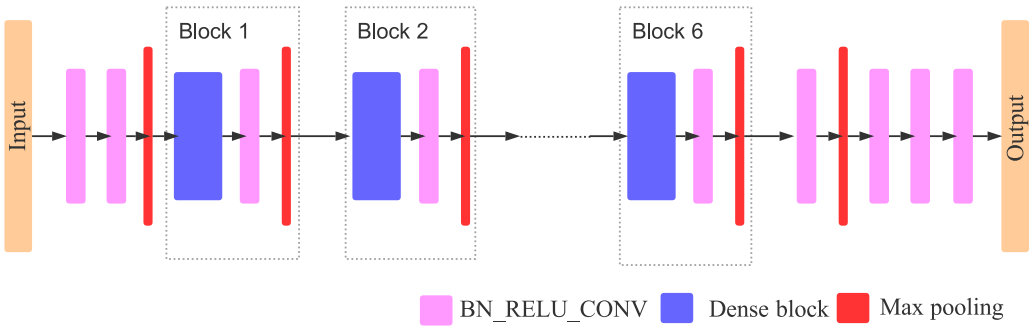


Fig. 4. Overview of BoNet+ architecture. It consists of six dense blocks, multiple convolutional layers, multiple batch normalization layers, and pooling layers. The input is a 256 × 256 gray image and the output is a neuron which represents the predicted bone age.

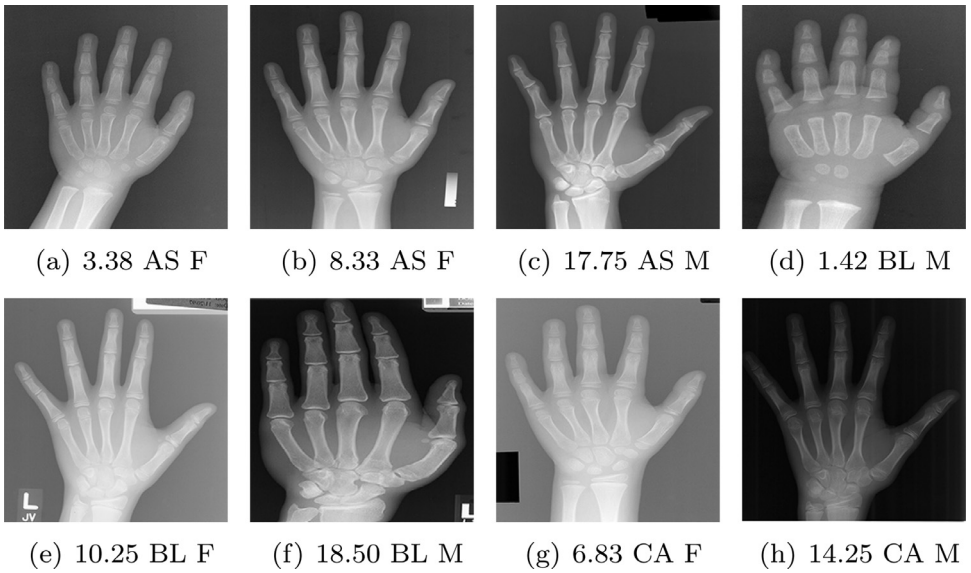


Fig. 5. Examples of X-ray images in the Digital Hand Atlas Database. AS: Asian, BL: Black, CA: Caucasian, HI: Hispanic; M: Male, F: Female.

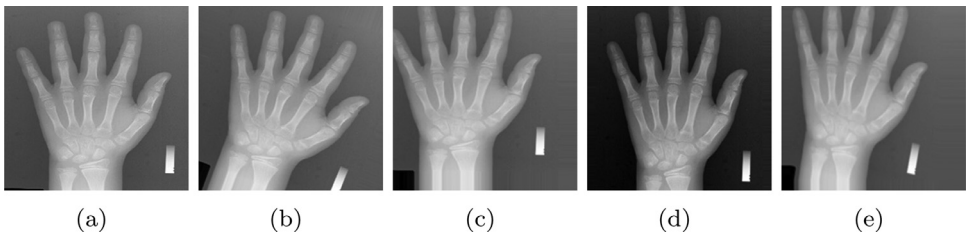


Fig. 6. Examples of data augmentation. a: origin image. b: random rotation. c: random shift. d: random zoom. e: augmentation combination.

Although employed widely, MAE has some inevitable drawbacks. When MAE value has been quite low, a little improvement can not make obvious changes. So, the authors in [21] proposed an evaluation metric called Cumulative Correct Score (CCS). This metric is defined as the total number of the evaluated images, whose error values are under a year threshold, which is defined as follows:

$$CCS(t) = \sum_{i=1}^{N} h(|\overline{y_i} - y_i| - t), \tag{3}$$

$$h(x) = \begin{cases} 1 & x \le 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

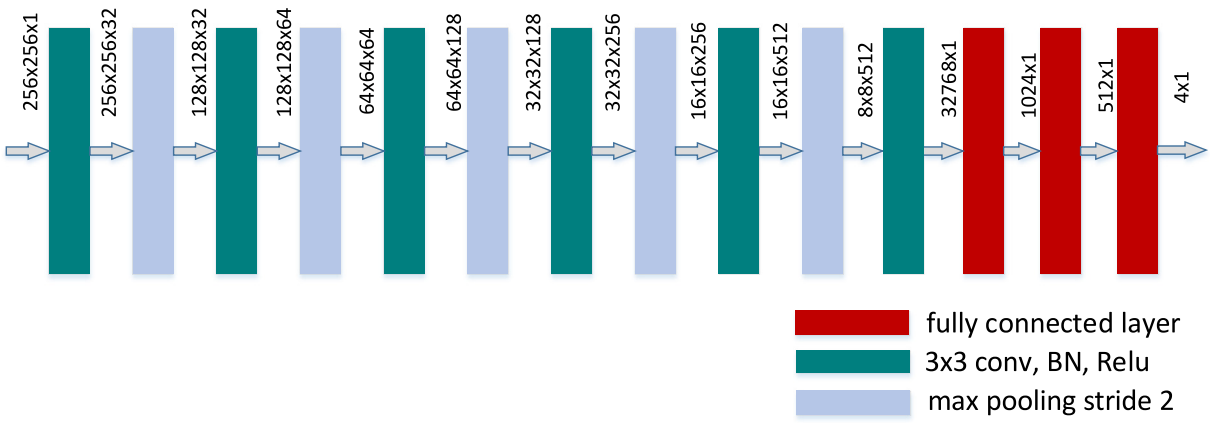


Fig. 7. Classification model based on VGG.

where N is the total number of evaluated X-ray images, $\overline{y_i}$ is the ith predicted bone age, y_i is the ith ground truth bone age, and t is the error threshold. Compared with MAE and MdAE, CCS can intuitively reflect the distribution of the error and show very minor improvement. Thus, we use CCS as an additional evaluation metric in this study.

3.2. Loss function

Loss function is one of the most significant factors in deep learning, because optimization can be regarded as a process of minimizing the target loss function. BAA problem in this study is regarded as a regression problem (Section 3.1.1) and in this kind of problems, MSE is the preferred loss function for its great mathematical properties (e.g., convexity, simplicity, and strictly differentiable), which is defined as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\overline{y_i} - y_i)^2. \tag{5}$$

To the best of our knowledge, all of previous BAA regression models used MSE loss, a smooth function and easy to optimize. In contrast, MAE loss is not fully smooth, leading to a great difficulty in optimization. So, there has been no research on the BAA model with MAE loss, while it has achieved great success in age estimation [21]. Therefore, in this work, we investigate whether MAE loss also performs well in BAA problem.

$$L_{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\overline{y_i} - y_i|. \tag{6}$$

3.3. Model architecture

In this subsection, we study several ways to design a model architecture to deal with both poor-quality and high-quality images together. The key point of BAA assessment of poor-quality images is how to improve the image quality. Generally speaking, it is a brilliant choice to handle different kinds of poor-quality images using different ways, which very likely make the model more accurate. In the following, we will verify whether it is correct.

3.3.1. Classification and quality improvement model architecture

As mentioned in Section 1, there are mainly three kinds of poor-quality images including images with Poisson noise, corrupted images with label, and low-resolution images. It is a natural way to process these images according to the kinds of poor quality. Therefore, we first use a classification model based on VGG [22] to judge the kinds of the input images. As shown in Fig. 7, the input of the classification model is a gray X-ray image with poor quality or high quality. And a softmax classification layer is finally applied to the last 4-D fully-connected layer with a cost function of cross entropy.

Then, different kinds of X-ray images are sent to corresponding quality improvement network based on U-Net [23] as shown in Fig. 8. U-Net architecture is first applied in biomedical image segmentation and researchers find that U-Net also has an excellent performance on image denoising, super resolution, etc. Therefore, we employ U-Net to improve the quality of X-ray images as shown in Fig. 8. Moreover, inspired by [24], the proposed model also adopts the residual learning formulation as shown in Fig. 9, where a single residual unit is employed to predict the residual images.

The quality improvement network, as shown in Fig. 8, is composed of a contracting path (left part) and an expansive path (right part) [23]. On one hand, the contracting path is similar to the classical architecture of a CNN model, consisting of repeated application of two 3×3 convolution layers, each followed by a 2×2 max pooling layer with stride 2.

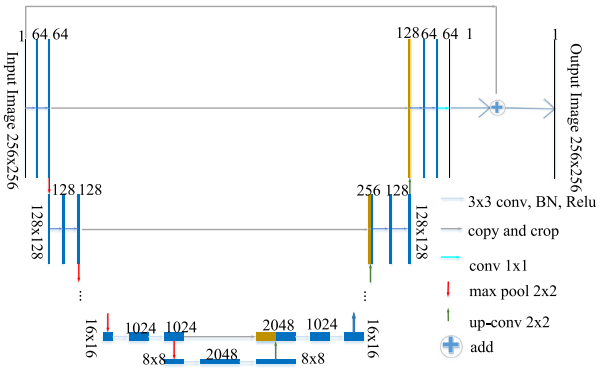


Fig. 8. Quality improvement network with residual U-Net architecture.

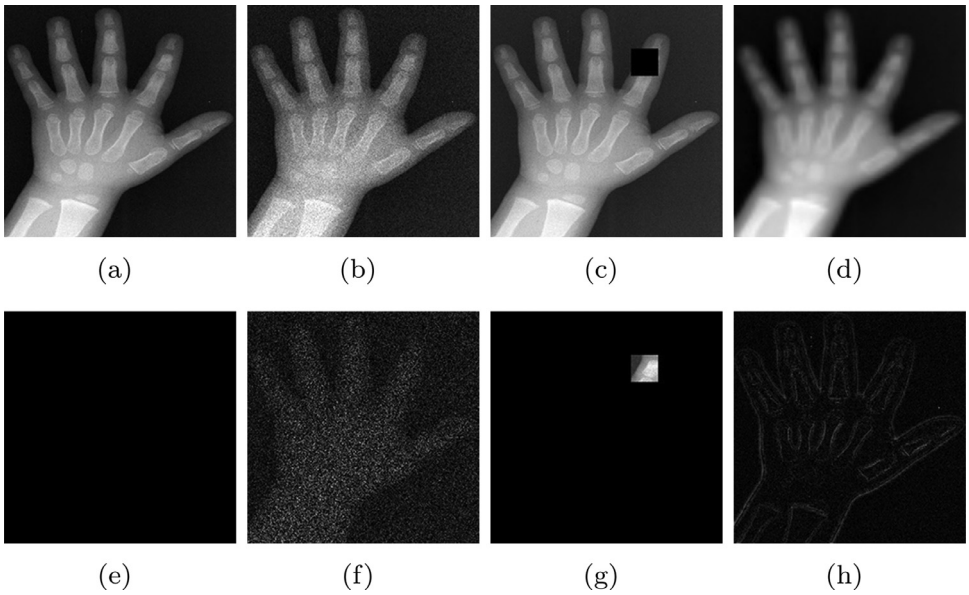


Fig. 9. Examples of origin and residual images. a: high-quality image. b: image with Poisson noise. c: corrupted image with label. d: low-resolution image. e~h: corresponding residual images.

Moreover, at each downsampling step, the number of feature channels are doubled. On the other hand, every step in the expansive path is composed of an upsampling of feature maps followed by a 2 × 2 convolution ("up-convolution") for halving the number of feature channels, a concatenation with the corresponding feature maps from the contracting path, and two 3 × 3 convolution layers. Finally, a 1 × 1 convolution layer is applied to map the 64-component feature vector to the desired residual gray images. In total, we use 23 convolutional layers, 5 max pooling layers, 5 up-convolution layers, and 5 concatenation layers in this network.

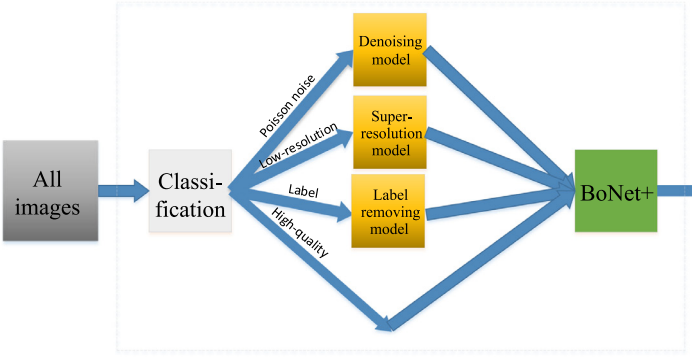
After quality improvement, processed images are sent to BoNet+ to assess the bone age. The whole architecture is illustrated in Fig. 10(a) and we call this model architecture 'BoNet+_CQ' (classification and quality improvement).

3.3.2. Direct quality improvement model architecture

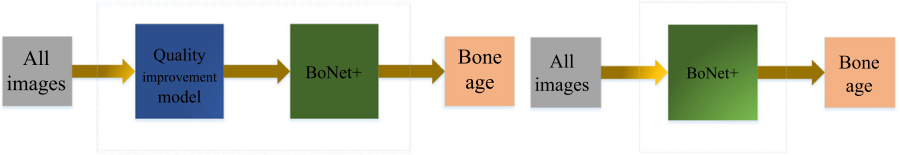
Although classification and quality improvement are a natural method, BoNet+_CQ model costs much time and needs a lot of computational power. In clinical scenarios, these requirements are difficult to be met. Besides, the accuracy will heavily depend on the performance of classification part. Thence, we wonder whether it is possible to only use a single quality improvement model to handle all kinds of X-ray images as illustrated in Fig. 10(b). In other words, whatever kinds of quality these images have, we directly send them to a quality improvement model and then regard the output as the input of BoNet+. We name this model architecture 'BoNet+_DQ' (direct quality improvement).

3.3.3. No quality improvement model architecture

The above two architectures (BoNet+_CQ and BoNet+_DQ) both consider the effect of poor-quality images on BAA problem. In order to show the importance of this consideration, we propose a simple model called 'BoNet+_NQ' (no quality



(a) BoNet+_CQ model architecture



(b) BoNet+_DQ model architecture (c) BoNet+_NQ model architecture

Fig. 10. Overview of different model architectures.

improvement), not considering the problem of poor-quality images and directly using all images as the input of BoNet+ as shown in Fig. 10(c).

4. Experiment results and analyses

In this section, we first describe the experiment setting details. Then, we report the experiment results of BoNet+ with different loss functions and compare the best result with the state-of-the-art result using the same dataset. Finally, we give simulation results of real-world BAA model using different model architectures.

4.1. Experiment settings

The implementation of all models has been done in TensorLayer toolbox, a high-level Python library extended from TensorFlow and on a NVIDIA GTX 1080 Ti GPU. We train all networks using the Adam optimizer by setting $\beta_1 = 0.9$, $\beta_2 = 0.9999$ and $\epsilon = 10^{-8}$ from scratch similar to [8].

When comparing the performance of different loss functions, the batch size is set to 20 and the learning rate is set to 0.0001. It takes about 30 s per epoch and the whole training process takes about 1.5 h.

When comparing the performance of different model architectures, the batch size of classification model is set to 16 and that of quality improvement model is set to 8. The learning rates of them are all set to 0.0002. The classification model takes about 17 s per epoch and the whole training process takes about 1 h. Each quality improvement model takes about 79 s per epoch and the whole training process takes about 1.7 h. Additionally, we trained BoNet+_CQ and BoNet+_DQ sequentially instead of end-to-end.

4.2. Analyses of BoNet+ using different loss functions

To find a better loss function for BAA problem, we train BoNet+_MAE and BoNet+_MSE from the baseline architecture. Experiment results of two models are shown in Table 2 and Fig. 11.

We can see that BoNet+_MAE outperforms BoNet+_MSE under all evaluation metrics. To find the reasons behind the experiment results, we analyze the properties of MAE and MSE loss. First, since MAE is one of the evaluation metrics,

Table 2
The results of BoNet+ with different loss functions.

Model	MAE	MdAE
BoNet+_MAE	0.762 ± 0.103	0.603 ± 0.082
BoNet+_MSE	0.872 ± 0.182	0.774 ± 0.117

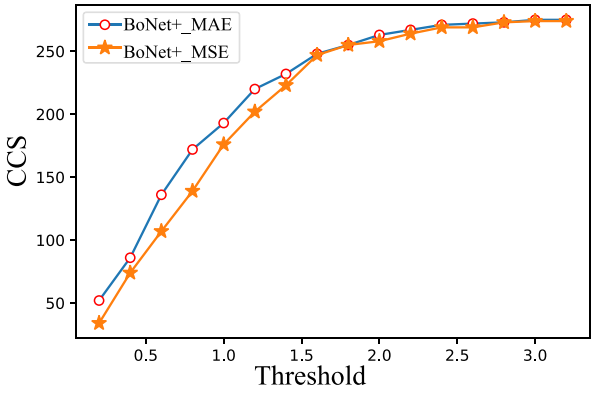


Fig. 11. CCS of BoNet+_MAE and BoNet+_MSE.

Table 3
Comparison with the state-of-the-art methods over the Digital Hand Atlas Database. (Part of the data is from [8].)

Method	Time	MAE
Hsieh et al. [4]	2007	2.57
Gertych et al. [5]	2007	2.10
Giordano et al. [7]	2016	1.82
Fine-tuned GoogLeNet [8]	2017	0.82
BoNet [8]	2017	0.79
BoNet+_MAE	**2018**	**0.76 ± 0.10**

optimizing MAE loss naturally improve the performance under MAE metric as shown in Table 2. Besides, MAE loss is not as sensitive as MSE loss to outliers. MSE loss squares the errors before they are averaged and hence models with MSE loss usually put more effort into reducing very big errors. However, in practice, label noises are unavoidable and hence sometimes model with MSE loss will focus on reducing few inaccurate label errors at the expense of other accurate labels. Therefore, as shown Fig. 11, performance gap between BoNet+_MAE and BoNet+_MSE with a low CCS threshold is quite large, while the one with a very high CCS threshold is very small.

4.3. Comparison with the state-of-the-art methods for high-quality X-ray images

In this subsection, we compare our proposed BoNet+_MAE with other BAA methods using the Digital Hand Atlas Database in Table 3. The first three methods were based on traditional machine learning and others were all based on deep learning.

BoNet+_MAE achieves the state-of-the-art accuracy, which is much higher than that of traditional methods and little higher than that of BoNet. However, our training speed is about 30 s per epoch (Section 4.1) while the training speed of BoNet was over 240 s per epoch using a similar GPU. The training speed of BoNet+ is much higher than that of BoNet. Besides, the number of parameters in BoNet+ is just less than 10 M while that of BoNet was about 40 M. These advantages are crucial in clinical application for the shortage of high-performance computers in most hospitals. Compared with traditional methods, BoNet+ also has a tremendous advantage over test speed. Although it takes much time to train deep learning-based models, these kinds of methods just take several milliseconds to assess an X-ray image while traditional methods usually need several seconds at least [8].

The contrast of the X-ray images is a fundamental problem when using traditional methods. Some references show that their system may reject the X-ray images for their poor contrast. We also notice this problem in our experiment. In the Digital Hand Atlas Database, there is an X-ray image (ID: 4185) with poor contrast as shown in Fig. 13. During the test period, the predicting error of this X-ray image is 0.83 years. It seems that the poor contrast does not affect the predicting accuracy. In [25], the authors provided an evaluation of 4 state-of-the-art deep neural network models for image classification under quality distortions. They considered five types of quality distortions: blur, noise, contrast, JPEG, and JPEG2000 compression. They also found neural networks shown resiliency with respect to contrast change while blur and noise affected classification seriously.

To show the generalization of our network, we use our trained network to predict the X-ray images in Chinese children's standard bone age map. The MAE of this test dataset is 1.017 years. When we trained our networks, we did not use any dataset from China children. However, it still achieved a good accuracy on this dataset. This shows the good generalization of our proposed network.

Table 4
The MAE performance of different model architectures for real-world BAA problem.

	High-quality image	Image with Poisson noise	Corrupted image with label	Low-resolution image
BoNet+	0.762	1.124	1.065	1.215
BoNet+_CQ	**0.762**	**0.909**	**0.804**	0.769
BoNet+_DQ	0.764	0.922	0.808	**0.767**
BoNet+_NQ	0.850	0.967	0.876	0.879

Table 5
The MdAE performance of different model architectures for real-world BAA problem.

	High-quality image	Image with Poisson noise	Corrupted image with label	Low-resolution image
BoNet+	0.603	0.828	0.774	1.071
BoNet+_CQ	**0.603**	0.788	**0.662**	0.714
BoNet+_DQ	0.627	**0.774**	0.663	**0.692**
BoNet+_NQ	0.697	0.825	0.752	0.799

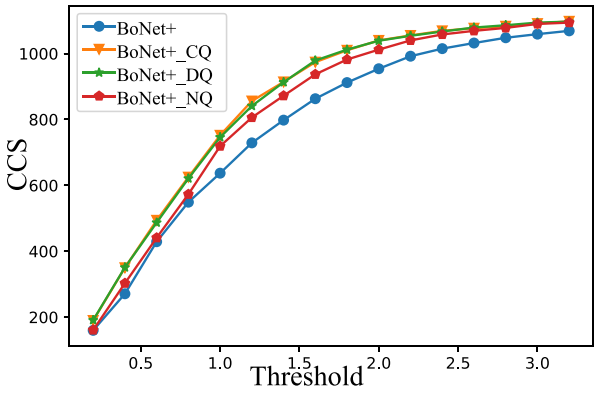


Fig. 12. CCS of different model architectures for real-world BAA problem.

4.4. Analyses of real-world BoNet+ using different model architectures

Since BoNet+ with MAE loss works much better than that with MSE loss, in this section, we compare the performance with different model architectures all based on BoNet+_MAE. Tables 4, 5 and Fig. 12 show the performance comparison of the proposed three model architectures and BoNet+ with different metrics.

The classification accuracy was over 98%. The errors incurred for the following reason. First, there are some poor-quality images in Digital Hand Atlas Database, and but we regarded all of them as high-quality images in classification part. In other words, sometimes the ground truth was wrong. Second and more commonly, the poor-quality images were created by simulation and labels were inevitably added to the black background. We regarded these images as corrupted images while in fact, they were high-quality.

Under all evaluation metrics, the proposed three model architectures consistently outperform single BoNet+. This underlines the necessity of building a system for real-world BAA problem. Moreover, the performance of BoNet+_CQ and BoNet+_DQ is much better than that of BoNet+_NQ, which does not improve the quality of poor-quality images. It suggests the necessity of quality improvement and if the input image quality is improved, the error will be much smaller. For the BoNet+_DQ, considering the case where all three types of poor quality are present in an image, we use X-ray images with all three types together to train BoNet+_DQ. Experiment results show that the MAE and the MdAE of performance of BoNet+_DQ there is 0.824 years and 0.733 years.

Under normal circumstances, it very likely works better if different kinds of poor-quality X-ray images are handled via different models. However, our experiment results are contrary to it and the performance of BoNet+_CQ and BoNet+_DQ is pretty near. Thus, we then study the performance of their quality improvement models. As can be seen in Fig. 14, the outputs of BoNet+_CQ and BoNet+_DQ are so similar that it is hard to distinguish them with the naked eye. Similarily, [24] also found that handling multiple denoising tasks (e.g., single image super-resolution, blind Gaussian denoising, and JPEG image deblocking) with only a CNN model could achieve a much better result than some state-of-the-art methods. We speculate that our proposed quality improvement model is of high expressivity for widening convolutional layers, and deepening the network, making it possible for a single model to handle multiple poor-quality X-ray images. Furthermore, CNN model may be more robust and of better performances because more training samples are sent to one network. The above reasons may

Fig. 13. X-ray image(ID: 4185) from Digital Hand Atlas Database.

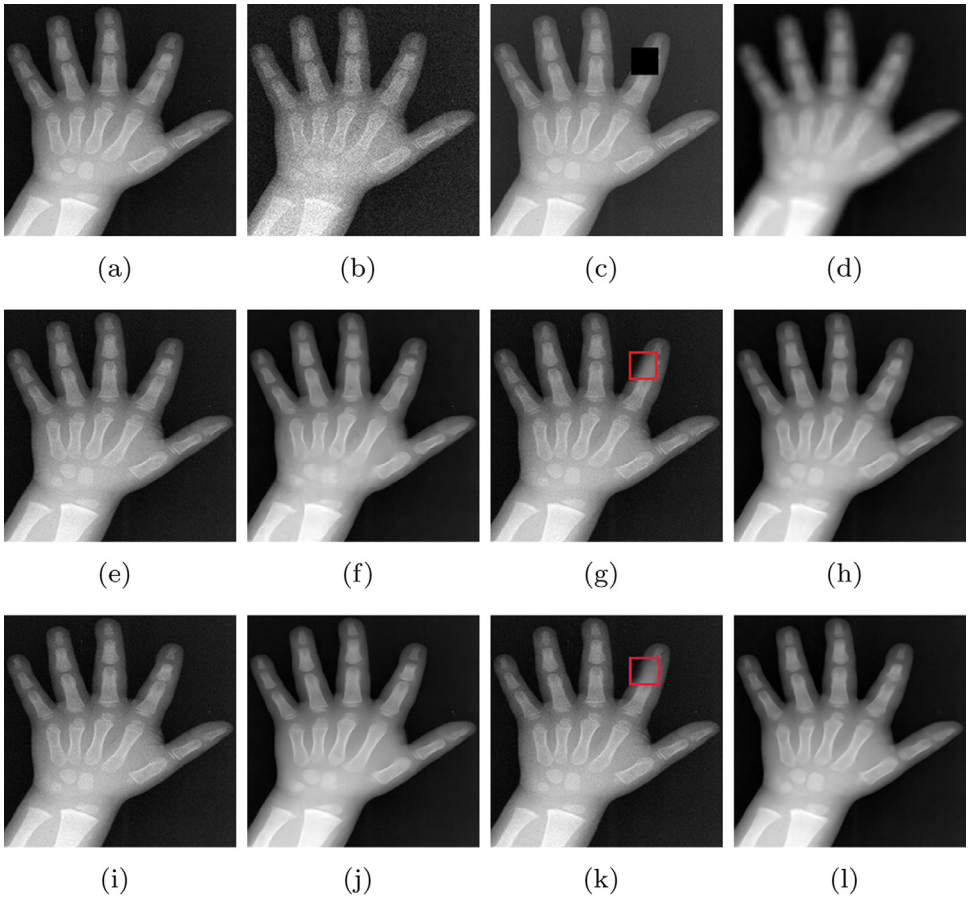


Fig. 14. Examples of quality improvement model outputs. Top row: input images; middle row: outputs of CQ model; bottom row: outputs of DQ model. a: high-quality image; b: image with Poisson noise; c: corrupted image with label; d: low-resolution image.

explain this interesting result. Ultimately, considering the performance of accuracy and complexity, BoNet+_DQ model is a better choice in real application than BoNet+_CQ model.

5. Conclusion

In this study, we have proposed a fully automated BAA system for real-world X-ray images. We first establish a model based on densely connected convolutional networks for high-quality X-ray images and investigate the performance discrepancy between MAE loss and MSE loss. Experiment result shows that MAE loss is more suitable in BAA problem than MSE loss. Moreover, our model achieves a state-of-the-art performance under MAE metric. Then, starting from BoNet+_MAE, we perform in-depth diagnosis of three model architectures for BAA of poor-quality X-ray images. The accuracy of BoNet+_DQ is as high as that of BoNet+_CQ, and both of them exceed others by a margin. In addition, we tentatively put forward that if the expressivity of CNN model is enough high, multiple tasks can be handled together just by a single model.

6. Limitation and future work

The most serious limitation of this paper is the dataset, which is fully generated via simulation. Though the generated synthetic images can partly exhibit the characteristic of the poor-quality X-ray images, they must be different from the real datasets. Therefore, the proposed system cannot be directly applied in real-world scenario. If this system is going to be deployed, the parameters in the neural networks need to be updated using real datasets. Besides, we here only take three types of poor quality into consideration, which is another major limitation of this paper. Three types of poor quality are not enough and the realistic scenarios are more complicated. In the future, we will collect as much real data as possible and conduct the experiment on the real data instead of simulation.

In this paper, we mainly focus on handling the problem of poor-quality X-ray images, ignoring the assessment model, which is fundamental to accurate bone age estimation. It is necessary to make great efforts in the bone age assessment part and perform ablation studies. Besides, the proposed BoNet+_CQ is too complex for extra VGG classification network. It must be a better choice to directly use features extracted by BoNet+ for classification.

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.

Acknowledgment

This work was supported by the National Key Scientific Instrument and Equipment Development Projects of China [81527802]. Jiajia Guo and Jianyue Zhu contributed equally.

References

[1] Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2017:170236.
[2] Wang S, Shen Y, Shi C, Yin P, Wang Z, Cheung PW-H, Cheung JPY, Luk KD-K, Hu Y. Skeletal maturity recognition using a fully automated system with convolutional neural networks. IEEE Access 2018;6:29979–93.
[3] Michael DJ, Nelson AC. HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs. IEEE Trans Med Imag 1989;8(1):64–9.
[4] Hsieh C-W, Jong T-L, Chou Y-H, Tiu C-M, et al. Computerized geometric features of carpal bone for bone age estimation. Chin Med J (Engl) 2007;120(9):767–70.
[5] Gertych A, Zhang A, Sayre J, Pospiech-Kurkowska S, Huang HK. Bone age assessment of children using a digital hand atlas. Comput Med Imaging Graph 2007;31(4):322–31.
[6] Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The bonexpert method for automated determination of skeletal maturity. IEEE Trans Med Imag 2009;28(1):52–66.
[7] Giordano D, Kavasidis I, Spampinato C. Modeling skeletal bone development with hidden markov models. Comput Methods Programs Biomed 2016;124:138–47.
[8] Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in x-ray images. Med Image Anal 2017;36:41–51.
[9] Mansourvar M. Bone age assessment using hand and clavicle x-ray images. University Malaya; 2014.
[10] Thodberg HH. An automated method for determination of bone age. J Clin Endocrinol Metab 2009;94(7):2239–44.
[11] Giordano D, Spampinato C, Scarciofalo G, Leonardi R. An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones. IEEE Trans Instrum Meas 2010;59(10):2539–53.
[12] Martin DD, Heil K, Heckmann C, Zierl A, Schaefer J, Ranke MB, Binder G. Validation of automatic bone age determination in children with congenital adrenal hyperplasia. Pediatr Radiol 2013;43(12):1615–21.
[13] Pinsker JE, Ching MSL, Chiles DP, Lustik M, Mahnke CB, Rooks VJ. Automated bone age analysis with lossy image files. Mil Med 2017;182(9-10):e1769–72.
[14] Thodberg HH, van Rijn RR, Jenni OG, Martin DD. Automated determination of bone age from hand x-rays at the end of puberty and its applicability for age estimation. Int J Legal Med 2017;131(3):771–80.
[15] O'Keeffe D. Denoising of carpal bones for computerised assessment of bone age. University of Canterbury; 2010.
[16] Huang G, Liu Z, Weinberger K.Q, van der Maaten L. Densely connected convolutional networks. 2016, arXiv:160806993.

[17] Zhang A, Gertych A, Liu BJ. Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones. Comput Med Imaging Graph 2007;31(4):299–310.
[18] Harmsen M, Fischer B, Schramm H, Seidl T, Deserno TM. Support vector machine classification based on correlation prototypes applied to bone age assessment. IEEE J Biomed Health Inform 2013;17(1):190–7.
[19] Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017:1–15.
[20] Štern D, Kainz P, Payer C, Urschler M. Multi-factorial age estimation from skeletal and dental MRI volumes. In: International workshop on machine learning in medical imaging. Springer; 2017. p. 61–9.
[21] Xing J, Li K, Hu W, Yuan C, Ling H. Diagnosing deep learning models for high accuracy age estimation from a single image. Pattern Recognit 2017;66:106–16.
[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv:14091556.
[23] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer; 2015. p. 234–41.
[24] Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans Image Process 2017;26(7):3142–55.
[25] Dodge S, Karam L. Understanding how image quality affects deep neural networks. In: IEEE QoMEX. IEEE; 2016. p. 1–6.

Jiajia Guo received the B.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2016, and the M.S. degree from University of Science and Technology of China, Hefei, China, in 2019. His areas of interests currently include deep learning and medical image processing.

Jianyue Zhu is currently pursuing the Ph.D. degree in information and communication engineering with the School of Information Science and Engineering, Southeast University, Nanjing. Her current research interests include machine learning and optimization theory

Hongwei Du received his Ph.D. degree in biomedical engineering from the University of Science and Technology of China in 2007. Currently, he is an associate professor at the University of Science and Technology of China. His current research interests include deep learning, compressive sensing, and biomedical imaging.

Bensheng Qiu is a professor at the University of Science and Technology of China. He is focusing on medical image processing, deep learning, image-guided minimally invasive therapy, and molecular imaging.