IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 27, 2021, accepted April 11, 2021, date of publication April 21, 2021, date of current version April 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074713

Fully Automatic Model Based on SE-ResNet for Bone Age Assessment

JIN HE AND DAN JIANG
School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China

Corresponding author: Jin He (hejin198809@126.com)

This work was supported by the Beijing University of Posts and Telecommunications (BUPT) Excellent Ph.D. Students Foundation under Grant CX2019318.

ABSTRACT Bone age assessment (BAA) based on hand X-ray imaging is a common clinical practice for investigating disorders and predicting the adult height of a child. However, the traditional manual method is time consuming and prone to obverse variability. There is an urgent need for a fully automatic framework based on deep learning with high performance and efficiency. We propose an end-to-end BAA model based on lossless image compression and a squeeze-and-excitation deep residual network (SE-ResNet). First, we apply the compression module to compress the raw image without losing important features. Second, the SE-ResNet-based model extracts features of the compressed images. Furthermore, the regression model with improved loss function predicts bone age. The experiments on a public dataset reveal that our method outperforms the baseline models. In conclusion, the presented method is a fully automatic and effective solution to process hand X-ray images for BAAs.

INDEX TERMS Bone age assessment, deep learning, convolutional neural network, residual network, regression.

I. INTRODUCTION

Bone age assessment (BAA) through radiographs of the left hand is widely used in the diagnosis, treatment and monitoring of endocrine, genetic and growth disorders in children [1]. Moreover, this work is often applied to the prediction of adult height in a child [2], which is potentially beneficial for sports medicine. Figure 1 shows some typical hand radiographs of different ages for BAA. The bone joint and cartilage of children's hands of different ages are obviously different. Male and female children of the same age have different hand bone structures.

Based on the subtle bone/cartilage development pattern in X-ray images, physicians can manually assess hand bones in accordance with the Greulich and Pyle (GP) method [3] or the Tanner-Whitehouse (TW) method [4]. The GP method compares a hand radiograph with a reference atlas, while the TW method is based on a scoring system. However, both methods are time consuming and rely on domain knowledge and the experience of radiologists. Thus, in this task, automated BAA methods by computers are necessary.

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

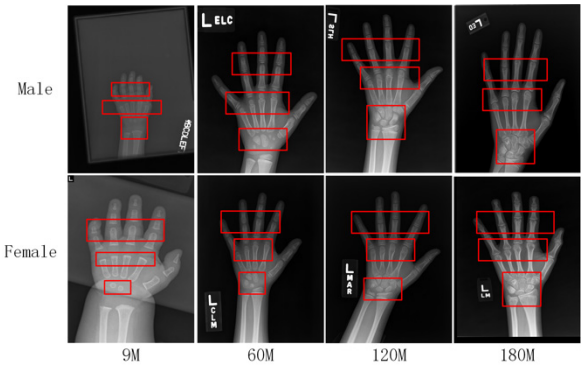

FIGURE 1. Left-hand X-ray images of male and female children at different ages. The red box indicates some areas that need to be focused on.

Automatic computer methods have played important roles in medical image processing, as well as in BAA [5]–[9]. Moreover, deep learning methods play an important role in BAA [10]–[17].

In this work, we proposed an automatic BAA model based on a convolutional neural network (CNN) and regression processing. We introduce a compression module to reduce

the size of raw images, solve the time-consuming problem in model training, and preserve image features. We first apply the squeeze-and-excitation deep residual network (SE-ResNet) for BAA. The experimental results prove that the SE-ResNet significantly outperforms other CNN based models on features extraction of hand bone images. We improve loss function with adding a Gaussian weight, which can make the model pay more attention on the small samples and mitigate the impact of data imbalance on results.

II. RELATED WORK

Conventional approaches of BAA rely on professional experience and manual work. All the procedures consume too much time and observer variability. However, they provide some design ideas for the automatic computer method. For example, when analyzing an X-ray image, radiologists first ignore the background and focus on the regions of interest (ROIs) in the image. Many researchers draw on this process for automatic BAA methods. Pietka *et al.* developed semantic features for BAA and suggested a multiple-step processing method: direction determination and background elimination, phalanx location and length measurement, ROIs extraction, global size and distance length determination [5]. According to the TW method, Bocchi *et al.* applied a preprocessing method to obtain features from important regions, and used neural networks to make the hone complex to a maturation stage [6].

As one of the image processing tasks, the BAA of hand X-ray images is generally defined as a classification or regression problem. In the last two decades, numerous machine learning methods have been used for BAA, such as the decision tree method [7], k nearest neighbor (kNN) classification [8], and support vector machine (SVM) [9]. However, the aforementioned works need to use prior knowledge to extract features manually, and these methods have low accuracy.

Recently, the success of CNNs for medical image processing has been reported [18]–[20]. Deep learning methods based on CNNs have been suggested for image preprocessing and feature extraction in BAA tasks, such as fine-tuned CNNs [10], methods based on visual geometry groups (VGGs) [11], UNets for segmentation [12], [13], deep residual network (ResNet)-based models [14], and CNNs with attention mechanisms [15]. Moreover, scholars have proposed new ideas for solving the problem of BAA. Liu *et al.* proposed a multiscale data fusion framework for BAA with X-rays based on a transform and CNNs [16]. Liu pioneered the application of ranking learning to BAA and presented a ranking CNN to predict bone age [17]. These works are successful in the task of BAA. However, a number of works have applied two- or three-stage methods to solve this problem of BAA. They segment the images as data preprocessing. The processed data are input into feature extraction and the BAA model for training. Furthermore, image segmentation or other preprocessing requires labeled data to evaluate the processing performance, which increases the complexity of training. Thus, a fully automatic end-to-end model for BAAs is necessary. Above all, we can learn from the success of these applications and optimize the network architecture for BAA.

III. MODEL

For hand X-ray images, we presented a fully automatic BAA model based on CNN. Figure 2 shows the architecture of the proposed model in this work. First, the raw images are input to the compression module for data preprocessing. In this process, the image is compressed to a specific size without losing important features, and the model initially extracts the image features. Moreover, the SE-ResNet module learns the features of the compressed images and the gender. Finally, the regression module obtains the bone age through feature images. Furthermore, we improve the loss function to train the model.

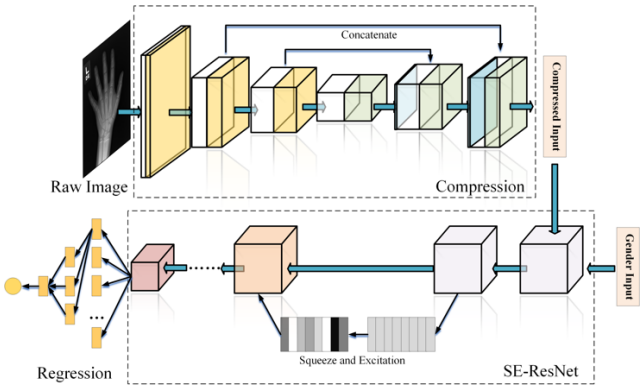

FIGURE 2. Overall view of the proposed method. The architecture contains raw image compression, SE-ResNet, and a regression module.

A. COMPRESSION

In the BAA task of hand X-ray image, there are several effective preprocessing methods, for example, bone segmentation [11], [12], convex-hull based method [21], and so on.

The proposed compression module is applied to compression and feature extraction of raw X-ray images. Due to the large size of the X-ray image, there are many blank areas in the image beside the hand area. These areas have no practical significance, and focusing on these areas will also cause a waste of computing resources. Moreover, using 23wrX-ray images to predict bone age usually focuses on a few specific areas. Compared with the size of the whole image, the size of these key areas is very small. If the image size is reduced blindly, it will lead to the loss of bone age features, thus reducing the accuracy rate.

To compare the effects of different input sizes on the model, we used images of different sizes to train the VGG-16 model [22] as Table 1.

In Table 1, the larger the size of the input images is, the better the result. The size of the input images can affect the final recognition accuracy. However, the large size of the input makes the training time too long, and the calculation of

TABLE 1. Results of different size training images in terms of the mean absolute error (MAE).

Image size	Time/s	MAE/m
128*128	281	12.9
256*256	373	11.2
512*512	764	10.2
1024*1024	2344	9.6

the model too large to train. Thus, the size of the data input to the feature extraction network must be within a certain range. If the raw features of the image can be retained more, the bone age prediction can be better completed. Thus, we apply the compression module based on CNN to compress the image size without losing important features. This module can preserve the image features and shorten the training time. The structure of the compression module is shown in Figure 3.

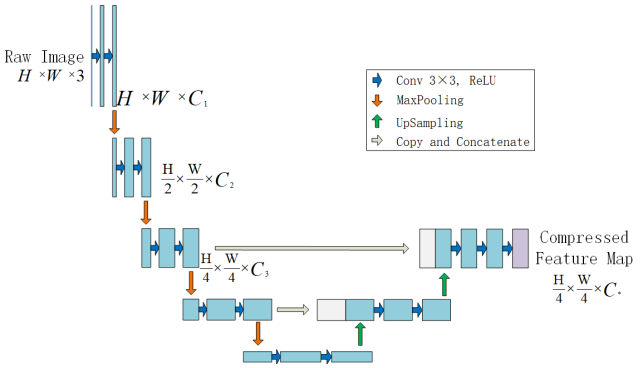

FIGURE 3. The structure of the compression module. Different colored arrows indicate different operations.

The compression module scales the length and width of the image by convolution and pooling and completes the preliminary feature screening by upsampling and shallow feature fusion. In Figure 3, through two convolutional layers with a kernel of 3×3, the size of the input data is transferred from $H \times W \times 3$ to $H \times W \times C_1$, where C_1 is the number of channels. Then, through max-pooling, the size is compressed to $\frac{H}{2} \times \frac{W}{2} \times C_2$. When compressed to a specific size (for example, $\frac{H}{4} \times \frac{W}{2}$), feature fusion and extraction are performed. The final size of the feature map is $\frac{H}{4} \times \frac{W}{4} \times C$.

This module retains as many key features as possible and removes the features of the blank area to reduce the size of the raw image.

B. SE-ResNet

CNNs have shown its superiority in computer visual tasks. The squeeze-and-excitation (SE) block can associate and learn the features between channels to improve the quality of feature representations using a feature recalibration strategy [23]. The SE method automatically obtains the weight of each channel by learning, and enhances the important features. Figure 4 shows the architecture of the SE method.

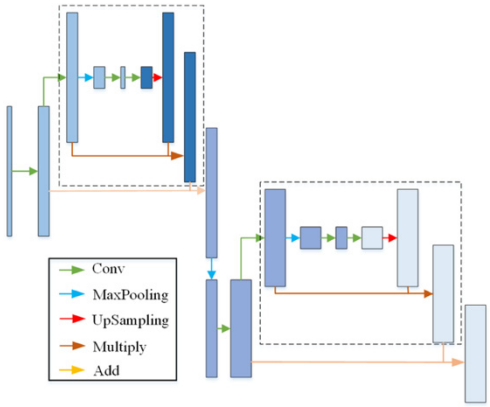

FIGURE 4. The architecture of two SE blocks, and one block contains two operators: squeeze and excitation. Different colored arrows indicate different operations.

The squeeze operator compresses the input information to generate a statistical channel by applying global average pooling. The excitation processing contains two fully connected layers around the nonlinearity and a ReLU. The excitation operator weights the input data to generate a weight channel. The size of the feature channel is maintained through the squeeze-and-excitation operators.

The residual block of ResNet can make good use of shallow features to obtain more key feature values, and it has often been applied as the main feature extraction structure in the task of image classification and recognition [24]. Thus, this paper used ResNet as the main structure of the feature extraction part. The SE-ResNet module can be obtained by combining the SE block with the ResNet model, and the calculation process the SE-ResNet module is shown in Figure 5.

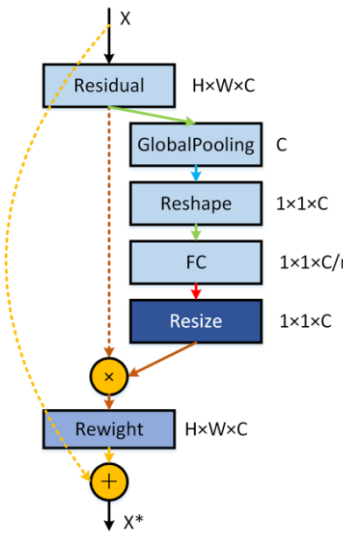

FIGURE 5. The structure of the SE-ResNet module.

As shown in Figure 5, squeeze and excitation work before the summation operation. The residual block of ResNet integrates the SE structure, which can not only make full use of

the shallow features but also further reweight each channel of the shallow features to enhance the key feature extraction. We gain the output of SE-ResNet as (1).

$$y = F\left(f_{se}\left(x\right), \left(\omega_i\right)\right) + x \tag{1}$$

where x and y are the input and output of the SE-ResNet, $f_{se}(.)$ is the function of the SE block, and ω_i is the weight of the network of the i-th input. However, in the process of the squeeze operation, we need to define the scale of the feature image, which greatly affects the value of reweight. Since the size of each input feature image is not the same, this paper proposes an adjustable scale according to the size of the feature channel. We define the output of the j-th SE-ResNet block as in (2).

$$y_j = F(f_{se}(x_j), (\omega_{ij})) + x_j \tag{2}$$

where y_j is the output of the j-th SE-ResNet structure.

C. REGRESSION AND IMPROVED LOSS FUNCTION

The linear regression module is composed of multiple fully connected layers, and takes the output feature vector of the upper layer as the input. One of the most common loss functions for regression task is the mean squared error (MSE) as in (3).

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - g_i|^2 \tag{3}$$

where y_i and g_i are the i-th output of the proposed method and ground truth, respectively.

Considering the problem of data imbalance, we improved the MSE loss by adding Gaussian parameter. This parameter can increase the weights of small classes. The improved loss function is obtained as in (4) and (5).

$$L = \frac{1}{n}\sum_{i=1}^{n} (1 + \alpha_i) |y_i - g_i|^2 \tag{4}$$

$$\alpha_i = \beta \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{c_i^2}{2\sigma^2}\right) \tag{5}$$

where α_i is the Gaussian weight, c_i is the count of samples per age, and β is the parameter to control the weight of every age and make sure $\alpha_i \in [0, 1]$.

IV. EXPERIMENTS

A. DATASET

For the task of BAA, public datasets are limited. Fortunately, the Pediatric Bone Age Challenge, organized by the Radiological Society of North America (RSNA), provided 12,611 left-hand X-ray images with corresponding bone ages under 19 years old of male and female. We randomly picked 2,323 images for validation and 200 for testing. The remaining images were used for training. Some examples from the randomly selected RSNA dataset are shown in Figure 6.

In Figure 6, the hand X-ray images vary considerably in intensity, contrast, and brightness. Moreover, the size and position of the hands in different images also vary greatly.

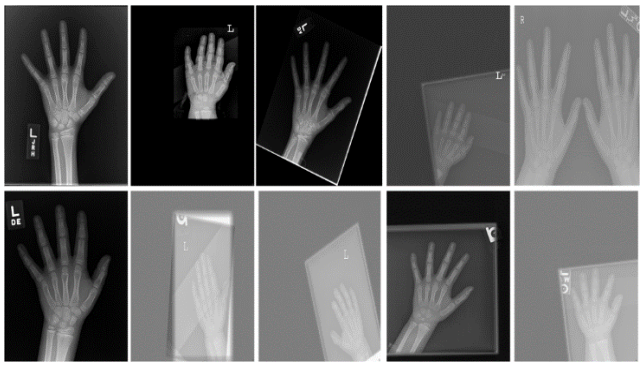

FIGURE 6. Examples of RSNA dataset.

This variance increases the difficulty of training an end-to-end model for BAA. Thus, it is necessary to preprocess the raw images of the RSNA dataset.

B. BASELINES

We applied the following models as the baselines for BAA task.

- kNN [9]: Harmsen *et al.* implemented a kNN model for the BAA task based on grouped data. The best experimental result for kNN was obtained by k = 5.
- SVM [9]: Harmsen *et al.* proposed an SVM model based on content-based image retrieval. They applied the radial basis function as a kernel for the multiclassification task.
- VGG [11]: Iglovikov *et al.* applied a U-Net as the segmentation model and a VGG-style network as BAA feature extraction.
- Mask R-CNN [25]: The proposed framework is composed of a Mask R-CNN subnet of segmentation and a residual attention network for BAA.
- U-Net + CNNs [12]: Pan *et al.* proposed an active learning segmentation model based on U-Net and applied deep CNNs with pretrained weights on ImageNet for BAA.

C. OBJECTIVE EVALUATION

The mean absolute error (MAE) between the output and the label annotated by expert is the most commonly used evaluation metric in BAA. In this paper, we also calculate the MAE as in (6).

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

where x_i and y_i are the label and estimated bone age in months.

D. EXPERIMENT SETTING

We set the hyperparameters of model training as Table 2.

As shown in Table 2, we trained our model by using Adam with gradient clipping [26]. A dropout rate of 0.5 was chosen before the output layer, and the batch size was set to 32. The initialized learning rate is set to 0.001. σ and β in (5) control

TABLE 2. Hyperparameters setting of our model.

Hyperparameter	value
Activation function of CNNs	ReLU
Optimizer	Adam
Dropout	0.5
Batch size	32
Learning rate	0.001
Epoch	100
σ in (5)	400
β in (5)	100

the sample weight of each age. The largest class in training set has 1,113 samples. Thus, we set $\sigma = 400$ and $\beta = 100$ to make the model work best.

We used Keras from https://keras.io as our deep learning framework, and all the models were trained on four NVIDIA Titan X GPUs.

V. RESULTS AND DISCUSSION

A. OVERALL PERFORMANCE

Table 3 shows the results achieved on the RSNA dataset. The proposed model offered relative improvements of 5.96 months compared with kNN and 3.92 and 2.02 lower than SVM and VGG based methods in terms of the MAE. Our method also outperformed the Mask R-CNN and the U-Net + CNNs even if they have similar feature extraction structures.

TABLE 3. Results of experiments in terms of MAE.

model	MAE/m
kNN [9]	12.00
SVM [9]	9.96
VGG [11]	8.08
Mask R-CNN [21]	7.38
U-Net + CNNs [12]	7.35
Ours	6.04

SVM and kNN are based on the artificial features for prediction, while CNN is based on the machine automatic acquisition of features to make better use of the influence features. Therefore, the effect of the VGG model is better than SVM and kNN in the experiment, and the VGG model will make the shallow features disappear gradually after multiple convolutions. ResNet can retain the shallow features by superimposing the residual structure in the training process, retain the global features, and improve the recognition accuracy. On the basis of ResNet, the SE structure was added, which can fuse two channel features on the basis of shallow features. For high-resolution images, directly reducing the size of images will lose the key features of the target. Thus, the proposed compression module can fully retain the image size features, which can better retain the image features and further improve the recognition accuracy.

In order to reflect the improvement of gender information on the results, we conducted a comparative experiment with and without gender information in the training of our model. We divided the test results into male, female and the whole set, and the results are shown in Figure 7.

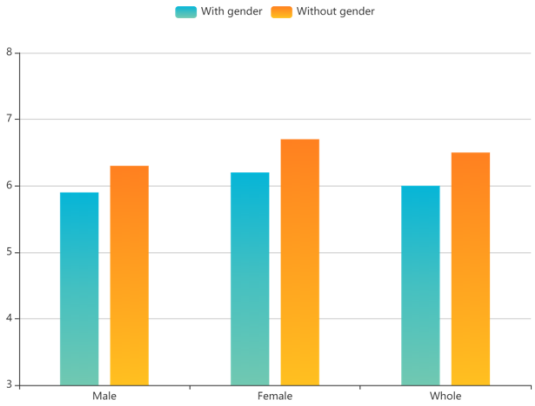

FIGURE 7. The result on male, female and the whole test set with and without gender information in training.

As shown in Figure 7, when considering gender information in model training, the experimental results are better than ignoring gender information.

B. FURTHER IDENTIFICATION

The attention mechanism is applied to add the weight of important areas in images. Figure 8 reports the focus areas of the SE-ResNet module on the RSNA. The key areas are concentrated in the joints of the hand, which is consistent with the analysis of experts. Moreover, the model can avoid the interference of noise, for example, other people's hand in the image.

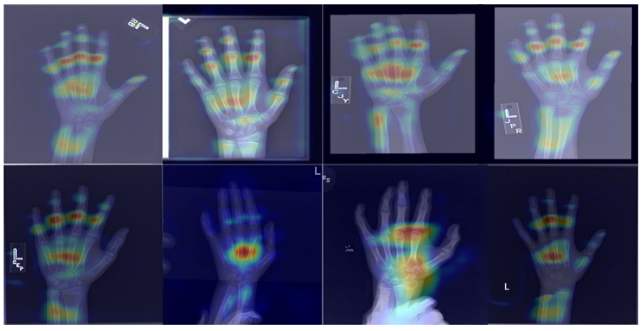

FIGURE 8. The focus areas of SE-ResNet module. The highlighted colors in the figure indicate the area that the model focuses on.

The ablation studies can ensure all the components of the proposed model are efficient. The main components of our model include compression module, SE-ResNet, and improved loss function. The proposed model can be expressed as C + S + I. The models depicted in Table 4 are as follows: S + I (SE-ResNet with improved loss function, and the size of input image is 256*256), C + I (the same

structure as our method but it applies the ResNet instead of SE-ResNet module), C + S (applying the mean squared error instead of the improved loss function), CH + S + I (using the convex hull instead of the compression module, and the size of input image is 1024*1024), U + S + I (using the U-Net instead of the compression module, and the size of input image is 1024*1024), and ours (the size of input images is 1024*1024, and the size of compressed images is 256*256). All the models applied the same hyperparameters as the proposed model.

In Table 4, we report the MAE and time consuming of different models, and the results show that every part of our model is necessary for the entire architecture. The role of the compression module is raw image compression without losing important features. Compared with convex hull and U-Net pre-processing method, the proposed compression module showed its improvement of results. The convex hull and other methods can cut the hand region and remove some useless image information. However, these methods lack the process of feature selection. The compression module can not only automatically adjust the extraction of the target region in the training process, but also make use of the feature information of the target area, and send more useful feature information into SE-ResNet structure. Moreover, the improvement of the SE structure to the result shows that the SE-ResNet module plays a role in feature extraction of bone age. Furthermore, the improved loss function is better than the mean squared error loss for BAA.

TABLE 4. Experimental results of ablation studies in terms of the MAE.

model	Time/s	MAE/m
S + I	781	8.63
C + I	945	8.39
C + S	1067	6.32
CH + S +I	1084	6.17
U + S + I	1683	6.07
Ours (C + S + I)	1080	6.04

The computing time of U + S + I and our model show that our model pays less time than U-Net based model. It proves that the proposed compression module can significantly reduce the amount of computing.

The goal of medical imaging is to fully display the characteristics of the lesions as much as possible, so it is easy to observe their state. Therefore, the size of medical images (whether X-ray or CT) is often large. However, in most cases, the aim area is smaller than the image size, such as nodules and tumors. Because of the difficulty of convolution image feature selection, it cannot make full use of convolution image features to identify the target. However, when processing large-scale images, to ensure the training efficiency, the image is scaled to a small size, which leads to the loss of target features and a reduction in recognition accuracy. For the actual needs of lesion analysis, the target features need to be retained as much as possible. Thus, the original size of the image needs to be retained as much as possible. The model proposed in this paper can make full use of image features to obtain more lesion features to improve the recognition of lesions. It can provide a new idea for medical image processing models based on convolutional neural networks.

VI. CONCLUSION

We propose a fully automatic deep neural network solution to process X-ray images of hands for BAA based on CNN. Our work presents a lossless compression module to compress large train data and integrates SE and ResNet to extract the features. The proposed method is proven to be effective in reading bone age. The experimental results show that our method is comparable to the state-of-the-art models with similar strategy in the RSNA dataset. The compression module and SE-ResNet can be applied in other fields of medical image decision problems.

In future work, we will focus on applying the structure of the compression module and SE-ResNet on the classification or recognition tasks of high-resolution training images.

REFERENCES

[1] D. D. Martin, J. M. Wit, Z. Hochberg, R. R. Rijn, O. Fricke, G. Werther, N. Cameron, T. Hertel, S. A. Wudy, G. Butler, H. H. Thodberg, G. Binder, and M. B. Ranke, "The use of bone age in clinical practice—Part 1," *Hormone Res. Paediatrics*, vol. 76, no. 1, pp. 1–9, 2011.

[2] L. L. Morris, "Assessment of skeletal maturity and prediction of adult height (TW3 method): Book review," *Australas. Radiol.*, vol. 47, no. 3, pp. 340–341, Sep. 2003.

[3] S. M. Garn, "Radiographic atlas of skeletal development of the hand and wrist," *Southern Med. J.*, vol. 53, no. 11, p. 1480, Nov. 1960.

[4] J. M. Tanner, R. H. Whitehouse, N. Cameron, W. A. Marshall, M. J. Healy, and H. Goldstein, *Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method)*. London, U.K: Academic, 1975.

[5] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. K. Huang, and V. Gilsanz, "Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction," *IEEE Trans. Med. Imag.*, vol. 20, no. 8, pp. 715–729, Dec. 2001.

[6] B. Bocchi, F. Ferrara, I. Nicoletti, and G. Valli, "An artificial neural network architecture for skeletal age assessment," in *Proc. Int. Conf. Image Process.*, 2003, pp. 1077–1080.

[7] S. Aja-Fernández, M. Á. Martán-Fernández, and C. Alberola-López, "A computational TW3 classifier for skeletal maturity assessment. A computing with words approach," *J. Biomed. Informat.*, vol. 37, no. 2, pp. 99–107, Apr. 2004.

[8] B. Fischer, P. Welter, C. Grouls, R. Guenther, and T. M. Deserno, "Bone age assessment by content-based image retrieval and case-based reasoning," in *Proc. Med. Imag. Comput.-Aided Diagnosis*, Lake Buena Vista, VA, USA, vol. 7963, Mar. 2011, pp. 1–8.

[9] M. Harmsen, B. Fischer, H. Schramm, T. Seidl, and T. M. Deserno, "Support vector machine classification based on correlation prototypes applied to bone age assessment," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 190–197, Jan. 2013.

[10] H. Lee, S. Tajmir, J. Lee, M. Zissen, B. A. Yeshiwas, T. K. Alkasab, G. Choy, and S. Do, "Fully automated deep learning system for bone age assessment," *J. Digit. Imag.*, vol. 30, no. 4, pp. 427–441, Aug. 2017.

[11] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets, "Paediatric bone age assessment using deep convolutional neural networks," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 300–308.

[12] X. Pan, Y. Zhao, H. Chen, D. Wei, C. Zhao, and Z. Wei, "Fully automated bone age assessment on large-scale hand X-ray dataset," *Int. J. Biomed. Imag.*, vol. 2020, pp. 1–12, Mar. 2020.

[13] S. Cao, Z. Chen, C. Li, C. Lv, T. Wu, and B. Lv, "Landmark-based multi-region ensemble convolutional neural networks for bone age assessment," *Int. J. Imag. Syst. Technol.*, vol. 29, no. 4, pp. 457–464, Dec. 2019.

[14] X. Chen, J. Li, Y. Zhang, Y. Lu, and S. Liu, "Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age," *Future Gener. Comput. Syst.*, vol. 110, pp. 795–801, Sep. 2020.
[15] X. Ren, T. Li, X. Yang, S. Wang, S. Ahmad, L. Xiang, S. R. Stone, L. Li, Y. Zhan, D. Shen, and Q. Wang, "Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2030–2038, Sep. 2019.
[16] Y. Liu, C. Zhang, J. Cheng, X. Chen, and Z. J. Wang, "A multi-scale data fusion framework for bone age assessment with convolutional neural networks," *Comput. Biol. Med.*, vol. 108, no. 4, pp. 161–173, May 2019.
[17] B. Liu, Y. Zhang, M. Chu, X. Bai, and F. Zhou, "Bone age assessment based on rank-monotonicity enhanced ranking CNN," *IEEE Access*, vol. 7, pp. 120976–120983, 2019.
[18] J. W. Smith, S. Thiagarajan, R. Willis, Y. Makris, and M. Torlak, "Improved static hand gesture classification on deep convolutional neural networks using novel sterile training technique," *IEEE Access*, vol. 9, pp. 10893–10902, 2021.
[19] T. Hassanzadeh, D. Essam, and R. Sarker, "An evolutionary DenseRes deep convolutional neural network for medical image segmentation," *IEEE Access*, vol. 8, pp. 212298–212314, 2020.
[20] M. Liang, Z. Ren, J. Yang, W. Feng, and B. Li, "Identification of colon cancer using multi-scale feature fusion convolutional neural network based on shearlet transform," *IEEE Access*, vol. 8, pp. 208969–208977, 2020.
[21] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Jul. 2013.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–8.
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, CA, USA, Jun. 2018, pp. 7132–7141.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2016, pp. 770–778.
[25] E. Wu, B. Kong, X. Wang, J. Bai, Y. Lu, F. Gao, S. Zhang, K. Cao, Q. Song, S. Lyu, and Y. Yin, "Residual attention based network for hand bone age assessment," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Venice, Italy, Apr. 2019, pp. 1158–1161.
[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

JIN HE was born in Shanxi, China, in 1988. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. His research interests include artificial intelligence and deep learning.

DAN JIANG was born in Heilongjiang, China, in 1986. She is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. Her research interests include nature language processing and deep learning.