IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 13, 2021, accepted August 15, 2021, date of publication August 26, 2021, date of current version September 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108219

Automatic Bone Age Assessment of Adolescents Based on Weakly-Supervised Deep Convolutional Neural Networks

KEXIN LI[1], JINGZHE ZHANG[2], YUNFEI SUN[3], XINWANG HUANG[3], CHUNXUE SUN[3], QIANCHENG XIE[3], AND SHIJIE CONG[3]

[1]School of Artificial Intelligence, Wuxi Vocational College of Science and Technology, Wuxi 214028, China
[2]School of Marxism, Wuxi Vocational College of Science and Technology, Wuxi 214028, China
[3]School of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China

Corresponding author: Kexin Li (dillonlkx@163.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities through the Ministry of Education under Grant 2572016CB03.

ABSTRACT Hand bone age, as the biological age of humans, can accurately reflect the development level and maturity of individuals. Bone age assessment results of adolescents can provide a theoretical basis for their growth and development and height prediction. In this study, a deep convolutional neural network (CNN) model based on fine-grained image classification is proposed, using a hand bone image dataset provided by the Radiological Society of North America (RSNA) as the research object. This model can automatically locate informative regions and extract local features in the process of hand bone image recognition, and then, the extracted local features are combined with global features of a complete image for bone age classification. This method can achieve end-to-end bone age assessment without any image annotation information (except bone age tags), improving the speed and accuracy of bone age assessment. Experimental results show that the proposed method achieves 66.38% and 68.63% recognition accuracy of males and females on the RSNA dataset, and the mean absolute errors are 3.71 ± 7.55 and 3.81 ± 7.74 months for males and females, respectively. The test time for each image is approximately 35 ms. This method achieves good performance and outperforms existing methods in bone age assessment based on weakly supervised fine-grained image classification.

INDEX TERMS Bone age assessment, deep learning, convolutional neural network, fine-grained image.

I. INTRODUCTION

The concept of bone age was first proposed and applied in the medical field to monitor the development and growth of children. Bone age is an abbreviation of skeletal age. X-ray images of the left hand and wrist are generally taken for bone age assessment (BAA) [1]–[5]. A doctor observes the development of the ossification center of the left metacarpal phalanx, carpal bone, and the lower end of the radius and ulna through X-ray images to determine bone age [6]. The hand bones at different stages have different morphological characteristics. Therefore, BAA can more accurately reflect an individual's growth and development level and maturity. It can not only determine the biological age of children but also understand the growth and development potential of children and the trend of sexual maturity through the bone age.

At present, BAA methods are divided into the TW scoring and G-P atlas methods [7]. The TW method evaluates and scores each part of the hand by analyzing relevant bone morphological characteristics and finally accumulates the scores of the different regions to obtain the final bone age. The G-P atlas method is an image comparison method. This method compares the evaluated image with a standard image, and the bone age of the standard image with the highest similarity to the evaluated image is its estimated value. The scoring method of the TW method is more objective than that of the atlas method, and therefore, the TW method is considered to have higher reproducibility than the G-P atlas method [1], [2]. Figure 1 shows the regions of interest (ROIs) of the entire

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

hand image in the TW method, where the red box is the metacarpophalangeal region, and the green box is the carpal region.

The above methods require an orthopedic expert to perform a manual evaluation. Manual reading is time-consuming and laborious, and the evaluation results are subject to the subjective influence of the evaluator. Therefore, a fast and real-time automatic BAA technology is required.

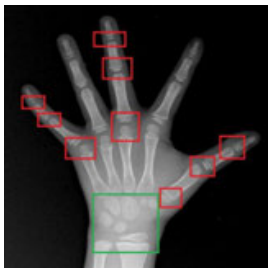

FIGURE 1. ROIs used in TW method.

Research on automatic BAA systems mainly focuses on two types of methods, that is, traditional machine learning methods and deep learning (DL)-based methods. Traditional BAA includes the CASAS system, CASMAS, the BoneXpert system [1], neural network analysis based on the epiphyses and carpal, and the royal orthopedic hospital skeletal aging system [8].

In recent years, DL technology has gradually been applied to the field of medical image analysis owing to its good performance, and convolutional neural network (CNN)-based image classification models have become popular. The task of automatic BAA can be regarded as a classification problem of dividing hand X-ray images into the corresponding bone age category, and CNNs have the advantages of feature extraction and classification, which can greatly simplify the entire BAA process [9]. Therefore, researchers have begun applying CNNs to automatic BAA. Stem *et al*. [10] used a traditional image processing method to extract the corresponding sub-region feature from hand MRI images and then feed each sub-region to CNNs for training, and finally, the results of all CNNs are synthesized to obtain the predicted bone age. However, this model requires manual extraction of ROIs and separate training, preventing it from being a fully automatic evaluation method. Lee *et al*. [11], [5] fine-tuned the pretrained GoogLeNet model on the ImageNet dataset to obtain a deep neural network model suitable for hand bone images. However, this is a transfer learning strategy, and the accuracy of bone age prediction is only 60%. Similar transfer learning methods have also been presented by Han *et al*. [12], Tang *et al*. [13], and Wenxiang [14]. They extracted SIFT and HOG features from an X-ray image, combined the two features, and fed them to a CNN for automatic BAA. This is a typical two-stage method. Another two-stage approach is a combination of U-Net and VGGNet [15], [16], where the U-Net network is used to segment the hand region, followed by a VGG network for feature extraction and classification. An attention mechanism is incorporated into a VGG network to improve the accuracy of the model, achieving a mean absolute error (MAE) of 6.1 months. B.Liu *et al*. [17] combined U-Net and an adversarial generation network (GAN) for BAA, achieving an MAE of 6.05 months. Ren *et al*. [18] proposed a regression CNN for automatic BAA. First, a faster R-CNN is used to detect the hand region, then a Hessian filter is used to filter the image and generate a fine-grained attention map, and finally, the coarse-fine attention map is feed to the regression network to obtain the bone age, achieving an MAE of 5.2 months. Wu *et al*. [19] first used Mask R-CNN to segment the hand region, and then, a residual attention network was used to estimate the bone age, achieving an MAE of 7.38 months. Li *et al*. [20] first uses an unsupervised method for the entire hand segmentation, then MobileNetV3 for feature extraction, and finally multi-layer perceptron for bone age estimation. The MAE is 6.2 months. Salim and Hamza [21] proposed a RidgeNet model, which uses a Mask-RCNN for hand segmentation, Ridge regression to complete bone age estimation, and VGG19 as the backbone.

The structure of two-stage models is complex: it requires not only preprocessing operations on the image but also training multiple deep networks, all of which require manual labeling, and algorithm execution efficiency and real-time performance are poor.

Another type of BAA method is the end-to-end method. Spampinato *et al*. [22] proposed a six-layer Bonet CNN model and used a deformed layer to calibrate the position of the hand bone. It is an end-to-end bone age prediction method, achieving an MAE of 0.79 years. Souza and Oliveira [23] proposed an end-to-end BAA method based on residual learning, achieving an MAE of 6.44 months. Ji *et al*. [24] proposed a network called PRSNet. The network consists of part relation and selection modules. The part relation model uses multi-scale context information to accurately find the underlying correlation between the different parts of the hand bone; the part selection model ranks the importance of the part relation and selects the most important part to assist in BAA, which is an end-to-end model. He and Jiang [25] used SE-ResNet and a regression model to estimate the bone age. A compression step was added to the beginning of the model, achieving an MAE of 6.04. However, this kind of method often requires an additional preprocessing step, such as removing characters, normalizing gray levels, and calibrating images.

Although the above methods have achieved relatively good results, they still have some shortcomings. The problems are mainly concentrated on two aspects. (1) Although the two-stage method uses a CNN, it needs to be combined with traditional machine learning methods and cannot achieve end-to-end BAA, and its real-time performance is poor. (2) Some related studies still require preprocessing operations, even when they achieve end-to-end functions, or their proposed networks cannot focus on informative regions; thus, the networks cannot extract deeper features.

Unlike natural image classification tasks, automatic BAA can be regarded as a classification problem of subcategory images (bone age) with the same category (hand bone images), which requires the classification model to be able to find more subtle local features to achieve more accurate results. The fine-grained classification network is divided into two types: strong supervision and weak supervision [26]. Strongly supervised classification requires the addition of image category labels and additional information such as object labeling boxes during model training [27], whereas weakly supervised classification networks do not need to add any labeling information other than image-level tags [28]. In this study, a weakly supervised CNN model based on NTS-Net [29] is proposed for the automatic assessment of adolescent bone age. This network can automatically locate the ROIs during the image recognition process. The local features of ROIs and global features of the entire hand image are fused, and the fused features are used to classify the hand bone age.

II. METHODOLOGY

The task of BAA can be regarded as the process of dividing the hand bone image into image categories of certain bone ages. The main difficulty in this task is the high similarity of hand bone images to different bone ages. The features used by CNNs for classification are all detailed features, and they only exist in local regions such as the wrist and metacarpal epiphyses. Thus, the primary problem to be solved is how to automatically locate local information-rich regions. This research is mainly based on the NTS-Net network, which is called the Navigator-Teacher-Scrutinizer Network. Figure 2 shows the network architecture of automatic BAA based on the NTS-Net network.

A. ARCHITECTURE OF BONE AGE ASSESSMENT

As shown in Figure 2, the hand bone image X is first sent to the feature extractor module (Resnet-50), then informative regions are detected by feature pyramid networks (FPN) in the navigator agent, and the informativeness of the regions is calculated. Subsequently, these informative regions are sent to the teacher agent to evaluate the confidence of these regions, and the confidence C of the regions and the informativeness I are kept in the same order to optimize the location of the informative regions. When the order is consistent, the features of the informative regions and that of the entire image are together sent to the scrutinizer agent, and it uses these fused features to make predictions.

1) NTS-NET MODEL

The model designs a new training paradigm, consisting of three parts: a navigator agent, a teacher agent, and a scrutinizer agent. The navigator agent can detect the most informative sub-regions under the guidance of the teacher agent. Then, the scrutinizer agent carefully scrutinizes the features of the sub-regions suggested by the navigator and uses them to make the prediction. Figure 3 is a schematic diagram of

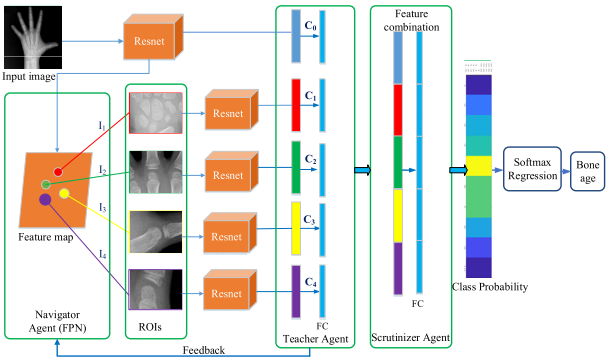

FIGURE 2. Architecture of bone age assessment based on NTS-net network.

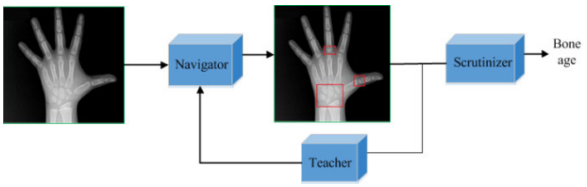

FIGURE 3. NTS-net framework.

the NTS-Net network description. The navigator agent is responsible for finding the most informative regions (ROIs, represented by a rectangular box), whereas the teacher agent evaluates and provides feedback on the ROIs proposed by the navigator agent. After that, the scrutinizer agent carefully scrutinizes these ROIs and uses them to make predictions.

2) NAVIGATOR AGENT

The main function of the navigator agent is to find the informative sub-regions from the hand bone image. The principle is based on the FPN. The network mainly solves the multi-scale problem, and the detection performance of small objects or details is better. Because the epiphysis at the joints of the hand bones is an important information for bone age assessment. The navigator agent uses top-down horizontal connection architecture to detect multi-scale regions. The convolutional layer is used to calculate the feature hierarchy layer by layer, followed by the ReLU activation function and maximum pooling. After the pooling operation, a series of feature maps with different spatial resolutions will be obtained. The network uses lateral connection to merge the up-sampling results and bottom-up feature maps, and uses the high-resolution of low-level features and semantic information of high-level features. Figure 4 is a diagram of the navigator agent. Using multi-scale feature maps with different hierarchies, informative regions can be generated with different scales. In the parameter setting of the module, a feature map with a size of $\{14 \times 14, 7 \times 7, 4 \times 4\}$ is used, which corresponds to a region with a size of $\{48 \times 48, 96 \times 96, 192 \times 192\}$, and the parameter in the module is expressed as W_I.

3) TEACHER AGENT

A confidence function is defined in the teacher agent, which calculates the confidence of each informative region. After

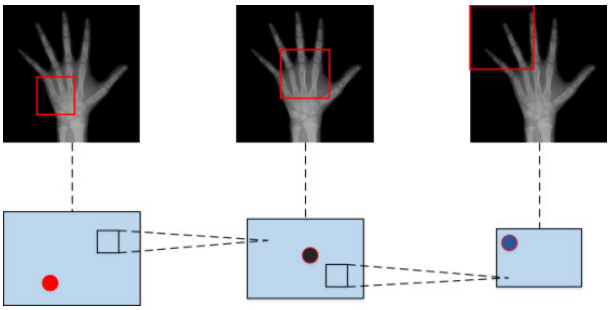

FIGURE 4. Navigator agent.

receiving the M scaled normalized informative regions $\{R_1, R_2, \ldots, R_M\}$ from the navigator agent, the teacher agent outputs the confidence to help the navigator learn. The teacher agent, in addition to the shared layer in the feature extractor, has a fully connected layer with 2048 neurons, and the parameter in this module is expressed as W_C.

4) SCRUTINIZER AGENT

After receiving the top-K informative regions from the navigator agent, the scrutinizer agent first resizes these regions to a predefined size (224 × 224 in our experiments) and then sends these regions into the feature extractor (Resnet-50), which generates the feature vectors of these regions, each of which has a length of 2048. Finally, the scrutinizer agent combines these K features with the feature of the entire image and feeds them into a fully connected layer with $2048 \times (k + 1)$ neurons. This module uses the function S to represent the composition of the transformations, and the parameter is denoted as W_S.

B. PRINCIPLE OF BONE AGE ASSESSMENT

1) PRINCIPLE OF THE MODEL

Unlike the traditional CNN, to focus on the informative regions, the main body of the model adopts the residual network baseline (Resnet-50), and a learning ranking model is developed. This method is based on the assumption that regions with larger informativeness improve the network's target recognition performance. If the features of regions with large informativeness are fused with the global features of the entire image, the fused features can be used for target recognition, improving classification performance. Therefore, the core function of the network model is to automatically locate the informative regions in the X-ray image and order these regions based on their confidence. For the input image **X**, it is assumed that all informative regions are rectangular areas, and **A** represents the set of all ROIs in a given hand bone image. $\mathbf{R} \in \mathbf{A}$ represents the ROIs in the set. This method uses the function **I** to calculate the informativeness of ROIs and uses the function **C** to evaluate the confidence that a region belongs to the ground truth. Generally, more informative regions should have higher confidence, so the following condition should hold:

$$\forall R_1, R_2 \in A, \quad \text{if } C(R_1) > C(R_2), \text{ then } I(R_1) > I(R_2). \tag{1}$$

In this model, the navigator agent is used to approximate the information function **I**, and the teacher agent is used to approximate the confidence function **C**. In summary, this method selects M regions A_M in the sub-region space A. Then, the navigator agent evaluates the informativeness $\mathbf{I}(R_M)$ of the selected M regions, and the teacher agent evaluates their confidence $\mathbf{C}(R_M)$. The network optimizes the navigator agent so that $\{I(R_1), I(R_2), \ldots, I(R_M)\}$ and $\{C(R_1), C(R_2), \ldots, C(R_M)\}$ have the same order.

In summary, the network represents the M most informative regions predicted by the navigator agent as $R = \{R_1, R_2, \ldots, R_M\}$, the informativeness corresponding to the region R is expressed as $I = \{I_1, I_2, \ldots, I_M\}$, and the confidence predicted by the teacher agent is expressed as $C = \{C_1, C_2, \ldots, C_M\}$. Finally, among the M regions with the most informativeness, we select the top-K regions with the most informativeness and feed the features of these K regions, along with the global features of the entire image, to the scrutinizer agent for classification, thereby completing the bone age prediction process.

2) LOSS FUNCTION

We need to optimize these informative regions **R** to make the confidence function **C** and the information function **I** have the same order. The loss function used in the optimization process includes the navigator and teacher loss functions, as shown in (2) and (3).

The navigator loss function is as follows:

$$L(I, C) = \sum_{(i,s):C_i<C_s} f(I_s - I_i), \tag{2}$$

where the function f is a non-increasing function. If $C_s > C_i$, courage $I_s > I_i$, and f is the hinge loss function.

The teacher loss function is as follows:

$$L_C = -\sum_{i=1}^{M} logC(R_i) - logC(X). \tag{3}$$

The loss function penalizes the reversed pairs between **I** and **C** and encourages **I** and **C** to keep the same order. When the navigator navigates to the most informative region $\{R_1, R_2, \ldots, R_K\}$, the scrutinizer agent provides a fine-grained recognition result P_i. $P_i = S(X, R_1, R_2, \ldots, R_K)$, and we use the cross-entropy loss as the classification loss.

$$L = -\sum_{i=1}^{N} y_i log(P_i), \tag{4}$$

where N represents the number of categories, y_i represents an indicator variable, and P_i is a prediction probability.

C. FAST SCANNING ALGORITHM

The dataset consists of hand and wrist radiographs with bone age labels. The radiograph shows high variability, including different acquisition methods and variations in brightness, contrast, resolution, and even aspect ratio. Most of the images in the dataset are obtained through computed radiography (CR) or digital radiography (DR), and only a small part of the images are obtained from the film, as shown in Figure 5.

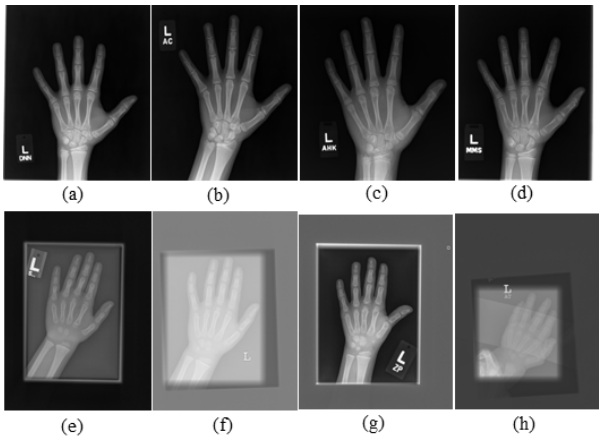

FIGURE 5. Sample radiographs available on the RSNA. (a–d) Images acquired through CR or DR. (e–h) Images digitized from traditional film (irregular samples).

These images obtained from the film have a large invalid region, except for the hand. They are called irregular samples in this study. These invalid regions are useless for bone age recognition. On the contrary, they will increase the burden of data processing. To remove these invalid regions, this study proposes a fast scanning algorithm based on scan lines to extract valid regions of the hand images. Valid regions are defined, as indicated by the green box in Figure 6. The steps of the scanning algorithm are given below and the principle is shown in Figure 7.

Procedure Fast Scan Algorithm

Input: An original image I and a threshold T;
Output: An image I_{vr} with a valid region;
Begin
1. Determine the scanning direction (Row or Column);
2. Start scan;
3. Compute intensity L of each pixel on the scan line;
4. Compute max intensity difference L_{lim} on this scan line;
5. Compare L_{lim} with T, if $L_{lim} >$ T to 6, else return to 2;
6. Stop scan;
7. Record the position of the scan line (X_{left}, X_{right}, Y_{top}, Y_{bottom});
8. Compute the center X_0, Y_0 and the height H, width W of the valid region;
9. Output the image I_{vr} with valid region;
End

The working process of the fast scanning algorithm is as follows: in the row direction of the image, the algorithm scans the pixels one by one from the top to the bottom until it reaches the high pixel value point on the top of the finger. In the same way, the algorithm scans the pixels one by one from the bottom to the top until it reaches the high pixel value point at the end of the wrist. In the column direction of the image, the algorithm scans the pixels one by one from left to right until it reaches the high pixel value point on the leftmost edge of the hand, and then scans from right to left

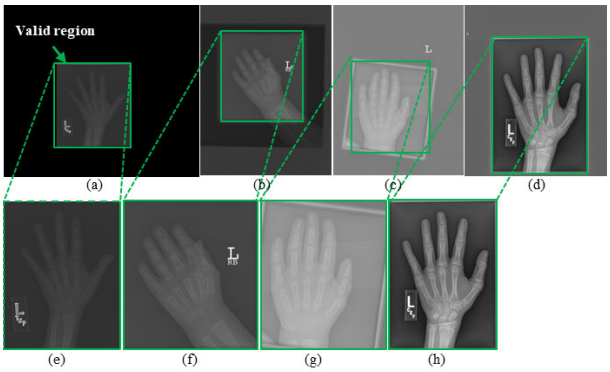

FIGURE 6. Valid region extraction. (a–d) Original images. (e–h) The valid region images by the scan algorithm.

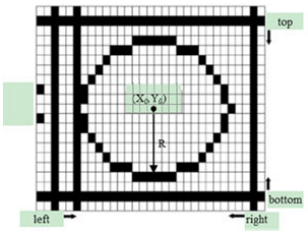

FIGURE 7. Illustration of the fast scan algorithm.

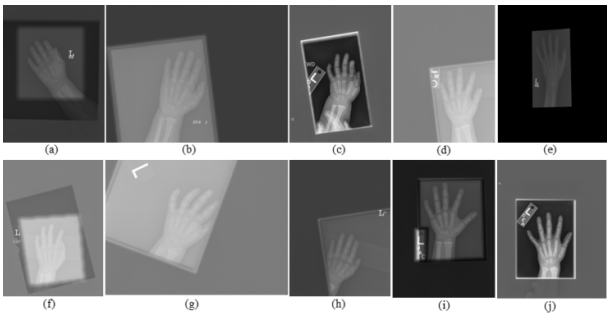

FIGURE 8. Irregular samples digitized from traditional film.

until it reaches the high pixel value point on the right edge of the hand. Where, L represents the gray value of the image, L_{lim} represents the maximum gray difference on a scan line (step 4).

The purpose of developing the fast scanning algorithm is to verify whether a simple preprocessing step can improve our model's performance.

III. RESULTS

A. DATA PREPROCESSING

1) DATA ACQUISITION

This research uses the RSNA 2017 Pediatric Bone Age dataset as the research dataset. The dataset contains 6833 male and 5778 female X-ray images of the hand, ranging in age from 1 to 228 months. Each image was manually labeled by experts. Some samples have invalid regions, as shown in Figure 8. We selected 150 male and 157 female irregular images for preprocessing necessity analysis. Therefore, the final dataset used for training and testing includes 6683 male and 5621 female hand bone images. The sample distribution of the dataset is shown in Figure 9.

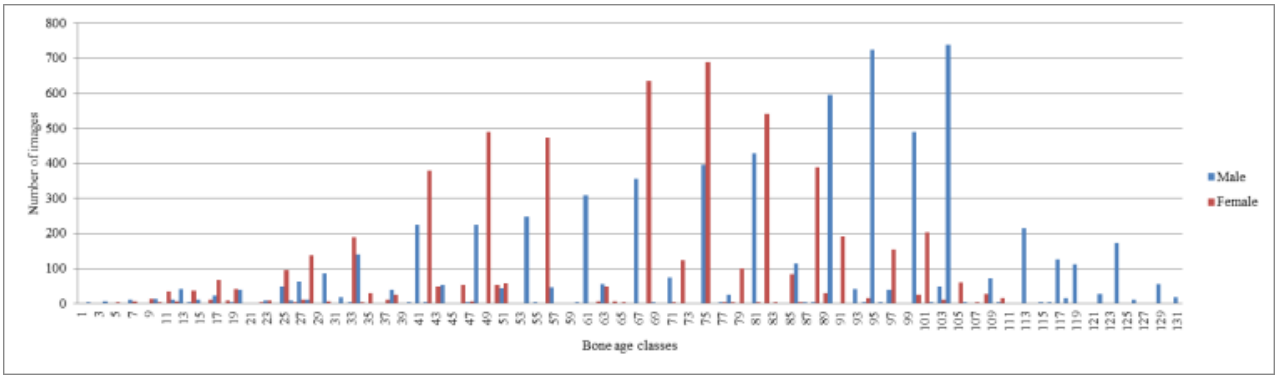

FIGURE 9. Age and gender distribution in the RSNA dataset.

Figure 9 shows that the number of hand images in the RSNA dataset is unbalanced in different age categories. The number of samples is only 1 in some bone age, and the maximum is 718 in others. Such a distribution will significantly reduce our model's classification accuracy. Therefore, the dataset needs to be augmented to balance these samples.

2) DATA AUGMENTATION

DL is a standard data-driven method. As a black-box deep network, it strongly relies on a large amount of data to solve a problem. The sample distribution of the dataset in this study is unbalanced, so the dataset needs to be augmented.

Data augmentation methods typically include affine transformation methods such as translation, rotation, mirroring, and elastic deformation [30]. Because hand radiographs have high consistency, such as the location of the hand and the background color in the image, if we choose the augmentation method arbitrarily, this consistency will be destroyed. In our experiments, five methods of scaling, translation, rotation, shearing, and elastic deformation are used to augment the samples to approximately 60 images for each class. Ultimately, 12,501 images of males and 10,461 images of females were obtained. Figure 10 shows the data augmentation method and the results.

We use 80% of the augmented data as the training set, and the remaining 20% of the samples are used for testing. A random strategy is used to generate the training/test set samples, and the result is the average of the results of five experiments (different training/testing samples). Data augmentation parameters include image rotation in the range of $-45°$ to $45°$, image translation ranging in the ratio of 0-0.1, image scaling ranging in the ratio of 0.8-1.1, image shear ranging from $-8°$ to $8°$, and image elastic deformation with alpha ranging from $90°$ to $105°$ and sigma of 20. Table 1 shows the distribution of the augmented dataset.

B. PARAMETERS SETTING

Our experiments were run on a machine with a 3-GHz CPU and 24-GB RAM, equipped with two Nvidia Titan XP GPUs. The system environment is Ubuntu 14.04, the programming language is Python3.6, and the DL framework used is PyTorch.

TABLE 1. Distribution of dataset images.

Split ratio	Male	Female
Training set（80%）	10001	8369
Test set（20%）	2500	2092
Total	12501	10461

TABLE 2. Experimental results.

Gender	Top1 accuracy	MAE (months)	RMSE
Male	66.38%	3.71 ± 7.55	7.56
Female	68.63%	3.81 ± 7.74	7.75

The parameters are set as follows. The number of local regions is $K = 4$; that is, the features of four local informative regions and those of the entire image are all fed to the final classification network. The pretrained model of ResNet-50 is loaded and set "pretrained = true," a batch size of 16 images was used for each step, and the weight decay rate of 10^{-4} was used to prevent overfitting. The learning rate is set to 0.001 at the start and then decays at a decay rate of 10% every 25 epochs. The pretrained ResNet-50 is used as a feature extractor, and stochastic gradient descent is used for optimization. The test model is saved every epoch during the training process. The test model size is approximately 113 MB. We trained the model for 100 epochs.

C. EXPERIMENT RESULTS

1) ACCURACY OF BONE AGE ESTIMATION

We used the top1 accuracy, MAE, and root-mean-square error (RMSE) as performance metrics. Their formulas are given below.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \tag{6}$$

where N represents the number of test data points, and y_i and $\hat{y}_i$ denote the ground truth and the predicted bone age, respectively.

The results of the prediction accuracy, MAE, and RMSE are shown in Table 2.

The data in Table 2 are the average value of the five experiments. The test model corresponding to the epoch

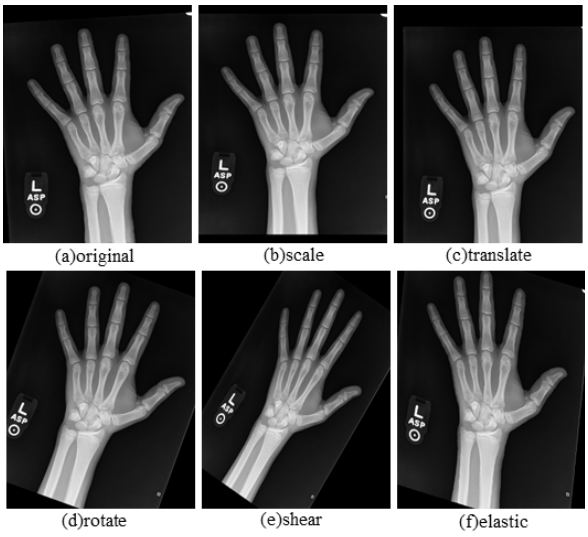

FIGURE 10. Data augmentation methods.

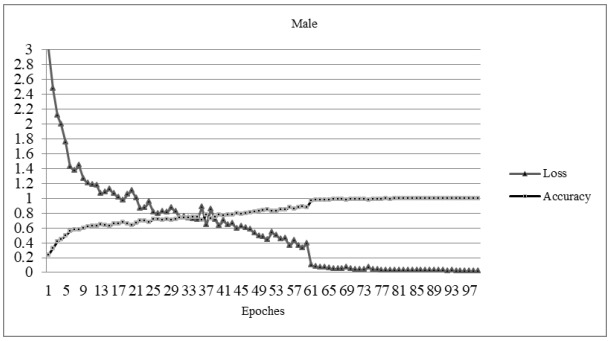

FIGURE 11. Loss and accuracy curve of training set (male).

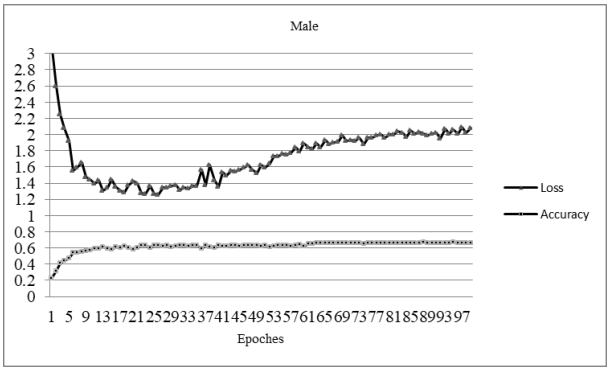

FIGURE 12. Loss and accuracy curve of test set (male).

with the highest accuracy during the training process was saved. The fluctuations in the loss and the test accuracy value during the training process of the experiment are shown in Figures 11, 12, 13 and 14 respectively. The line with the triangle mark represents the loss rate, and the line with the square mark represents the top1 accuracy. The curve value is the average of five training/testing experiments.

For the male data set, the number of iterations for each epoch is 625 (batch size = 16), and the model training process requires a total of 62500 iterations. For the female data set, the number of iterations is 523 per epoch, and a total of 52,300 iterations are required for the model training process.

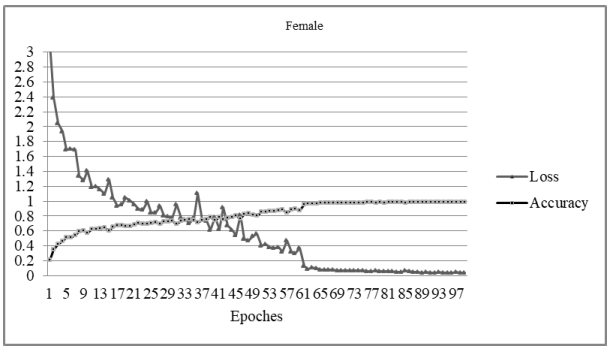

FIGURE 13. Loss and accuracy curve of training set (female).

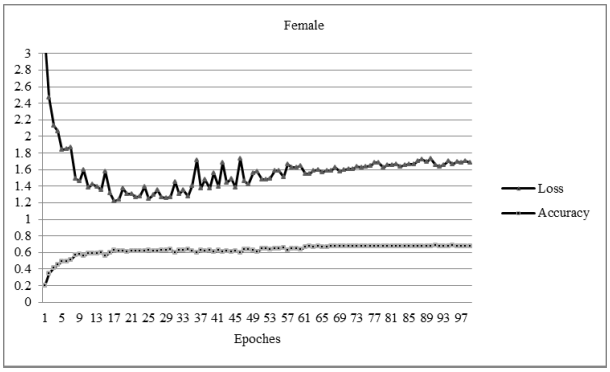

FIGURE 14. Loss and accuracy curve of test set (female).

Our method achieved a top1 accuracy of 66.38% for males and 68.63% for females on the test set, and an MAE of 3.71 and 3.81 months for males and females, respectively, was obtained. Compared with previous studies, our model shows the best performance in the fine-grained weakly supervised object classification algorithm, which also illustrates the effectiveness of the self-supervision mechanism.

2) BONE AGE PREDICTION ACCURACY OF IRREGULAR SAMPLES

The irregular samples mentioned in section II.C may be caused by an unsuitable angle of film shooting. These samples have large invalid areas, which significantly degrade bone age recognition. This study uses the fast scanning algorithm to remove these invalid regions and then estimates the bone age of valid regions to verify the necessity of the image preprocessing step. In the study, irregular samples with and without preprocessing were tested, and the results are shown in Table 3. The data in the table are also the average of the same five test models in section C.

TABLE 3. Bone age estimation results of irregular samples (MAE (months)).

Gender	Original	Preprocessing	Samples
Male	13.59 ± 21.05	11.71 ± 18.12	147
Female	11.81 ± 18.54	9.74 ± 13.95	150

As illustrated in Table 3, compared with that of the original images, the estimated bone age MAE of the preprocessed images for males and females increased by 13.83% and 17.53%, respectively. This study shows that performing

TABLE 4. Comparison of different deep learning-based methods.

Author	methods	Preprocess	Type	Dataset size	Results (months)	
					MAE	RMSE
Our method	NTS-Net	No	End to end	12501	3.71	7.56
Xuhua Ren et al[18]	Faster R-CNN+ Regression Network	Yes	Two-stage	12480	5.20	5.10
Bo Liu et al[17]	VGG-U-Net+ VGG	Yes	Two-stage	12811	5.98	9.72
Spampinato et al[22]	CNN+ Regression Network	Yes	End to end	1391	9.48	---
Daniel Souza et al[23]	Residual networks	Yes	End to end	12500	6.44	---
Iglovikov et al[15]	U-Net+ VGG	Yes	Two-stage	12600	6.16	---
Lee et al[11]	CNN+ GoogLeNet	Yes	Two-stage	8325	---	9.84
E. Wu et al[19]	MaskRCNN+ Residual Attention Network	Yes	Two-stage	12500	7.38	---

only a simple preprocessing step (the scan line scanning algorithm) can improve our model's bone age estimation accuracy.

D. ALGORITHM COMPARISON

For algorithm comparison, this study's results are compared with the results of DL-based BAA algorithms published in the literature. The comparison results are shown in Table 4.

Based on the comparison results in Table 4, our method outperforms the existing methods. The method proposed in this study is an end-to-end method. The model directly receives an original image as input and outputs the bone age. The method in this study is a typical end-to-end bone age estimation algorithm based on weak supervision. The method lso proves the effectiveness of this self-supervised training paradigm for fine-grained image classification.

IV. DISCUSSION

In this study, a fine-grained image recognition network was used to automatically estimate the bone age, and relatively satisfactory results were obtained on the RSNA public dataset.

We use the attention map of a neural network to illustrate the superiority of the algorithm proposed in this study, as shown in Figure 15.

The figure shows that our algorithm is divided into two branches. The first branch (denoted as ①) extracts the features of the complete image directly using ResNet, and the second branch (denoted as ②) generates four part images, with features extracted for each part. Finally, the feature maps of the two branches are concatenated (denoted as ③), and the

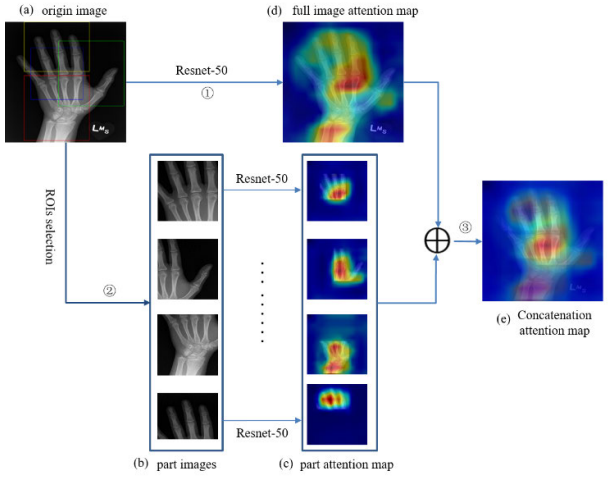

FIGURE 15. Illustration of the merit of our proposed method using attention map: (a) input image; (b) part images proposed by the navigator module; (c) attention map of the part images; (d) attention map of the original image; and (e) concatenation attention map.

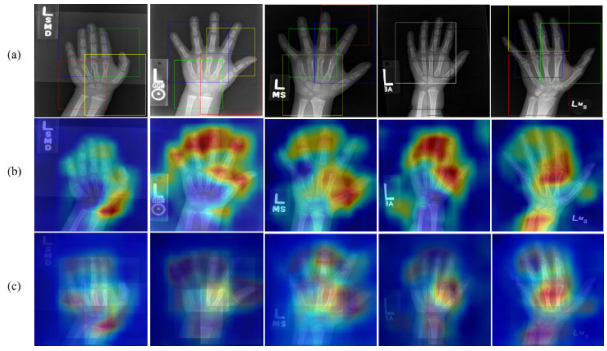

FIGURE 16. Attention map of images: (a) original image and the attention map; (b) attention map of full images; and (c) attention map of the fused features.

fusion feature map of the hand bone image is obtained. The algorithm uses the combined features to classify the hand bone age. From the attention map in Figure 15, the combined feature map (e) pays attention to more regions and detailed information than the image (d). The red area indicates a high attention degree. In addition, Figure 15 (e) pays more attention to the carpal region, which mainly contributes to the part images. Compared with the phalangeal region, the carpal region has significant differences in features between different bone ages. Therefore, clinicians typically pay more attention to this region when reading radiography. This is also the reason for the higher accuracy of our algorithm because the algorithm pays more attention to the carpal regions.

Figure 16 shows the comparison of the combined feature visualization (c) and the original image feature visualization (b).

The first row (a) represents the original image. The second row (b) represents the class activation map (CAM) of the full image extracted by ResNet-50, and the bottom row (c) represents the visualization CAM results of the fused feature. Figure 16 shows that the red area for each image in row c is greater than that in row b, which means that the fused feature pays more attention to hand bone information.

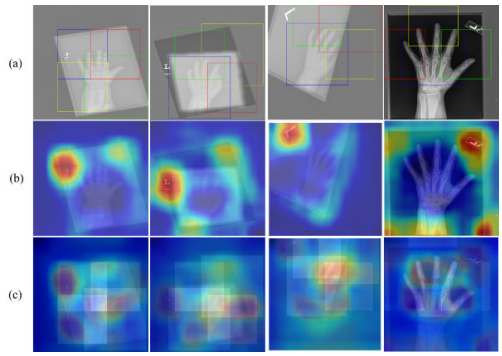

FIGURE 17. Attention map of single feature and fusion feature: (a) original image; (b) full image attention map in a; (c) attention map of fusion features.

In addition, when the hand bone image is sampled irregularly, the method can still focus on the hand region, as shown in Figure 17 (row c). If we use only the features from the full image, the attention areas will be located outside the hand region, completely deviating from the target, as shown in row b in Figure 17.

Our algorithm also has some disadvantages. First, in the algorithm execution process, the selection process of the part images (ROIs) is complicated. It needs to complete a region proposal network process, compute the amount of information for each region, and calculate each region's confidence and the ranking score of regions, so the computational cost is high. A research topic to be studied in the future is how to develop a simpler method for locating the key regions (such as the carpal area).

Second, different bone age estimation methods use different datasets, making a fair comparison difficult. Even when using the same dataset (RSNA2017), it is difficult to achieve a completely fair comparison owing to differences in data augmentation methods, training/test split ratios, and abnormal sample preprocessing steps. As shown in Table 4, in the case of a slight difference in the data size, the method in this study can still achieve a smaller MAE value than can studies in the literature [8]–[11] and [12].

Finally, although the residual network (ResNet) used in the model has good feature extraction capabilities, the number of parameters generated by the 50-layer residual network during training is greater than that of the DenseNet network, which also has good feature extraction capabilities. We will consider replacing the residual network with a DenseNet network in the future.

V. CONCLUSION

In this study, a new deep CNN based on fine-grained image recognition is used to predict the bone age of adolescents. This network can automatically extract the local features of hand radiographs and fuse these features with those of the full image to estimate bone age. It is an end-to-end BAA method. The model in this study uses a random strategy under a fixed training/test ratio (8:2) to generate training/test sets, and a total of five test models under different training/test samples are obtained. The mean value is calculated by performing five experiments to ensure the reliability of the experimental results. Ultimately, the model achieved a prediction accuracy of 66.38% for males and 68.63% for females on the RSNA public dataset. MAE values of 3.71 and 3.81 months were obtained for males and females, respectively. The results outperform those in existing literature [5], [8]–[12]. The algorithm in this study has a fast convergence speed and a high bone age estimation accuracy. The average time for bone age estimation is approximately 35 ms per image. From an application viewpoint, this model can be used to develop a BAA system, with rapid real-time BAA, meeting clinical requirements.

ACKNOWLEDGMENT

The authors would like to thank the RSNA for providing the dataset.

REFERENCES

[1] M. Satoh, "Bone age: Assessment methods and clinical applications," *Clinical Pediatric Endocrinol.*, vol. 24, no. 4, pp. 52–143, 2015, doi: 10.1297/cpe.24.143.

[2] A. M. Mughal, N. Hassan, and A. Ahmed, "Bone age assessment methods: A critical review," *Pakistan J. Med. Sci.*, vol. 30, no. 1, pp. 211–215, Dec. 1969, doi: 10.12669/pjms.301.4295.

[3] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, "The RSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, Nov. 2018.

[4] 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, pp. 161–173, May 2019.

[5] S. H. Tajmir, H. Lee, R. Shailam, H. I. Gale, J. C. Nguyen, S. J. Westra, R. Lim, S. Yune, M. S. Gee, and S. Do, "Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability," *Skeletal Radiol.*, vol. 48, no. 2, pp. 275–283, Feb. 2019.

[6] Z. Yuqing, "Talking about the application of bone age in the study of children's physique," *Phys. Educ.*, vol. 3, pp. 55–56, Mar. 1988.

[7] W. W. Greulich, S. I. Pyle, and T. Wingate, *Todd Radiographic Atlas of Skeletal Development of the Hand and Wrist*, vol. 2. Stanford, CA, USA: Stanford Univ. Press, 1959.

[8] M. Mansourvar, M. A. Ismail, T. Herawan, R. G. Raj, S. A. Kareem, and F. H. Nasaruddin, "Automated bone age assessment: Motivation, taxonomies, and challenges," *Comput. Math. Methods Med.*, vol. 2013, Oct. 2013, Art. no. 391626.

[9] C. Wu, *Neural Network and Deep Learning*. Beijing, China Electronic Industry Press, 2016, pp. 30–84.

[10] D. Stern, C. Payer, V. Lepetit, and M. Urschler, "Automated age estimation from hand MRI volumes using deep learning," in *Proc. MICCAI*, Athens, Greece, 2016, pp. 194–202.

[11] 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, 2017.

[12] J. Han, Y. Jia, C. Zhao, and F. Gou, "Automatic bone age assessment combined with transfer learning and support vector regression," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 61–66.

[13] W. Tang, G. Wu, and G. Shen, "Improved automatic radiographic bone age prediction with deep transfer learning," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–6.

[14] Z. Wenxiang, "Automatic evaluation of bone age based on X-ray images," M.S. thesis, Dept. Electron. Eng., Univ. Electron. Sci. Technol., Chengdu, China, 2018.

[15] V. Iglovikov, A. Rakhlin, A. Kalinin, and A. Shvets, "Pediatric bone age assessment using deep convolutional neural networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 300–308.

[16] Y. Gao, T. Zhu, and X. Xu, "Bone age assessment based on deep convolution neural network incorporated with segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 12, pp. 1951–1962, Dec. 2020.
[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] 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.
[19] 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)*, Apr. 2019, pp. 1158–1161.
[20] S. Li, B. Liu, S. Li, X. Zhu, Y. Yan, and D. Zhang, "A deep learning-based computer-aided diagnosis method of X-ray images for bone age assessment," *Complex Intell. Syst.*, pp. 1–11, Apr. 2021, doi: 10.1007/s40747-021-00376-z.
[21] I. Salim and A. B. Hamza, "Ridge regression neural network for pediatric bone age assessment," *Multimedia Tools Appl.*, pp. 1–18, May 2021, doi: 10.1007/s11042-021-10935-8.
[22] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," *Med. Image Anal.*, vol. 36, pp. 41–51, Feb. 2017.
[23] D. Souza and M. M. Oliveira, "End-to-end bone age assessment with residual learning," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 197–203.
[24] Y. Ji, H. Chen, D. Lin, X. Wu, and D. Lin, "PRSNet: Part relation and selection network for bone age assessment," in *Proc. MICCAI*. Shenzhen, China, 2019, pp. 413–421.
[25] J. He and D. Jiang, "Fully automatic model based on SE-ResNet for bone age assessment," *IEEE Access*, vol. 9, pp. 62460–62466, 2021.
[26] J. W. Luo, "Review of fine-grained image classification based on deep convolution features," *Acta Automatica Sinica*, vol. 43, no. 8, pp. 1306–1318, Aug. 2017.
[27] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 834–850.
[28] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. ECCV*, Zurich, Switzerland, Sep. 2014, pp. 834–849
[29] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 1–45.
[30] J. Li, D. Wang, S. Li, M. Zhang, C. Song, and X. Chen, "Deep learning based adaptive sequential data augmentation technique for the optical network traffic synthesis," *Opt. Exp.*, vol. 27, no. 13, pp. 18831–18847, Jun. 2019.

KEXIN LI was born in Rongcheng, Hebei, China, in 1977. He received the B.S. degree in agricultural electrification and automation from Agriculture University of Hebei, in 2001, the M.S. degree in precision instruments and machinery from Harbin University of Science and Technology, Harbin, in 2004, and the Ph.D. degree in optical engineering from Harbin Institute of Technology, Harbin, Heilongjiang, in 2009.

From 2009 to 2011, he was a Lecturer with the Electrical Engineering Department, Northeast Forestry University, where he has been an Assistant Professor with the Control Science and Engineering Department, since 2011. His research interests include artificial intelligence, pattern recognition, medical image segmentation, object detection, and plant phenotype detection.

JINGZHE ZHANG was born in Harbin, China, in 1984. She received the B.S. degree in clinical medicine from the Medical School, Jiamusi University, in 2008. Her current research interests include medical imaging and nursing.

YUNFEI SUN was born in Chaoyang, Liaoning, China, in 1995. He received the B.S. degree in engineering from the School of Information and Control Engineering, Shenyang Jianzhu University, Liaoning, in 2018. He is currently pursuing the master's degree with the School of Mechanical and Electrical Engineering, Northeast Forestry University. His current research interests include computer vision and deep learning.

XINWANG HUANG was born in Xiaogan, China, in 1994. He received the B.S. degree from College of Mechatronics and Control Engineering, Hubei Normal University, Huangshi, China, in 2018. He is currently pursuing the M.S. degree with the School of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, China. His current research interests include medical image segmentation and deep learning.

CHUNXUE SUN was born in Baiquan, Qiqihar, Heilongjiang, China, in 1997. She received the B.S. degree from the School of Electronic Engineering, Heilongjiang University, Harbin, in 2015. She is currently pursuing the master's degree with the School of Mechanical and Electrical Engineering, Northeast Forestry University. Her current research interests include computer vision, deep learning, and hyperspectral images processing.

QIANCHENG XIE was born in Qitaihe, Heilongjiang, China, in 1994. He received the B.S. degree from the College of Information Engineering, Shenyang University of Chemical Technology, in 2017. He is currently pursuing the master's degree with the School of Mechanical and Electrical Engineering, Northeast Forestry University. His current research interests include computer vision and deep learning.

SHIJIE CONG was born in Weihai, Shandong, China, in 1997. He received the B.S. degree from the School of Information Engineering, Nanchang University, Nanchang, in 2019. He is currently pursuing the master's degree with the School of Mechanical and Electrical Engineering, Northeast Forestry University. His current research interests include artificial intelligence, computer vision, and deep learning.