IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 January 2024, accepted 19 January 2024, date of publication 26 January 2024, date of current version 6 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358821

RESEARCH ARTICLE

KDALDL: Knowledge Distillation-Based Adaptive Label Distribution Learning Network for Bone Age Assessment

HAO-DONG ZHENG[1], **LEI YU**[1], **YU-TING LU**[2], **WEI-HAO ZHANG**[1], **AND YAN-JUN YU**[1]

[1]College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
[2]School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China

Corresponding author: Lei Yu (ylcqnu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 72071019, and in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX0185.

ABSTRACT Deep learning-based bone age assessment (BAA) approaches have certain drawbacks, such as ignoring the correlation of age labels and simply assuming that bone development is linearly related to bone age, which can affect the accuracy of predictions. To solve these problems, a knowledge distillation-based adaptive label distribution learning method called KDALDL is proposed. The KDALDL framework comprises a teacher model and a student model, both consisting of modules for multi-scale feature extraction, feature refinement, and label distribution learning. First, a multi-scale feature extraction module is designed based on the swin transformer to extract feature information at various scales. Subsequently, these features are fed into the feature refinement module to capture the optimal image features. Then, the discrete labels obtained from the age labels through the Gaussian formula are used to train the teacher model. Finally, the teacher model's outputs are used to train the student model through the knowledge distillation technique, which enables the student model to achieve improved results by learning from the teacher model. The proposed method is validated using the Radiological Society of North America (RSNA) dataset, which exhibits outstanding results.

INDEX TERMS Bone age assessment, knowledge distillation, label distribution learning, swin transformer.

I. INTRODUCTION

Bone age assessment (BAA) is a method used to estimate a person's level of development by analyzing radiographs of their hands. In clinical medicine, BAA can provide significant clinical information in children with endocrine problems [1] and development abnormalities [2], [3]. In addition, it can predict the adult height of children and provide a significant basis for the selection of athletes [4].

Traditional manual methods for BAA mainly include the Greulich-Pyle (GP) method [5] and the Tanner-Whitehouse (TW) method [6]. The GP approach involves the use of a set of standard hand-bone X-ray pictures. In actual practice, radiologists choose the sample from the atlas that most closely reflects the patient's X-ray, and then they use that sample's age to determine the evaluation result. The TW method is a skeletal maturity rating scheme that rates 20 regions of interest (ROIs) based on maturity and calculates an overall maturity score by adding the scores of these regions. Traditional methods in clinical practice have significant limitations, such as manual BAA that is time-consuming and is subjected to subjective judgments, resulting in inconsistent results among doctors for the same X-ray image.

In recent years, the remarkable progress of deep learning has facilitated its extensive incorporation in diverse domains. Fields such as object detection [7], laser point cloud segmentation [8], and medical image processing [9] have all benefited from deep learning techniques. Notably, deep learning has made significant strides in the field of BAA, surpassing the performance of even seasoned experts. However, these methods define BAA as a multi-classification

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

or regression task. Multi-classification methods assume that age labels are mutually independent, and the goal is to learn the features of various bone ages and predict the likelihood of each category. Nevertheless, age labels are an ordered set, so classification methods ignore the correlation between labels. Regression methods presume a linear relationship between bone characteristics and bone age, while the development of bones does not follow a linear pattern [10].

Inspired by the knowledge distillation (KD) technique [11], [12] and label distribution learning (LDL) [13], we propose a knowledge distillation-based adaptive label distribution learning (KDALDL) network to address these problems. In the KDALDL framework, there is a teacher model and a student model, which are similar in structure and both have three steps: feature extraction, feature enhancement, and label distribution learning. Firstly, image features at different scales are extracted using a multi-scale feature extraction module, and then these features are processed using a feature refinement module to capture the optimal image features. Subsequently, we train the teacher model using the discrete labels obtained from age labels through a Gaussian distribution formula, along with the image features and gender. Finally, the KD technique is employed to transfer the teacher model's adaptive label distribution to the student model, which can further refine the labels and improve the accuracy of BAA.

The following is a summary of KDALDL's major contributions:

1) In the proposed KDALDL, the KD technique allows for the learning of adaptive label distributions. This approach can effectively capture the correlation between labels and enhance the utilization of labels.
2) The swin transformer network structure is optimized by adding skip connections, which are beneficial for the extraction of multi-scale features. Furthermore, a feature enhancement module is specifically designed to fuse and improve these extracted features.
3) The effectiveness of the proposed method for BAA is validated on the RSNA dataset. The experimental results show a good performance, achieving a mean absolute error (MAE) of 4.45 months.

II. RELATED WORKS

A. BONE AGE ASSESSMENT

In the field of BAA, numerous machine learning methods have been applied over the past decades, including k-nearest neighbor classification [14], decision tree methods [15], and support vector machines [16], [17]. However, these approaches are time-consuming and often lead to low accuracy.

The emergence of deep learning-based methods for BAA has been facilitated by the advancements in deep neural networks and the availability of a substantial dataset provided by the RSNA [18]. These methods fall into two major categories: classification-based methods and regression-based methods. In classification-based methods, the X-ray image is inputted into a convolutional neural network (CNN), which subsequently generates predictions for the age category. For example, Chen [19] employed VGGNet with transfer learning as the classification model and applied the GP method for their study. Lee et al. [20] introduced an automated BAA system that involves segmenting ROIs, normalization and preprocessing of images, and age classification. Bian and Zhang [21] improved GoogleNet for age classification and expanded datasets to prevent overfitting problems. Mao et al. [22] regarded BAA as a fine-grained image classification task. Larson et al. [23] used ResNet as the basic framework to out the most probable age categories. Bui et al. [24] combined the TW3 method with a deep convolutional network using Faster R-CNN and InceptionV4 for detection and classification. Gao et al. [25] first used U-Net to eliminate the extraneous background from the images and then used VGGNet with attention mechanism for age classification.

The regression-based methods predict bone age by fitting a linear relationship between image features and bong age. For example, BoNet is proposed by Spampinato et al. [26] which contains a feature extraction network and an age regression network. Iglovikov et al. [27] combined active mining techniques with a U-Net allowing for quick hand masking for image segmentation. Then, VGGNet was used as a regression network to estimate bone age. Koitka et al. [28] built an automated system for BAA, which contains a detection network for identifying the ossification areas and a region-specific regression network for estimating bone age. Nguyen et al. [29] applied transfer learning techniques to sex determination and demonstrated that gender provides important information for BAA. Escobar et al. [30] introduced a new BAA framework that includes hand posture estimation and hand detection for local feature extraction. Ji et al. [31] used a Feature Pyramid Network to locate the informative ROIs and subsequently employed them for bone age regression. Liu et al. [32] introduced a multi-scale data fusion framework using the Non-Subsampled Contourlet Transform (NSCT). This method first transmits NSCT coefficient maps at different scales to the neural network and then fuses the information at different scales to predict bone age, which demonstrates that incorporating multi-scale information leads to a more robust estimation. Wang et al. [33] proposed a novel method using a dual-path network (DPN) with attention mechanisms. The DPN incorporates residuals and dense connectivity, which helps to extract deeper and more efficient features. In addition, two different attentional mechanisms were utilized to further enhance the feature extraction process. In addition, a two-stage convolutional transformer network was proposed by Mao et al. [34]. In the first stage, the ROIs were extracted using YOLOv5, next in the second stage, gender information was integrated with image features to replace the positional encoding of the transformer. The transformed features were finally used to predict age. Jian et al. [35] proposed a TENet for bone age regression. This model consists of a topology module, which focuses on extracting information about the ROIs, and an edge enhancement module, which enhances the hand bone edge features. Li et al. [36] proposed a two-stage network for BAA. In the first stage, a visual heat map is used to sequentially identify the two most discriminatory ROIs. Then, in the second stage, a feature fusion strategy is employed to combine the features of these selected ROIs with gender information for more precise prediction. The existing methods used to assess bone age have certain limitations. One limitation is seen in classification methods, which overlook the correlation between different labels. On the other hand, regression methods tend to oversimplify the skeletal image features and bone age, assuming a linear relationship between them.

B. LABEL DISTRIBUTION LEARNING

Label ambiguity refers to the uncertainty that exists in assigning labels to instances. This often increases the difficulty of object detection and classification tasks. To tackle this challenge, Geng proposed a method called label distribution learning (LDL) [13]. The LDL method optimizes the training process of the model by learning from multiple label sources and exploiting the interrelation among labels. This approach significantly reduces the impact of label ambiguity on the results. Hence its introduction, LDL has gained significant attention in the realm of deep learning and has been successfully implemented in diverse domains, such as facial expression recognition [37], traffic prediction [38], and facial age estimation [39], [40]. Zhou et al. [37] presented an innovative approach to facial expression recognition by introducing emotional distribution learning. They highlighted that a facial expression is influenced by multiple emotions, rather than just a single emotion. The core concept of emotion distribution learning involves understanding the specific contributions of different emotions in describing facial expressions. Moreover, it aims to establish a connection between facial expression images and the distribution of emotions. Zeng et al. [38] proposed a method for traffic speed prediction based on conditional distribution learning. They aimed to overcome the limitations imposed by conditional distribution during the training process. The authors first modeled the speed class using conditional distribution learning and then utilized the learned speed distribution to predict the speed results effectively. He et al. [39] introduced a data-dependent LDL method for facial age estimation. They not only emphasized that facial age estimation is influenced by a single age label but also highlighted the impact of the label context. The label context refers to the set of labels associated with visually similar facial samples.

Indeed, the issue of label ambiguity is prevalent in BAA tasks. The bone age labels often come in a discrete form (in months), which can result in instances where the actual bone age for an image falls between two labels. Additionally, multiple labels may hold similar importance for a particular instance. Furthermore, in datasets like RSNA, the bone age labels are determined jointly by multiple medical experts, leading to variations in their interpretation of images and further exacerbating label ambiguity. To enhance the accuracy of BAA, scholars have explored the use of LDL technology. For example, Chen et al. [41] attempted to transform bone age labels into fixed age distribution labels using the normal distribution formula. They then combined LDL with expectation regression to predict bone age. However, this approach has limitations due to the uneven speed of bone development at different stages, making the fixed age distribution inadequate in representing the complex bone development process. To address these issues, we propose a novel method for adaptive label distribution learning based on knowledge distillation. Our approach involves training a teacher model with a fixed age distribution, which serves as a source of rich bone age information. Through knowledge distillation, we distill an adaptive label distribution from the teacher model, enabling a more accurate representation of the bone development process at different stages. Finally, we utilize the distilled knowledge to train a student model, thereby improving the accuracy of bone age evaluation.

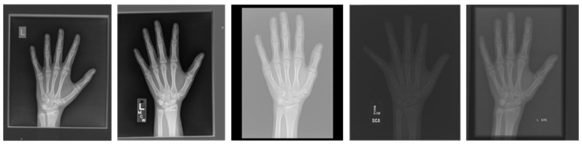


FIGURE 1. The X-ray image samples in the RSNA database.

III. MATERIALS AND METHODS

A. DATASET

The dataset utilized in this experiment is derived from the RSNA [18], which is recognized as the largest publicly available dataset in the field of BAA. It encompasses a vast collection of over 14000 digital radiography images of pediatric hand radiographs, which is structured with 12611 training images, 1425 validation images, and 200 testing images. These images were captured using digital radiography equipment sourced from various institutions, including the University of Colorado, Stanford University, and the University of California - Los Angeles. All images are stored in PNG format and are accompanied by skeletal age annotations. These annotations were meticulously generated through a manual process conducted by two pediatric radiologists from each participating institution. The annotations are presented in a spreadsheet format, providing comprehensive information about the estimated skeletal age in months, as well as gender information for each image. For a visual representation, Fig. 1 displays a selection of image samples from the RSNA database.

B. METHOD OVERVIEW

To capture the detailed and optimized features from the bone images and make the most of the relationships between labels in BAA, we propose a novel network employing multi-scale

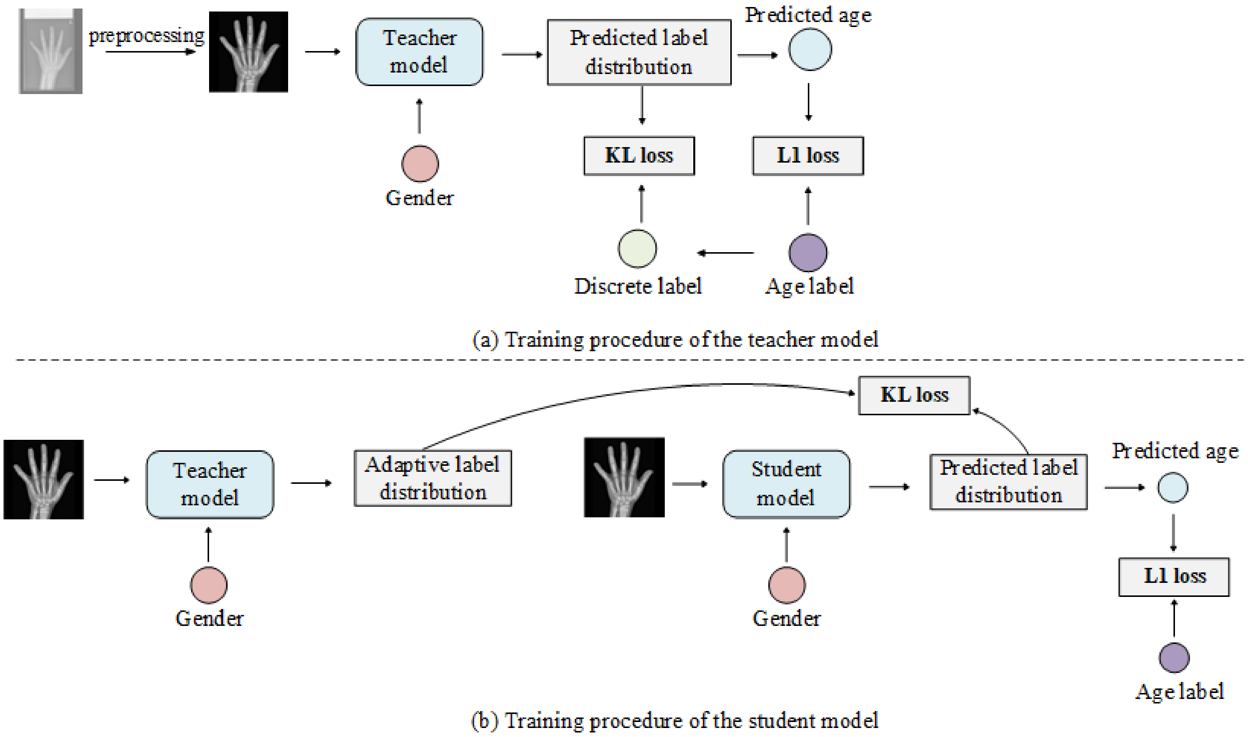

preprocessing
Teacher model
Predicted label distribution
Predicted age
Gender
KL loss
L1 loss
Discrete label
Age label
(a) Training procedure of the teacher model
KL loss
Teacher model
Adaptive label distribution
Student model
Predicted label distribution
Predicted age
L1 loss
Gender
Gender
Age label
(b) Training procedure of the student model

FIGURE 2. The overall framework of the proposed method. (a)Training procedure of the teacher model; (b)Training procedure of the student model.

feature extraction, attention mechanisms, and KD techniques. Fig. 2 illustrates the overall framework of our proposed method.

The entire framework can be split into data preprocessing, teacher model training, and student model training. The first step involves hand segmentation and generating discrete labels based on age labels. In the second step, as depicted in Fig. 2a, we input the preprocessed images together with the gender information into the teacher model and train it. Afterward, the model outputs the predicted label distributions and preliminary estimated age. Finally, as shown in Fig. 2b, the student model inherits the adaptive label distribution learned by the teacher model through the KD technique and outputs the final predicted age.

C. DATA PREPROCESSING

First, a UNet++ is employed to segment the hand from the X-ray image to minimize the influence of irrelevant noise on the experimental results. Then, the contrast of the images is improved using the histogram equalization technique, which enhances the details of the bones.

Furthermore, to improve the performance of the teacher model, we incorporate the concept of LDL into the training. Specifically, we generate the discrete distribution labels based on age labels using the Gaussian distribution formula. The normal distribution formula is shown in (1).

$$p_{ij} = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - j)^2}{2\sigma^2}), \tag{1}$$

where y_i is the real age label, p_{ij} is the probability that the bone age of sample i is j, σ is a hyperparameter used to control the level of discreteness of the label values.

D. THE TEACHER MODEL

As depicted in Fig. 3, the proposed teacher model comprises a multi-scale feature extraction module, a feature refinement module, and a label distribution learning module.

1) FEATURE EXTRACTION

Due to the swin transformer introducing a multi-level feature representation, it is used as the basic structure in the multi-scale feature extraction module. Further details can be observed in Fig. 4.

To utilize the pre-trained weights of the swin transformer on ImageNet, we convert the single-channel grayscale images to three-channel RGB images. The 512×512×3 image is first divided into patches of size 4×4 pixels by the Patch Partition layer. Subsequently, these patches are flattened along the channel direction, resulting in an image shape of 128×128×48. Finally, four stages are stacked to construct feature maps at various scales. Each stage (except for Stage 1, where Linear Embedding is applied first to decrease the number of channels in the feature map) starts with a Patch Merging layer that reduces the size of the feature map while simultaneously doubling the number of channels.

To obtain multi-scale information from the image, we have modified the swin transformer structure. Specifically, skip

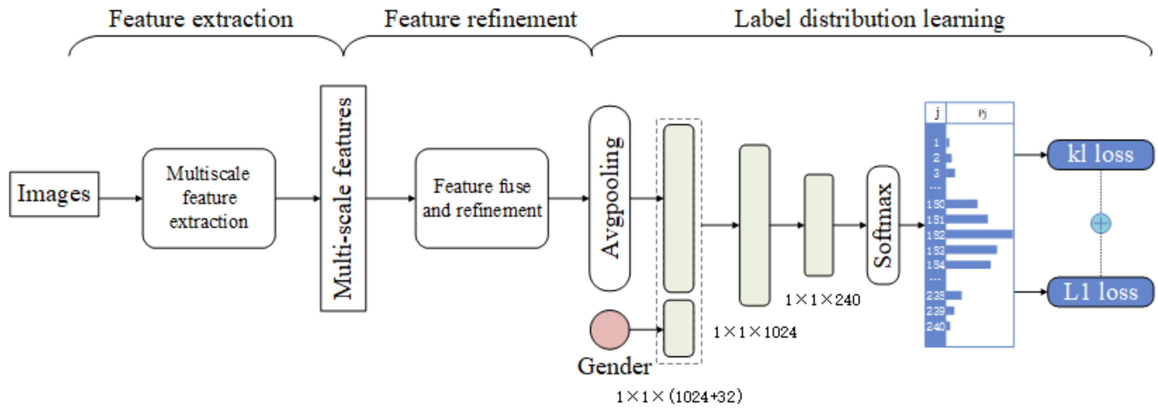


FIGURE 3. The network architecture diagram.

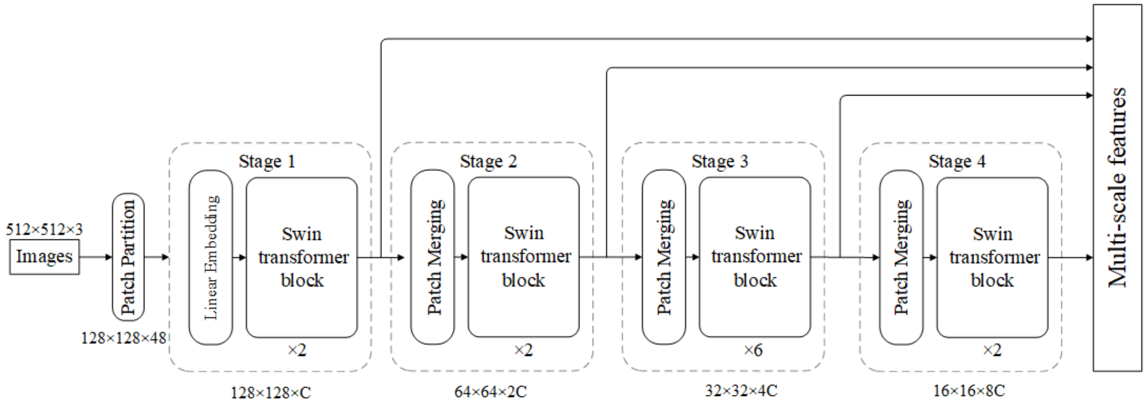


FIGURE 4. The multi-scale feature extraction module.

connections are added after each stage to extract feature maps with different dimensions. This modification provides several advantages. Firstly, it allows the model to better capture local information in the images, such as bone edges and textures, through the extraction of the small-scale features. Secondly, it enhances the model's perception of global information by extracting large-scale features from the images. Furthermore, the inclusion of skip connections increases the diversity of features, allowing information from feature maps at different scales to complement one another. This enhances the model's overall perception and understanding capabilities, providing a more comprehensive representation.

2) FEATURE REFINEMENT

The feature refinement module is designed to obtain the optimal image feature. Fig. 5 illustrates the structure of the module.

In this module, the feature maps extracted from the multi-scale feature extraction module are first up-sampled to 1/4 of the original image size using bilinear interpolation. Then, they are connected along the channel dimension. Ultimately, the optimal image features are captured by using the Squeeze-and-Excitation channel attention mechanism, this mechanism enables the model to be more focused on the discriminative features and effectively suppress noisy or less relevant information by adaptively assigning weights to the different channels.

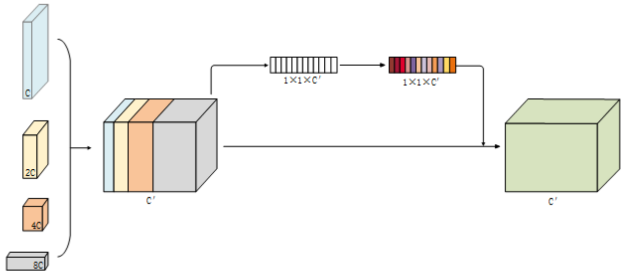


FIGURE 5. The feature refinement module.

3) LABEL DISTRIBUTION LEARNING

The purpose of the label distribution learning module is to learn the relationship between age labels. The structure of the module is shown on the right of Fig. 3.

In BAA, in addition to the image features, gender information is also crucial as female development tends to occur at a faster pace compared to males. To make the most of this information, we map gender to 32 neurons and then concatenate it with the feature vector that is obtained from the global average pooling of the image feature map. Subsequently, the combined feature vector is fed into two fully connected layers with 1024 and 240 neurons, respectively. Finally, the model outputs the logits for each month, denoted as $z_i \in R^{240}$, which is then transformed into the predicted age distribution using a softmax activation function. The softmax formula is defined as (2).

$$\hat{p}_{ij} = \frac{\exp(z_{ij})}{\sum_j \exp(z_{ij})}, j = 1, 2, \ldots, 240, \tag{2}$$

where $\hat{p}_{ij}$ denotes the probability of the i-th sample having a predicted age of j. The following formula produces the final estimated age:

$$\hat{y}_i = \sum_{j=1}^{240} j \cdot \hat{p}_{ij}. \tag{3}$$

We take into account two criteria when training the teacher model. The first criterion is the Kullback-Leibler (KL) divergence, which quantifies the disparity between the discrete label p_{ij} and the predicted distribution $\hat{p}_{ij}$. Following is a definition of the KL loss function:

$$L_{KL} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{240} j \ln \frac{p_{ij}}{\hat{p}_{ij}}. \tag{4}$$

And the second is the MAE, which measures the difference between the real age label y_i and the expected age $\hat{y}_i$. The MAE can be expressed as follows:

$$L_{MAE} = \frac{1}{N} \sum_{i=1}^{N} || \hat{y}_i - y_i || . \tag{5}$$

Finally, the joint loss function is obtained by combining the two loss functions:

$$L = \lambda L_{MAE} + (1 - \lambda) L_{KL}, \tag{6}$$

where λ is hyperparameters to balance the two losses.

E. THE STUDENT MODEL

The adaptively distributed labels that the teacher model produces after training are then passed along to the student model as new knowledge. The transferred knowledge can be represented as

$$\hat{p}^t_{ij} = \frac{\exp(\frac{z_{ij}}{\tau})}{\sum_j \exp(\frac{z_{ij}}{\tau})}, j = 1, 2, \ldots, 240, \tag{7}$$

where the variable "τ" represents the distillation temperature, which controls how soft the labels will be. With an increase of τ, the distribution will become softer. As shown in Fig. 2, the training procedure of the student model is similar to that of the teacher model, the only distinction lies in the KL loss. Specifically, the KL loss in the student model measures the discrepancy between the predicted distribution $\hat{p}^s_{ij}$(student model's logits after distillation function) and the transferred knowledge $\hat{p}^t_{ij}$. Finally, in the bone age prediction stage, only the student model is used, and the distillation temperature is set to $\tau = 1$.

TABLE 1. Parameter settings.

Parameters	Value
Total epoch	55
Batch size	8
Learning rate	0.0001
σ	14
λ	0.5
τ	2

IV. EXPERIMENT RESULTS AND DISCUSSIONS

A. EVALUATION METRICS AND EXPERIMENTAL SETTING

The performance of the proposed KDALDL is evaluated by calculating the MAE between the predicted age and the true age from the test set. The MAE is the most widely used and authoritative evaluation metric in the BAA domain.

Table 1 lists the parameter settings of our experiments.

B. COMPARISON WITH OTHER METHODS

We test the proposed method on the RSNA dataset to confirm its effectiveness, and the results are displayed in Fig. 6. As can be seen, the KDALDL can achieve good results. As demonstrated in Fig. 6a, the results indicate a high degree of consistency between the predicted and the true labels. Fig. 6b shows that the absolute deviation is mainly less than 15 months. Overall, the proposed network can achieve high accuracy and reliability in the BAA task.

Next, we compared KDALDL with other state-of-the-art BAA methods on the RSNA test set. These methods can be grouped into two categories: classification-based methods [22], [24], [25] and regression-based methods [27], [28], [29], [33], [34], [35], [36], [42], [43], [44]. Among these methods, [22], [24], [27], [28], [34], [35], [43] trained the model through ROI annotation.

Comparing the MAE indicators in Table 2, it can be concluded that: (1) Regression-based methods are superior to classification-based methods because age is an ordered continuous variable, rather than mutually independent discrete labels. Regression-based methods can provide a coherent and smooth prediction across the entire age range, rather than simply assigning the prediction to discrete age labels. (2) ROI-based methods usually achieve better results compared to ROI-free methods. This is because ROIs not only reduce the redundant information in the images but also guide the model to extract crucial features to enhance prediction accuracy. (3) The proposed KDALDL method achieves the best results without ROI annotation. In comparison to the classification-based methods, the KDALDL method combines the LDL technology to transform the classification task into an LDL task, which makes full use of label correlation and achieves better performance. Additionally, in comparison to the regression-based methods, the KDALDL method outperforms by leveraging the KD technology to learn the adaptive label distribution at different stages of development. Furthermore, the proposed method not only extracts multi-scale information from the images but also utilizes attention mechanisms to reduce redundancy and

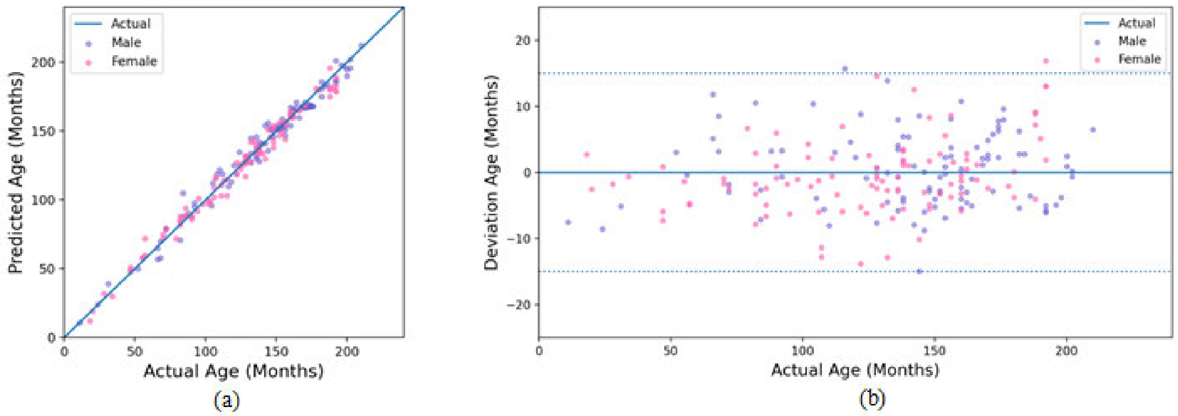


FIGURE 6. Performance of KDALDL in RSNA. (a) Correlation between the actual and predicted bone ages; (b) Correlation between the actual ages and deviation.

TABLE 2. Comparisons with the state-of-the-art methods on the RSNA dataset.

	Method	ROI annotation	MAE (in months)
Classification-based	Bui et al. [24]	✓	7.08
	Mao et al. [22]	✓	6.65
	Gao et al. [25]	✗	9.99
Regression-based	Iglovikov et al. [27]	✓	4.97
	Liu et al. [43]	✓	4.97
	Koitka et al. [28]	✓	4.56
	Mao et al. [34]	✓	4.585
	Jian et al. [35]	✓	5.35
	Nguyen et al. [29]	✗	4.68
	Wang et al. [33]	✗	4.76
	Li et al. [36]	✗	5.45
	Tang et al. [42]	✗	5.53
	Ozdemir et al. [44]	✗	5.75
Our	KDALDL	✗	**4.45**

noise, which achieves similar effects to ROIs by accurately capturing key features related to the bone structure.

C. VISUALIZATION AND INTERPRETATION

To visualize the results of feature learning, we perform visualization experiments using the Grad-CAM [45]. This technique generates attention maps on input images, highlighting the significance of individual local regions in making predictions. By analyzing these attention maps, we can understand the significance of different regions that contribute to the decision-making process of the models.

Fig. 7 shows some visualization examples from the RSNA test set. These images can be categorized into four age groups: (a) toddler stage, (b) pre-puberty stage, (c) mid-puberty stage, and (d) late-puberty stage. As shown in Fig. 7, there is a very significant difference in the areas that the model focuses on at different age groups. For toddlers, the carpal bones and metacarpals are highlighted. For pre-puberty, the highlighted area is the carpal bones. During the mid-puberty stage, the highlighted region shifts to the junction between the phalanges and metacarpals. Finally, during the late-puberty stage, the phalanges become the most important highlighted area when the carpal bones overlap. This is consistent with human a priori knowledge that the carpal bones provide an important basis for BAA when the carpal bones do not overlap, whereas when the carpal bones overlap, the carpal bone information becomes unreliable and analysis of the other bones can lead to more accurate results [10]. These results further demonstrate the credibility and reliability of the proposed method.

Fig. 8 illustrates the differences between KDALDL and InceptionV3 (Ozdemir et al. [44]) in terms of attention maps. By comparing the figure, we can observe distinct behaviors between the two models. Firstly, the attention maps of InceptionV3 exhibit a highly localized focus, which is primarily centered around the wrist or palm region and can simultaneously disregard information from other areas. This localized attention may result in overlooking other crucial features and contextual information when processing hand images. In contrast, the attention maps of our KDALDL model can cover the entire hand region, indicating that our model can focus on the global features of the hand. In addition, the global attention of KDALDL enables it to capture crucial information from various hand regions, and also allow for prioritization of different focus areas. This capability can further improve the effectiveness of KDALDL in extracting multi-scale features in image processing.

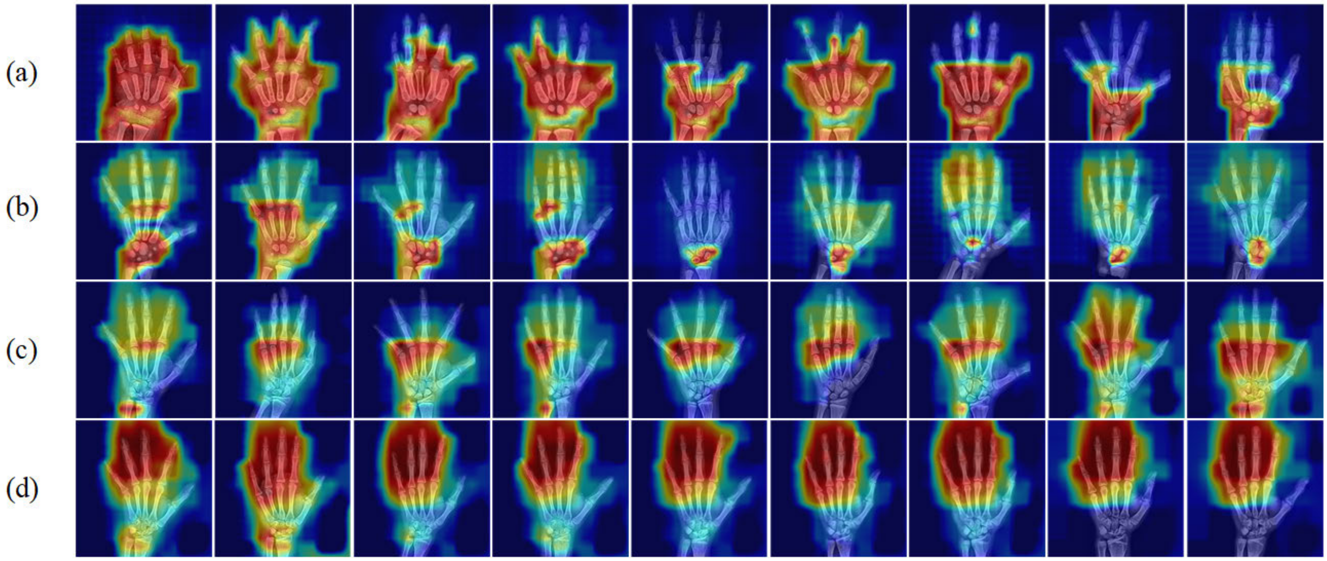


FIGURE 7. Examples of visualization with attention maps. (a) toddler stage; (b) pre-puberty stage; (c) mid-puberty stage; (d) late-puberty stage.

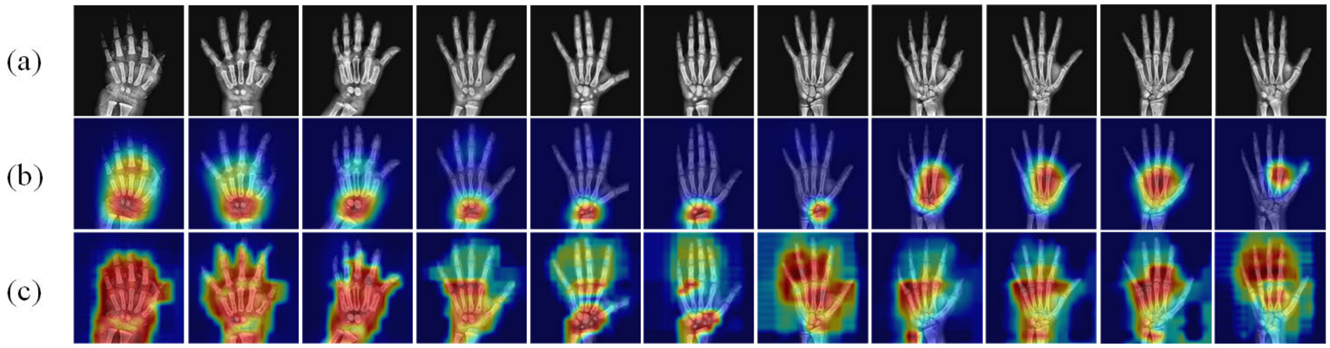


FIGURE 8. Experiments on attention maps. (a) input images; (b) the attention maps of InceptionV3; (c) the attention maps of our KDALDL model.

TABLE 3. Comparison with regression, regression combined with LDL, and KDALDL (Regression with LDL and KD) on the RNSA dataset.

Method	MAE (in months)
Reg	4.97
Reg + LDL	4.63
KDALDL (Reg+ LDL + KD)	**4.45**

TABLE 4. Comparison of results with different teacher models based on swin transformer.

Models	MAE (in months)
Swin-Tiny	4.72
Swin-Small	**4.63**
Swin-Base	4.69
Swin-Large	4.87

D. ABLATION STUDIES

1) THE SUPERIORITY OF KDALDL

To demonstrate that LDL can learn the relationship between age labels, we compare the performance of regression combined with LDL versus regression. Meanwhile, to demonstrate the effectiveness of KDALDL, we compare it with regression combined with LDL.

The results of the experimental comparison are presented in Table 3. It is evident from the results that incorporating LDL can improve network performance, which confirms that LDL can mitigate the impact of label ambiguity. In addition, KDALDL can further enhance the performance of the network, indicating that the adaptive label distribution learned by the teacher model is beneficial for training the student model.

2) THE IMPACT OF DIFFERENT TEACHER OR STUDENT MODELS

Table 4 lists the MAEs of different teacher models based on swin transformer with different sizes on the RSNA dataset. The table reveals that a larger swin transformer is not always superior and Swin-Small gets the best performance.

Table 5 displays the results using the different student models. The results indicate that the best performance

TABLE 5. Comparison of results with different student models.

Models	MAE (in months)	
	Without KD	With KD
Densent	5.37	4.92
Resnet	5.38	5.23
Efficientent	5.46	5.16
MobileNet	5.79	5.44
InceptionV3	5.43	5.17
Swin-Small	**4.63**	**4.45**

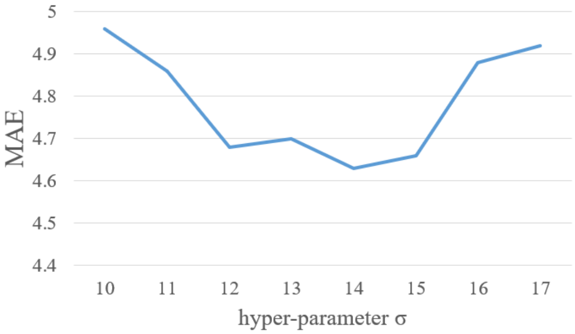


FIGURE 9. The influence of the hyper-parameter σ on teacher model training.

TABLE 6. The influence of the hyper-parameter τ on student model training.

Hyper-parameter τ	MAE (in months)
1	4.65
2	**4.45**
3	4.54
4	4.73

obtained employing Swin-Small and KD can further improve the accuracy of BAA. From a different perspective, they also demonstrate the superiority of adaptive label distribution learning in the BAA task.

3) INFLUENCE OF THE HYPER-PARAMETER σ

In (1), the parameter σ controls the width of the Gaussian distribution. Specifically, the larger the value of σ, the more dispersed the distribution. Conversely, the smaller the value of σ, the narrower the distribution.

We conduct a comparison of MAEs on the RSNA dataset for various σ values. With a fixed λ value of 0.5. The results for σ values ranging from 10 to 17 are presented in Fig. 9. When $\sigma = 14$, we achieve the best MAE of 4.63. Therefore, for our experiments, we opt to use $\sigma = 14$.

4) INFLUENCE OF THE HYPER-PARAMETER λ

The hyper-parameter λ is used in our model to balance the significance of the LDL and regression. Specifically, we fix σ to 14 and adjust the parameter λ from 0 to 1. Finally, we find that setting the parameter λ to 0.5 yields the best performance.

5) INFLUENCE OF THE HYPER-PARAMETER τ

In knowledge distillation, the parameter τ is used to adjust the softness for the predicted label distribution from the teacher model. Then, the student model is trained to match these soft targets. Table 6 presents the results of the experiment. As can be seen, the student model performs best when τ is set to 2.

6) INFLUENCE OF THE GENDER

To investigate the impact of gender in BAA, we conduct an experiment by removing the gender feature. The results indicate a significant decline in model performance after the removal of the gender feature. The MAE with the gender information is 4.45 while the MAE without it is 6.50. This indicates that gender feature is an important factor for BAA, which is consistent with recent research [29].

V. CONCLUSION

In this paper, a BAA method based on knowledge distillation for adaptive label distribution learning (KDALDL) is proposed. In the KDALDL framework, LDL can mitigate the influence of label ambiguity and enhance the utilization of labels, so it is more suitable for BAA tasks compared to classification and regression methods. In addition, we employ the Grad-CAM technique to generate heat maps, which help to visualize the importance of each local area in the model prediction, thereby can increase the model's credibility. It is interesting to note that the regions of interest identified by our model are consistent with prior knowledge, which provides further support for the validity of our results. Our proposed method achieves an MAE of 4.45 on the RSNA test set, which outperforms some recent methods.

Although the proposed model has achieved remarkable performance, it is crucial to acknowledge that there is still room for further improvement. The two-stage training model presented in this paper may introduce additional complexities, such as computational and memory overhead. In our future research endeavors, we aim to explore lighter and more efficient end-to-end models that can be seamlessly integrated into clinical practice. Furthermore, the implications and economic advantages of this research in clinical practice are substantial. It not only reduces the workload of physicians but also enhances the efficiency of the healthcare process. By aiding in the diagnosis and treatment of disorders related to bone development, it has the potential to greatly improve patient care outcomes. Additionally, the findings of this research have the potential to expand the capabilities of telemedicine, allowing for remote assessment and monitoring of bone age, and further improving accessibility to healthcare services.

REFERENCES

[1] M. Phillip, O. Moran, and L. Lazar, "Growth without growth hormone," *J. Pediatr. Endocrinol. Metab.*, vol. 15, pp. 1267–1272, Dec. 2002.

[2] U. Hägg and J. Taranger, "Skeletal stages of the hand and wrist as indicators of the pubertal growth spurt," *Acta Odontologica Scandinavica*, vol. 38, no. 3, pp. 187–200, Jan. 1980.

[3] W. A. Marshall, "Interrelationships of skeletal maturation, sexual development and somatic growth in man," *Ann. Hum. Biol.*, vol. 1, no. 1, pp. 29–40, Jan. 1974.

[4] R. Vanderwilde, L. T. Staheli, D. E. Chew, and V. Malagon, "Measurements on radiographs of the foot in normal infants and children," *J. Bone Joint. Surg. Amer.*, vol. 70, no. 3, pp. 407–415, Mar. 1988.

[5] N. Bayley and S. R. Pinneau, "Tables for predicting adult height from skeletal age. Revised for use with the Greulich–Pyle hand standards," *J. Pediatr.*, vol. 40, no. 4, pp. 423–441, 1952.

[6] L. L. Morris, "Assessment of skeletal maturity and prediction of adult height (TW2 method)," *Amer. J. Hum. Biol.*, vol. 14, no. 6, pp. 788–789, Oct. 1976.

[7] Y. Wu, S. Zhao, Z. Xing, Z. Wei, Y. Li, and Y. Li, "Detection of foreign objects intrusion into transmission lines using diverse generation model," *IEEE Trans. Power Del.*, vol. 38, no. 5, pp. 3551–3560, Oct. 2023.

[8] Z. Xing, S. Zhao, W. Guo, F. Meng, X. Guo, S. Wang, and H. He, "Coal resources under carbon peak: Segmentation of massive laser point clouds for coal mining in underground dusty environments using integrated graph deep learning model," *Energy*, vol. 285, Dec. 2023, Art. no. 128771.

[9] S. J. Lewis, Z. Gandomkar, and P. C. Brennan, "Artificial intelligence in medical imaging practice: Looking to the future," *J. Med. Radiat. Sci.*, vol. 66, no. 4, pp. 292–295, Nov. 2019.

[10] V. Gilsanz and O. Ratib, *Hand Bone Age: A Digital Atlas of Skeletal Maturity*. Berlin, Germany: Springer, 2005.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 115–117, Mar. 2015.

[12] A. Amirkhani, A. Khosravian, M. Masih-Tehrani, and H. Kashiani, "Robust semantic segmentation with multi-teacher knowledge distillation," *IEEE Access*, vol. 9, pp. 119049–119066, 2021.

[13] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[14] B. Fischer, P. Welter, R. W. Günther, and T. M. Deserno, "Web-based bone age assessment by content-based image retrieval for case-based reasoning," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 7, no. 3, pp. 389–399, May 2012.

[15] S. Aja-Fernández, R. de Luis-Garcia, M. Á. Martin-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.

[16] 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.

[17] K. Somkantha, N. Theera-Umpon, and S. Auephanwiriyakul, "Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression," *J. Digit. Imag.*, vol. 24, no. 6, pp. 1044–1058, Feb. 2011.

[18] 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, Feb. 2019.

[19] M. Chen, "Automated bone age classification with deep neural networks," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2016.

[20] 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, Mar. 2017.

[21] Z. Bian and R. Zhang, "Bone age assessment method based on deep convolutional neural network," in *Proc. 8th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jun. 2018, pp. 194–197.

[22] K. Mao, W. Lu, K. Wu, J. Mao, and G. Dai, "Bone age assessment method based on fine-grained image classification using multiple regions of interest," *Syst. Sci. Control Eng.*, vol. 10, no. 1, pp. 15–23, Dec. 2022.

[23] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, Apr. 2018.

[24] T. D. Bui, J.-J. Lee, and J. Shin, "Incorporated region detection and classification using deep convolutional networks for bone age assessment," *Artif. Intell. Med.*, vol. 97, pp. 1–8, Jun. 2019.

[25] Y. Gao, T. Zhu, and X. Xu, "Bone age assessment based on deep convolution neural network incorporated with segmentation," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 15, no. 12, pp. 1951–1962, Sep. 2020.

[26] 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.

[27] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets, "Paediatric bone age assessment using deep convolutional neural networks," in *Proc. Int. Workshop Deep Learning Med. Image Anal.* (Lecture Notes in Computer Science), vol. 11045, Sep. 2018, pp. 300–308.

[28] S. Koitka, M. S. Kim, M. Qu, A. Fischer, C. M. Friedrich, and F. Nensa, "Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101743.

[29] Q. H. Nguyen, B. P. Nguyen, M. T. Nguyen, M. C. H. Chua, T. T. Do, and N. Nghiem, "Bone age assessment and sex determination using transfer learning," *Exp. Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 116926.

[30] M. Escobar, C. González, F. Torres, L. Daza, G. Triana, and P. Arbeláez, "Hand pose estimation for pediatric bone age assessment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 11769, Oct. 2019, pp. 531–539.

[31] Y. Ji, H. Chen, D. Lin, X. Wu, and D. Lin, "PRSNet: Part relation and selection network for bone age assessment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 11769, Oct. 2019, pp. 413–421.

[32] 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.

[33] S. Wang, S. Jin, K. Xu, J. She, J. Fan, M. He, L. S. Stephen, Z. Gao, X. Liu, and K. Yao, "A pediatric bone age assessment method for hand bone X-ray images based on dual-path network," *Neural Comput. Appl.*, pp. 1–16, Oct. 2023, doi: 10.1007/s00521-023-09098-4.

[34] X. Mao, Q. Hui, S. Zhu, W. Du, C. Qiu, X. Ouyang, and D. Kong, "Automated skeletal bone age assessment with two-stage convolutional transformer network based on X-ray images," *Diagnostics*, vol. 13, no. 11, p. 1837, May 2023.

[35] K. Jian, S. Li, M. Yang, S. Wang, and C. Song, "Multi-characteristic reinforcement of horizontally integrated TENet based on wrist bone development criteria for pediatric bone age assessment," *Int. J. Speech Technol.*, vol. 53, no. 19, pp. 22743–22752, Jul. 2023.

[36] Z. Li, W. Chen, Y. Ju, Y. Chen, Z. Hou, X. Li, and Y. Jiang, "Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction," *Front. Artif. Intell.*, vol. 6, Mar. 2023, Art. no. 1142895.

[37] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1247–1250.

[38] Z. Zeng, W. Zhao, P. Qian, Y. Zhou, Z. Zhao, C. Chen, and C. Guan, "Robust traffic prediction from spatial–temporal data based on conditional distribution learning," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13458–13471, Dec. 2022.

[39] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.-H. Yang, and Y. Zhuang, "Data-dependent label distribution learning for age estimation," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3846–3858, Aug. 2017.

[40] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[41] C. Chen, Z. Chen, X. Jin, L. Li, W. Speier, and C. W. Arnold, "Attention-guided discriminative region localization and label distribution learning for bone age assessment," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1208–1218, Mar. 2022.

[42] H. Tang, X. Pei, X. Li, H. Tong, X. Li, and S. Huang, "End-to-end multi-domain neural networks with explicit dropout for automated bone age assessment," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3736–3749, Feb. 2023.

[43] Z.-Q. Liu, Z.-J. Hu, T.-Q. Wu, G.-X. Ye, Y.-L. Tang, Z.-H. Zeng, Z.-M. Ouyang, and Y.-Z. Li, "Bone age recognition based on mask R-CNN using xception regression model," *Frontiers Physiol.*, vol. 14, Feb. 2023, Art. no. 1062034.

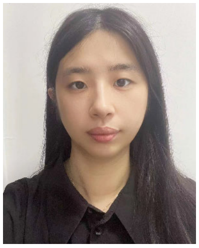
[44] C. Ozdemir, M. A. Gedik, and Y. Kaya, "Age estimation from left-hand radiographs with deep learning methods," *Traitement Du Signal*, vol. 38, no. 6, pp. 1565–1574, Dec. 2021.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

HAO-DONG ZHENG received the B.S. degree from the College of Computer Science and Technology, Nanyang Normal University, China, in 2017. He is currently pursuing the M.S. degree with the College of Computer and Information Science, Chongqing Normal University, China. His current research interests include medical image analysis, computer vision, and knowledge distillation.

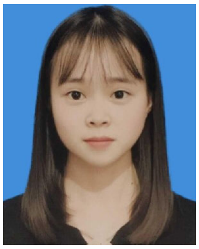
LEI YU received the B.S. and M.S. degrees in information and computing science and the Ph.D. degree in computer science from Chongqing University, China, in 2003, 2006, and 2009, respectively.

From January 2019 to January 2020, he was selected by the Government as a Visiting Scholar with the University of Alberta. He is currently a Professor with the College of Computer and Information Science, Chongqing Normal University. His research interests include pattern recognition, image processing, and machine learning. He is a member of the China Computer Federation (CCF) and the Chinese Association for Artificial Intelligence (CAAI).



YU-TING LU received the B.S. degree in information and computing science from Yangzhou University, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Big Data and Software Engineering, Chongqing University, China. Her current research interests include medical image analysis, computer vision, and machine learning.

WEI-HAO ZHANG received the B.S. degree from the College of Electronic and Information Engineering, Anhui Jianzhu University, China, in 2017. He is currently pursuing the M.S. degree with the College of Computer and Information Science, Chongqing Normal University, China. His current research interests include medical image processing, image fusion, and computer vision.



YAN-JUN YU received the bachelor's degree in computer science and technology from Huainan Normal University, in 2017. She is currently pursuing the master's degree in computer technology with Chongqing Normal University. Her research interest includes bone age prediction.