Postprint: Pediatric Bone Age Assessment Using Dual Attention Network Fusion

Authors: Zhang Xin, Zhang Junhua, Zhang Shuai

Date: 2022-05-18T16:08:25Z

Abstract

Bone age assessment is a commonly used method for detecting endocrine and growth abnormalities in children; however, low-quality hand X-ray images in deep learning methods reduce the final assessment accuracy. To address this issue, we propose an alignment network that increases the region of interest (ROI) area in hand X-ray images. This network employs the Swin Transformer structure as its backbone to learn hand similarity from images and obtain affine coefficients, and does not require large-scale hand annotations during training. In the bone age assessment network, we propose improvements to the Efficient Channel Attention and Spatial Attention mechanisms, specifically a Dual-Pooling Efficient Channel Attention and an Asymmetric Convolution Spatial Attention method. These two methods are combined with the Xception network in a dual-attention manner to propose DA-Xception. Tested on the RSNA dataset, this bone age assessment method achieves a mean absolute error of 5.37 months, and compared with other deep learning methods, it can extract features more comprehensively and optimize assessment results.

Full Text

Preamble

Vol. 39 No. 10
Application Research of Computers
ChinaXiv Partner Journal

Pediatric Bone Age Assessment Method Combined with Dual Attention Network

Zhang Xin, Zhang Junhua†, Zhang Shuai
(School of Information Science & Engineering, Yunnan University, Kunming

650500, China)

Abstract: Bone age assessment is a common method for detecting endocrine and growth abnormalities in children. However, low-quality hand X-ray images in deep learning methods reduce final evaluation accuracy. To address this problem, this paper proposes an alignment network that increases the region of interest (RoI) area in hand X-ray images. This network uses the Swin Transformer structure as the backbone to learn hand similarity and obtain affine coefficients, and does not require large-scale hand annotation during training. In the bone age assessment network, we propose improvements to efficient channel attention and spatial attention mechanisms: dual-pooling efficient channel attention and asymmetric convolution spatial attention. Combining these two methods with the Xception network in a dual-attention form, we propose DA-Xception. Tested on the RSNA dataset, this bone age assessment method achieves a mean absolute error of 5.37 months. Compared with other deep learning methods, it can extract features more fully and optimize evaluation results.

Keywords: bone age assessment; X-ray image alignment; dual attention; deep learning

0 Introduction

Human developmental age can be divided into chronological age and biological age, with biological age more objectively reflecting actual human growth and development. The primary basis for biological age is the degree of maturation of hand bones [?] and tooth growth [?].

Hand bone age assessment is widely used in modern pediatric clinical diagnosis. Physicians analyze X-ray images of the subject’ s non-dominant hand to determine the corresponding bone age. By comparing this with chronological age, they can assess children’ s growth potential and skeletal maturity. Additionally, bone age assessment provides a reference for children’ s height development [?].

Current traditional bone age assessment methods include the atlas method and scoring method. The atlas method compares a subject’ s hand X-ray image with a standard atlas, using the label of the closest matching standard image as the subject’ s bone age. The common Greulich-Pyle atlas method [?] has an average error of 11.5 months [?]. The scoring method evaluates several representative bones in the hand X-ray separately, then calculates a total score and converts it to the corresponding bone age using a formula. Common scoring methods include the TW scoring method [?] and the China-05 scoring method [?]. However, both methods have obvious drawbacks: the atlas method’ s results are subjectively influenced by the evaluator, leading to large errors, while the scoring method is time-consuming and inefficient due to the need to score wrist bones, epiphyses, and other hand regions separately.

With the rise of computer vision technology, automated bone age assessment methods have developed. Early automated methods automatically extracted features used in manual assessment. For example, Thodberger et al. [?] designed the BoneXpert system as commercial bone age assessment software, which analyzes the region of interest (RoI) in the hand and scores the RoI area to obtain bone age results.

In recent years, Spampinato et al. [?] adopted deep learning methods for automated bone age assessment, proposing an end-to-end BoNet system built with convolutional neural networks, achieving a final error of 9.5 months. Since then, various deep learning-based bone age assessment methods have been proposed [?], with many evaluated on the publicly available hand X-ray image dataset provided by the Radiological Society of North America.

In 2019, Wu et al. [?] first used a Mask R-CNN network to segment hand X-ray images, removing interfering noise, then employed a residual attention subnet (RAS) for bone age regression, achieving a final mean error of 7.38 months. However, this method required extensive manual annotation of hand regions during image segmentation, affecting assessment efficiency. Liu et al. [?] proposed the VGG-U-Net network for hand segmentation, replacing U-Net [?] downsampling layers with a VGG16 [?] pretrained network model to improve segmentation accuracy on small-sample datasets, subsequently leveraging inter-label correlations to achieve 6.05 months error on the dataset.

In 2020, Hao et al. [?] proposed the OCNet bone age multi-classification method instead of regression, and based on the continuity of hand development, obtained three different bone age range values through the bone age assessment model, finally calculating the overlapping range to determine bone age, achieving a final error of 5.84 months. In 2021, He et al. [?] first performed lossless compression of hand X-ray images, then used an SE-ResNet network to extract features for bone age assessment, achieving a mean absolute error of 6.04 months. This method increased the hand proportion in input images but did not process interfering information, leaving image quality unimproved. Salim et al. [?] added a ridge regression layer to the bone age assessment model on the basis of image segmentation, proposing the ridge regression network (RRNet), which achieved an absolute error of 6.38 months.

From these methods, we learn that the quality of hand X-ray images affects final bone age assessment accuracy, while processing large-scale low-quality image segmentation reduces assessment efficiency. To address this problem, this paper processes dataset images to reduce image noise and unify contrast. Given that different images have varying hand RoI, we introduce an alignment network to make original dataset images have consistent hand structure with standard hand X-ray images, reducing the impact of different hand sizes and angles in the dataset. To strengthen the feature extraction capability of the bone age assessment network, we design a dual-attention Xception network (DA-Xception) in bone age regression assessment. This network learns hand RoI in image space and channels through parallel dual branches, finally performing feature fusion regression to obtain bone age assessment results.

Compared with current bone age assessment methods, our main contributions are: 1) Introducing a hand image alignment network that, based on hand structure similarity, ensures consistent hand RoI regions in bone age assessment. Compared with image segmentation, hand alignment does not require large-scale image annotation and increases the effective RoI area in bone age assessment. 2) Designing the DA-Xception network in the bone age regression network, proposing dual-pooling efficient channel attention (DPECA) to strengthen overall and texture features in image channels. 3) Proposing asymmetric convolution spatial attention (ACSA) to extract fine-grained features in image space. Dataset experiments demonstrate that our method outperforms other bone age assessment methods.

1 Research Methods

Our bone age assessment method consists of two main backbone networks, as shown in Figure 1. The first is a hand alignment network that improves original image quality and calibrates hand position, followed by a bone age regression network that extracts hand RoI information. Inspired by the excellent performance of Vision Transformer (ViT) [?] networks in computer vision tasks in recent years, we introduce the Swin Transformer network [?] as the backbone in the alignment network to extract features. The network then reduces image feature dimensions and connects to a fully connected layer to finally obtain the affine relationship between the original input image and the standard image. Using affine coefficients, the original image is transformed to have standard hand structure, reducing the impact of hand size and angle variations in X-ray images on final assessment results. In the regression network, hand images sized 299×299 are input into the DA-Xception network to extract hand RoI feature information. After global average pooling reduces channel dimensions, features are fused with encoded gender features, then connected to a fully connected layer with 512 neurons and a dropout layer. Finally, a single-neuron fully connected layer regresses to obtain the image bone age result.

1.1 Hand Alignment Network

Hand regions in original X-ray images from bone age datasets typically exhibit rotation, translation, and scaling variations, which also exist in other X-ray datasets [?]. To address these issues, we introduce a hand alignment network to reduce morphological variations of hands in original images and enhance feature extraction capability in subsequent bone age regression. The alignment network workflow is shown in Figure 2. First, original images are input into the Swin Transformer network to extract hand features for training, with standard images as labels, outputting image affine coefficients. The original input images are then aligned to standard images through affine transformation.

For a given input image I and standard image T, the alignment network obtains affine coefficients ϕ such that the affine-transformed image $\phi(I)$ has a standard hand structure similar to the standard image. During network training, the structural loss can be defined as $L_S = f(\phi(I), T)$.

Hand X-ray images input to the alignment network yield affine coefficients ϕ consisting of five parameters: $\phi = (t_x, t_y, s_x, s_y, \theta)$. Here, t_x and t_y represent displacement in horizontal and vertical directions, s_x and s_y represent scaling in horizontal and vertical directions, and θ is the rotation angle. The affine transformation relationship for original images using these coefficients is shown in Equation (1):

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} s_x \cos\theta & -s_y \sin\theta & t_x \\ s_x \sin\theta & s_y \cos\theta & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

where G represents image rasterization and B represents bilinear interpolation, both reducing feature information loss during affine transformation.

Convolutional Neural Networks (CNNs) extract image features hierarchically, enhancing feature extraction capability and reducing network parameters through local connectivity and parameter sharing, but they lack attention to global image information. ViT networks introduce multi-head attention from Natural Language Processing (NLP) to compensate for this deficiency. However, due to differences between natural language and image input scales, ViT networks have high computational complexity. The Swin Transformer network uses different sampling values hierarchically to reduce computational load and proposes a shifted window method to allow cross-window connections of local features, improving efficiency and increasing network receptive field.

Our Swin Transformer backbone network consists of 4 stages, each comprising 2, 2, 6, and 2 Swin Transformer Blocks respectively. Figure 3 shows two successive basic blocks, containing shifted window Multi-head Self Attention (MSA), Multilayer Perceptron (MLP), and LayerNorm (LN) layers.

We introduce cosine similarity as the structural loss for the alignment network. Cosine similarity between images converts two images into vectors and calculates the cosine of the angle between vectors as the image similarity relationship. Therefore, the alignment network loss function is shown in Equation (2):

$$L_S = 1 - \frac{\sum_{i=1}^{n} I_i T_i}{\sqrt{\sum_{i=1}^{n} I_i^2} \sqrt{\sum_{i=1}^{n} T_i^2}}$$

where I is the original image vector and T is the standard image vector.

1.2 Bone Age Regression Network

Our bone age regression network is DA-Xception, proposed by combining the Xception convolutional neural network [**?**] with a dual attention mechanism.

The original Xception structure is shown in Figure 4.

1.2.1 Dual Attention Xception The original Xception network model consists of multiple depthwise separable convolutional layers. Except for the network’ s beginning and end, all other convolutional modules use linear residual connections. According to the processing order of data input to Xception, the network structure in Figure 4 is mainly divided into three modules: entry flow, middle flow, and exit flow. Compared with the Inception V3 network [?], both have similar parameter counts, but Xception uses depthwise separable convolutions to reduce computational load. This architecture uses model parameters more effectively and achieves better feature extraction performance.

Although the original Xception structure has residual connection modules, it is overall a linear neural network where the model can only extract features sequentially. This approach cannot fully extract channel and spatial information from features. Fu et al. [?] proposed Dual Attention Network (DA-Net) to strengthen feature correlations in semantic segmentation, using parallel spatial and channel attention structures to simultaneously extract features in image channels and positions. Lin et al. [?] proposed Bilinear models for image classification, dividing the network into two branches and using different branches to extract image features in parallel, finally using bilinear pooling to fuse features from both branches and output classification results. Liu et al. [?] proposed Center Boundary Dual Attention Network (CBDA-Net) for object detection in remote sensing images, generating center region attention and boundary region attention in a dual structure to eliminate background noise interference and detect key objects.

Inspired by these works and combining Xception network structure characteristics, we designed the DA-Xception network shown in Figure 5.

In DA-Xception, the original Xception network’ s middle flow structure, which repeats 8 linear residual blocks, is replaced with a bilinear structure where left and right branches each repeat 4 times. DPECA and ACSA modules are added multiple times to residual blocks in each branch. In the left branch, DPECA strengthens feature information between image channels, while the right branch learns spatially correlated features. The bilinear structure in middle flow extracts channel and spatial feature information in parallel, with features from both branches fused in the exit flow structure before connecting to a fully connected layer to obtain bone age assessment results.

Compared with DA-Net, which adds dual attention only once before the output structure, our proposed DA-Xception network adds channel and spatial attention mechanisms multiple times on dual branches in middle flow, guiding the two branches to extract channel and spatial features separately. Bilinear models [?] use dual-branch learning, but both branches have identical structures, which may cause redundancy during final feature fusion. Our proposed bilinear fusion method effectively extracts hand region features during bone age assessment and eliminates background noise interference.

1.2.2 Dual Pooling Efficient Channel Attention Attention mechanisms increase the weight of feature tensors in regions of interest during training while reducing the weight of meaningless backgrounds to improve network feature extraction capability.

Wang [**?**] proposed the Efficient Channel Attention (ECA) module shown in Figure 6(a), which avoids reducing channel dimensions when focusing on inter-channel attention and captures cross-channel interactions in a lightweight manner.

However, ECA only performs Global Average Pooling (GAP) on input features, extracting only overall background information while losing texture feature extraction. Based on this, we propose the DPECA structure shown in Figure 6(b), performing both GAP and Global Maximum Pooling (GMP) on input features to strengthen the bone age assessment network' s ability to fully extract both overall and texture features from hand regions in images.

For DPECA input features $F_{in} \in \mathbb{R}^{W \times H \times C}$, the image features after GAP and GMP are defined as F_{gap} and F_{gmp}, with corresponding formulas (3) and (4):

$$F_{gap} = \frac{1}{W \times H} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} F_{in}(i, j)$$

$$F_{gmp} = \max_{i,j} F_{in}(i, j)$$

where $\max(\cdot)$ represents taking the maximum value in the corresponding channel of the feature map, and W, H, and C represent the length, height, and number of channels of input features respectively. The pooled features are concatenated and pass through the same 1D convolutional layer to share weight parameters, as shown in Equation (5):

$$M = \sigma(\text{Conv1D}([F_{gap}, F_{gmp}]))$$

where Conv1D represents 1D convolution operation and σ is the HSigmoid nonlinear activation function. HSigmoid uses piecewise fitting to implement the Sigmoid function but computes faster during network training, with the specific formula shown in Equation (6):

$$\text{HSigmoid}(x) = \begin{cases} 0, & x \leq -3 \\ 0.5 \times x + 1.5, & -3 < x < 3 \\ 1, & x \geq 3 \end{cases}$$

Finally, the two features are aggregated through a fully connected layer to generate inter-channel attention relationships.

To achieve appropriate cross-channel interactions, the convolution kernel size k is adaptively selected based on the number of channels C in the feature map, with their relationship shown in Equation (7):

$$k = |\psi(C)|_{\text{odd}} = |\log_2(C) + \gamma|_{\text{odd}}$$

where $|\cdot|_{\text{odd}}$ represents taking the closest odd number to the operation result.

1.2.3 Asymmetric Convolution Spatial Attention Sanghyun proposed the Convolutional Block Attention Module (CBAM) [**?**]. Based on the Spatial Attention (SA) in CBAM structure, we propose the ACSA structure shown in Figure 7.

In ACSA, average pooling and max pooling operations are first performed on image features to aggregate spatial information, generating two weight matrices $F_{avg} \in \mathbb{R}^{W\times H\times 1}$ and $F_{max} \in \mathbb{R}^{W\times H\times 1}$. Then, 1×7 and 7×1 convolutional kernel groups are used instead of the 7×7 convolutional kernel in CBAM, maintaining the same receptive field while extracting fine-grained information in image space and reducing computational load. The two weights are combined to obtain spatial attention $M_s(F)$, with specific calculation shown in Equation (8):

$$M_s(F) = \sigma(f_{7\times 7}([\text{AvgPool}(F); \text{MaxPool}(F)]))$$

where $f_{7\times 7}$ represents the asymmetric convolutional kernel group. This spatial attention mechanism selectively aggregates spatial feature similarity with weighted attention. Attention weights for features at different spatial positions are determined by feature similarity between two positions, with similar features showing stronger correlation.

1.3 Model Loss Function and Evaluation Metrics

Bone age assessment is a regression task where the model’ s final output is a specific real value. Therefore, we select Root Mean Square Error (RMSE) as the loss function, with calculation formula shown in Equation (9):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where N is the number of samples, $\hat{y}_i$ is the model’ s predicted bone age result, and y_i is the corresponding ground truth annotation. As shown in Equation (9), as the RMSE value decreases, the model optimizes assessment results.

The RMSE loss function exhibits nonlinear loss value reduction in regression compared to Mean Absolute Error (MAE) loss. When the loss is large, the network model' s gradient descent is fast, enabling rapid convergence. When the loss is small, RMSE and MAE values are similar, and the network model reduces loss linearly.

In bone age assessment, MAE is used as the metric, as shown in Equation (10):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

2 Experimental Results and Analysis

2.1 Dataset

Our bone age assessment dataset is taken from the publicly available dataset of the 2017 Pediatric Bone Age Challenge competition held by the Radiological Society of North America (RSNA). The dataset contains 12,811 images, including 6,933 male and 5,878 female hand bone X-ray images. Each image corresponds to skeletal age divided with monthly precision. The bone age distribution in the RSNA dataset is shown in Figure 8.

During bone age assessment result optimization, the training set is used for model weight update and optimization, the validation set monitors the training process and provides real-time performance feedback, and the test set provides final evaluation of model generalization capability. In our experiments, 800 images were randomly selected from the dataset for validation, 200 images for testing, and the remaining images for network model training.

2.2 Experimental Configuration

All network model training was completed on hardware with an Intel(R) Core(TM) i7-10700KF CPU, NVIDIA GeForce RTX 3070 GPU, and 4×8GB memory. The software environment used TensorFlow 2.5.0 as the deep learning framework and Keras 2.6.0 API for model training, with Python 3.8.12 as the programming language. We used the Adam [**?**] optimizer with an initial learning rate of 0.001, training the model for 100 epochs. During training, we monitored validation loss and reduced the learning rate if no loss optimization appeared for multiple epochs. Batch size was set to 16, network input image size was 299×299, RMSE was used as the network model loss, and MAE was used as the evaluation metric to measure the gap between bone age assessment values and ground truth.

2.3 Image Preprocessing

Due to differences in acquisition equipment and exposure methods, hand X-ray images have varying resolutions and non-uniform grayscale distributions, leading to larger assessment errors. We performed contrast unification and denoising on X-ray images. First, histogram equalization was applied to adjust image grayscale distribution to an appropriate range, enhancing local contrast without affecting overall contrast. Then, adaptive gamma transformation was applied to stretch contrast, increasing low-brightness pixel values and suppressing high-brightness pixels. To eliminate noise interference, bilateral filtering was used for smoothing, as shown in Figure 9.

2.4 Image Alignment Results

Dataset images were affine-transformed after alignment to obtain standard images, as shown in Figure 10. Figure 10(a) shows that original images have differences in hand position and size, causing inconsistent hand RoI regions in each image. Figure 10(b) shows hand X-ray images after alignment. It can be seen that aligned images adjust hand tilt angles and hand region proportions in images, making hand regions of interest more obvious and consistent, reducing interference from erroneous information in subsequent bone age regression, and enabling more effective extraction of hand features.

2.5 Baseline Network Comparison

To select an appropriate network structure for bone age assessment, we chose five classic network structures: EfficientNetB4, ResNet101, DenseNet201, Inception ResNet V2, and Xception for bone age assessment and observed evaluation metrics. In baseline network evaluation, no processing was applied to network structures or datasets. Images were uniformly resized to 299×299 and input into the five networks, yielding the bone age assessment results shown in Table 1. Using the original Xception network achieved the best performance among the five networks with a mean absolute error of 7.41 months and moderate parameter count. Therefore, subsequent bone age assessment work selected the Xception network structure for improvement to optimize final regression accuracy.

Table 1. MAE and Parameters of Different Baseline Networks

Network	MAE (months)	Parameters (10^7)
ResNet101		
DenseNet201		
EfficientNetB4		
Inception ResNet V2		
Xception	7.41	

2.6 Ablation and Gender Experiments

To verify the effectiveness of our bone age assessment method, we conducted ablation experiments on various modules in the assessment process to measure the role of different structures. First, we verified the effectiveness of different network structure modules; then we explored and compared our dual attention method with other attention improvement mechanisms; finally, we investigated the impact of gender factors on bone age regression results.

2.6.1 Results of Different Module Ablation Experiments Our bone age assessment work mainly consists of three parts: image preprocessing, Swin Transformer network alignment, and DA-Xception network for bone age regression. Ablation experiments were conducted on these three modules, with comparison methods including: 1) Using only Xception network for bone age regression; 2) Adding image preprocessing; 3) Adding image alignment; 4) Adding DPECA and ACSA modules to the regression network. The bone age assessment accuracy of each experiment is shown in Table 2.

Table 2. MAE of Ablation Experiments

				Dual	MAE
Method	Xception	Preprocessing	Alignment	Attention	(months)

In Table 2, the mean absolute error after image preprocessing is 6.67 months. After further adding image alignment, the bone age assessment error decreases to 5.72 months. Finally, feeding aligned images into the DA-Xception network structure combining Xception network and dual attention yields a final error result of 5.37 months. The three modules reduce error by 0.74 months, 0.95 months, and 0.35 months respectively. Therefore, in the bone age assessment process, image preprocessing effectively reduces noise in original images, aligning hand images of different angles and sizes increases the proportion of effective hand regions in images and ensures consistent hand RoI. Adding dual attention to the Xception network structure enables the network to focus on richer key features in images, improving final bone age assessment accuracy.

Neural network iteration reflects model performance. We selected the optimal model with a bone age assessment error of 5.37 months and plotted its training process curves in Figure 11, where the blue curve represents training set MAE and the red curve represents validation set MAE. It can be seen that as the number of epochs increases, the mean absolute error values for both training and validation sets continuously decrease, slowing down after 20 epochs and gradually stabilizing. When training reaches 100 epochs, the training set loss curve decays slowly, indicating the network has fully extracted effective features and bone age regression results have stabilized.

2.6.2 Dual Attention Effectiveness Validation Experiment To verify the effectiveness of our proposed dual attention mechanism, this subsection compares our method with other attention improvement mechanisms, as shown in Table 3.

Table 3. Results of Different Attention Mechanism Methods

Methods	MAE (months)
Xception	
Xception + Bilinear	
Xception + DA-Net	
Xception + ECA + SA	
Xception + DPECA + ACSA (Ours)	

It can be seen that in the second experiment, the bilinear structure has insufficient attention to hand regions in bone age assessment, thus increasing the error compared to the original Xception network structure. The third experiment combines Xception network with DA-Net, and its heatmap shows that the network extracts not only hand region features but also hand edge background information, with background features reducing bone age assessment results.

The fourth and fifth experiments compare original dual attention with our improved dual attention in the Xception network structure. In Figure 12, both network structures show similar attention regions, but the fifth group focuses on larger hand region areas and more critical regions. The final bone age regression results show that our proposed method reduces error by 0.28 months. Therefore, using DPECA and ACSA modules can extract hand features more effectively.

2.6.3 Gender Factor Comparison Experiment In human growth and development, male and female hand development maturity differs at the same age. The impact of gender factors on bone age assessment results is shown in Figure 13.

The experiment is divided into four parts: 1) Bone age assessment on male hand X-ray images alone from the RSNA dataset; 2) Bone age assessment on female images alone; 3) Bone age assessment after removing gender information from the RSNA dataset; 4) Bone age assessment with added gender information. The mean absolute errors for male-only and female-only assessments are 5.43 months and 5.65 months respectively. The error without gender information is 6.52 months, while adding gender information yields an MAE of 5.37 months. Compared with no gender information, single-gender assessment reduces errors by 1.09 months and 0.87 months respectively, while adding gender information reduces error by 1.15 months. Therefore, adding gender information in bone age assessment can effectively reduce error values and improve regression accuracy.

2.7 Comparative Analysis of Different Deep Learning Methods

To better demonstrate the advancement of our method in bone age assessment, we compared it with other recent bone age assessment methods. Table 4 shows the mean absolute error values of bone age assessment under different methods.

Table 4. Results of Different Bone Age Assessment Methods

Methods	MAE (months)
VGG16 [?]	9.97
RAS [?]	7.38
RRNet [?]	6.39
Ranking CNN [?]	6.05
SE-ResNet [?]	6.04
OCNet [?]	5.84
DA-Xception (ours)	5.37

In Table 4, reference [?] divides bone age assessment into two stages: first segmenting images to remove background interference from hand X-ray images, then using VGG16 network for bone age regression, achieving a final regression error of 9.97 months. References [?, ?] adopt the same approach of first segmenting hand regions to remove interfering label information, then inputting segmented images into bone age regression networks, achieving final regression errors of 7.38 months, 6.05 months, and 6.39 months respectively. Image segmentation methods effectively remove background interference and optimize bone age assessment results but require extensive manual annotation of hand regions in dataset images. Compared with these segmentation methods, our proposed image alignment method only requires weak annotation of a small number of images, using the alignment network to automatically extract hand RoI features and make hand RoI regions consistent across images. Additionally, our image preprocessing work can suppress noise interference in images.

Reference [?] proposes adding a lossless compression module to the bone age assessment network to maintain stable image quality when reducing image size, then inputs images into a bone age regression network combining SE attention mechanism and ResNet network, achieving a final bone age assessment error of 6.04 months. Reference [?] adds residual attention to make the network focus more on RoI regions. Compared with single attention mechanisms, we adopt a dual-branch approach with parallel channel and spatial attention mechanisms, achieving a bone age assessment error of 5.37 months, superior to other methods in Table 3, further improving assessment accuracy.

In summary, our method in bone age assessment requires only lightweight image annotation to complete alignment work, without excessive manual processing, making it more clinically feasible. Moreover, using a dual attention structure for bone age regression can extract hand feature information more fully than linear networks with attention mechanisms, further improving accuracy and reducing bone age assessment error.

3 Conclusion

Addressing the issues of low-quality hand X-ray images and variations in hand region size and angle in current bone age assessment, this paper improves X-ray image quality, aligns hand images, and innovatively proposes the DA-Xception network that combines Xception network with two attention mechanisms through a dual-branch parallel structure.

Our bone age assessment method consists of two parts: first, preprocessing X-ray images to unify contrast and brightness, and using Swin Transformer network to extract features to align hand region of interest in X-ray images; second, using the DA-Xception network to extract hand RoI features for bone age regression to obtain assessment results. Experiments demonstrate that our method can effectively reduce the impact of image quality on assessment results. Compared with current methods, it achieves higher precision and provides important reference value for subsequent bone age assessment work, better helping prevent adolescent growth and development diseases.

References

[1] Fishman L S. Radiographic evaluation of skeletal maturation: a clinically oriented method based on hand-wrist films [J]. The Angle Orthodontist, 1982, 52 (2): 88-112.

[2] Liversidge H M, Molleson T I. Developing permanent tooth length as an estimate of age [J]. Journal of Forensic Science, 1999, 44 (5): 917-920.

[3] Martin D D, Wit J M, Hochberg Z, et al. The use of bone age in clinical practice-part 1 [J]. Hormone Research in Paediatrics, 2011, 76 (1): 1-9.

[4] Bayer L M. Radiographic atlas of skeletal development of the hand and wrist [J]. California Medicine, 1959, 91 (1): 53.

[5] King D G, Steventon D M, Osullivan M P, et al. Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods [J]. The British Journal of Radiology, 1994, 67 (801): 848-851.

[6] Tanner J M, Oshman D, Bahhage F, et al. Tanner-Whitehouse bone age reference values for North American children [J]. Journal of Pediatrics, 1997, 131 (1): 34-40.

[7] Zhang Shaoyan, Liu Lijuan, Wu Zhenlie, et al. The skeletal development standards of hand and wrist for Chinese Children-China 05 I. TW3-CRUS, TW3-C Carpal, and RUS-CHN Methods [J]. Chinese Journal of Sports Medicine, 2006, 25 (5): 509-516.

[8] Thodberg H H, Kreibor S, Juul A, et al. The BoneXpert method for automated determination of skeletal maturity [J]. IEEE Trans on Medical Imaging, 2009, 28 (1): 52-66.

[9] Spampinato C, Palazzo S, Giordano D, et al. Deep learning for automated skeletal bone age assessment in X-ray images [J]. Medical Image Analysis, 2017, 36: 41-51.

[10] Igiovikov V I, Rakhlin A, Kalinin A A, et al. Paediatric bone age assessment using deep convolutional neural networks [M]. Cham: Springer, 2018: 300-308.

[11] Liang Baoyu, Zhai Yunkai, Tong Chao, et al. A deep automated skeletal bone age assessment model via region based convolutional neural network [J]. Future Generation Computer Systems, 2019, 98 (9): 54-59.

[12] Gao Yunyuan, Zhu Tao, Xu Xiaohua. Bone age assessment based on deep convolution neural network incorporated with segmentation [J]. International Journal of Computer Assisted Radiology and Surgery, 2020, 15 (12): 1951-1962.

[13] Wu, Eric, Kong Bin, Wang Xin, et al. Residual attention based network for hand bone age assessment [C]// Proc of the 16th International Symposium on Biomedical Imaging. New York: IEEE Access, 2019.

[14] Liu Bo, Zhang Yu, Chu Meicheng, et al. Bone age assessment based on Rank-Monotonicity enhanced ranking CNN [J]. IEEE Access, 2019, 7.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10). https://arxiv.org/pdf/1409.1556.pdf.

[16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[17] Hao Pengyi, Xie Xuhang, Han Tianxing, et al. Overlap classification mechanism for skeletal bone age assessment [C]// Proc of the 2nd ACM International Conference on Multimedia in Asia. New York: Association for Computing Machinery, 2021: 1-7.

[18] He Jin, Jiang Dan. Fully automatic model based on SE-ResNet for bone age assessment [J]. IEEE Access, 2021, 9: 62460-62466.

[19] Salim I, Hamza A B. Ridge regression neural network for pediatric bone age assessment [J]. Multimedia Tools and Applications, 2021: 1-18.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2021-06-03). https://arxiv.org/pdf/2010.11929.pdf.

[21] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: hierarchical vision transformer using shifted windows [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021.

[22] Liu Jingyu, Zhao Gangming, Fei Yu, et al. Align, attend and locate: chest X-Ray diagnosis via contrast induced attention network with limited supervision [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press. 2019: 10632-10641.

[23] Chollet F. Xception: deep learning with depthwise separable convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2017: 1251-1258.

[24] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2016: 2818-2826.

[25] Fu Jun, Liu Jing, Tian Haijie, et al. Dual attention network for scene segmentation [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press. 2019: 3146-3154.

[26] Lin T Y, Roy Chowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015.

[27] Liu Shuai, Zhang Lu, Lu Huchuan, et al. Center-boundary dual attention for oriented object detection in remote sensing images [J]. IEEE Trans on Geoscience and Remote Sensing, 2021, 60: 1-14.

[28] Wang Qilong, Wu Banggu, Zhu Pengfei, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 11534-11542.

[29] Woo S, Park J, Lee JY, et al. CBAM: convolutional block attention module [C]// Proc of the 15th European Conference on Computer Vision. Cham: Springer, 2018: 3-19.

[30] Kingma D, Jimmy B. Adam: a method for stochastic optimization [EB/OL]. (2017-01-30). https://arxiv.org/pdf/1412.6980.pdf.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.