Original Article

Bone age assessment by multi-granularity and multi-attention feature encoding

Bowen Liu[1], Yulin Huang[1], Shaowei Li[2], Jinshui He[2], Dongxu Zhang[1]

[1]State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University, Xiamen, China; [2]Department of Pediatrics, Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou, China

Contributions: (I) Conception and design: B Liu; (II) Administrative support: S Li, J He, D Zhang; (III) Provision of study materials or patients: S Li, J He; (IV) Collection and assembly of data: B Liu, Y Huang; (V) Data analysis and interpretation: B Liu, Y Huang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dongxu Zhang, PhD. State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University, No. 4221-117, Xiang'an South Road, Xiamen 361102, China. Email: zhangdongxu@xmu.edu.cn.

Background: Bone age assessment (BAA) is crucial for the diagnosis of growth disorders and the optimization of treatments. However, the random error caused by different observers' experiences and the low consistency of repeated assessments harms the quality of such assessments. Thus, automated assessment methods are needed.

Methods: Previous research has sought to design localization modules in a strongly or weakly supervised fashion to aggregate part regions to better recognize subtle differences. Conversely, we sought to efficiently deliver information between multi-granularity regions for fine-grained feature learning and to directly model long-distance relationships for global understanding. The proposed method has been named the "Multi-Granularity and Multi-Attention Net (2M-Net)". Specifically, we first applied the jigsaw method to generate related tasks emphasizing regions with different granularities, and we then trained the model on these tasks using a hierarchical sharing mechanism. In effect, the training signals from the extra tasks created as an inductive bias, enabling 2M-Net to discover task relatedness without the need for annotations. Next, the self-attention mechanism acted as a plug-and-play module to effectively enhance the feature representation capabilities. Finally, multi-scale features were applied for prediction.

Results: A public data set of 14,236 hand radiographs, provided by the Radiological Society of North America (RSNA), was used to develop and validate 2M-Net. In the public benchmark testing, the mean absolute error (MAE) between the bone age estimates of the model and of the reviewer was 3.98 months (3.89 months for males and 4.07 months for females).

Conclusions: By using the jigsaw method to construct a multi-task learning strategy and inserting the self-attention module for efficient global modeling, we established 2M-Net, which is comparable to the previous best method in terms of performance.

Keywords: Bone age assessment (BAA); computer-aided diagnosis; multi-task learning; self-attention

Submitted Jun 05, 2023. Accepted for publication Oct 27, 2023. Published online Feb 19, 2024.
doi: 10.21037/qims-23-806
View this article at: https://dx.doi.org/10.21037/qims-23-806

Introduction

Bone age assessment (BAA) of X-ray images is a common clinical examination used to examine the level of children's growth and development (1,2). Additionally, BAA can be used to reliably diagnose growth disorder diseases and ensure the rational use of pediatric somatropin-containing

medicines (3,4). Thus, achieving an accurate and objective BAA result is crucial in pediatric departments.

The Greulich-Pyle (G-P) atlas method is one of the most commonly employed assessment methods in clinical procedures. It considers the whole hand scan and compares the radiograph with a series of standard atlases (each representing a standard developmental pattern for a certain age range) (5). Due to its simplicity and speed, this approach is used by most radiologists. However, the reliability of this clinical method is greatly limited by intra-observer differences (which have been reported to range from 0.11 to 0.89 years) and inter-observer differences (which have been reported to range from 0.07 to 1.25 years) (6). This means that the significant role of objective and stable BAA in the diagnosis of growth disorders and optimization of treatments is negatively affected.

To meet the need for high-quality and objective BAAs and drive the development of novel machine-learning applications, the Radiological Society of North America (RSNA) launched the Pediatric Bone Age Machine Learning Challenge in 2017 and published a large-scale data set containing 14,236 images (hereafter referred to as the RSNA data set) (7). Using this data set, a large number of automated BAA methods have been proposed. The representative approaches are briefly described below, and the mean absolute error (MAE) is used as the uniform test metric (8).

Larson *et al.* (9) first established a regression model based on a convolutional neural network (CNN), sufficiently demonstrating the feasibility and superiority of deep learning for BAA. A MAE of 6.00 months was achieved, and this timeframe served as a performance benchmark in the above-mentioned challenge. A report on the challenge also outlined the top three solutions, with MAEs of 4.2, 4.4, and 4.4 months respectively (7). In the first solution, the Inception V3 network was used for feature extraction on the input gray image (input size: 500×500), and gender input (in which 0 and 1 correspond to female and male, respectively) was transformed into a 32-dimensional feature vector through the fully connected layer. Finally, the concatenation of the two feature vectors allowed the network to learn from images and gender information to accurately predict bone age. The data augmentation strategy was the key step in achieving better performance. In the final testing phase, ensemble prediction with multiple models of good performance also improved the final test results of BAA.

The second solution first designed a preprocessing pipeline for image enhancement, including image cropping, image size adjustment (to 560×560), and contrast enhancement. Next, 49 overlapping local regions (patches), 224×224 in size, were extracted from the processed image, and each patch was used as input to the deep regression network. The final prediction result was the median of all predictions. The deep network used in this solution was ResNet50, which was first pre-trained using the ImageNet data set and then fine-tuned on the bone age prediction task. Similar to the first solution, a data augmentation strategy and a model ensemble strategy (for nine models) were used to avoid overfitting and improve generalizability. In the third solution, the training set was divided into five parts, each used to train a model, and the poorest performing model was removed based on the results of the validation set. The final model prediction result was the average of multiple model predictions. The feature vector in the model comprised two parts: (I) the feature vector corresponding to image information; and (II) the feature vector corresponding to gender information. The deep neural network used in this model was designed to be lightweight, with a much smaller parameter count than that commonly used in large deep neural networks.

All of the methods mentioned above adopted appropriate data augmentation strategies. In addition, all of these methods used model ensemble strategies to avoid overfitting and improve model generalization ability. However, in practical clinical applications, the trade-off between speed and accuracy must be considered. For example, multiple large-scale deep neural networks were used for ensemble prediction in both the first and second solutions, but the sacrificed speed may not be proportional to the accuracy improvement obtained (8).

In addition to the solutions proposed in response to the challenge, many other representative deep-learning methods have been proposed. Spampinato *et al.* (10) first established an end-to-end regression model, which included a CNN and a multi-layer perceptron. Most importantly, a spatial transform network (STN) was inserted between the input image and the CNN to learn adaptive geometric transformations. The STN first learned a set of six-dimensional parameters and then used these parameters to perform geometric transformations on the input image, including translation, rotation, shearing, and scaling. This module automatically learned suitable image preprocessing operations in an end-to-end manner, reducing noise interference caused by hand posture, and helping improve model performance. The model was trained and validated on a data set containing 1,391 wrist X-ray images, and had

a MAE of 0.8 years. However, the STN in this method could only make slight corrections and could not provide enough discriminative information. Escobar *et al.* (11) used the Faster R-CNN bounding-box detection model and key-point estimation model to emphasize specific bones. The model did not crop specific regions further but instead input the original image and its key-point heatmap into a CNN to obtain prediction results. The model had a MAE of 4.14 months. To better characterize local changes and assist in diagnosis and understanding disease progress, Jia *et al.* (12) built a data set of 2,129 cases and a proposed method that incorporated image preprocessing, palm segmentation, fine-grained small region of interest detection using a recursive feature pyramid, maturity assessment of each bone through transfer learning, and bone age calculation based on the percentile curve of bone maturity. The MAE reached 0.61 years on the data set and had a labeling precision of 1 year. Ren *et al.* (13) first converted the original image into coarse-grained and fine-grained attention maps through an attention module, then concatenated them into two-channel inputs and used a deep regression network for prediction. The generation of coarse-grained attention maps was achieved by cropping the image using the Faster R-CNN, and the generation of fine-grained attention maps was obtained using Hessian filtering on the coarse-grained attention map. The model had a MAE of 5.2 months. Rather than using additional manual annotations, Chen *et al.* (14) obtained significant regions in a weakly supervised way. Specifically, a deep classification network was trained and combined with class activation mapping to obtain a saliency map. Next, a sub-saliency map was obtained via masking. After multiple iterations, the multiple salient regions could be obtained. Finally, the three saliency maps were merged into a three-channel input, and a deep regression network was used. The best MAE achieved on the test set was 4.7 months. Liu *et al.* (15) also introduced a novel weakly supervised mechanism to locate discriminative regions, and the MAE of the model was 3.99 months.

In summary, recognizing the subtle differences between images with similar hand structures is important in improving performance. The commonly used approach is to design a localization module to emphasize part regions in the further feature extraction phase. In this study, we used a novel multi-task learning approach to enable the network to focus on multi-granularity part regions without the need for the localization module. Additionally, the self-attention mechanism acted as a plug-and-play module to effectively enhance feature representation capabilities. Specifically, we used the jigsaw method to create related tasks emphasizing regions with different granularities and made these tasks help each other by using a hierarchical sharing mechanism. The scheme of the self-attention mechanism in this work consisted of capturing spatially aware non-local (NL) dependencies and reinforcing higher-level features via channel-wise negative correlations. The proposed solution was named the "Multi-Granularity and Multi-Attention Net (2M-Net)". The main contributions of this work are as follows:

(I) We replaced the localization module with the jigsaw method to efficiently emphasize key part regions with different granularities in a multi-task learning fashion.
(II) Using the self-attention mechanism, we directly modeled spatially aware long-range relationships between any positions in the whole hand-wrist structure to achieve a global understanding and reinforce higher-level features via channel-wise negative correlations for better representation. Notably, both of these calculations were achieved in a plug-and-play way.
(III) Without strong or weak annotations, 2M-Net was comparable to the previous best method in terms of its performance against the public benchmark.

Methods

In this section, we present the proposed 2M-Net for automated BAA. We begin with a summary of how 2M-Net is encouraged to perform discriminative feature learning and model long-range relationships between distant positions in the whole hand-wrist context. Next, a detailed illustration of 2M-Net and its optimization is introduced. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

The motivation of 2M-Net

We observed that regions that largely reflect skeletal development, such as the phalangeal epiphysis and carpal bones, have multiple granularities. These salient regions are not adjacent in the hand radiograph, and any final assessment requires a comprehensive analysis of these regions. Thus, we sought to encourage the network to effectively emphasize salient parts in the hand-wrist structure and efficiently model relationships between distant regions.

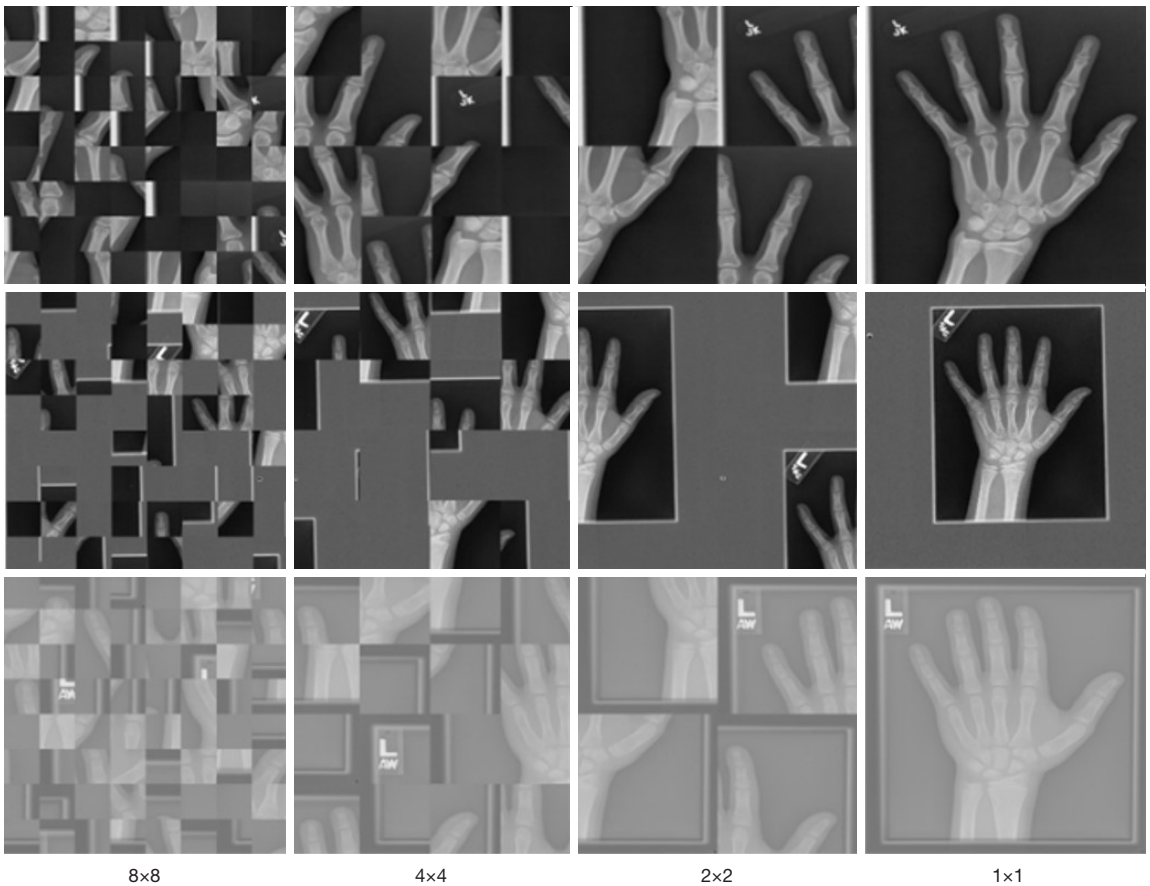

Figure 1 Examples of applying the jigsaw method. *N*=8, 4, 2, 1, respectively.

The jigsaw method for multi-task learning

The jigsaw method was first proposed by Noroozi *et al.* (16) with the aim of establishing a supervisory signal from the data itself to learn transferable semantic features. This method first divides the image into $N \times N$ equally sized regions and then shuffles their relative positions (see *Figure 1*).

To learn important regions at different granularities, the jigsaw method is used to generate a set of input images with different granularities (including the original image without being scrambled by the jigsaw method), and these inputs are arranged in descending order of *N*. Using the CNN as the backbone for learning representations on images, the output feature for the input with the larger *N* is lower level. This means that the input corresponding to smaller granularity outputs a feature map with a smaller receptive field and does not introduce excessive noise in the learning process. Thus, for these related tasks emphasizing regions with different granularities, the shared representation in the CNN enables these tasks to help each other. Due to the relatedness between tasks, the above approach can be regarded as an implicit data augmentation strategy that can improve generalization performance.

In this work, the residual neural network (ResNet) is used as the backbone (17). The following is a detailed description of the training procedure. *Figure 2A*,*2B* illustrate two types of ResNet blocks, respectively. The ResNet Basic Block shown in *Figure 2A* consists of two 3×3 convolutional layers, while the ResNet Bottleneck Block shown in *Figure 2B* consists of two 1×1 convolutional layers and a 3×3 convolutional layer. For the output *y*, a shortcut is introduced to transform the mapping into a residual function $F(x) = y - x$. *Figure 2C* shows the simplified ResNet structure, where feature extraction is mainly accomplished through *K* convolution groups. Each *k*-th convolution group is composed of L_k ResNet bottleneck blocks, where $k \in 1, \ldots, K$. Finally, the feature maps are processed differently according to different tasks. The down-sampling operation between different convolution groups is

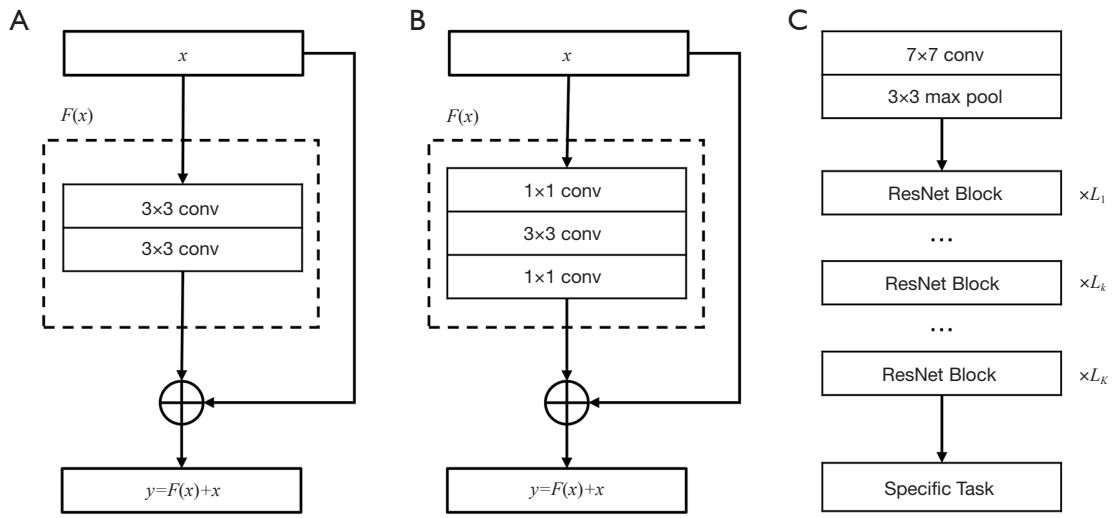

Figure 2 Residual connect in the ResNet Block (A,B) and the structure of ResNet (C). conv, convolution; ResNet, residual neural network.

Table 1 Receptive fields in common residual neural networks

Model	Check point	Receptive field (pixels)
ResNet18	Stage1	43
	Stage 2	99
	Stage 3	211
	Stage 4	435
ResNet50	Stage 1	35
	Stage 2	99
	Stage 3	291
	Stage 4	483
ResNet101	Stage 1	35
	Stage 2	99
	Stage 3	835
	Stage 4	1,027

ResNet, residual neural network.

implemented by a 1×1 convolution, which is not shown in *Figure 2C*. Compared with other deep neural networks that approximate the objective function, ResNet makes it easier to learn optimized residual functions by adding a shortcut to the non-linear convolutional layers. It also solves the problem of gradient vanishing that occurs with an increase in network depth and improves the efficiency of information propagation in deep networks.

As mentioned above, a set of input images with different granularities $\{I_1, \ldots, I_t, \ldots, I_T\}$ is obtained using the jigsaw method, where I_T is the original image (N=1). For the input I_T, the final output is obtained through the multi-scale features of the ResNet. The input I_{T-1} corresponds to the output feature of the K-th convolution group of the ResNet. Continuing in this manner, the input I_t and the corresponding output feature depth k satisfy $(T - 1) - t = K - k$ and $K \geq T - 1$. The information of the receptive fields in the common ResNets is shown in *Table 1*. Notably, in ResNet18, the receptive field of any scale feature is approximately half the size of that of a deeper layer feature. Thus, ResNet18 is selected as the ideal network structure. The input I_t corresponds to the granularity of $N = 2^{T-t}$. When K=4 and T=3, the model training process is shown in *Figure 3*.

By combining the jigsaw method with the hierarchical processing property of CNNs, the model focuses on regions with different granularities in the different training stages. Additionally, the progressive intermediate layer supervision is helpful for learning the higher-level features.

Modeling long-range relationships

Long-range relationships need to be modeled in many visual tasks. The self-attention mechanism provides a powerful and simple solution for long-range relationship modelling. Most notably, Wang *et al.* (18) used NL image processing and proposed NL neural networks that provide solid improvements on the tasks of image/video classification, object detection, etc. Many variants of NL

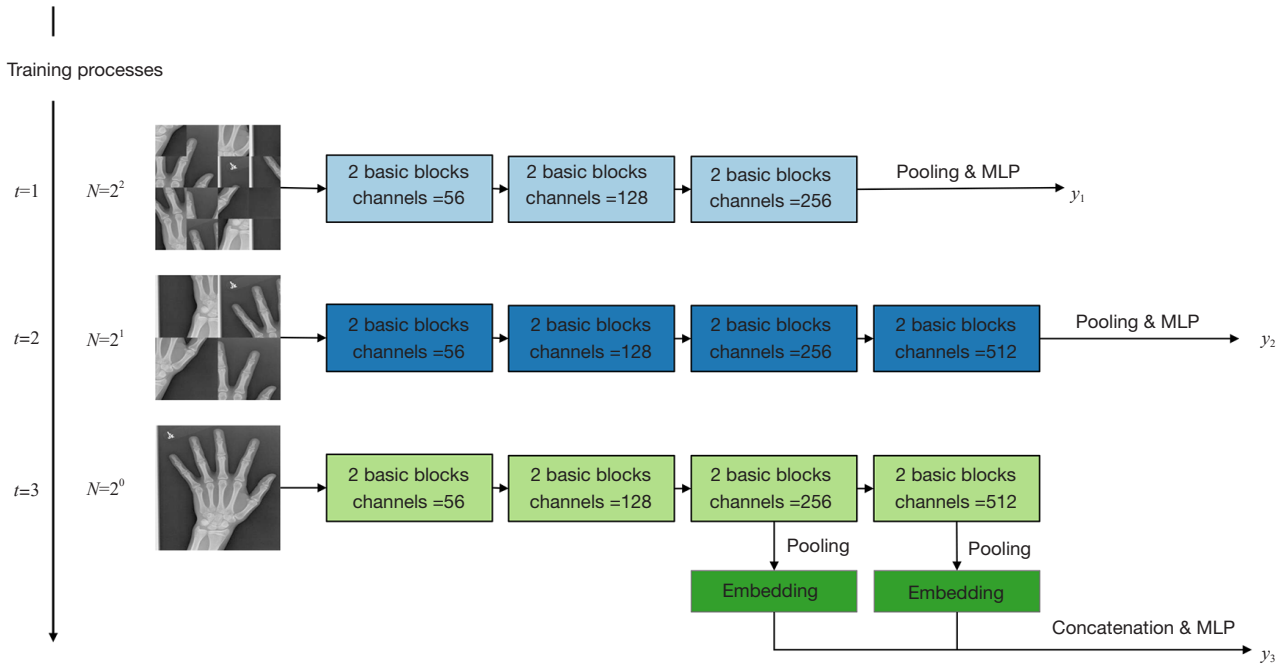

Figure 3 Training time of ResNet18 based on the jigsaw method. In this example, there are K=4 convolution groups in the ResNet and t=3 steps in the training time. The jigsaw method generates a set of input images emphasizing different granularities (N=2^2, 2^1, 2^0, respectively) that correspond to different steps. For the input in the last step, I_T, the final output is obtained through multi-scale features. MLP, multi-layer perceptron; ResNet, residual neural network.

modules also focus on employing cross-channel clues based on NL operations (19) or finding nearly equivalent replacements with lower computational complexity (20). These approaches are helpful in directly leveraging NL operations to efficiently model relationships between distant regions.

The convolution operation, as the main calculation in ResNet18, encodes information locally, considering only the dependency within a local region of the input. As the areas that significantly reflect skeletal development are not adjacent, modeling long-range dependencies in the feature map is crucial to gain a better global understanding of the feature representation. Thus, in this study, we introduced self-attention mechanisms to efficiently incorporate relationships among arbitrary local regions. The weights between different connections in the self-attention mechanism are dynamically generated and can handle inputs of arbitrary dimensions. As a result, the self-attention module can be inserted in any position of a CNN. The calculation steps in the self-attention mechanism are similar to those in the attention mechanism: first, the relationship between any position in the input feature map is calculated; and second, the relationship is used as the attention distribution to weigh the input feature (21).

In Wang *et al.* (18), a convolution-based NL block was used to achieve self-attention computation. This module directly calculates information relationships between any positions while keeping the dimensions of input and output unchanged. Below, we introduce the general formula of the NL module. In which we assume that N groups of input information are represented as $X = [x_1, \ldots, x_N] \in R^{D \times N}$, where $x_n \in R^D$ represents one group of input information, and the output information is $H = [h_1, \ldots, h_N] \in R^{D \times N}$. Then the general formula of the NL module is as follows:

$$h_i = \frac{1}{C(X)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right), \quad C(X) = \sum_{\forall j} f\left(x_i, x_j\right) \qquad [1]$$

where i and j represent the positions of any two groups in the input information, function g is a linear mapping function achieved by a 1×1 convolutional layer, while function C is for normalization, and function f represents the attention scoring function, which is used to calculate the similarity between any two groups in the input information. The output information of the NL module is the weighted

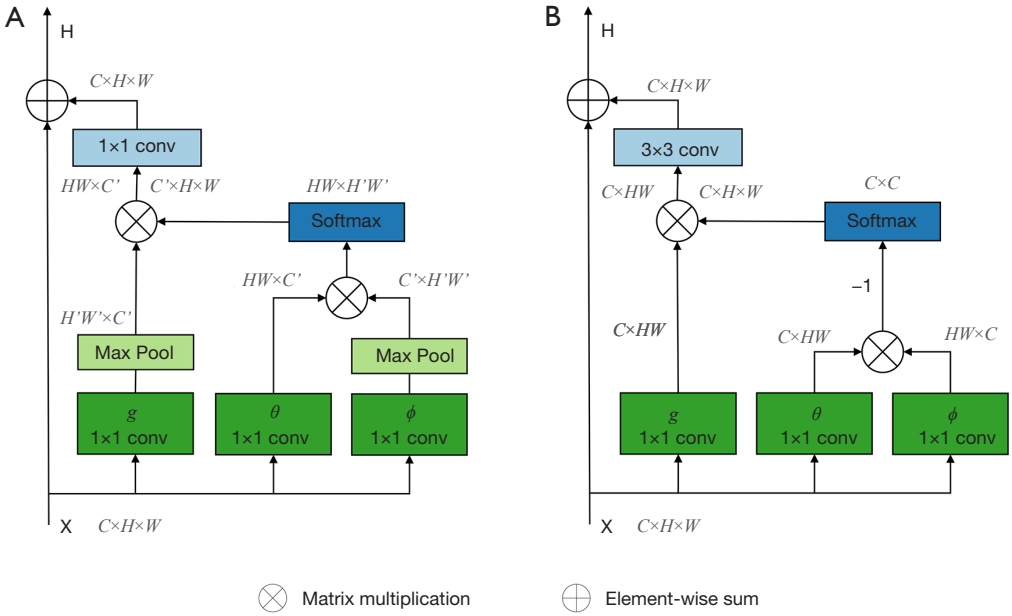

Figure 4 Illustration of the calculation and dimension change in self-attention modules. The left and the right sub-figure correspond to the spatial dimension (A) and the channel dimension (B), respectively. As detailed in Section "Results", we separately inserted spatial and channel self-attention modules into different network stages to show their effectiveness. H (in the y-axis), the output feature; X (in the x-axis), the input feature; C, the number of channels; H, height; W, width; conv, convolution.

sum of all the positions, making it a self-attention mechanism based on the global context.

Based on the general formula of the NL module, we calculate long-range relationships in both spatial and channel dimensions to effectively select and enhance feature representation. The specific implementation methods are described below.

Assume the input feature map is $F \in R^{c \times h \times w}$, where c represents the number of channels, and h and w represent the height and width of the feature map. Given the spatial dimension, the feature map is divided into $h \times w$ c-dimensional input vectors, and the Embedded Gaussian is used to calculate the attention distribution as shown in the following equation:

$$f\left(x_i, x_j\right) = \exp\left(\theta\left(x_i\right)^T \phi\left(x_j\right)\right), \quad x_n \in R^c, n \in \{1,2,\ldots,hw\} \quad [2]$$

where functions θ and φ are linear functions achieved through 1×1 convolutional layers. The spatial calculation and dimension change are shown in *Figure 4A*, which includes max-pooling to down-sample and increase computational efficiency, as well as a shortcut.

For the channel dimension, the feature map is divided into c ($h \times w$)-dimensional input vectors. As shown in the following equation, the negative correlation between different input vectors is calculated in Gaussian form to mine complementary semantic information:

$$f\left(x_i, x_j\right) = \exp\left(-x_i^T x_j\right), \quad x_n \in R^{hw}, n \in \{1,2,\ldots,c\} \quad [3]$$

The calculation process is shown in *Figure 4B*.

During the experiment, the self-attention module was inserted into different layers of ResNet18 to verify the effectiveness of the above two modules. Based on the experimental results, which will be discussed next, the spatial self-attention module is inserted in the middle stage to enhance the feature maps with higher resolution as they provide more spatial information. The channel self-attention module is used for high-level feature maps, which have more semantic information in the channel dimension.

Optimization objective

In training step t, the corresponding objective function is the L1 loss, which can be described as:

Table 2 Summary information for the RSNA data set

Data set	Male	Female
Training set	6,833	5,778
Validation set	773	652
Test set	100	100

RSNA, Radiological Society of North America.

$$Loss_t\left(x_m, y_m^t\right) = \min \frac{1}{M}\sum_{m=1}^{M}\left|x_m - y_m^t\right| \quad [4]$$

where M is the number of samples, m is the index, y^t denotes the output of step t, and x is the ground truth label. In the last step (analyzing the original image), multi-scale features are applied for prediction, which is also used as the assessment result in the inference phase. During the training process, the sampling operation of the jigsaw method in each epoch continuously changes the model input, and the local structure cannot always be fully preserved. This is a data augmentation strategy that plays a regularization role. At the same time, the main task for the original input image can selectively use the features learned in other tasks to enhance feature representation. Thus, the training signals for the extra tasks that are created by the jigsaw method serve as an inductive bias. Finally, the model performance is evaluated based on MAE.

Results

In this section, we describe how the RSNA data set was used to evaluate the proposed 2M-Net. Additionally, we compare the method with existing methods and examine the impact of the jigsaw method and two self-attention modules.

The RSNA data set

The RSNA data set comprised 14,236 hand radiographs, including 12,611 images in the training set, 1,425 images in the validation set, and 200 images in the testing set. Each hand radiograph was labeled with the ground truth bone age (0, 1, …, 228 months) and gender. The bone age labels were from six independent assessment results that used the G-P atlas method (7). The summary information for the RSNA data set is shown in *Table 2*. Several samples in the data set are shown in *Figure 5*, and the bone age distribution is shown in *Figure 6*. In most existing methods, the RSNA data set is regarded as the benchmark. To conveniently compare the performance with other methods, we followed the original training/validation/testing split of the RSNA data set. Because there are significant differences in the development pattern of skeletons in different genders, the proposed method was developed on males and females separately.

Implementation details

Pre-trained ResNet18 was employed as the backbone of the 2M-Net and was from the PyTorch library (22). We trained the network on an input resolution of 448×448. Specifically, for the male data set, we first resized the input image into 512×512 and then cropped the image at the center, and for the female data set, we directly resized the image to 448×448. The above differences depended on the experimental results. Random rotation, random flip, and random gamma contrast adjustment were used in the data augmentation pipeline. All the experiments were developed with a total batch size of 64 on 4 2080Ti graphics processing units (GPUs). The Stochastic Gradient Descent (SGD) optimizer with a weight decay of 0.0001 and a momentum of 0.9 was employed to minimize the regression loss for 130 epochs. In the first five epochs, the learning rate is warmed up from 0.001. The learning rate was then set to 0.0025 and 10× drops per 30 epochs.

Ablations on the jigsaw method

To examine the effectiveness of the jigsaw method, we fed different input queues into the ResNet18 without any self-attention layer. The results are shown in *Table 3*. In the 'Input Queue' column, the element in the list denotes the corresponding granularity and the order of elements is the input order. Based on the experimental results, the relevant hyperparameters of the jigsaw method were determined (K=4, T=3, and N in each step is 4, 2, 1 in order).

Ablations on self-attention modules

Based on the jigsaw method, we separately inserted spatial and channel self-attention modules into different network stages to show their effectiveness. The results are shown in *Tables 4,5*. According to the experimental results, the best performance was achieved when the spatial self-attention module was inserted into the middle layer. It may be that the low-level semantic features were more sensitive to changes in the input image, while the high-level semantic

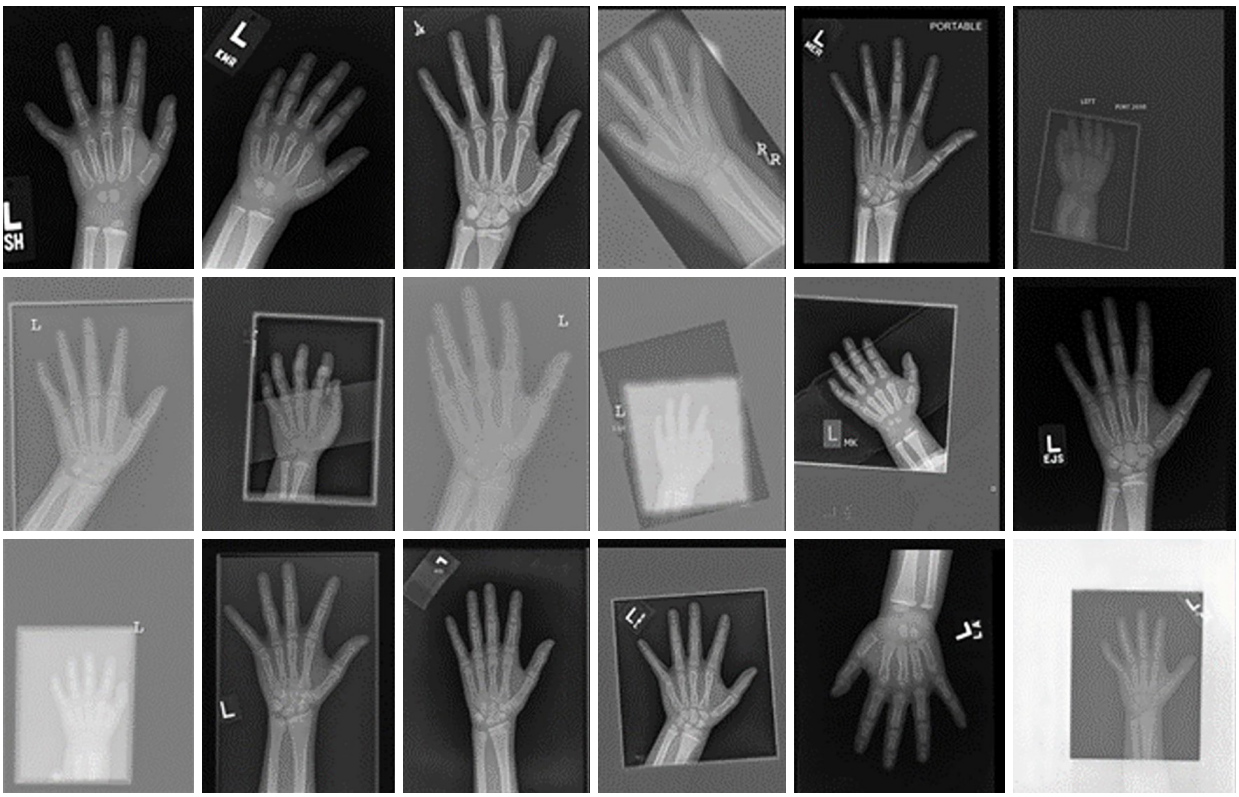

Figure 5 Samples in the RSNA data set. RSNA, Radiological Society of North America.

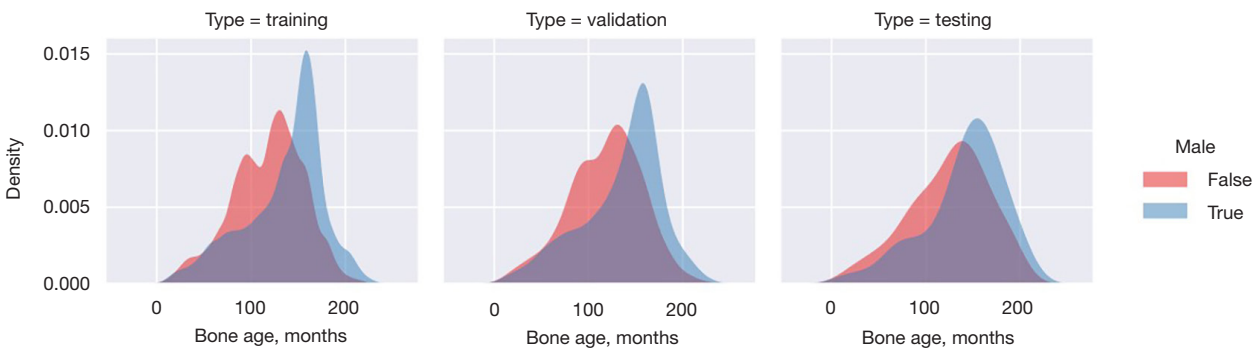

Figure 6 Bone age distribution in the RSNA data set ("True" and "False" refer to whether the individual is male or not). RSNA, Radiological Society of North America.

features lacked spatial information, and the middle layer had sufficient spatial information and a certain degree of robustness. Similarly, the model performed the best when calculating the channel-wise self-attention for high-level semantic features. Finally, by combining the jigsaw method with both of the self-attention modules, the MAE predicted by the model was 3.89 months (male) and 4.07 months (female).

Comparison

We compared the proposed 2M-Net with existing representative methods using the RSNA data set, and the results are summarized in *Table 6*. The column "Estimated parameters" in *Table 6* represents the number of parameters in the corresponding feature extraction network of each method, with M indicating a million.

Table 3 Jigsaw method with different input queues

Input queue	MAE/months	
	Male	Female
(1)	5.55	4.73
(2, 1)	4.89	5.11
(4, 2, 1)	4.64	4.49
(8, 4, 2, 1)	4.76	4.62

MAE, mean absolute error.

Table 4 Only adding the spatial self-attention module into different stages

Model setting	MAE/months	
	Male	Female
Baseline	4.64	4.49
Stage 2	4.26	4.27
Stage 3	4.03	4.09
Stage 4	4.30	4.21
Stage 5	4.48	4.42

MAE, mean absolute error.

Table 5 Only adding the channel self-attention module into different stages

Model setting	MAE/months	
	Male	Female
Baseline	4.64	4.49
Stage 4	4.43	4.28
Stage 5	4.34	4.19
Stages 4 & 5	4.29	4.12

MAE, mean absolute error.

The first row shows the range of the MAE for the six independent assessment results in the RSNA data set. The model established in this study showed better performance than the manual assessment method. The BoNet includes a bounding-box detection model, a key-point estimation model, and a regression model based on CNNs, and the first two models rely on extra manual annotations. The PEAR-Net model was designed with a weakly supervised localization module to locate key local regions and then to combine global image features with local image features for bone age prediction. It can be seen that without strong or weak annotations, 2M-Net was comparable to the previous best method in terms of its performance against the public benchmark. Additionally, by showing that 2M-Net can achieve better performance than other models even with a relatively smaller number of parameters, we showed its efficiency in learning and capturing key features from input data.

Discussion

Feature visualization

The attention map in the spatial self-attention module is visualized to show the relationship between local regions. The results are shown in *Figure* 7, and the red-marked positions in the image represent query positions. The attention weight specific to the query position in the self-attention module represents the importance of the corresponding position to the query position. Notably, the model achieved arbitrary relationships between positions in the spatial direction during the optimization process to achieve better context understanding.

Error analysis

Figure 8 shows the confusion matrices for the RSNA training

Table 6 Comparisons with other methods using the RSNA testing set

Method	Manual annotations	Estimated parameters	MAE/months
Doctors (7)	No	None	5 to 7
BoNet (10)	Yes	>50 M	4.14
PEAR-Net (15)	No	25.6 M	3.99
Ours	No	11.8 M	3.98

RSNA, Radiological Society of North America; MAE, mean absolute error; M, million.

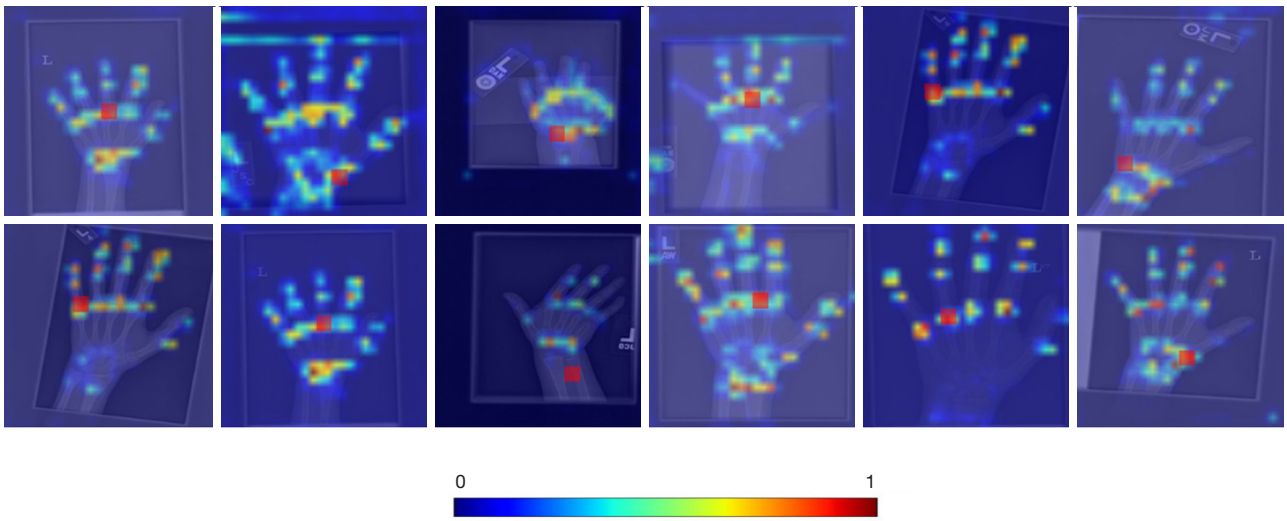

Figure 7 Visualization of attention maps in the spatial self-attention module.

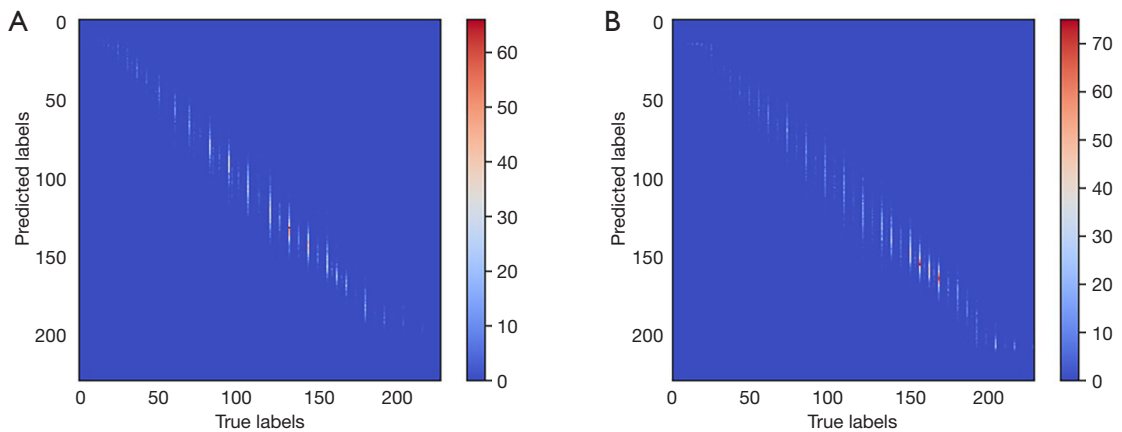

Figure 8 Confusion matrices for the RSNA training data set (A) female, and (B) male. RSNA, Radiological Society of North America.

set; *Figure 8A* corresponds to the female and *Figure 8B* corresponds to the male. The horizontal axis represents the ground truth labels, the vertical axis represents the predicted results. Notably, the main misclassifications were concentrated near the diagonal because the closer the bone age labels (the higher the similarity between image samples), the more challenging it is for the model to make accurate predictions. As the bone age in the RSNA data set was measured in months for each image, the differences between images with bone ages that were too close were extremely subtle. Additionally, there might have been interference from factors, such as image quality and variations in hand pose. Even for doctors, it is difficult to make accurate judgments. These facts reflect the main research challenges in learning fine and comprehensive skeletal development patterns.

Conclusions

In summary, we established an automated BAA method using multi-granularity and multi-attention feature encoding. Unlike localization-based methods, we encouraged 2M-Net to perform discriminative feature learning efficiently and directly model long-range relationships between distant positions in the whole hand-wrist context. We achieved this by: (I) applying the jigsaw method to generate related tasks emphasizing regions with different granularities, and training these tasks using

a hierarchical sharing mechanism; and (II) modelling spatially aware long-range relationships directly between key positions in a global context for high-resolution feature maps, and capturing channel-wise negative correlations as a complement to model representation for low-resolution feature maps. Notably, both of these calculations were implemented in a plug-and-play way. Without strongly or weakly supervised annotations, 2M-Net was comparable to the previous best method in terms of its performance against the public benchmark and is much more lightweight. In a further study, we will continue to explore the fine-grained representation of skeletal development patterns, and we will comprehensively analyze the inadequacy of the model from the perspective of model uncertainty.

Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China (No. 62003284), the Study on the Correlation between Vitamin A and E Levels and Respiratory Tract Infections in Children (School Collaboration: 20223160A0463), and Doctoral Studio Ascent Project Fund (A) (No. PDA202105)

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-806/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Lee BD, Lee MS. Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment. Korean J Radiol 2021;22:792-800.
2. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol 1996;167:1395-8.
3. Loder RT, Farley FA, Herring JA, Schork MA, Shyr Y. Bone age determination in children with Legg-Calvé-Perthes disease: a comparison of two methods. J Pediatr Orthop 1995;15:90-4.
4. Jang WY, Ahn KS, Oh S, Lee JE, Choi J, Kang CH, Kang WY, Hong SJ, Shim E, Kim BH, Je BK, Jung HW, Lee SH. Difference between bone age at the hand and elbow at the onset of puberty. Medicine (Baltimore) 2022;101:e28516.
5. Mansourvar M, Ismail MA, Raj RG, Kareem SA, Aik S, Gunalan R, Antony CD. The applicability of Greulich and Pyle atlas to assess skeletal age for four ethnic groups. J Forensic Leg Med 2014;22:26-9.
6. Berst MJ, Dolan L, Bogdanowicz MM, Stevens MA, Chow S, Brandser EA. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am J Roentgenol 2001;176:507-10.
7. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT, Abdala N, Kitamura FC, Thodberg HH, Chen L, Shih G, Andriole K, Kohli MD, Erickson BJ, Flanders AE. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019;290:498-503.
8. Siegel EL. What Can We Learn from the RSNA Pediatric Bone Age Machine Learning Challenge? Radiology 2019;290:504-5.
9. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology 2018;287:313-22.
10. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017;36:41-51.
11. Escobar M, Gonzalez C, Torres F, Daza L, Triana G, Arbelaez P, editors. Hand Pose Estimation for Pediatric Bone Age Assessment. 10th International Workshop on Machine Learning in Medical Imaging (MLMI)/22nd International Conference on Medical Image Computing

and Computer-Assisted Intervention (MICCAI); 2019 Oct 13-17; Shenzhen, China; 2019.
12. Jia Y, Zhang X, Du H, Chen W, Jin X, Qi W, Yang B, Zhang Q, Wei Z. Fine-grained precise-bone age assessment by integrating prior knowledge and recursive feature pyramid network. Eurasip J Image Video Process 2022;2022:12.
13. Ren X, Li T, Yang X, Wang S, Ahmad S, Xiang L, Stone SR, Li L, Zhan Y, Shen D, Wang Q. Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment From Hand Radiograph. IEEE J Biomed Health Inform 2019;23:2030-8.
14. Chen C, Chen Z, Jin X, Li L, Speier W, Arnold CW. Attention-Guided Discriminative Region Localization and Label Distribution Learning for Bone Age Assessment. IEEE J Biomed Health Inform 2022;26:1208-18.
15. Liu C, Xie H, Zhang Y. Self-Supervised Attention Mechanism for Pediatric Bone Age Assessment With Efficient Weak Annotation. IEEE Trans Med Imaging 2021;40:2685-97.
16. Noroozi M, Favaro P, editors. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. 14th European Conference on Computer Vision (ECCV); 2016 Oct 08-16; Amsterdam, the Netherlands; 2016.
17. He K, Zhang X, Ren S, Sun J, editors. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Seattle, WA; 2016:770-8.
18. Wang X, Girshick R, Gupta A, He K, editors. Non-local Neural Networks. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Jun 18-23; Salt Lake City, UT; 2018.
19. Yue K, Sun M, Yuan Y, Zhou F, Ding E, Xu F, editors. Compact Generalized Non-local Network. 32nd Conference on Neural Information Processing Systems (NIPS); 2018 Dec 02-08; Montreal, Canada; 2018:6511-20.
20. Cao Y, Xu J, Lin S, Wei F, Hu H, editors. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27-Nov 02; Seoul, South Korea; 2019.
21. Woo S, Park J, Lee JY, Kweon IS, editors. CBAM: Convolutional Block Attention Module. 15th European Conference on Computer Vision (ECCV); 2018 Sep 08-14; Munich, Germany; 2018:3-19.
22. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch. NIPS 2017 Workshop on Autodiff. Long Beach, California, USA; 2017.

Cite this article as: Liu B, Huang Y, Li S, He J, Zhang D. Bone age assessment by multi-granularity and multi-attention feature encoding. Quant Imaging Med Surg 2024. doi: 10.21037/qims-23-806