End-to-end bone age assessment based on Attentional Region Localization

Chunde Yang\textsuperscript{1,2}, Qingshui Liu\textsuperscript{1}, Zhangyong Li\textsuperscript{2}, Wei Wang\textsuperscript{2} and Xinwei Li\textsuperscript{2,*}

\textsuperscript{1}School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China
\textsuperscript{2}School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing, China

*Corresponding author’s e-mail: lixinwei@cqupt.edu.cn

Abstract. Pediatric bone age assessment (BAA) is an extremely important clinical method to investigate endocrinology, genetic and growth disorders of adolescents. The current deep learning-based BAA scheme generally feeds the images into the training model, but ignores the local details of the skeleton images and does not exclude the noise around the images. Many methods train segmentation or detection networks to exploit local information, but requires a lot of manual annotation and additional costs. In this paper, we proposed an attentional region localization method for BAA to automatically localize the hand region and local regions without any additional annotations. First, an attentional hand location module (AHLM) was used to obtain a clearer hand region, which eliminates the interference noise of the original image. Then an attentional region generation module (ARGM) was used to extract the local attentional regions with high discriminant features, which can help to optimize the entire network framework. We integrate the entire network into an end-to-end structure by jointly optimizing the network through a shared backbone and fully connected layers. The effectiveness of the proposed attentional region localization method was evaluated on an open dataset Radiological Society of North America (RSNA) with an average absolute error (MAE) of 6.14, which performs better than most existing methods.

1. Introduction
Bone age is an important physiological characteristic of children’s growth and development, and can be used as the most reliable indicator to evaluate whether children’s development is abnormal. Therefore, bone age assessment (BAA) is a common clinical method to study adolescent growth disorders and endocrine conditions\cite{1,2}. Currently, internationally recognized bone age assessment methods are Greulich-Pyle(GP) and Tanner-Whitehouse(TW). The GP method compares the evaluated images with a standard bone age atlas to obtain the result. The TW method is the most commonly used method in routine clinical practice, where radiologists comparing multiple fixed regions of interest with regional bone atlases at different growth stages, and then using the average of all regions as the final assessment of bone age. At present, BAA is mainly manually diagnosed by professional radiologists. However, reading X-ray images of hand bones is a time-consuming and laborious task that require a lot of experience and is prone to subjective errors. Therefore, it is of great significance to automate BAA evaluation, which can not only save resources but also promote the standardization of the assessment.
1.1. Related work
To overcome the defects of automatic BAA, many computer vision and image processing methods have been proposed to realize automatic BAA. The BAA methods based on deep learning can be divided into two categories. The first category is the end-to-end/one-stage scheme, which directly inputs the complete original image into the convolutional neural network (CNN) to obtain the predicted bone age. For example, Spampinato et al.[3] proposed BO-Net which consists of five convolutional layers and one pooling layer. Lee et al.[4] used GoogleNet with pre-training weights on ImageNet to obtain bone age and achieved good results. The second type is the multi-stage scheme that relies on additional manual annotations or image preprocessing. For example, Iglovikov et al.[5] first trained a UNET model to segment the complete hand bone to prevent background noise, and then aligned the hand area and feed it into the VGG16 network for bone age prediction. Gao et al.[6] added an attention mechanism to the network framework to get a better evaluation effect. Wang et al.[7] used the Faster-RCNN object detection method to extract the ROIs of the image, and used the extracted ROIs to predict the bone age. Assessing bone age is a fine-grained problem, and many important details are hidden in specific ROIs[8]. However, the single-stage BAA methods often ignore the attention to the local key areas of the image, and the results cannot be explained in clinical applications. Besides, the multi-stage BAA methods require a lot of additional labels and do not support end-to-end training.

1.2. Contributions
To solve the above problems, this paper proposed an end-to-end network framework based on the localization of attentional regions for BAA. This method can obtain the hand region and three local regions without manual annotation. The hand region was used to evaluate the bone age, while the local regions were used to jointly optimize the entire network by sharing the backbone network and the fully connected layers.

Our contributions of this paper are as follows:
- The region extraction modules proposed in this paper can accurately locate the hand region and local attentional regions, which can eliminate the noise interference of the original image and utilize the tiny features of the hand bones for training.
- The entire network was integrated into an end-to-end architecture that combines the advantages of multi-stage and single-stage BAA schemes to achieve accurate bone age assessment.

2. Method
2.1. Overall Framework
Figure 1 illustrates the overall framework of our proposed BAA method. First, the original image was fed into the backbone network, which is ResNet50 in this paper, to get the feature maps. Then, the feature maps were fed into an attentional hand location module (AHLM) to obtain the mask of the complete metacarpal bone through the Selective Convolutional Descriptor Aggregation (SCDA)[9]. The mask was used to cut the original image to obtain the hand region, which was then fed into the backbone network for further training, and the resulting feature maps were fed into an attentional region generation module (ARGM) to obtain three key regions. These three key regions were then used to further train the backbone network. The backbone was followed by a Global Average Pooling (GAP) layer and a fully connected layer, then softmax was used to get probability scores for each month. The same color layer in Figure 1 indicate that their parameters are shared across the network framework. The ARGM in the dashed frame was used only in the training phase but not required in the testing phase.
2.2. Attentional Hand Location Module

We can treat the hand region extraction as weakly supervised region localization. Let $F \in \mathbb{R}^{C \times H \times W}$ represent the feature maps extracted from the original image through the backbone network, where $C$ is the number of feature channels, $H$ and $W$ are the height and width of the feature map, respectively. Sum the values of $F$ in the channel dimension to get the aggregate feature map $A$, as shown in Eq.1, where $f_i$ represents the i-th feature map. Then the average value $\hat{a}$ of the aggregated feature map $A$ was calculated as shown in Eq.2.

$$A = \sum_{i=0}^{C-1} f_i \quad \text{\textdollar MERGEFORMAT (1)}$$

$$\hat{a} = \frac{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} A(x, y)}{H \times W} \quad \text{\textdollar MERGEFORMAT (2)}$$

Next, $\hat{a}$ was used as the threshold for the aggregated feature map $A$ to get a hand mask, which was denoted as $M_1$, as shown in Eq.3. However, as discussed in[10], there is still a lot of noise in the region extracted only by the last convolution module of the backbone. Therefore, we also used the activation map generated in the penultimate convolution module of the backbone network to get another mask, denoted as $M_2$. The final mask $M$ was the intersection of $M_1$ and $M_2$ as shown in Eq.4. Finally, we used $M$ to crop the hand region from the original image. Figure 3 shows the hand regions obtained by the AHLM.

$$Mask_{(x,y)} = \begin{cases} 1 & \text{if } A_{(x,y)} > \hat{a} \\ 0 & \text{otherwise} \end{cases} \quad \text{\textdollar MERGEFORMAT (3)}$$
\[ M = M_1 \cap M_2 \]  

\section*{2.3. Attention Region Generation Module}

In addition, we proposed the ARGM to obtain more distinguishable bone regions, which can improve the robustness of the model. It has been proven that the areas with higher activation values in the activation map are usually the key regions that we want to focus on[11]. The ARGM utilized the idea of sliding window in the Regional Proposal Network (RPN)[12] to locate the key regions. In detail, we fed the feature maps generated by the hand region into the ARGM to generate a set of candidate rectangular regions \( \{x_1, x_2, ..., x_n\} \). Each rectangular region has a score \( S_\alpha \), namely the activation value generated by the convolution of the ARGM. The scales of the anchor points were set to \( \{48, 96, 192\} \) and the ratios were set \( \{1:1, 2:3, 3:2\} \). After generating a large number of anchor points, the resulting boxes will have a lot of overlap. Therefore, we further use the Non Maximum Suppression (NMS) to de-redundant the boxes, where the threshold was set to 0.1. The NMS process is illustrated in Figure 2.

![Figure 2. Remove redundant local regions based on NMS.](image)

Then, according to the existing experience[13], we selected the top three boxes with the highest scores. Crop out these boxes on the hand image, and then fed into the backbone to optimize the entire network through a shared structure. The visualization results in Figure 4 show that the extracted attentional regions were concentrated in the wrist bones and metacarpal joints, which was consistent with the focus of the actual clinical practice.

\section*{2.4. Loss Function}

The whole network has three branches, including the original image branch, the hand region branch and the attentional region branch. Each branch has a cross-entropy loss function, as shown in Eq. 5, 6 and 7 respectively, where \( G \) represents the real age labels, \( N \) represents the number of selected attentional regions, \( P_r \), \( P_h \), and \( P_p \) represent the predicted age month probabilities of the three branches. The total loss \( L_{\text{total}} \) was the sum of these three losses, as shown in Eq.8.

\begin{align*}
    L_{\text{raw}} &= - \log (P_r(G)) \\
    L_{\text{hand}} &= - \log (P_h(G)) \\
    L_{\text{parts}} &= - \sum_{n=0}^{N-1} \log (P_{a(n)}(G)) \\
    L_{\text{total}} &= L_{\text{raw}} + L_{\text{hand}} + L_{\text{parts}}
\end{align*}
calculations without affecting performance. This may be because the shared convolutional layers and fully connected layers have been optimized by the three attentional regions. In the end, we only used the hand region to assess bone age. Let $p_i$ represents the probability of age obtained from the $i$-th sample in the hand region branch. The predicted value of age $y_i$ is the age corresponding to the maximum value in $p_i^i$. The mean absolute error (MAE) between the ground truth ages $\hat{y}_i$ and the predicted ages $y_i$ was used to evaluate the model, as shown in Eq.9.

$$MAE = \sum_i \| \hat{y}_i - y_i \|$$ \hspace{1cm} /* MERGEFORMAT (9)

3. Experiment

3.1. Data set
In this paper, we used the dataset provided by the children’s bone age prediction competition held by the Radiological Society of North America (RSNA) in 2017. This dataset provides bone pictures and corresponding children's bone age months (1-240). Bone age was assessed by six professional doctors. The dataset also provides the gender label of each image. There are 12,611 pictures in the dataset, including 6,833 male images and 5778 female images. Because the growth of boys and girls are inconsistent, we trained the network structure for boys and girls separately. We take out 500 images of boys and girls respectively as the test set, and the rest was used as the training set.

3.2. Details and Experimental environment
We resized the original images to (448, 448) and fed them into the backbone. The hand regions obtained by AHLM were also resized to (448, 448) and fed into the backbone. In addition, the three attentional regions obtained by ARGM were resized to (224, 224) and fed into the model for training. The number of nodes in the fully connected layer was 240, which represented the number of months of bone age distribution. Our backbone was Resnet-50, and the pre-training weights on ImageNet were loaded to get more accurate activation maps. The size of activation map is $14 \times 14$. Throughout the training process, we only used the bone age labels, and no additional manual annotations were required. We used SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. There were a total of 200 training epochs, with the initial learning rate set at 0.001, multiplied by 0.1 for every 50 epochs. Our network was trained on an Ubuntu workstation equipped with 4 NVIDIA RTX TITAN GPUs. The batch size was set to 6. We trained and tested the model on Pytorch 1.5.0 in the Python environment.

3.3. Results
To prove that the AHLM and ARGW we used can effectively improve BAA accuracies, we compared the results in the absence of additional modules and in the use of a single module, see Table 1. The results confirmed the importance of ARGW and AHLM in BAA. We found that using AHLM was more effective than using ARGW. We believe that this is because the entire hand region contains more complete information, and the use of ARGW alone may cause the network to ignore many details that are not sufficiently focused. In addition, we found that bone age was more accurately predicted for boys than for girls. This may be because girls grow and develop faster than boys, so the bones and joints close first, making it harder for the network to distinguish the joint details.
Table 1. Ablation study on modules

|        | Baseline | ARGM | AHLM | ARGM+AHLM |
|--------|----------|------|------|-----------|
| Male   | 7.23     | 6.97 | 6.45 | 6.07      |
| Female | 7.41     | 7.13 | 6.73 | 6.21      |
| Average| 7.32     | 7.05 | 6.59 | 6.14      |

We also compared the average MAE values of males and females with other existing BAA methods, as shown in Table 2. The results show that our end-to-end attentional region localization network can achieve a small bone age estimation error, even more accurate than some multi-stage methods that require additional manual annotations.

Table 2. Comparison results with other methods.

| Method         | Stage       | MAE  |
|----------------|-------------|------|
| Spampinato[3]  | Single-stage| 9.12 |
| Gao[6]         | Multi-stage | 9.99 |
| Wibisono[14]   | Multi-stage | 6.97 |
| Larson[15]     | Single-stage| 6.24 |
| Iglovikov[5]   | Multi-stage | 6.16 |
| ours           | Single-stage| **6.14** |

3.4. Visualization

Our BAA framework can easily produce visual results to observe the attentional regions extracted from the network. Figure 3 shows the hand regions extracted from the original images using AHLM. We can see that the automatically obtained hand regions are relatively complete, which can not only reduce the noise, but also extract key regions for better prediction of the network.

![Figure 3. Visualization experiments for hand region extraction.](image)

Figure 4 shows the obtained three attentional regions using ARGM. The attentional regions focused on the skeletal joints in the carpal and metacarpal regions, which are indeed important areas that doctors pay attention to in clinical BAA. This demonstrated the clinical effectiveness of our proposed method.
4. Conclusion
In order to improve the performance of BAA, the common method is to train the segmentation or detection networks using additional labels to extract important regional information that is important to BAA task. These regions contain fine-grained features of the image, which can effectively improve the evaluation results. In this work, the hand region and local attentional regions were obtained through attention guidance with only image-level labels, which is more practical and objective. Our method can effectively reduce the interference of background noise, fuse the local features during the assessment process, and improve the accuracy of BAA. The whole process was integrated into an end-to-end network. In addition, our results visualized key areas of BAA for clinicians. In summary, our method has shown good performance and fully meets the requirements of clinical use.

References
[1] Gertych A, Zhang A, Sayre J, Pospiech-Kurkowska S, Huang HK. Bone age assessment of children using a digital hand atlas. 2007;31(4-5):322-31.
[2] Stern D, Ebner T, Bischof H, Grassegger S, Urschler M, editors. Fully Automatic Bone Age Estimation from Left Hand MR Images. International Conference on Medical Image Computing & Computer-assisted Intervention; 2014.
[3] Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi RJ. Deep learning for automated skeletal bone age assessment in X-ray images. 2017.
[4] Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully Automated Deep Learning System for Bone Age Assessment. 2017.
[5] Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks. 2018.
[6] Gao Y, Zhu T, Xu X. Bone age assessment based on deep convolutional neural network incorporated with segmentation. 2020;1-12.
[7] Wang S, Shen Y, Zeng D, Yong H, editors. Bone age assessment using convolutional neural networks. 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD); 2018.
[8] Chen C, Zhihong, Jin X, Li L, Corey W. Attention-Guided Discriminative Region Localization for Bone Age Assessment. 2020.
[9] Wei XS, Luo JH, Wu J, Zhou ZH. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. 2017;26(6):2868.
[10] Yao H, Zhang S, Hong R, Zhang Y, Xu C, Tian QJ. Deep representation learning with part loss for person re-identification. 2019;28(6):2860-71.
[11] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, editors. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
Acknowledgments
This paper is sponsored by the Science and Technology Research Program of Chongqing Municipal Education Commission, Grant/Award Number: KJQN201800622.