Automated Bone Age Assessment: A New Three-Stage Assessment Method from Coarse to Fine

Xinzheng Xu *, Huihui Xu and Zhongnian Li

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; ts2017006a31@cumt.edu.cn (H.X.); zhongnianli@cumt.edu.cn (Z.L.)
* Correspondence: xxzheng@cumt.edu.cn

Abstract: Bone age assessment (BAA) based on X-ray imaging of the left hand and wrist can accurately reflect the degree of the body’s physiological development and physical condition. However, the traditional manual evaluation method relies too much on inefficient specialist labor. In this paper, to propose automatic BAA, we introduce a hierarchical convolutional neural network to detect the regions of interest (ROI) and classify the bone grade. Firstly, we establish a dataset of children’s BAA containing 2518 left hand X-rays. Then, we use the fine-grained classification to obtain the grade of the region of interest via object detection. Specifically, fine-grained classifiers are based on context-aware attention pooling (CAP). Finally, we perform the model assessment of bone age using the third version of the Tanner–Whitehouse (TW3) methodology. The end-to-end BAA system provides bone age values, the detection results of 13 ROIs, and the bone maturity of the ROIs, which are convenient for doctors to obtain information for operation. Experimental results on the public dataset and clinical dataset show that the performance of the proposed method is competitive. The accuracy of bone grading is 86.93%, and the mean absolute error (MAE) of bone age is 7.68 months on the clinical dataset. On public dataset, the MAE is 6.53 months. The proposed method achieves good performance in bone age assessment and is superior to existing fine-grained image classification methods.

Keywords: bone age assessment; deep learning; object detection; convolutional neural network; fine-grained image classification.

1. Introduction

Human growth and development age can be divided into calendar age and biological age. Biological age is generally judged by bone age medically. Different bones of the human body show continuity and stages in the development process. Figure 1 shows typical X-ray pictures of hands and wrists at different ages; it can be seen in the figure that there are obvious differences in bone morphology at different ages. Bone age [1] is determined by observing the developmental status of the left hand bone, such as its shape, size, location, and degree of closure. Bone age assessment (BAA) was first applied to medical research to help understand and promote children’s growth retardation, prevent precocious puberty, predict children’s adult height, and assist in diagnosing endocrine diseases. Subsequently, the application scenarios were widely expanded, such as determining the actual age of athletes and forensic science as a criterion for no conviction [2].

The original BAA was based on a left hand X-ray, and bone age was assessed by physicians by manual reading by the Greenich and Pyle (GP) [3] method or the Tanner–Whitehouse (TW) method. The GP atlas method compares the image to the reference atlas and selects the most similar profile for its estimated bone age. The TW method uses cumulative scores for the developmental morphology of different bones to assess bone age. The latest version of the TW3 method is more objective than the GP mapping method, so it is considered to have higher accuracy and reproducibility than the GP mapping method [4,5]. However, the BAA method by orthopedic specialists is not only cumbersome...
and time-consuming, but also has differences between observers. Therefore, it is of great significance to research and develop an automatic accurate, stable, simple, inexpensive, and rapid method for end-to-end automated BBA.

![X-ray images of hand](image)

Figure 1. Left hand X-ray images of male and female children at different ages. The red boxes are the ROIs.

To overcome the above difficulties, we propose a novel BAA framework to achieve end-to-end BAA. The three-stage hierarchical assessment method cascades the convolutional neural network of object detection and fine-grained classification. Firstly, to improve the accuracy of the assessment, the YOLOv5 object detection model is used to extract 13 ROIs, which is the R (radial, ulnar, metacarpal, and phalangeal bones) series in the TW3 method. Secondly, because the difference between different bone ages of the same bone is small, the use of a simple classifier for classification has a low accuracy rate, and the fine-grained classification method solves this problem very well. Finally, estimates are made using the recognized TW3 method to enable the end-to-end automated BAA system. Experiments on our clinical datasets verified the superiority of the proposed three-stage hierarchical BAA method.

The contributions of the proposed methods can be summarized as follows:
1. We design a novel framework and propose a three-stage hierarchical assessment method based on object detection and fine-grained classification for BAA.
2. We use fine-grained image classification based on context perception to grade 13 bones, resulting in higher accuracy in bone age classification.
3. Experiments on clinical datasets prove that the proposed three-stage hierarchical model based on object detection and fine-grained classification has excellent evaluation performance compared with the most-advanced methods.

2. Related Works

In recent years, deep learning has gradually been applied to the field of medical imaging due to its powerful properties and, in some cases, has reached or even exceeded the expert level. Therefore, researchers have begun to apply convolutional neural networks to BAA to improve the efficiency and accuracy of BAA.

2.1. BAA via Deep Learning

The existing end-to-end BAA detection methods in China generally only output an estimated bone age value, which lacks the interpretability of the model. Spampinato et al. [6]
proposed a lightweight convolutional neural network model BoNet consisting of five convolutional layers, one pooling layer, and one fully connected layer that could be assessed for bone age across races, age ranges, and sexes, yielding an average difference of 9.48 months. Lee et al. [7] normalized the images using pre-processing, followed by a BAA using fine-tuned GooleNet, and generated structured radiology reports based on the doctor’s final decision, with an accuracy rate of 61.40% for males and 57.32% for females, respectively. In another work, Son et al. [8] used faster R-CNN to segment 13 ROIs from left hand X-ray images and then trained the VGG network using ROIs extracted from 3344 images, which differed from the experts by an average of 5.52 months.

Alexander et al. [9] won first place in the Pediatric Bone Age Machine Learning Challenge organized by the North American Radiological Society (RSNA) for the Inception V3 [10] architecture for pixel information in tandem with gender information, with an additional dense layer after concatenation to enable the network to learn the relationship between pixel and gender information, obtaining a mean absolute error (MAE) of 4.27–4.5 months on the RSNA dataset. Yang et al. [11] proposed to perform fine-grained precision assessment by integrating prior knowledge and a pyramid network of recursive features, and the MAE of 1–18 years old was 7.32 months. Liu et al. [12] constructed a two-stage automated assessment method that cascaded coarse-to-fine hands and an integrated convolutional neural network cascade based on stacking, and 5.42/6.58 months of MAE for both males and females was achieved on the RSNA dataset. The above study shows that the automatic BAA method based on deep learning can achieve considerable accuracy. Radiologists focus not only on bone age values, but also on detailed information about bone development. However, the above studies only provide bone age values, which cannot meet the needs of doctors.

2.2. Object Detection

Convolutional neural networks (CNNs) are the simplest and most widely used deep learning algorithms for object detection, which are mainly divided into one-stage and two-stage methods. One-stage algorithms extracts feature directly from the network to predict object classification and location. Redmon et al. [13] proposed YOLO’s first-stage deep learning detection algorithm in 2016, currently available in five versions, which are optimal in terms of speed and results. The SSD algorithm, proposed by Liu et al. [14], raises the speeds by eliminating the need for a regional recommendation network, while improving detection accuracy by applying multi-scale features and default boxes.

The two-stage first generates a region proposal (a pre-selection box that may contain the object to be detected) and then classifies the samples through a convolutional neural network. For example, Uijlings et al. [15] proposed a region-based convolutional neural network (R-CNN) in 2013, which is mainly divided into four steps: candidate region generation, feature extraction, category judgment, and position refinement, which is the pioneering work of introducing the CNN into object detection. This was followed by Fast R-CNN [16] and Faster R-CNN [17], which became the leaders in object detection. The two methods can be combined: Zhang et al. [18] proposed a two-way parallel feature pyramid network to effectively capture multi-scale spatial information from all levels of the feature pyramid network (FPN), achieving better performance than other FCNs.

Wang et al. [19] proposed end-to-end object detection based on a fully convolutional network, which introduces predictive perception one-to-one label assignment for classification to achieve end-to-end detection. In addition, a simple three-dimensional maximum filter (3DMF) was proposed to improve the observance of local regional convolution by using multi-scale features, and it showed good performance on coco and other datasets. Now, target detection technology has become very mature and is gradually being used in the field of medical image processing.
2.3. Fine-Grained Image Classification

The problem of fine-grained image classification is to identify sub-classes under large categories, and the granularity of the categories to which fine-grained images belong is finer, while the differences within classes are large; however, the differences between classes are small, so classification is more challenging [20]. At present, fine-grained classification methods are mainly divided into two categories: strongly supervised classification methods and weakly supervised classification methods.

Strong supervised fine-grained classification requires global and local information for fine-grained classification, that is not only category labels, but also local annotation information such as the bounding box and part annotation. Zhang et al. [21] used the R-CNN algorithm to detect the object level (such as all kinds of birds) and its local areas (head, body, and other parts) of fine-grained images. Their proposed model, partly based on the R-CNN, uses both object-level features and local features in the classification process, so the classification accuracy is relatively ideal. Wei et al. [22] proposed the Mask CNN model, which is divided into two modules. The first is part localization, and the second is feature learning of global and local image blocks. The difference from the previous method is that this method proposes that the FCN learn a part-based segmentation model by using the fully connected network. This method achieved the highest classification accuracy of 87.3% for fine-grained image classification at that time.

The fine-grained classification model with weak supervision information uses the attention mechanism, clustering, and other means to enable the model to automatically find distinctive regions and only uses classification tags to complete training. Lin et al. [23] proposed a recognition structure called “bilinear CNN”, which uses bilinear model fusion features for classification. Yang et al. [24] proposed that the navigator–teacher–scrutinizer network (NTS-Net) be a weakly supervised network that integrates the overall features with the region with the largest amount of information extracted from the original image without using only category labels. Liu et al. [24] proposed a mixed-attention network (SMA-Net), which mainly uses the attention network to locate discriminant regions and feature extraction, which effectively improves the performance of the network.

3. Method

In this section, we introduce our novel BAA framework, a hierarchical network model based on object detection and fine-grained classification; the overall architecture of the entire model is shown in Figure 2, and the ultimate goal is to achieve end-to-end BAA.

Figure 2. Hierarchical network structure. YOLOv5 is used to predict the bone category, select the corresponding classifier according to the output results, classify the bone grade, obtain the classification results, and evaluate the bone age Y. There are 13 classifiers in total, and one classifier is taken as an example in the figure.

3.1. Structure of Bone Age Assessment Model

We divided the bone identification and grades into a three-level tree structure (including the root nodes), as shown in Figure 3. The network first takes the left hand X-ray image as the input, selects the YOLOv5 model for recognition, obtains the prediction result
of the bone category, and saves the detection results by cropping. Then, according to the corresponding category, the output images of the previous network are transferred to the corresponding classifier for classification through a selector. In the test phase, the prediction result of the input image bone category is obtained through the detection model, and then, the corresponding branch structure is selected by the selector based on the result. Finally, the bone age is estimated according to the TW3 method based on the 13 bone grade results, and the result is output.

Figure 3. The tree structure of bone categories and grades. The thirteen colors, such as blue, yellow, green, orange, and gray, correspond to the thirteen categories, respectively (ulna, radius, the 1st proximal and distal phalanges, the 3rd proximal, middle, and distal phalanges, the 5th proximal, middle, and distal phalanges, and the 1st, 3rd, and 5th metacarpals). Different shades of each color represent different grades (A, B, C, D, E, F, G, H, I (ulna does not have an I-th grade)).

3.2. Extraction of 13 ROIs

We used YOLOv5 to extract the ROIs from the left hand X-ray image, and the structure diagram of YOLOv5 we used is shown in Figure 4. We input an image and performed mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling processing to improve the accuracy and inference speed of small object detection. In the feature extraction stage, the slicing operation and CSP structure are used to strengthen the ability of network feature fusion so that more fine-grained features can be extracted. The network uses the GIOU_Loss [25] function and uses the non-maximum suppression (NMS) method to filter the target frame, outputs 13 bone category detection frames, and saves the results. The algorithm description of GIOU_Loss is shown in Algorithm 1.

Figure 4. YOLOv5 structure diagram; the black box is the prediction box.
Algorithm 1  Generalized intersection over union [25]

**Input:** Two arbitrary convex shapes: \( A, B \subseteq S \subseteq \mathbb{R}^n \)

**Output:** \( GIoU \)

1. For \( A \) and \( B \), find the smallest enclosing convex object \( C \), where \( C \subseteq S \subseteq \mathbb{R}^n \)
2. \( IoU = \frac{|A \cap B|}{|A \cup B|} \)
3. \( GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|} \)

### 3.3. Hierarchical Classification of Hand Bones

Since the training of classifiers differs only in the dataset categories, let us take one of the classifiers as an example. Given an image \( P \), input the basic convolutional neural network (Xception) [26], perform feature extraction, and send the output features to the CAP [27] module. Learning the importance of adjusting different areas of features, the module correlates the contextual information of the picture at the pixel level to obtain relevant attention features. To make the classification of the model more accurate, the learnable pooling operation is selected, the feature information is integrated by combining hidden layers with similar responses, the response values of hidden layers are assigned to different class clusters according to the weight of softmax, and all the weights are normalized to output the bone grade classification results.

The loss function is an indicator to measure the performance of the expected results of the prediction model. Denote \( X \) as the input space of the training instance and \( Y = \{1, 2, \ldots, K\} \) as the output label space. Let \( D = \{(x_i, y_i)\}_{i=1}^{n} \) be the training instance image, where \( x_i \subseteq X \) denotes each training instance image, and \( y_i \subseteq Y \) denotes the ground-truth label of \( x_i \). Classifier \( f : X \rightarrow Y \) from \( D \), which aims to predict ground-truth labels from a set of candidate labels. The basic CELoss function is given by:

\[
CELoss(x, y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log(Y_{i,k})
\]

where \( N \) is the number of samples, \( K \) is the number of categories, \( y_{i,k} \) is the actual value, and \( Y_{i,k} \) is the predicted value.

### 4. Experiments

#### 4.1. Datasets

The datasets used in this study include a public dataset and a clinical dataset. The public dataset came from the 2017 Pediatric Bone Age Challenge organized by the Radiological Society of North America (RSNA), which includes 12.6 k images. Among them, there are 6833 images for males and 5778 images for females.

The clinical dataset came mainly from Xuzhou Central Hospital, which provides 2518 left hand X-ray maps of children under the age of 18. Firstly, we desensitized each image to remove sensitive information from the image content and protect the patients’ privacy. Secondly, a number of professional radiologists marked the pre-treated images with bone categories and bone grades to ensure the accuracy of these data. As shown in Figure 5, the labeling of each picture in the object detection dataset is composed of the category and coordinate information of 13 bones, including the ulna, the radius, the 1st, 3rd, and 5th metacarpal bones, the 1st, 3rd, and 5th proximal phalanges, the 3rd and 5th middle phalanges, and the 1st, 3rd, and 5th distal phalanges. On this basis, we cropped 13 ROIs and labeled them with image-level bone levels, resulting in 32,734 fine-grained classification datasets.
4.2. Training Details

The structure diagram of YOLOv5 we used is shown in Figure 4, in the bone class classification experiment. Using 2518 images labeled by the doctor, the training set and the test set were divided into 8:2, and Figure 6 shows the amount of data for each type of dataset. YOLOv5 with 100 epochs was trained using the pre-trained model YOLOv5s, and the parameters of the model were as follows: the learning rate was 0.01, and the batch size was 16.

The training used an NVIDIA Tesla K80 GPU, 12 GB of graphics memory, and 8 GB of running memory. The CPU model was an Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50 GHz. The operating systems were Ubuntu 18.04, CUDA 10.0, and cuDNN 7.6. The software versions were Python 3.5, Tensorflow 1.13.1, and Keras 2.2.

The classifier uses fine-grained classification based on context-aware attention pooling, trains 13 classifiers according to different bone categories, uses a dataset of 2518 images, uses the Xception pre-training network, and selects stochastic gradient descent (SGD) [28] to train the model. The training cycle was 100, the learning rate 0.0001, the momentum 0.9, and the batch size 8.
4.3. Results
4.3.1. Extraction of 13 ROIs

We selected 13 skeletons that conform to human prior knowledge as recognition positions. By observing the visual results of the training process, it was shown that YOLOv5 is capable of extracting ROIs, which can ensure stable age assessment without additional foreground–background segmentation.

Figure 7 shows some of the losses during model training. The box_loss is the loss function of the prediction box, and the smaller the value, the more accurate the target box it; Obj_loss is the target detection, and the smaller the target detection, the more accurate the detection is; cls_loss is the classification loss, and the smaller the classification, the more accurate the classification is. Each of the results shown converged appropriately to 0; therefore, it can be concluded that the proposed model acquired as many ROIs as possible.

Figure 7. The loss of the model. The first line is the training loss, and the second line is the validation loss. The box_loss is the loss function of the prediction box; the obj_loss is the target detection; the cls_loss is the classification loss.

The mean average precision (mAP) results of the training are shown in Figure 8. Precision represents how many of the data predicted as positive samples are true positive samples. The recall is how many positive samples the model retrieves out of the total positive samples. The AP on the PR graph is the area below the PR curve, and the value range is from 0–1; for the accuracy of the model, the higher the recall, then the better the performance of the model, that is, the larger the AP, the better the performance of the model is. As shown in the figure, the mAP for the 13 categories of the model was 0.993, which indicates that the recommended model performs well.
4.3.2. Hierarchical Classification of Hand Bones

To verify the validity of the fine-grained classification model we used, we compared it with other classification models and output bone maturity level results. Table 1 shows the classification accuracy of bone grade classification using different classifiers on our clinical dataset. The results show that, compared with other classification methods, the model we used achieved the highest classification accuracy (the average was 86.93%) in bone grade, which further verifies the superiority of the proposed three-stage automatic pediatric bone age assessment method.

Table 1. Accuracy comparisons of different classification models on the clinical dataset.

| ROI No. | MLP-Mixer [29] | PMG [30] | WS-DAN [31] | CAP |
|---------|----------------|----------|--------------|-----|
| 1       | 66.59%         | 79.12%   | 80.07%       | 98.79% |
| 2       | 66.25%         | 75.50%   | 73.28%       | 94.15% |
| 3       | 51.30%         | 71.20%   | 69.52%       | 97.17% |
| 4       | 46.73%         | 62.00%   | 63.59%       | 83.87% |
| 5       | 59.69%         | 66.63%   | 64.57%       | 91.80% |
| 6       | 54.50%         | 49.19%   | 59.30%       | 87.60% |
| 7       | 55.32%         | 64.56%   | 62.61%       | 82.54% |
| 8       | 54.10%         | 63.35%   | 62.23%       | 85.74% |
| 9       | 58.52%         | 63.91%   | 61.85%       | 84.38% |
| 10      | 51.05%         | 60.39%   | 61.70%       | 83.13% |
| 11      | 52.96%         | 61.92%   | 60.41%       | 81.05% |
| 12      | 55.12%         | 60.32%   | 61.81%       | 79.23% |
| 13      | 58.07%         | 61.92%   | 59.37%       | 80.65% |
| AVG     | 56.17%         | 64.62%   | 64.64%       | 86.93% |

4.3.3. End-to-End Automated BAA

The evaluation index MAE of the bone age assessment method: The MAE formula is defined as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |R_i - P_i|}{N}$$

(2)

where $N$ is the number of pictures in the test set, $R_i$ represents the true bone age of the $i$-th picture, and $P_i$ represents the predicted bone age of the $i$-th picture.
The results of this study were compared with those of the BAA algorithm published in the existing literature and are shown in Table 2. The results showed that the three-stage hierarchical assessment method based on object detection and fine-grained classification proposed by us is superior to the existing methods, and the accuracy is much higher than that of the traditional manual method of assessing bone age.

Table 2. Comparison with different methods on the RASA dataset.

| Evaluation Index       | MAE (Months) |
|------------------------|--------------|
| Wu et al. [32]         | 7.38         |
| Steenkiste et al. [33] | 6.80         |
| Iglovikov et al. [34]  | 6.89         |
| BoneXpert [35]         | 8.16         |
| YOLOv5+CAP [35]        | 6.53         |

5. Discussion

In this study, our proposed three-stage BAA method achieved relatively satisfactory results on both the clinical dataset and the public dataset RSNA. In the first stage, YOLOv5 was used to extract 13 ROIs, and the detection accuracy as high as 98.93%. Most of the studies used Faster R-CNN in the extraction of ROIs, and these studies all had the phenomenon of missing ROI detection [36,37]. In the second stage, CAP was used to perform 13 bone grade staging, and the average accuracy was 86.93%, which was much higher than the 63.15% accuracy achieved by the recursive feature pyramid network proposed by Jia et al. [11]. In the third stage, the TW3 method was used to estimate bone age, which is more accurate than building a regression model of bone age.

At present, most BAA studies based on deep learning only provide an end-to-end bone age assessment result. However, we learned through communication with doctors that radiologists pay more attention to the details of hand bone changes than a simple bone age. Our BAA system not only provides bone age, but also provides the detection results of 13 ROIs, which is convenient for doctors to further observe the information and provide help for the next diagnosis.

In our research, we found that in the dataset samples provided by the hospital, the sample size of 0–5 years old is much lower than that of 6–18 years old, which leads to uneven sample distribution. The imbalance of the sample is likely to affect the results of bone age detection, making the bone age assessment value of 0–6 years old inaccurate. Our next plan is to design a more suitable framework to solve the bone age assessment of imbalanced samples based on the existing basis, so as to make the bone age assessment more accurate.

6. Conclusions

Bone age assessment (BAA) based on X-ray imaging of the left hand and wrist can accurately reflect the degree of the body’s physiological development and physical condition. However, the traditional manual evaluation method relies too much on inefficient specialist labor. In this paper, we designed a three-stage hierarchical approach (consisting of object detection, fine-grained classification, and TW3 computation) to implement end-to-end BBA. The framework can obtain detailed information such as bone age values, 13 ROIs, and the maturity level of each bone, which is very good for the needs of doctors. The BAA system has high bone age estimation accuracy and computational speed block and achieved 86.93% classification accuracy and an MAE of bone age of 7.68 months on the clinical dataset, as well as a 6.53-month MAE on the public dataset, to meet clinical needs.

Author Contributions: Conceived of and designed the study, X.X. and H.X.; collected the data, X.X. and H.X.; analyzed the data and involved in writing the manuscript, X.X., H.X. and Z.L. All authors have read and agreed to the published version of the manuscript.
Funding: This work was supported by the National Natural Science Foundation of China (No. 61976217), the Fundamental Research Funds of Central Universities (No. 2019XKQYMS87), and the Science and Technology Planning Project of Xuzhou (No. KC21193).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Lu, Y.; Zhang, X.; Jing, L.; Fu, X. Data Enhancement and Deep Learning for Bone Age Assessment using The Standards of Skeletal Maturity of Hand and Wrist for Chinese. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico City, Mexico, 1–5 November 2021; pp. 2605–2609.

2. Siebler, T.; Robson, H.; Shalet, S.M.; Williams, G.R. Glucocorticoids, thyroid hormone and growth hormone interactions: Implications for the growth plate. *Horm. Res. Paediatr.* **2001**, *56*, 7–12. [CrossRef] [PubMed]

3. Gann, S.M. Radiographic atlas of skeletal development of the hand and wrist. In *Am. J. Hum. Genet.;* 1959, 11, 282.

4. Satoh, M. Bone age: assessment methods and clinical applications. *Clin. Pediatr. Endocrinol.* **2015**, *24*, 143–152. [CrossRef] [PubMed]

5. Mughal, A.M.; Hassan, N.; Ahmed, A. Bone age assessment methods: A critical review. *Pak. J. Med. Sci.* **2014**, *30*, 211. [CrossRef] [PubMed]

6. 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. [CrossRef] [PubMed]

7. Lee, H.; Tajmir, S.; Lee, J.; Zissen, M.; Yeshiwas, B.A.; Alkasab, T.K.; Choy, G.; Do, S. Fully automated deep learning system for bone age assessment. *J. Dig. Imaging* **2017**, *30*, 427–441. [CrossRef] [PubMed]

8. Son, S.J.; Song, Y.; Kim, N.; Do, Y.; Kwak, N.; Lee, M.S.; Lee, B.D. TW3-based fully automated bone age assessment system using deep neural networks. *IEEE Access* **2019**, *7*, 33346–33358. [CrossRef]

9. Larson, D.B.; Chen, M.C.; Lungren, M.P.; Halabi, S.S.; Stence, N.V.; Langlotz, C.P. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* **2018**, *287*, 313–322. [CrossRef] [PubMed]

10. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

11. 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. [CrossRef]

12. Liu, R.; Zhu, H.; Wang, L.; Han, B.; Du, J.; Jia, Y. Coarse-to-fine segmentation and ensemble convolutional neural networks for automated pediatric bone age assessment. *Biomed. Signal Process. Control* **2022**, *75*, 103532. [CrossRef]

13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.

15. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]

16. Girshick, R. Fast r-cnn. In *Proceedings of the 2015 IEEE Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Montreal, QC, Canada, 7–12 December 2015.

18. Zhang, Z.; Zhang, L.; Wang, Y.; Feng, P.; Sun, B. Bidirectional Parallel Feature Pyramid Network for Object Detection. *IEEE Access* **2022**, *10*, 49422–49432. [CrossRef]

19. Wang, J.; Song, L.; Li, Z.; Sun, H.; Sun, J.; Zheng, N. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 20–25 June 2021; pp. 15849–15858.

20. Wang, Y.; Wang, Z. A survey of recent work on fine-grained image classification techniques. *J. Vis. Commun. Image Represent.* **2019**, *59*, 210–214. [CrossRef]

21. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 834–849.

22. Wei, X.S.; Xie, C.W.; Wu, J.; Shen, C. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recogn.* **2018**, *76*, 704–714. [CrossRef]

23. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
24. Liu, C.; Huang, L.; Wei, Z.; Zhang, W. Subtler mixed attention network on fine-grained image classification. *Appl. Intell.* 2021, 51, 7903–7916. [CrossRef]

25. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.

26. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.

27. Behera, A.; Wharton, Z.; Hewage, P.R.; Bera, A. Context-aware attentional pooling (cap) for fine-grained visual classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 929–937.

28. Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 928–936.

29. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* 2021, 34, 24261–24272.

30. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 153–168.

31. Hu, T.; Qi, H.; Huang, Q.; Lu, Y. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv* 2019, arXiv:1901.09891.

32. Wu, E.; Kong, B.; Wang, X.; Bai, J.; Lu, Y.; Gao, F.; Zhang, S.; Cao, K.; Song, Q.; Lyu, S.; et al. Residual attention based network for hand bone age assessment. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1158–1161.

33. Van Steenkiste, T.; Ruyssinck, J.; Janssens, O.; Vandersmissen, B.; Vandecasteele, F.; Devolder, P.; Achten, E.; Van Hoecke, S.; Deschrijver, D.; Dhaene, T. Automated assessment of bone age using deep learning and Gaussian process regression. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 674–677.

34. Iglovikov, V.I.; Rakhlin, A.; Kalinin, A.A.; Shvets, A.A. Paediatric bone age assessment using deep convolutional neural networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 300–308.

35. Rijn, R.V.; Thodberg, H. Bone age assessment: automated techniques coming of age? *Acta Radiol.* 2013, 54, 1024–1029. [CrossRef] [PubMed]

36. Bai, T.D.; Lee, J.J.; Shin, J. Incorporated region detection and classification using deep convolutional networks for bone age assessment. *Artif. Intell. Med.* 2019, 97, 1–8. [CrossRef] [PubMed]

37. Koitka, S.; Demircioglu, A.; Kim, M.S.; Friedrich, C.M.; Nensa, F. Ossification area localization in pediatric hand radiographs using deep neural networks for object detection. *PLoS ONE* 2018, 13, e0207496. [CrossRef] [PubMed]