applied sciences

MDPI

Article

A Global-Local Feature Fusion Convolutional Neural Network for Bone Age Assessment of Hand X-ray Images

Qinglei Hui [1,†], Chunlin Wang [2,†], Junwei Weng [3], Ming Chen [4] and Dexing Kong [1,*]

[1] School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China; qlhui@zju.edu.cn
[2] The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310030, China; hzwangcl@zju.edu.cn
[3] School Hospital, Zhejiang University, Hangzhou 310027, China; wjw88919212@zju.edu.cn
[4] Zhejiang Cancer Hospital, Chinese Academy of Sciences, Hangzhou 310022, China; chenming@zjcc.org.cn
* Correspondence: dxkong@zju.edu.cn
† These authors contributed equally to this work.

Abstract: Bone age assessment plays a critical role in the investigation of endocrine, genetic, and growth disorders in children. This process is usually conducted manually, with some drawbacks, such as reliance on the pediatrician's experience and extensive labor, as well as high variations among methods. Most deep learning models use one neural network to extract the global information from the whole input image, ignoring the local details that doctors care about. In this paper, we propose a global-local feature fusion convolutional neural network, including a global pathway to capture the global contextual information and a local pathway to extract the fine-grained information from local patches. The fine-grained information is integrated into the global context information layer-by-layer to assist in predicting bone age. We evaluated the proposed method on a dataset with 11,209 X-ray images with an age range of 4–18 years. Compared with other state-of-the-art methods, the proposed global-local network reduces the mean absolute error of the estimated ages to 0.427 years for males and 0.455 years for females; the average accuracy rate is within 6 months and 12 months, reaching 70% and 91%, respectively. In addition, the effectiveness and rationality of the model were verified on a public dataset.

Keywords: bone age assessment; deep learning; convolutional neural network; feature fusion; pediatric

Citation: Hui, Q.; Wang, C.; Weng, J.; Chen, M.; Kong, D. A Global-Local Feature Fusion Convolutional Neural Network for Bone Age Assessment of Hand X-ray Images. *Appl. Sci.* **2022**, *12*, 7218. https://doi.org/10.3390/app12147218

Academic Editors: Cristina Portalés Ricart, João M. F. Rodrigues and Pedro J. S. Cardoso

Received: 2 April 2022
Accepted: 25 June 2022
Published: 18 July 2022

1. Introduction

Bone age assessment (BAA) is a common method used for assessing skeletal maturity in pediatric radiology, which can be used to assess human growth potential by comparing the skeletal age with the chronological age. In a pediatric clinical, through a skeletal bone age analysis, combined with physical examination and testing, it is possible to find out the causes of the child's growth, development, and diseases (in time), and take effective intervention measures to obtain a good prognosis. BAA has been commonly used in clinical medicine, preventive medicine, sports science, biology, and forensic anthropology [1–3].

The previous clinical BAA approaches can be divided into two categories, including the Greulich-Pyle atlas method (GP) [4] and the Tanner-Whitehouse scoring method (TW) [5,6]. GP is a simple technique, which compares hand X-ray images with a list of standard bone age maps to predict bone age. However, this relies heavily on the physician's intuition and experience and may lead to different results from the same X-ray image at different times with poor consistency. In contrast, the TW scoring method is more objective and accurate in prediction, it can assess the maturity level of the region of interest (the ROI is a specific region selected from the image, delineating this region can reduce the processing time and improve task accuracy) of the hand and wrist. Scores are assigned to individual

ROIs according to maturity level; finally, all scores are combined to estimate bone age by means of a pre-defined strategy.

In the medical field, with the developments of computer vision technology in image analysis, deep learning methods particularly have powerful automatic feature extraction capabilities, which can model the input data and approach complex functions through the deep nonlinear network. Deep learning BAA techniques have developed rapidly and achieved considerable progress compared to traditional methods. These methods generally regard BAA task as a classification or regression problem, involving hand segmentation [7,8], ROI detection, feature extraction [9], and the design of regressors [10] or classifiers [11].

Despite some methods yielding accurate results in this field, BAA suffers from the following problems: (1) Reading these images is time-consuming, labor-intensive, and mainly depends on the experiences of radiologists. Moreover, since different experts may have different results, robust evaluations cannot be obtained, and they are prone to significant inter- and intra-observer variability. (2) Medical images are expensive to acquire compared to natural images and the quality of X-ray images may be poor for various reasons during the acquisition process. The images need to be labeled by professional radiologists. Therefore, the dataset of high-quality labels for BAA is very limited. (3) It is difficult for the performances of traditional assessment methods to meet the clinical needs; most of them require manual extraction of feature information, failing to satisfy automation. The manually extracted features are often not accurate enough or do not represent the essence of things well, making it difficult to improve the learning effect. (4) While deep learning methods can achieve leading results in BAA, they have a low ROI focus on hand bone images, and the prediction performance can still be further improved by adding some knowledge to the deep learning framework.

To address the above issues, we further studied the automatic feature extraction of hand bone images and propose a global-local feature fusion framework, including ROI detection, global-local fusion, and a bone age prediction network. The detection network involves extracting 18 key patches of hand bones; in the fusion network, global-local feature information is extracted by a convolutional neural network (CNN), followed by two fully connected layers to finally determine the age of the bone. In this paper, we aimed to mimic the workflow of radiologists and address the problem that deep learning methods for bone age assessments pay less attention to ROIs. The contributions of this paper are as follows:

1. Histogram equalization was adopted to improve image contrast and reduce the influence of low-quality X-ray images; label smoothing is proposed to reduce overfitting.
2. Object detection is combined with bone age assessment and YOLOv5 is used for detection to locate and extract ROIs in real-time. These ROIs are the ossification areas that a radiologist's workflow focuses on.
3. The global-local fusion network can effectively and efficiently integrate global context and local fine structures, yielding high-quality prediction. Either global or local information is proven to be indispensable.

The remainder of the paper is organized as follows. Section 2 briefly summarizes several CNN-based BAA systems. Section 3 explains the details of materials and resources. Section 4 describes the proposed model in detail. Experimental results are provided in Section 5 and further discussions are made in Section 6.

2. Related Work

Our review of the related research primarily focuses on deep learning-based approaches. In 2016, Stern et al. [12] firstly proposed a deep convolutional neural network (DCNN) BAA method using hand magnetic resonance imaging, which followed TW2 to obtain age information by integrating the ossification stages of 13 bones. Chen [11] reported the first CNN-based BAA method with transfer learning, which proposed a VGGNet-based BAA classification model, adopting the GP framework. Kashif [13] used sparse and dense feature points for key point selection. For each type, five feature descriptors

were extracted within the epiphyseal region of interest and classified using a support vector machine. Spampinato et al. [14] tested several CNN models, including OverFeat [15], GoogleNet [16], OxfordNet [17], and the proposed BoNet, consisting of five convolutional and pooling layers; they performed feature extraction for the first time for specific parts of the hand bones. Moreover, the results show an average error of 0.8 years in males. In 2017, Lee et al. [18] proposed a fully automated BAA system. First, it was segmented to remove the background interference, and then the bone age was recognized by GoogLeNet [16] with the whole picture as the input. The experimental results show that the effect of transfer learning is slightly worse than that of the BoNet network, mainly because the features of a hand bone X-ray have less commonality with natural images. For example, translation invariance and deformation invariance in natural images do not exist in X-ray hand images, and the corresponding bone age will change if local deformation occurs in the hand.

The following research studies focused on specific patches of hand bones, providing new thoughts for BAA. Iglovikov et al. [19] used U-Net [20] to segment the hand and a VGG [17] was used to detect three key points in order to remove the hand deformation and rotation interference. Taking the whole image as input, two network structures based on classification and regression were designed to evaluate bone age. Through the integration of the two networks, the mean absolute error (MAE) was obtained as 0.508 years. Wang et al. [21] proposed an automated BAA method that took the distal radius and ulnar regions as ROIs, extracted their features separately, and predicted the final bone age by classification. Bui et al. [22] combined the TW3 with DCNN, using Faster R-CNN [23] and Inception-v4 networks [24] for detection and classification, respectively. Mining expert knowledge from TW3 and engineering features from DCNN to improve the accuracy of BAA, the final MAE was 0.59 years. Liang et al. [25] took full consideration of the regions and proposed a novel deep automated skeletal BAA model via the region-based convolutional neural network. It transferred Faster R-CNN from object detection to bone age regression in order to detect the ossification centers of the epiphysis and carpal, causing MAE to be 0.51 years. Liu [26] proposed a multi-scale feature fusion framework based on a non-sub-sampled contour wave transform and convolutional neural networks. In particular, the method had significant advantages compared to the corresponding spatial domain methods, the MAE was 0.519 years for males and 0.553 years for females. Wu et al. [27] proposed a residual attention network, which used a Mask R- CNN subnet to segment the hand region and then used the residual attention network to focus on the key components of the image to generate the final prediction results, the MAE was 0.615 years. Chen et al. [28] extracted local binary pattern features and modifier subunit features of glutamate-cysteine ligase from an image. A support vector machine was used to classify the features and the author proposed a multi-dimensional data features fusion model for BAA. Its MAE was 0.455 years. Koitka [29] built an automated system for pediatric BAA that mimicked and accelerated the workflow of the doctor without breaking it. The MAE obtained by this method was 0.38 years. The system provided self-explanatory results for radiologists. Han [9] automatically extracted key features of the bone age of the left joint based on the residual network (ResNet) model [30], and automatically assessed the bone age using a convolutional neural network; the average absolute error was 0.455 years.

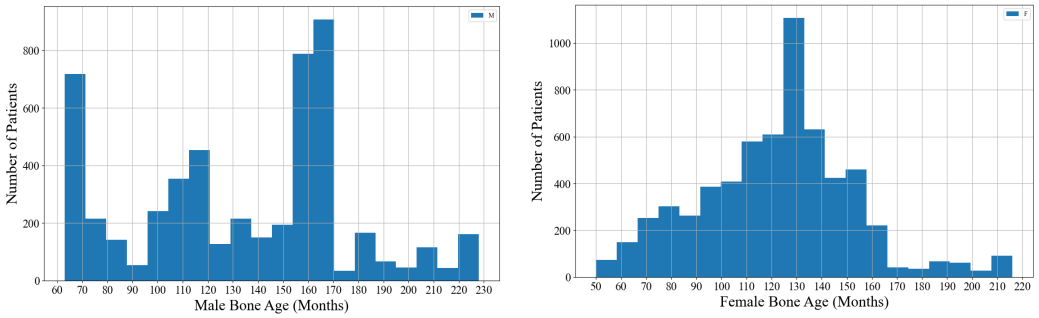
While deep learning has already achieved a leading performance in bone age assessment, prior methods usually took the whole X-ray image or cropped image as the input, and radiologists observed some key patches during the practical process of manual BAA. Inspired by this heuristic knowledge, we integrated object detection into BAA to detect these key patches and present a global-local feature fusion network; higher accuracy could be achieved by the combination of object detection and feature fusion.

3. Materials and Resources

3.1. Data Sources

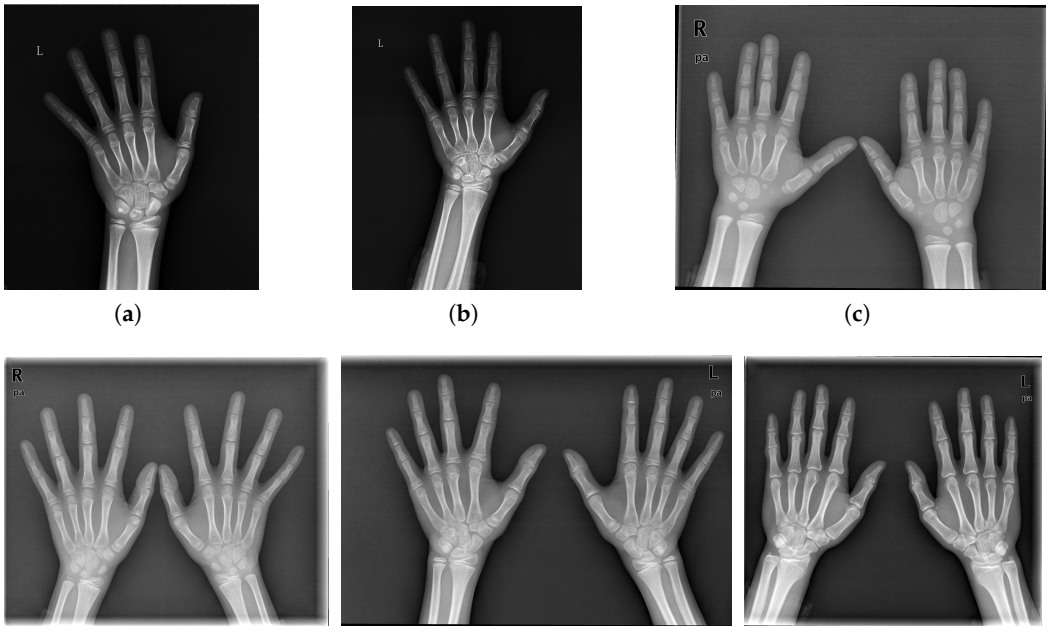
We used an in-house dataset to validate the performance in our own clinical environment and acquisition procedure. The dataset, from the First Affiliated Hospital of

Zhejiang University School of Medicine, contains 11,209 hand radiographs (male—5021, female—6188) from 4 to 18 years old (the ground truth is the average of two expert readings independently using TW3). The distribution of images among these categories is shown in Figure 1. The dataset contains two types of images, one is only the left hand, and the other is both hands. For several pictorial examples of the data, see Figure 2. Moreover, the reading labels are smoothed, e.g., if the label is 120 months, 120 ± 3 is used for training; all belonged to prediction accuracy in 117–123 months. Only the labels of the training set were softened, while the test set was not processed accordingly. Moreover, we used the Radiological Society of North America (RSNA) Pediatric Bone Age Challenge dataset, which included 14,236 hand radiographs to validate the model.



(**a**) (**b**)

Figure 1. The distribution of examples in this dataset based on the ground truth bone age. (**a**) Age distribution of the male group. (**b**) Age distribution of the female group.



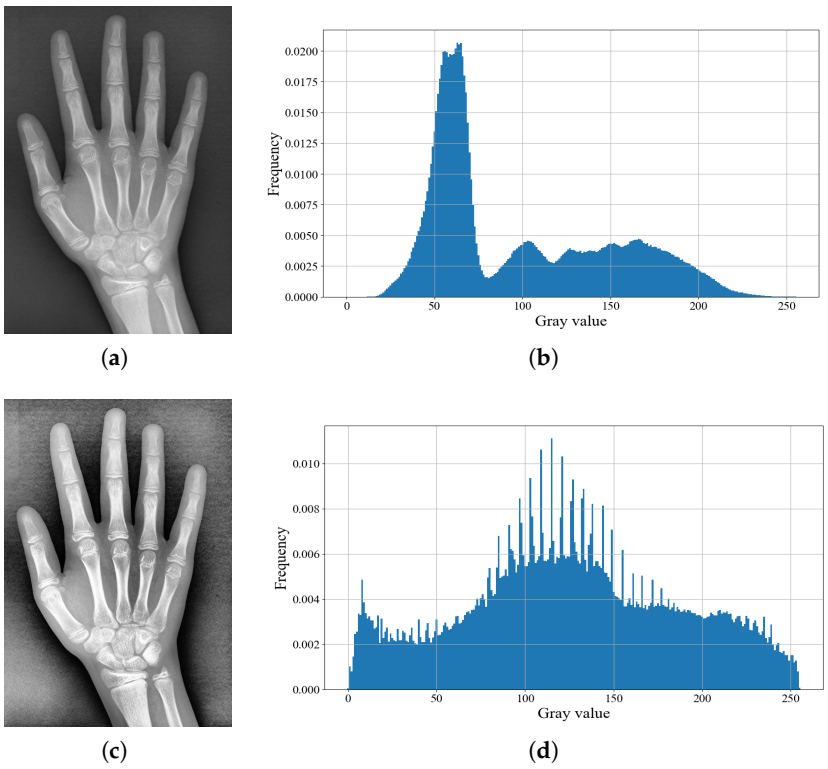
(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 2. Examples of raw datasets. "L" marks the left hand position, "R" marks the right hand position; (**a**) 10-year-old; (**b**) 13.5-year-old; (**c**) 5-year-old; (**d**) 6-year-old; (**e**) 13-year-old; (**f**) 18-year-old.

By analyzing the gray value of images, it was found that the gray distribution was inhomogeneous, which led to poor feature extraction and had a negative impact on training results. Therefore, this paper uses the histogram equalization method [31] to enhance the contrast of images. Figure 3 shows the comparison before and after histogram equalization; it can be seen that the gray value of the image is more uniformly distributed in the entire gray range, and the overall contrast is significantly enhanced, which is convenient for fea-

ture extraction. After equalization, data augmentation flips, mirrors, and rotates operations on the images.



(**a**) (**b**)

(**c**) (**d**)

Figure 3. Effect of the histogram equalization operation. (**a**) The original image; (**b**) original histogram; (**c**) the improved image by histogram equalization; (**d**) histogram after processing.

3.2. Deep Learning Libraries and Computing Resources

All the experiments on network training conducted via the deep learning framework TensorFlow. The hardware environment mainly includes an Intel Core i7-8700 CPU and an NVIDIA RTX 3090 GPU.

4. Methodology

The entire process framework is shown in Figure 4. After equalization and data augmentation, an ROI detection network is used to identify ossified regions of the hand bone. These local patches are leveraged as input to the local network. Convolutional neural networks are then used for global and local networks to extract deep features from the whole image and local patches. The following sections describe each step in more detail. In comparative experiments, we also used global and local features separately for bone age prediction.

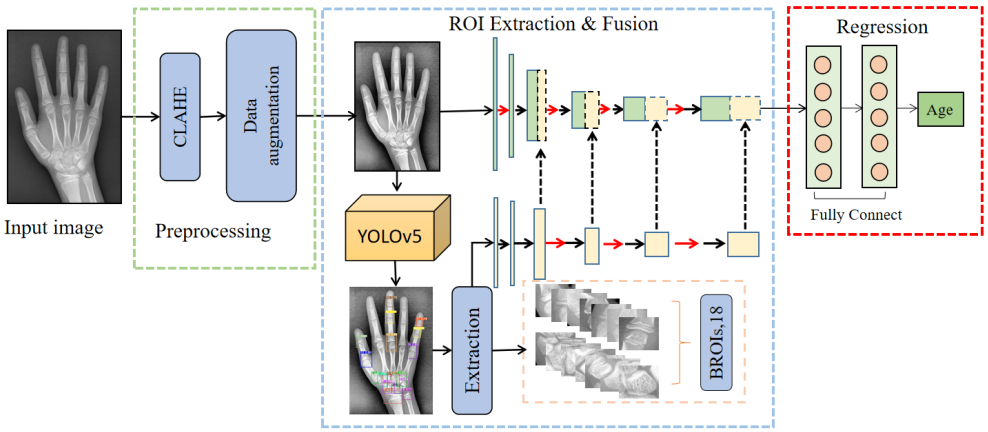
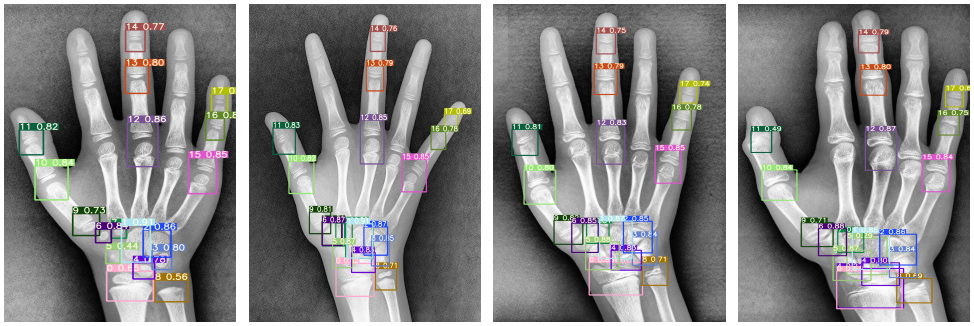


Figure 4. The overall process framework. It contains three major parts: (1) data preprocessing; (2) ROI extraction and feature fusion; (3) bone age prediction.

4.1. ROI Extraction-Detection Network

Object detection is a major direction of computer vision and is important in the fields of face recognition, unmanned vehicles, and security. Object detection algorithms based on deep learning can be divided into two categories. One is the two-stage detection paradigm, which divides the detection problem into two stages, first generates candidate regions, and then performs classification and location refinements. This class is represented by R-CNN algorithms, such as R-CNN [32], Fast R-CNN [33], and Faster R-CNN [23]. The other is the one-stage detection algorithm. Unlike the former, it directly generates the class category and location coordinate values of the object. Typical algorithms include Single Shot MultiBox Detector (SSD) [34] and You Only Look Once (YOLO) [35]. YOLO series algorithms [35–38] have a high detection speed; the latest YOLOv5 exceeds 140 fps, which make it possible for real-time object detection. In this paper, we chose YOLOv5 for hand bone ROI detection.

ROIs in TW3 can focus on information that is highly correlated with bone maturity. Therefore, based on TW3 evaluation criteria, 18 ROIs are detected and extracted by YOLOv5, which can ensure the trade-off of speed and accuracy. In this process, 500 images were stratified from the dataset according to the proportion of each age group, manually labeled, as well as 400 for training, and 100 for testing. The precision, recall, and F1 scores in the test set were 0.997, 0.993, and 0.996, respectively. The results of the ROI detection are shown in Figure 5. The number above the box indicates the confidence level of detection. It can also be seen that YOLO can still detect 18 ROIs even though the hand bones are immature at the age of 4. Figure 6 presents the ROI example for a 12-year-old hand bone image. Since there are images that contain both left and right hands in the dataset, and some images contain only the left hand, we performed the following operations: the raw data were passed through the detection network, and then expanded by 5 pixels according to the boundary coordinates of the 11th and 17th ROIs to crop out the entire left-hand skeleton, which was used as the input of the global network, and the corresponding 18 ROIs were used as the input of the local network.



(**a**) (**b**) (**c**) (**d**)

Figure 5. The features of 18 key regions based on (**a**) 4-year-old; (**b**) 8-year-old; (**c**) 12-year-old; (**d**) 15.5-year-old left-hand X-rays.

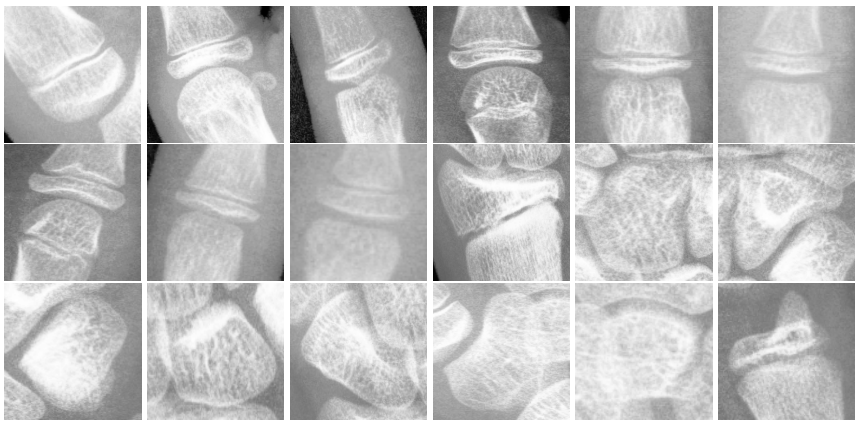


Figure 6. 18 ROI extraction thumbnails.

4.2. Global-Local Fusion Framework

In this section, our purpose is to extract and fuse local information into the global CNN and predict bone age. Figure 7 shows the architecture, which mainly consists of three modules: an input layer containing two types of inputs, a fusion network containing both global and local information, and a simple regression network. For the input layer, the entire left skeleton image is the global input while the local patches are the local input. In the fusion network, the global pathway and the local pathway are kept at the same convolution depth to ensure the consistency of information. After pooling and two fully connected layers, the final bone age result is the output.

Specifically, for the input layer, the left-hand image is scaled to 224 × 244 as the input of the global network; the local ROIs form a 64 × 64 × 18 3D matrix as the input of the local network. For both global and local networks, three-time convolution operations were first performed so that the feature size became 56 × 56 × 128. Afterward, the first feature fusion was performed to obtain a size of 56 × 56 × 256. Similarly, three more fusions were performed to make it to 7 × 7 × 2048, and then two more convolutions were performed to obtain 2 × 2 × 2048. The convolutional architecture in the fusion framework is shown in Table 1; it contains 11 convolutional layers with the corresponding batch normalization and activation layers (in total). In the regression prediction, via global max pooling, two dense layers were connected, including 512 neuron units, followed by a single-neuron layer that provided bone age prediction.

Table 1. The details of the fusion network architecture for BAA.

Convolution	Global	Local
Convolution 1	$7 \times 7 \times 64$, str = 2, pad = same	$5 \times 5 \times 64$, str = 1, pad = valid
Convolution 2	$7 \times 7 \times 128$, str = 2, pad = same	$5 \times 5 \times 128$, str = 1, pad = valid
Convolution 3	$7 \times 7 \times 128$, str = 1, pad = same	$3 \times 3 \times 128$, str = 1, pad = same
	Feature concatenation, Feature size: $56 \times 56 \times 256$	
Convolution 4	$3 \times 3 \times 256$, str = 2, pad = same	$3 \times 3 \times 256$, str = 2, pad = same
Convolution 5	$3 \times 3 \times 256$, str = 1, pad = same	$3 \times 3 \times 256$, str = 1, pad = same
	Feature concatenation, Feature size: $28 \times 28 \times 512$	
Convolution 6	$3 \times 3 \times 512$, str = 2, pad = same	$3 \times 3 \times 512$, str = 2, pad = same
Convolution 7	$3 \times 3 \times 512$, str = 1, pad = same	$3 \times 3 \times 512$, str = 1, pad = same
	Feature concatenation, Feature size: $14 \times 14 \times 1024$	
Convolution 8	$3 \times 3 \times 1024$, str = 2, pad = same	$3 \times 3 \times 1024$, str = 2, pad = same
Convolution 9	$3 \times 3 \times 1024$, str = 1, pad = same	$3 \times 3 \times 1024$, str = 1, pad = same
	Feature concatenation, Feature size: $7 \times 7 \times 2048$	
Convolution 10	$3 \times 3 \times 2048$, str = 2, pad = valid	/
Convolution 11	$3 \times 3 \times 2048$, str = 1, pad = valid	/

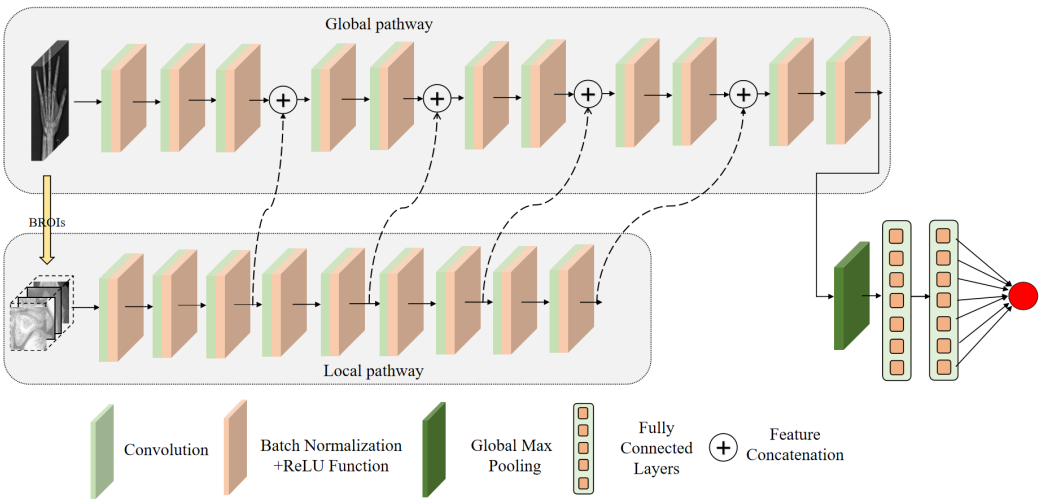


Figure 7. The framework of the global-local fusion network. The top is the global pathway to extract global information. The bottom is the local pathway to obtain local fine-grained details.

5. Experimental Results

5.1. Training Details

The distribution of training, validation, and testing set data is shown in Table 2. For the ROI detection network, batch size, IoU threshold, and epoch were set to 64, 0.5, and 100, respectively, the initial learning rate was 0.01, and the oneCycleLR strategy was used; the final learning rate was 0.001. There were three sizes of anchors by default, which were [10, 13, 16, 30, 33, 23], [30, 61, 62, 45, 59, 119], and [116, 90, 156, 198, 373, 326]. For the global-local fusion network, the batch size and maximum epoch were set to 64 and 100, respectively. Adam was selected as the optimizer; the initial learning rate was 0.001 with 1/3 decay for every 10 epochs and the activation function was ReLU. Early stopping with patience of 10 epochs was used, which means that the training was terminated automatically if the loss on the validation set did not improve for 10 epochs.

Table 2. Age distribution of training, validation, and testing datasets.

Age	Training Data			Validation Data			Testing Data		
(Years)	Male	Female	Total	Male	Female	Total	Male	Female	Total
[4, 6]	726	394	1120	69	32	101	73	50	123
(6, 8]	377	755	1132	37	95	132	35	96	131
(8, 10]	724	1289	2013	61	139	200	71	153	224
(10, 12]	414	1760	2174	41	200	241	35	217	252
(12, 14]	1627	572	2199	116	66	182	146	70	216
(14, 16]	224	146	370	19	17	36	23	18	41
(16, 18]	162	96	258	22	8	30	19	15	34
Total	4254	5012	9266	365	557	922	402	619	1021

For quantitative evaluation, the mean absolute error (MAE) and root mean square error (RMSE) are used in this paper, which are calculated as follows:

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y_i}|^2} \tag{2}$$

where y_i and $\hat{y_i}$ are the estimated bone age and the ground truth in the month, respectively.

Table 3 shows the MAE and RMSE of the proposed BAA system according to age groups. Most importantly, our model is able to effectively assess bone age with high accuracy for all genders and age ranges from 4 to 18 years old.

Table 3. Bone age prediction accuracy per age group.

Age	MAE			RMSE		
(Years)	Male	Female	Total	Male	Female	Total
[4, 6]	0.195	0.455	0.301	0.291	0.554	0.418
(6, 8]	0.54	0.424	0.455	0.69	0.542	0.585
(8, 10]	0.521	0.445	0.469	0.736	0.591	0.641
(10, 12]	0.8	0.43	0.482	0.992	0.557	0.635
(12, 14]	0.368	0.519	0.418	0.536	0.639	0.572
(14, 16]	0.518	0.616	0.561	0.637	0.818	0.722
(16, 18]	0.412	0.639	0.484	0.587	0.765	0.649
AVG.	0.427	0.455	0.444	0.617	0.588	0.599

Meanwhile, the identity lines for males and females are shown in Figure 8. The performance is quite steady for all age ranges and performs quite well on images centered around the mean but struggles with the lower and higher ranges.

In addition, we also introduced another effective curve-based evaluation method for BAA, named cumulative percentage accuracy curve, to calculate the cumulative accuracy of prediction error within 0 to 24 months. An example is shown in Figure 9, where the horizontal axis represents the month within the prediction error, and the vertical axis represents the accuracy rate. The prediction performance of the method can be seen more intuitively from the curve. The average accuracy rate within 6 months and 12 months reached 70% and 91%.

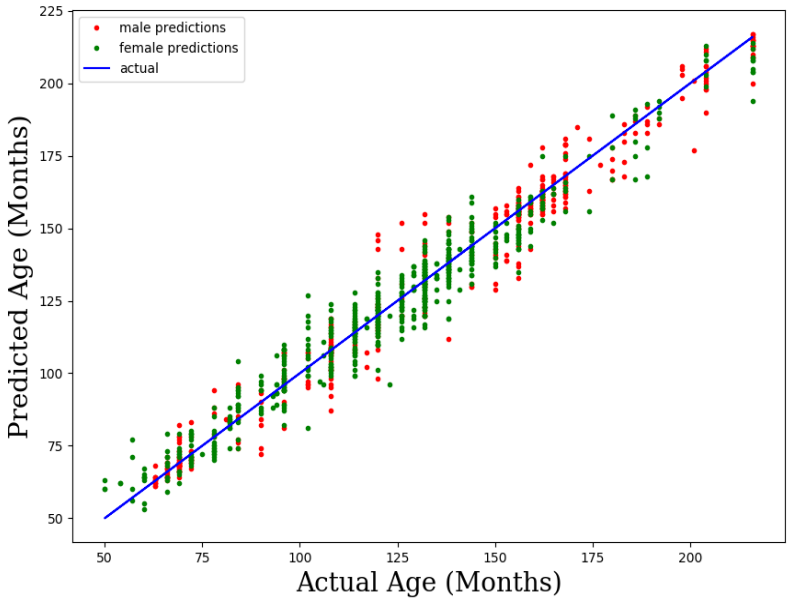


Figure 8. Identity lines for males and females.

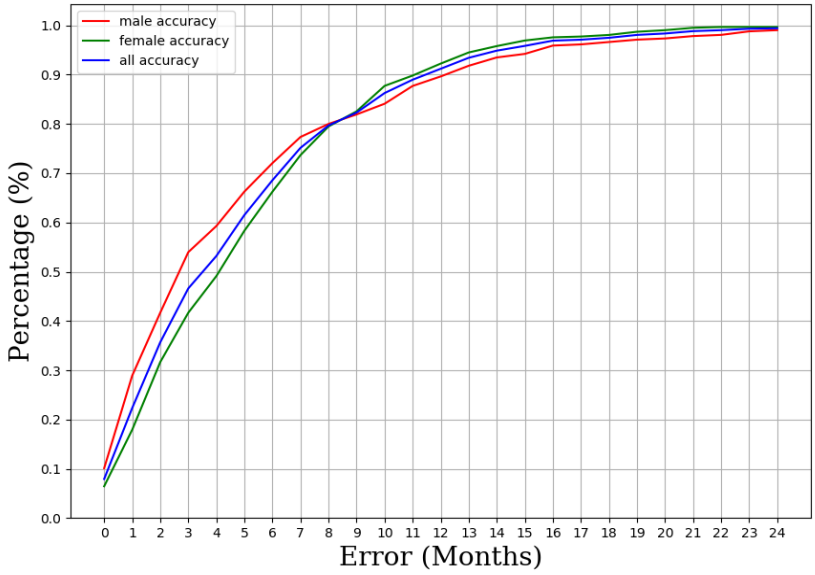


Figure 9. The cumulative percentage accuracy curve to evaluate BAA performance.

5.2. Comparison of Various Models

In order to evaluate the effects of histogram equalization and the global-local fusion strategy, we performed ablation experiments and gradually compared the performances of the model in these two aspects, see Table 4 for specific results. On the other hand, over the past two decades, many automated BAA methods have been proposed with accuracies (in MAE) ranging from 0.37 to 2.63 years. However, these methods are either tested on private datasets or their source codes are not available; thus, their results are not reproducible or usable as baselines. Here, we also selected several well-known bone age works, such as Spampinato [14], the RSNA children's bone age machine learning challenge [39], and the currently popular transformer method [9] for comparison. Moreover, we tested the performances of DenseNet, ResNet, and VGGNet in bone age assessment, respectively. Table 5 shows the comparison results of these methods.

Table 4. The ablation study results in terms of histogram equalization, global and local.

Histogram Equalization	Global	Local	RSNA	In-House
	✓		0.854	0.813
		✓	0.735	0.68
	✓	✓	0.712	0.658
✓	✓		0.68	0.6
✓		✓	0.53	0.467
✓	✓	✓	0.512	0.444

Table 5. Comparison in terms of average years between our method and state-of-the-art ones.

References	Methods	Number of Images	MAE
Spampinato [14]	BoNet + CNN	1390	0.8
Wang [21]	Faster R-CNN + Voting ensemble	3300	0.46
Bui [22]	RCNN + Inception-v4	1391	0.59
Liu [26]	Multi-scale fusion + VGG	1400	0.511
Halabi [39]	Inception-v3 + Dense	14,236	0.37
Thodberg [40]	Active appearance model plus principal component analysis	1559	0.42
Han [9]	ResNet plus spatial transformer	5876	0.455
DenseNet	/	/	0.85
ResNet	/	/	0.69
VGG	/	/	0.53
Ours	Global-local fusion plus CNN	11,209	0.444

6. Discussion

Bone age assessment is a specific and direct task for machine learning. In this study, we attempted to create a system for the fully automated estimation of bone age using pediatric X-ray hand bone images. Moreover, we showed that with a small labeled dataset, an ossification area detection network could be trained, which is stable enough to pave the way for the next step of local feature extraction. Meanwhile, the fusion of a global context and a local fine structure could effectively enhance bone age prediction (the final performance was 0.444 years).

Table 3 shows the prediction results of bone age for males and females in different age groups. It can be seen that the variance in the results in the male group is greater than that in the female group. In the experimental data, females obeyed the normal distribution, while the distribution of males was not balanced, indicating that the balance of the data was conducive to improving the stability of the model. In terms of data volume, the peaks of the male and female data volumes were at 13 years old and 11 years old, respectively. This is because boys start puberty at 12 years old while girls start at 10 years old. After about 1 to 2 years of growth and development, parents pay attention to their child's height. Since the experimental data in this paper were clinically-derived, the imbalance of distribution was unavoidable. So, in future research, we will continue to track the data to make them evenly distributed across age groups and improve the generalization ability of the model. In addition, it is worth noting that in each age group, the ages of (4, 6] were the best, and the ages of (14, 16] were the worst. From the image analysis, there were 8 ROIs concentrated in the wrist of the hand bone. In the early stage of child growth, due to the low degree of ossification development, the overlap between these blocks is low, the distinction is obvious, and it is easy for the network to learn local features, so the evaluation error is small. With growth and development, especially after puberty, skeletal ossification tends to mature, the degree of overlap increases, and some boundaries appear in two or more blocks. Meanwhile, the feature learning efficiency is low, and some feature redundancy exists, resulting in increased errors.

Models trained with global images or local patches may yield different results as to whether the image performs histogram equalization. This is because the image contrast changes and the models have different receptive fields, resulting in different suitable

training choices. Therefore, we carefully compared the models trained by these two ablation approaches (each using histogram equalization, global image, or local patches) and picked the best result. According to Table 4, after histogram equalization, the MAE improves from 0.813 to 0.6 for the global network, and the average improvement is about 0.2 years for the local and fusion networks. When the global-local feature fusion strategy is adopted, the predicted result improves from 0.6 to 0.444, an improvement of about 0.16 years. This ablation study proves that equalization, global, and local fusion can effectively collaborate to improve bone age assessment task results. The overall X-ray image is dark and the contrast is low, and histogram equalization can enhance the local contrast without affecting the overall contrast. After processing, the bone structure can be better displayed to obtain better details, which is friendly to network learning features. These local ROI features refer to the degree of development of the ossification centers of the phalanges, carpals, radius, and ulna, the criteria for TW3, which pediatricians focus on in the film reading.

From Table 5, Spampinato [14] was the first automated BAA work, they tested several network architectures for BoNet and then chose the best one as the final model for evaluation. The structure was mainly aimed at the extraction of middle and low-level visual features and had a non-rigid deformation denoising layer, but for X-rays, the noise was not a very fatal factor. Moreover, after many convolutions in deep learning, the feedback was more of deep-level image features; thus, the model lacks the learning and utilization of deep features, and cannot guarantee sufficient accuracy. Halabi [39] pointed out that all five winning algorithms used a preprocessing step, in which images were normalized or important anatomic areas were selected prior to algorithm training. Preprocessing seems to be an important component of the generalizability of the algorithm. The first place achieved an amazing effect of 0.37 years in the test set of 200 cases. This approach combined pixel and gender information and used an ensemble method to filter multiple high-performance models to improve the overall performance. However, the contestants claimed that there was a certain correlation in the testing set. Thoddberg [40] reconstructed bones using an active appearance model and predicted bone age based on shape, strength, and texture scores from the principal component analysis. It lacked robustness for the Chinese. Moreover, it requires high-quality X-ray images to obtain reliable results. In fact, it rejects images of poor quality or abnormal bone structure, for which cases, the analysis needs to be manual. It is a great challenge in clinical applications. Compared with the reconstructed bone image, the information of the raw image is more fidelity. Han [9] segmented hand bones based on morphological watersheds (downsampled by using bilinear interpolations on image pyramids); they used a spatial transformer to locate ROIs and retain positional parameters (attention mechanisms), and finally used ResNet as a backbone to predict the bone age. There was an error of 0.455 years, which reflected the strong learning ability of the transformer. The attention mechanism in BAA is less studied in the existing research, and Han's work illustrates the effectiveness of the attention mechanism in this task. Moreover, we tested DenseNet, ResNet, and VGGNet separately, see Table 5; the complex deep networks DenseNet and ResNet do not perform well; in contrast, the simple structured VGGNet turned out to be good.

The prediction results of the public dataset have also been improved to some extent, but the overall effect is lower than that of the in-house dataset. There are obvious racial differences between the two datasets. The public dataset is from the Children's Hospital Stanford and Colorado, involving many races, while the in-house dataset is for Chinese people. In terms of labels, the public dataset provides the real bone age by the GP atlas method, the latter adopts the TW3-Chinese version, and the real labels have methodological differences. These reasons lead to inevitable deviations in the model results.

Since our approach is designed to mimic how radiologists work, this of course has the inherent disadvantage that our method is limited to a specific workflow and image information of selected regions of the hand. It cannot be ruled out that an end-to-end network design with no assumptions at all will lead to more accurate results, for example,

by considering other regions of the image. Future work should study this in detail and integrate possible improvements through different network designs; we can also consider the fusion of multimodal data of X-rays, ultrasounds, and MRIs of hand bones; the self-attention mechanism of the transformer can be combined with CNN architecture to improve the accuracy of bone age prediction.

7. Conclusions

We presented a global-local feature fusion convolutional neural network for BAA. It has the advantages of short diagnosis times, it is labor-saving, and has high accuracy. This greatly reduces the workload of the physician and is highly reproducible. Our work integrates global and local information, which are shown to be effective at improving the accuracy of reading the bone age. Instead of directly employing the image as network input, 18 ROIs were firstly generated for feature pre-extraction. These ROIs were taken as the local features, then the global features of the whole image were extracted by convolution, overcoming the shortcomings of low utilization of the TW3 standard in deep learning. Experiments on an in-house dataset demonstrated the effectiveness of the proposed method. In comparison to other state-of-the-art deep learning networks, it exhibited generalization and adaptability. In future work, we will explore the potential of skeletal maturity and train the network according to the ranking criteria of each TW3 block of bones and present a visually interpretable form for radiologists to better understand the reasoning of the system. In addition, we will also consider the interpretability of deep learning.

Author Contributions: Conceptualization, C.W. and D.K.; methodology, Q.H., C.W. and D.K.; software, Q.H. and D.K.; validation, Q.H. and C.W.; formal analysis, D.K. and M.C.; investigation, Q.H. and C.W.; resources, C.W.; data curation, Q.H. and C.W.; writing—original draft preparation, Q.H.; writing—review and editing, J.W. and D.K.; visualization, Q.H. and J.W.; supervision, J.W.; project administration, M.C. and D.K.; funding acquisition, M.C. and D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Project of Zhejiang Province (grant number 2019C03003) and China Postdoctoral Science Foundation (grant number 2021M692834).

Institutional Review Board Statement: The study was approved by the Ethics Review Committee of the First Affiliated Hospital of Zhejiang University School of Medicine.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data sharing not applicable.

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

References

1. Martin, D.D.; Wit, J.M.; Hochberg, Z.; Sävendahl, L.; van Rijn, R.R.; Fricke, O. The Use of Bone Age in Clinical Practice—Part 1. *Horm. Res. Paediatr.* **2011**, *76*, 1–9. [CrossRef] [PubMed]
2. Widek, T.; Genet, P.; Ehammer, T.; Schwark, T.; Urschler, M.; Scheurer, E. Bone age estimation with the Greulich-Pyle atlas using 3T MR images of hand and wrist. *Forensic Sci. Int.* **2021**, *319*, 110654. [CrossRef] [PubMed]
3. Remy, F.; Saliba-Serre, B.; Chaumoitre, K.; Martrille, L.; Lalys, L. Age estimation from the biometric information of hand bones: Development of new formulas. *Forensic Sci. Int.* **2021**, *322*, 110777. [CrossRef] [PubMed]
4. Greulich, W.W.; Pyle, S.I. Radiographic atlas of skeletal development of the hands and wrists. *Am. J. Med. Sci.* **1959**, *238*, 393. [CrossRef]
5. Tann, E.J. *Assessement of Skeletal Maturity and Predicting of Adult Height (TW2 Method)*; Academic Press: Cambridge, MA, USA, 1983.
6. Morris, L.L. Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method): BOOK REVIEW. *Australas. Radiol.* **2003**, *47*, 340–341. [CrossRef]
7. Lin, H.; Shu, S.; Lin Y.; Yu, S. Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation. *Pattern Recognit.* **2012**, *45*, 322–332. [CrossRef]
8. Simu, S.; Lal, S. A study about evolutionary and non-evolutionary segmentation techniques on hand radiographs for bone age assessment. *Biomed. Signal Process. Control* **2017**, *33*, 220–235. [CrossRef]

9. Han, Y.; Wang, G. Skeletal bone age prediction based on a deep residual network with spatial transformer. *Comput. Methods Programs Biomed.* **2020**, *197*, 105754. [CrossRef]
10. Guo, J.; Zhu, J.; Du, H.; Qiu, B. A bone age assessment system for real-world X-ray images based on convolutional neural networks. *Comput. Electr. Eng.* **2020**, *81*, 106529. [CrossRef]
11. Chen, M. *Automated Bone Age Classification with Deep Neural Networks*; Stanford University: Stanford, CA, USA, 2016.
12. Štern, D.; Payer, C.; Lepetit, V.; Urschler, M. Automated Age Estimation from Hand MRI Volumes Using Deep Learning. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 194–202.
13. Kashif, M.; Deserno, T.M.; Haak, D.; Jonas, S. Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment. *Comput. Biol. Med.* **2016**, *68*, 67–75. [CrossRef]
14. 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]
15. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; Lecun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229v4.
16. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D. Going deeper with convolutions. *arXiv* **2015**, arXiv:1409.4842.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Lee, H.; Tajmir, S.; Lee, J.; Zissen, M.; Yeshiwas, B.A.; Alkasab, T.K. Fully Automated Deep Learning System for Bone Age Assessment. *J. Digit. Imaging* **2017**, *30*, 427–441. [CrossRef]
19. 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 International Publishing: Cham, Switzerland, 2018; pp. 300–308.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
21. Wang, S.; Shen, Y.; Shi, C.; Yin, P.; Wang, Z.; Cheung, P.W. Skeletal Maturity Recognition Using a Fully Automated System With Convolutional Neural Networks. *IEEE Access* **2018**, *6*, 29979–29993. [CrossRef]
22. Bui, T.D.; Lee, J.; Shin, J. Incorporated region detection and classification using deep convolutional networks for bone age assessment. *Artif. Intell. Med.* **2019**, *97*, 1–8. [CrossRef]
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
24. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
25. Liang, B.; Zhai, Y.; Tong, C.; Zhao, J.; Li, J.; He, X. A deep automated skeletal bone age assessment model via region-based convolutional neural network. *Future Gener. Comput. Syst.* **2019**, *98*, 54–59. [CrossRef]
26. Liu, Y.; Zhang, C.; Cheng, J.; Chen, X.; Wang, Z.J. A multi-scale data fusion framework for bone age assessment with convolutional neural networks. *Comput. Biol. Med.* **2019**, *108*, 161–173. [CrossRef]
27. Wu, E.; Kong, B.; Wang, X.; Bai, J.; Lu, Y.; Gao, F. 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.
28. Chen, X.; Li, J.; Zhang, Y.; Lu, Y.; Liu, S. Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age. *Future Gener. Comput. Syst.* **2020**, *10*, 795–801. [CrossRef]
29. Koitka, S.; Kim, M.S.; Qu, M.; Fischer, A.; Friedrich, C.M.; Nensa, F. Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. *Med. Image Anal.* **2020**, *64*, 101743. [CrossRef] [PubMed]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
31. Zuiderveld, K. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*; Academic Press: Cambridge, MA, USA, 1994; pp. 474–485.
32. Girshick, R.; Donahue, J.; Darrell, T. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 580–587.
33. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
34. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
36. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
37. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.

38. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
39. Halabi, S.S.; Prevedello, L.M.; Kalpathy-Cramer, J.; Mamonov, A.B.; Bilbily, A.; Cicero, M. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **2018**, *290*, 498–503. [CrossRef]
40. Thodberg, H.H.; Kreiborg, S.; Juul, A.; Pedersen, K.D. The BoneXpert Method for Automated Determination of Skeletal Maturity. *IEEE Trans. Med. Imaging* **2009**, *28*, 52–66. [CrossRef]