Automated Bone Age Assessment with Image Registration Using Hand X-ray Images

Mohd Asyraf Zulkifley 1,*,†, Siti Raihanah Abdani 1,† and Nuraisyah Hani Zulkifley 2

1 Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Selangor 43600, Malaysia; raihanah.abdani@siswa.ukm.edu.my
2 Community Health Department, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Selangor 43400, Malaysia; GS52834@student.upm.edu.my
* Correspondence: asyraf.zulkifley@ukm.edu.my; Tel.: +603-8921-6699
† These authors contributed equally to this work.

Received: 28 September 2020; Accepted: 12 October 2020; Published: 16 October 2020

Abstract: One of the methods for identifying growth disorder is by assessing the skeletal bone age. A child with a healthy growth rate will have approximately the same chronological and bone ages. It is important to detect any growth disorder as early as possible, so that mitigation treatment can be administered with less negative consequences. Recently, the most popular approach in assessing the discrepancy between bone and chronological ages is through the subjective protocol of Tanner–Whitehouse that assesses selected regions in the hand X-ray images. This approach relies heavily on the medical personnel experience, which produces a high intra-observer bias. Therefore, an automated bone age prediction system with image registration using hand X-ray images is proposed in order to complement the inexperienced doctors by providing the second opinion. The system relies on an optimized regression network using a novel residual separable convolution model. The regressor network requires an input image to be 299×299 pixels, which will be mapped to the predicted bone age through three modules of the Xception network. Moreover, the images will be pre-processed or registered first to a standardized and normalized pose using separable convolutional neural networks. Three steps image registration are performed by segmenting the hand regions, which will be rotated using angle calculated from four keypoints of interest, before positional alignment is applied to ensure the region of interest is located in the middle. The hand segmentation is based on DeepLab V3 plus architecture, while keypoints regressor for angle alignment is based on MobileNet V1 architecture, where both of them use separable convolution as the core operators. To avoid the pitfall of underfitting, synthetic data are generated while using various rotation angles, zooming factors, and shearing images in order to augment the training dataset. The experimental results show that the proposed method returns the lowest mean absolute error and mean squared error of 8.200 months and 121.902 months², respectively. Hence, an error of less than one year is acceptable in predicting the bone age, which can serve as a good supplement tool for providing the second expert opinion. This work does not consider gender information, which is crucial in making a better prediction, as the male and female bone structures are naturally different.

Keywords: skeletal bone age prediction; regression network; image registration; X-ray image

1. Introduction

It is important to identify growth disorder as early as possible, as it is normally treatable in the early stage, such as in the case of Lionel Messi, the world’s best footballer [1]. Through systematic hormonal therapy, he manages to avoid the short stature problem, which is a critical factor in being a
successful athlete. One of the methods used by the pediatricians to assess the growth level of a child is by using skeletal bone age assessment [2]. This test is used to identify any growth disorder, in which the skeletal bones might be overdeveloped or underdeveloped with regards to the child chronological age. The primary cause of this disorder can be attributed to the lack of nutrition, genetics-related disease, or problem in hormonal discretions [3]. Generally, bone age assessment is diagnosed through a radiography X-ray image of the left-hand region, which is the non-dominant hand. X-ray is a powerful imaging modality where it is even used in astronomy-related observations [4]. The age differences can be observed through an X-ray image, especially regions with the bone growth plates, where they will become thinner as a child grows older and totally disappears once he becomes an adult. In early 1960, the most popular assessment is done through the Greulich and Pyle (G&P) method [5], which subjectively compares the captured hand X-ray with a set of hand atlas images. In general, the reliability of this method is low due to the high bias in inter and intra-observers, where clinical experience plays a crucial factor in reporting the right assessment. Rather than looking at the whole X-ray image, the method by Tanner–Whitehouse (TW) [6] is looking at the specific regions on the X-ray image in order to reduce the assessment subjectivity [7]. Figure 1 shows the regions of interest of the TW method, where the primary areas are the epiphysis and metaphysis of carpal and phalanges bones. To be specific, only phalanges from three fingers are analyzed, which are thumb, middle and pinky fingers. The final bone age assessment is calculated through scores summation of all regions of interest.

Even though the observer bias is reduced through observing dedicated regions using the TW method, it still requires extensive experience for the doctors to accurately predict the bone age. Hence, a computer-assisted system will be a good supplement tool in helping them to do the prediction. In the early development of automated bone age assessment systems, conventional machine learning approaches such as neural networks (NN) [8] and support vector regression (SVR) [9] have been extensively used to predict the bone age while using the handcrafted features. These types of features are manually engineered, which are not optimally designed with regard to the problem. Hence, a deep learning approach has been proposed in [10] to optimally learn the unique features, which are then utilized for the skeletal bone age prediction based on hand X-ray images. The deep learning approach
mainly uses convolutional neural networks to extract the spatial information from various filters of various network levels, which is then classified or regressed while using dense neural network layers. This architecture has been successfully implemented in many biomedical applications that include eye disease detection [11], Alzheimer diagnosis [12], COVID-19 screening [13], physiotherapy [14], and cardiac analysis [15]. However, the input images will come in various sizes and conditions, where some images will be relatively small for the newborn baby and vice versa for the late teen case. Moreover, there is no standardization in the hand pose while capturing the X-ray images, where the hand might tilt at certain angles.

Therefore, we propose an image registration approach while using a composite function of separable convolutional neural network-based segmentation and keypoints detector to realign the images into a standard representation. The main advantage of the proposed approach is the lightweight nature of the networks that use separable convolution as the core operator, where the computational burden is lesser when compared to the full convolution architecture. The hand segmentation network is based on DeepLab V3 plus [16] and the keypoints regressor network is based on MobileNet V1 [17]. In the end, we apply a residual separable convolution scheme using Xception [18] network to train a regressor for an accurate bone age prediction that produces a mean absolute error value of fewer than 0.7 years. This network uses a unique three-layer residual separable convolution that reduces the probability of diminishing gradient issues during the training phase. Moreover, it is also in tune with the findings presented in [19], where the network should be designed, such that the depth and width of the filters are proportionally balanced to give the best classification or regression performance. The proposed system that utilized separable convolution extensively has been tested on X-ray images database that covers an age range of 1 to 228 months old. A large dataset allows for the network to learn better in generalizing the mapping between the X-ray images to the predicted age. Thus, several works [20,21] that are based on a small dataset will not be considered as the benchmark methods and, hence, all performance comparisons are benchmarked with the state-of-the-art deep learning networks. Data augmentation has also been applied to further improve the training process by using shearing, flipping, and contrast variation operations. Note that the proposed method does not utilize gender information in predicting the bone age, so that a more general prediction network can be produced. In some cases, gender information is not available to health practitioners due to privacy reasons. Therefore, a general network that is solely based on the X-ray image alone is more desirable. According to the report in [22], the best-performed methods have used both the X-ray images and gender information in designing their regression networks, which has also been validated on a small testing dataset of just 200 images. Contrary to this approach, this work divides the large dataset of 12,811 X-ray images into a training, validation, and testing set according to the ratio of 8:1:1, which is the standard data division strategy [23], which results in sizable testing data of 1281 images.

This work is organized into five sections. Section 2 discusses some conventional machine learning networks that have been applied to bone age assessment, followed by the convolutional neural networks approach. The radiography X-ray dataset of the hand used for training and testing the algorithm is described in Section 3. Subsequently, Section 4 explains the proposed image registration used to standardized and normalized the input X-ray images. Section 5 explains the applied Xception network regressor, which uses residual separable convolution to make an accurate prediction of the bone age. After that, Section 6 is dedicated to results discussion, where the proposed method performance is compared to the other state-of-the-art deep learning models. Concise conclusions and suggestions for future work are given in the last section.

2. Related Work

A recent systematic review conducted by Dallora et al. [24] shows that many of the automated bone age assessment methods still rely on conventional machine learning methods, where 22 out of the 26 methods have applied hand-crafted features extractor as an input to a regressor model. In [25], optimal fusion rules are analyzed in order to find the best combination weights for the 17 regions of
interest, where the extracted features are then passed to a least-square classifier. Rather than using a single method to extract the features, the work presented in [26] combines three feature extractors, which are histogram of oriented gradients, local binary pattern, and scale-invariant feature transform, which is then fed to a support vector machine (SVM) classifier. Although the extracted features cover various unique traits of the X-ray images, the computational burden is too high. In [9], the combination of SVM and SVR are used in order to complement the training process, where cross-correlation function between the tested regions of interest is used to produce the similarity scores. On the other hand, Gertych et al. [27] utilize wavelet features coupled with an 11-class fuzzy classifier to predict the bone age. The method that is presented in [28] combines both Fuzzy and neural networks (NN) classifiers to better predict the bone age using a combination of features that were extracted from each region of interest based on TW protocol. Contrary to the work in [28], Kashif et al. [8] use a NN classifier to identify the best-handcrafted features for bone age assessment among these five keypoints detector; SIFT, SURF, BRIEF, BRISK, and FREAK. They found out that SIFT detector produces the lowest mean error when it is coupled with a NN classifier. Instead of using iterative training process in the NN classifier, the work in [29] opts to use an extreme learning machine to train a single hidden NN using Moore–Penrose generalized inverse matrices. The networks between the input and hidden nodes are randomly assigned and remain fixed, while only the network between the hidden nodes and output is trained while using a single step update.

Spampinato et al. [10] is one of the earliest works that uses convolutional neural networks (CNN) in predicting the bone age by introducing their BoNet architecture. They have experimented with various compact CNN architectures, where the best prediction is obtained using a set of five layers CNN of 96, 2048, 1024, 1024, and 1024 filters. Usually, a compact CNN network consists of three to five CNN layers at most [30], and it is usually first pre-trained in a different domain to reduce the possibility of overfitting [31]. BoNet does not apply any pre-processing step, which will face difficulty if the input images are not captured in a standard orientation and scale. Therefore, the work in [32] has applied U-Net architecture to segment the hand regions from the background. The VGG-16 network is then used to perform regression on the bone age assessment. They have assessed three variants of regions of interest, which are the whole X-ray image, carpal region, and phalanges regions, where the whole X-ray image mode returns the lowest mean absolute error. Similarly, the work in [33] has also applied a combination of U-Net segmentation and VGG-16 architecture for bone age prediction. However, the VGG-16 architecture is applied in the form of classifiers instead of regressors, where the network is trained for a large number of output classes (240 classes). Zhao et al. [34] then improves on the network training process, where they use paced transfer learning where the regressor network is tuned from the top layer first until the last bottom layer. The performance of several deep learning architectures that include AlexNet, GoogleNet, and VGG-16 are analyzed in [35] by using a normalized and resized input images. They have generated synthetic data to augment their training dataset using geometric transformations, photometric transformations, noise injections, and color jittering. Mask R-CNN is used in the work by Wu et al. [36] in order to segment the hand from an X-ray image, where a residual attention network is then used to predict the bone age. It is a unique end-to-end approach, where the segmentation and regression networks are trained together, not separately. Instead of looking at the whole hand region, the work in [37] locates each of the 13 regions in the TW protocol and classifies them into one of the age categories using VGG-16 architecture. The final bone age prediction is obtained by adding up the scores from all 13 regions.

3. X-ray Image Dataset

The dataset used to validate the automated bone age prediction system is obtained from the Radiological Society of North America (RSNA) Pediatric Bone Age Machine Learning Challenge [22]. The total number of hand X-ray images is 14,236, which has been arranged into three classes of training, validation, and testing sets with 12,611, 1425, and 200 images for each respective class. All of the X-ray images are stored in the Portable Network Graphic format, where the highest resolution
The dataset was collected from two children hospitals, which are Children’s Hospital Colorado and Lucile Packard Children’s Hospital in the United States of America. The age range for the dataset starts from one month old until 228 months old. The number of male samples is slightly higher as compared to the female samples where male samples make up around 54.13% out of the total samples. Gender information is annotated for each of the X-ray image samples, but we omit this information in our research, so that a more general regressor can be developed. Six medical practitioners have been tasked to annotate the ground truth age, where the weighted mean is used to calculate the final age values. This RSNA dataset poses various challenges for an automated system because of the heterogeneous contrast values and non-standard pose images, as shown in Figure 2. Furthermore, some sample images are also much smaller as compared to the others, as the size is not normalized beforehand.

![Figure 2](image)

*Figure 2.* The first-row images depict some hand X-ray images taken with various contrast values. The second-row images depict some hand X-ray images taken from various tilting pose angles.

Because our proposed method utilizes image registration to standardize the X-ray images, the original validation dataset will be omitted as it is used to create the training set for our hand segmentation and four keypoints detector networks. Thus, the total number of X-ray images that are utilized in our experiment is reduced to 12,811 images, which is then divided into three classes of training, validation, and testing sets according to the ratio of 8:1:1, respectively. The original 200 testing images, as proposed by the RSNA challenge, are not ideal for performance comparison, as it is too small when compared to the total number of images in the training dataset.

4. Image Registration

Image registration is the transformation process to geometrically align all the images into a standardized form [38], where the segmented hand region will be positioned in the middle of the image with an upright position, where the angle between the middle finger and the horizontal axis is 90°. It is an important step so that a regressor can be trained using a standardized pose which will put more emphasis on the age differences, rather than focusing on the variations of the X-ray images. In this work, three operations are performed in order to register the image into a standardized representation, which are segmentation, rotation, and alignment, as shown in Figure 3. Let define a set of input images, X to be \( \{x_1, \ldots, x_N\} \in \mathbb{R}^4 \), where an image of \( x_i \in \mathbb{R}^3 \) is a vector comprises of its width (\( w_i \)), height (\( h_i \)), and channel (\( c_i \)) information. One of the novelties of this work is the application of separable convolutional neural networks in performing the image segmentation and rotation. It is a factorized form of the standard convolution into two operators, which are depthwise
and pointwise convolutions. A standard two-dimensional convolutional neural network operation with a kernel $K \in \mathbb{R}^3$ with a size of $k_x \times k_y \times k_z$ can be written as

$$V(m, n, o_{out}) = \sum_{m', n', o_{in}} X(m', n', o_{in})K(m', n', o_{in})$$  \hspace{1cm} (1)$$

where $k_x = k_y$ is used throughout this work and $(m, n) = \{0, \ldots, k_x\} \times \{0, \ldots, k_y\}$. $o$ represents the number of channel either for input or output layer with $o_{in} \in \{0, \ldots, k_z\}$. Subsequently, a two-dimensional separable convolution can be formulated, as follows

$$P(m, n, o_{in}) = \sum_{m', n'} X(m + j, n + k, o_{in})K(m, n, o_{in})$$  \hspace{1cm} (2)$$

and

$$Q(m, n, o_{out}) = \sum_{m', n'} X(m, n, o_{in})K(0, 0, o_{out})$$  \hspace{1cm} (3)$$

where $P$ represents the separable convolution operation and $Q$ represents the pointwise convolution operator, which can be linked to the standard convolution in the form of

$$\hat{V}(m, n, o_{out}) = P(m, n, o_{in})Q(m, n, o_{out})$$  \hspace{1cm} (4)$$

Hence, the registered image, $\hat{X} \in \mathbb{R}^4$ is a composite function of three operations, which are masked segmentation ($I_{segment}$), rotation alignment through keypoints regression ($I_{rotate}$), and positional alignment ($I_{align}$).

$$\hat{X} = I_{align}(I_{rotate}(I_{segment}(X)))$$  \hspace{1cm} (5)$$

The first step goal, which is a masked segmentation, is to extract the hand region from the background. The output of $I_{segment}$ function $X_1$ is an element-wise multiplication between the input image $X$ and the masked label $X_{mask}$ that was generated from the segmentation network.

$$X_1 = X \odot X_{mask}$$  \hspace{1cm} (6)$$

Figure 3. Algorithm flow for the proposed image registration module.

This operation will remove all background objects that include identification plates, medical tubes and equipment, noises, supporting tools, etc. from affecting the bone age prediction accuracy. DeepLab V3 plus [16] is used as the segmentation network, which uses Xception-65 [18] as its backbone. The full segmentation architecture follows the encoder-decoder format, where the encoder network comprises of Xception-65 and atrous spatial pyramid pooling (ASPP) module. On the other hand, the decoder network utilizes two up-sampling steps through a composite operation of separable convolutional layer and bilinear interpolation. The ASPP module consists of three parallel down-pooling branches with three different dilation rates, $D = 6, 12, 18$ and a residual network as used in [39]. This network requires a total number of 25,574,952 parameters that is trained using Adam optimizer [40] with a fixed learning rate of 0.001. Note that we have used the original validation dataset in [22] to train the segmentation network. Some of the output samples of the masked operation are shown in Figure 4.
Figure 4. Segmented hand region while using DeepLab V3 plus; the first-row samples depict the original input images, the second-row samples depict the segmented mask, and the third-row samples depict the masked hand regions.

Because the segmentation goal is to divide each image into two classes of background and foreground, a binary cross-entropy loss function is used as the objective function to optimize the segmentation accuracy. Let us define the predicted label from the segmentation network as $l_{\text{predict}}$ and the ground truth label as $l_{\text{gt}}$, then the loss function $L_{\text{segment}}$ is formulated, as follows.

$$L_{\text{segment}}(l_{\text{predict}}, l_{\text{gt}}) = -l_{\text{gt}} \cdot \log(l_{\text{predict}}) - (1 - l_{\text{gt}}) \cdot \log(1 - l_{\text{predict}})$$  \hspace{1cm} (7)

Subsequently, the second part of the image registration is to rotate the masked image $X_1$ into a standardized upright position of $X_2$ while using a rotation operation based on angle $\theta$ extracted from the tip of the middle finger and the bottom middle point of the carpal region, as shown in Figure 5a. The goal of this rotation operation is to align the image, such that the segmented hand region will be positioned in an upright form, where the middle finger is in parallel with the vertical axis. Only two keypoints will be utilized for the angle calculation, while all four keypoints will be utilized for the later positional alignment. The positional alignment is performed using rigid translation transformation, so that any sensitive part that is crucial for bone age prediction will not be affected. The four keypoints are the tips of the thumb ($KP_1$), middle ($KP_2$), and pinky ($KP_3$) fingers, as well as the bottom middle point ($KP_4$) of the carpal region. Each $KP_i \in \mathbb{R}^2$ of point $i$ represents the coordinate information in the form of $(m, n)$. The angle is directly calculated, as follows.

$$\theta = \cos^{-1}\left(\frac{KP_3(n) - KP_4(n)}{\sqrt{(KP_2(n) - KP_4(n))^2 + (KP_2(m) - KP_4(m))^2}}\right)$$  \hspace{1cm} (8)
Figure 5. (a) shows the angle \( \theta \) used for the rotation operation and (b) shows the four keypoints used in training the regressor.

The four keypoints are automatically detected while using a separable convolutional neural networks regressor. MobileNet V1 architecture \[17\] has been chosen to be the regressor by replacing the last softmax activation function with a linear activation function. The network output will be in the form of a vector with eight elements, which represent the four coordinates of the points of interest. It is chosen because of its lightweight design that produces good regression performance at a relatively low computational burden. This selection is in line with the whole design of this work, where a separable convolution scheme is used in both of the steps in the image registration module, while the final age prediction network also employs this scheme extensively. This network utilizes a set of 13 layers of separable convolution that takes input images size of 224 \( \times \) 224 pixels. The regressor will be trained using mean squared error loss function, where \( Kp^{\text{predict}} \) is the predicted output coordinate from the network, while \( Kp^{\text{gt}} \) is the ground truth coordinate. Similar to the previous segmentation module, the regressor network is trained while using the original validation dataset from Halabi et al. \[22\]. Hence, the loss function for the keypoints regressor network can be written as

\[
\mathcal{L}_{\text{rotate}}(Kp^{\text{predict}}, Kp^{\text{gt}}) = \frac{1}{4} \sum_{i=1}^{4} ((Kp^{\text{predict}}_i(m) - Kp^{\text{gt}}_i(m))^2 + (Kp^{\text{predict}}_i(n) - Kp^{\text{gt}}_i(n))^2) \tag{9}
\]

Once again, an Adam optimizer with a fixed learning rate of 0.001 is used to train the network for 500 epochs. The resultant \( X_2 \) images will be passed through to a positional alignment process, so that the position of the hand will be mostly located in the middle region, without skewing towards the right or left side of the image. The alignment is done to ensure that all of the keypoints are symmetrically placed, where the horizontal alignment is done by pivoting the middle points between \( KP_1 \) and \( KP_3 \), such that it will overlap with the middle point of the image. Likewise, the vertical alignment is done by pivoting the middle point between \( KP_2 \) and \( KP - 4 \), so that it will overlap with the middle point of the image. Zero paddings will be added to the image so that the registered image will have the same size as the input image. Figure 6 shows some of the output samples from \( \hat{X} \) that have been transformed through the image registration procedures.
Figure 6. Some output samples of the X-ray image that have passed through the image registration process. The first-row samples illustrate the original input images, while the second-row samples illustrate the image that has been transformed.

5. Xception-41 Regressor

In this section, the bone age will be predicted by a deep learning regressor while using a set of X-ray input images \( \hat{X} \) that has been pre-processed into a standardized form. The chosen network will be based on the Xception architecture \([18]\) of the 41 CNN layers version. The primary goal of this work is to design a network that can generalize the bone age prediction by deploying a large number of training data. Because the total number of training, validation, and testing data is large in number, which is 12,811 images, a deep network is suitable to be implemented without worrying much about the overfitting problem. Therefore, a residual network scheme is crucial in overcoming the challenge of diminishing gradient while training a deep network. Furthermore, a deep network will use a large number of parameters if a standard convolution is used and, hence, an alternative of separable convolution scheme will be implemented in order to reduce memory consumption. Given the two criteria, Xception-41 architecture fits these needs perfectly, where it extensively applies residual separable convolution scheme through eight repeating modules.

The separable convolution is a factorized version of the standard convolution using a composite function of depthwise and pointwise convolutions. The bone age will be regressed by using the RSNA dataset that is divided according to the ratio of 8:1:1 for training, validation, and testing, respectively. The network will be trained using Adam optimizer with variable learning rates, \( LR \) using the following piecewise function that depends on the epoch value \( (ep) \), where its value is fixed for the first 100 epochs, and it will then linearly decrease for another 50 epochs.

\[
LR = \begin{cases} 
0.001 & \text{if } 0 < ep \leq 50 \\
0.0005 & \text{if } 50 < ep \leq 100 \\
LR - \delta_{LR} & \text{if } ep > 100 
\end{cases}
\]  
(10)

where \( \delta_{LR} \in \mathbb{R}^1 \) is the reduction gradient of the learning rate. The full Xception architecture consists of three modules, which are entry \((\psi_1)\), middle \((\psi_2)\), and exit modules \((\psi_3)\) as shown in Figure 7. Let the regressor output, \( R_{BA} \in \mathbb{R}^1 \) be formulated, as follows
Figure 7. The full architecture of the proposed method with three sub-networks, which are entry, middle, and exit modules. The middle module consists of eight units of residual separable convolution module.

\[ R_{BA}(\hat{X}) = \psi_3(\psi_2(\psi_1(\hat{X}))) \]  

(11)

The entry module \( \psi_1 \) takes an input image of 299 × 299 pixels, where it comprises two layers of standard convolution and three layers of separable convolution. Carrying over from the previous subsection definitions, a single-layer residual convolution can be formulated as

\[ \hat{V}(m, n, o_{out}) = \hat{V}(m, n, o_{in}) + X(m, n, o_{in}) \]  

(12)

Therefore, a two-layer residual convolution can be written as

\[ V(m, n, o_{out}) = \hat{V}(\hat{V}(m, n, o_{in})) + X(m, n, o_{in}) \]  

(13)

Using the combination of \( V \) and \( \hat{V} \), the output feature maps \( \hat{X} \) of \( \psi_1 \) can be represented by

\[ \hat{X}_{\psi_1} = \left( \prod_{i=1}^{2} V_i \prod_{j=1}^{3} M(V_j) \right) \hat{X} \]  

(14)

where \( M \) is a maximum pooling operator. The output feature maps \( \hat{X}_{\psi_1} \) will be reduced to \( \frac{1}{4} \) of the original \( \hat{X} \) size. Taking over from the entry module, the middle module consists of eight repetitive three-layer residual convolution networks. Using similar convention, as in Equation (13), the network can be formulated, as follows

\[ \hat{V}(m, n, o_{out}) = \hat{V}(\hat{V}(m, n, o_{in})) + X(m, n, o_{in}) \]  

(15)

The residual convolution network \( \hat{V} \) for the middle module will be using the same number of filters for all three separable convolutions, which is contrary to the popular approach in [41], where a bottleneck scheme is used with the middle layer will have the most number of filters. The residual or skip connection will maintain the same size as its input, which will be combined with the main branch using the addition operator, as shown in Figure 8. Thus, the output feature maps \( \hat{X}_{\psi_2} \) after processing the middle module \( \psi_2 \) is as follows

\[ \hat{X}_{\psi_2} = \left( \prod_{i=1}^{8} \hat{V} \right) \hat{X}_{\psi_1} \]  

(16)
Figure 8. The architecture of a three-layer residual separable convolution network.

Finally, the last exit module $\psi_3$ will predict the bone age using a linear activation function. Global average pooling operator $G$, which is introduced in [42] is used to reduce the feature maps size to $1 \times 1 \times 2048$. These 2048 feature vectors are then passed to a fully connected layer $N$, where the output directly represents the predicted bone age $R_{BA}$. The exit module is then formulated, as follows

$$R_{BA} = (M(V_j)\cdot(\prod_{i=1}^{2}V_i)\cdot G(N))\hat{X}_{\psi_2} \tag{17}$$

6. Results and Discussion

The regressor is trained for 150 epochs using an additional augmented dataset, which was produced using Keras generator. The synthetic data are generated by varying the zoom ranges, rotation angles, and translational positions, as well as flipping and shearing process. Table 1 shows the hyper-parameter setup of the regressor. Two performance metrics are used to validate the experimental test on 1281 images, which are mean absolute error ($MAE$) and mean squared error ($MSE$).

Table 1. Hyper-parameters setting of the regressor network.

| $\delta_{LR}$ | Batch Size | Rotation Range of Synthetic Data | Zoom Range of Synthetic Data | Shearing Range of Synthetic Data | Translational Shift Factor |
|---------------|------------|----------------------------------|------------------------------|----------------------------------|---------------------------|
| 0.00001       | 16         | $20^\circ$                       | 0.15                         | 0.15                             | 0.2                       |

$$MAE = \frac{\sum_{i=1}^{T_{test}} |l_{regress}^t - l_{predict}^t|}{T_{test}} \tag{18}$$

$$MSE = \frac{\sum_{i=1}^{T_{test}} (l_{regress}^t - l_{predict}^t)^2}{T_{test}} \tag{19}$$

where $T_{test}$ is the total number of test images, while $l_{regress}^t$ is the annotated ground truth bone age by the pediatricians and $l_{predict}^t$ is the predicted bone age by the regressor network. For performance comparison, 10 other state-of-the-art deep learning networks have been tested while using the same image registration technique that includes MobileNet V1 [17], MobileNet V2 [43], MobileNet V3 [44], ShuffleNet V1 [45] and ShuffleNet V2 [46], VGG-19 [47], Inception V3 [48], ResNet [49], and DenseNet [42]. Basically, six of these models can be regarded as a lightweight model because of the total parameters used are less than 10 million. All of the networks have been trained until convergence, as shown in the training graph of Figure 9. Note that the training graphs have been capped at 40
epochs for better visualization. The proposed method has obtained the lowest mean squared error at the end of the 40 epochs, as shown by the red line graph.

**Table 2.** Experimental results of the proposed method and the benchmarked networks.

| Method          | Mean Absolute Error | Mean Squared Error | Total No. of Parameters |
|-----------------|---------------------|--------------------|-------------------------|
| ShuffleNet V1   | 15.728              | 372.575            | 936,457                 |
| SqueezeNet      | 14.164              | 311.783            | 735,939                 |
| VGG-19          | 14.028              | 307.416            | 38,911,041              |
| MobileNet V3    | 13.541              | 282.157            | 1,662,939               |
| ShuffleNet V2   | 12.010              | 226.951            | 5,380,761               |
| MobileNet V2    | 11.394              | 213.454            | 2,260,417               |
| MobileNet V1    | 10.886              | 190.349            | 3,229,889               |
| DenseNet        | 10.557              | 190.105            | 31,049,921              |
| ResNet          | 10.283              | 264.660            | 23,563,201              |
| Inception V3    | 9.774               | 191.696            | 18,783,649              |
| Proposed method | 8.200               | 121.902            | 20,863,529              |

Table 2 shows the experimental results of the bone age prediction of the proposed method and its benchmarked networks. Basically, our method which is based on the Xception regressor has produced the lowest MAE of 8.200 months and MSE of 121.902 using a total number of 20,863,529 parameters. The main advantage of the Xception network is its three-layer residual separable convolution unit with 728 filters. Besides, when a down-pooling operator is applied, the skip connection layer will also be down-pooled using a convolution operator and, thus, reducing the probability of diminishing gradient problem. Moreover, this network also applied a unique composite function of the rectified linear unit, convolution layer, and batch normalization operator, while the standard flow for the other networks is convolution layer, rectified linear unit, and batch normalization operator. The second best
performance is produced by the Inception V3 model, followed by ResNet and DenseNet with MAE values of 9.774, 10.283, and 10.557, respectively. Note that the proposed method prediction is around 1.5 months better when compared to the second best method, Inception V3. However, the prediction performance gap between Inception V3 with ResNet is just around 0.5 months. Thus, our method has produced relatively significant improvement as compared to the other networks. Besides, ResNet has produced a better MAE compared to the DenseNet, but its MSE is significantly big as compared to the DenseNet. This is because ResNet generally returns good bone age prediction, but it also returns a few bad predictions that are quite far when compared to the annotated ground truth age. Hence, its squared error is big as this metric punishes big error heavily as compared to the MAE metric.

It is also interesting to note the top four methods are produced by networks with a total number of parameters that are bigger than 10,000,000. The best performed lightweight model is MobileNet V1, with MAE equal to 10.886 and MSE equal to 190.349. Moreover, MobileNet V1 returns the best performance compared to the other MobileNet family architectures. There is a trend among the MobileNet family models where a bigger number of parameters used in the network will produce a better prediction performance. Similarly, the performance of ShuffleNet V2 is also better when compared to ShuffleNet V1, where it uses more parameters, which is 5,380,761 as compared to the ShuffleNet V1, with just 936,457 parameters. However, the convolution architecture also plays an important role in determining the network performance, where SqueezeNet with expand and squeeze scheme has produced a lower MAE when compared to ShuffleNet V1, even though it uses less memory of just 735,939 parameters. Nevertheless, the lightweight model performance is generally lower as compared to the standard deep learning models, except for the VGG-19 network. In fact, it is the biggest model with 38,911,041 parameters, but its architecture consists of just a single network flow without applying any residual or branching operations. Thus, it cannot learn well the regression features and produces relatively weak predictions with MAE of just 14.028. The best performed model without applying a residual scheme is Inception V3, but it utilizes a wide network architecture with four parallel branches of different convolution sizes.

7. Conclusions

In conclusion, the proposed method has produced the best bone age prediction with the lowest MAE and MSE of 8.200 and 121.902, respectively. Therefore, the prediction error of the bone age is well below the 12 months threshold, where prediction tolerance under a one-year gap is still considered acceptable. The algorithm utilizes a novel image registration technique using separable convolution to automatically segment the hand and locate the keypoints for image rotation correction. Moreover, the regressor network, which is used to predict the bone age has utilized three-layer residual separable convolution units to produce a deep network, but maintain an acceptable model size, which is around 20,000,000 parameters. The network has also been trained using variable learning rates where its value is linearly decreasing with respect to the training epoch. However, the proposed method does not consider gender role in making the prediction, where this factor will surely affect the prediction accuracy. Another limitation that has not been considered in this work is the relative size of the captured X-ray image, where some samples are small in size when compared to the others. For future work, the proposed network can be further improved by adding a wider set of convolution filters instead of just a single network flow in the main branch. Atrous convolution can also be considered to capture larger feature maps while using a small convolution kernel size.

Author Contributions: Conceptualization, M.A.Z., S.R.A. and N.H.Z.; methodology, M.A.Z., S.R.A. and N.H.Z; software, M.A.Z., S.R.A. and N.H.Z; validation, M.A.Z., S.R.A. and N.H.Z; formal analysis, M.A.Z., S.R.A. and N.H.Z; investigation, X.X.; resources, M.A.Z., S.R.A. and N.H.Z; data curation, M.A.Z., S.R.A. and N.H.Z; writing—original draft preparation, M.A.Z., S.R.A. and N.H.Z; writing—review and editing, M.A.Z., S.R.A. and N.H.Z; visualization, M.A.Z., S.R.A. and N.H.Z; supervision, M.A.Z., S.R.A. and N.H.Z; project administration, M.A.Z., S.R.A. and N.H.Z; funding acquisition, M.A.Z., S.R.A. and N.H.Z.. All authors have read and agreed to the published version of the manuscript.
Funding: This work was supported in part by the Universiti Kebangsaan Malaysia through Research University Grant Scheme under Grant GUP-2019-008, and in part by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme under Grant FRGS/1/2019/ICT02/UKM/02/1.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Ethical Statement: All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of National Institutes of Health, USA (U24CA180927).

Abbreviations

The following abbreviations are used in this manuscript:

- G&P: Greulich and Pyle
- TW: Tanner-Whitehouse
- NN: Neural networks
- SVR: Support vector regression
- SVM: Support vector machine
- RSNA: Radiological Society of North America
- CNN: Convolutional neural networks
- ASPP: trou spatial pyramid pooling

References

1. Himmelman, J. The Burden of Being Messi. *The New York Times Magazine*, 5 June 2014.
2. Alshamrani, K.; Hewitt, A.; Offiah, A. Applicability of two bone age assessment methods to children from Saudi Arabia. *Clin. Radiol.* 2020, 75, 156.e1–156.e9. [CrossRef]
3. Sanctis, V.D.; Maio, S.D.; Soliman, A.T.; Raiola, G.; Elalaily, R.; G., M. Hand X-ray in pediatric endocrinology: Skeletal age assessment and beyond. *Indian J. Endocrinol. Metab.* 2014, 18, S63–S71. [CrossRef]
4. Nazria, N.; Annuar, A. X-ray sources population in ngc 1559. *J. Kejuruter.* 2020, 3, 7–14. [CrossRef]
5. Greulich, W.W.; Pyle, S.I. Radiographic atlas of skeletal development of the hand and wrist. *Am. J. Med Sci.* 1959, 238, 393. [CrossRef]
6. Carty, H. Assessment of skeletal maturity and prediction of adult height. *J. Bone Jt. Surg.* 2002, 84-B, 310–311. [CrossRef]
7. Fernandez, J.R.; Zhang, A.; Vachon, L.; Tsao, S. Bone age assessment in Hispanic children: Digital hand atlas compared with the Greulich and Pyle (G&P) atlas. In *Medical Imaging 2008: PACS and Imaging Informatics*; Andriole, K.P., Siddiqui, K.M., Eds.; International Society for Optics and Photonics: Washington, DC, USA, 2008; Volume 6919, pp. 370–375. [CrossRef]
8. 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]
9. Haak, D.; Simon, H.; Yu, J.; Harmsen, M.; Deserno, T.M. Bone Age Assessment Using Support Vector Machine Regression. In *Bildverarbeitung für die Medizin 2013*; Meinzer, H.P., Deserno, T.M., Handels, H., Toldorff, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 164–169.
10. Spampinato, C.; Palazzo, S.; Giordano, D.; Aldinucci, M.; Leonardi, R. Deep learning for automated skeletal bone age assessment in X-ray images. *Med. Image Anal.* 2017, 36, 41–51. [CrossRef] [PubMed]
11. Zulkifley, M.A.; Abdani, S.R.; Zulkifley, N.H. Pterygium-Net: A deep learning approach to pterygium detection and localization. *Multimed. Tools Appl.* 2019, 78, 34563–34584. [CrossRef]
12. Yamanakkanavar, N.; Choi, J.Y.; Lee, B. MRI Segmentation and Classification of Human Brain Using Deep Learning for Diagnosis of Alzheimer’s Disease: A Survey. *Sensors* 2020, 20, 3243. [CrossRef] [PubMed]
13. Loey, M.; Smarandache, F.; Khalifa, M.E.N. Within the Lack of Chest COVID-19 X-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning. *Symmetry* 2020, 12, 651. [CrossRef]
14. Zulkifley, M.A.; Mohamed, N.A.; Zulkifley, N.H. Squat Angle Assessment Through Tracking Body Movements. *IEEE Access* 2019, 7, 48635–48644. [CrossRef]
15. Guo, F.; Ng, M.; Goubran, M.; Petersen, S.E.; Piechnik, S.K.; Neubauer, S.; Wright, G. Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach. *Med. Image Anal.* 2020, 61, 101636. [CrossRef]

16. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 803–851.

17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* 2017, arXiv:1704.04861.

18. 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. 1800–1807.

19. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the Conference on Machine Learning, Long Beach, CA, USA*, 10–15 June 2019; pp. 1–10.

20. Booz, C.; Yel, I.; Wichmann, J.L.; Boettger, S.; Al Kamali, A.; Albrecht, M.H.; Martin, S.S.; Lenga, L.; Huizinga, N.A.; D’Angelo, T.; et al. Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the greulich-pyle method. *Eur. Radiol. Exp.* 2020, 4. [CrossRef]

21. Kim, J.R.; Shim, W.H.; Yoon, H.M.; Hong, S.H.; Lee, J.S.; Cho, Y.A. Kim, S. Computerized bone age estimation using deep learning based program: Evaluation of the accuracy and efficiency. *Am. J. Roentgenol.* 2017, 209, 1374–1380. [CrossRef]

22. Halabi, S.S.; Prevedello, L.M.; Kalpathy-Cramer, J.; Mamonov, A.B.; Bilbily, A.; Cicero, M.; Pan, I.; Pereira, L.A.; Sousa, R.T.; Abdala, N.; et al. The RSNA pediatric bone age machine learning challenge. *Radiology* 2019, 290, 498–503. [CrossRef]

23. Krizhevsky, A.; Sutskever, I.; Hinton G.E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA*, 3–6 December 2012; pp. 1097–1105.

24. Dallora, A.L.; Anderberg, P.; Kvist, O.; Mendes, E.; Diaz Ruiz, S.; Sanmartin Berglund, J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. *PLoS ONE* 2019, 14, e0220242. [CrossRef]

25. Seok, J.; Kasa-Vubu, J.; DiPietro, M.; Girard, A. Expert system for automated bone age determination. *Expert Syst. Appl.* 2016, 50, 75–88. [CrossRef]

26. Dehghani, F.; Karimian, A.; Sirous, M. Assessing the Bone Age of Children in an Automatic Manner Newborn to 18 Years Range. *J. Digit. Imaging* 2020, 33, 399–407. [CrossRef]

27. Gertych, A.; Zhang, A.; Sayre, J.; Pospiech-Kurkowska, S.; Huang, H. Bone age assessment of children using a digital hand atlas. *Comput. Med. Imaging Graph.* 2007, 31, 322–331. [CrossRef]

28. Lin, H.H.; Shu, S.G.; Lin, Y.H.; Yu, S.S. Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation. *Pattern Recognit.* 2012, 45, 322–332. [CrossRef]

29. Mansourvar, M.; Shamshirband, S.; Raj, R.G.; Gunalan, R.; Mazinani, I. An Automated System for Skeletal Maturity Assessment by Extreme Learning Machines. *PLoS ONE* 2015, 10, 1–14. [CrossRef]

30. Zulkifley, M.A. Two Streams Multiple-Model Object Tracker for Thermal Infrared Video. *IEEE Access* 2019, 7, 32383–32392. [CrossRef]

31. Zulkifley, M.A.; Trigoni, N. Multiple-Model Fully Convolutional Neural Networks for Single Object Tracking on Thermal Infrared Video. *IEEE Access* 2018, 6, 42790–42799. [CrossRef]

32. 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*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradly, A., Papa, J.P., Belagianins, V., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 300–308.

33. Chu, M.; Liu, B.; Zhou, F.; Bai, X.; Guo, B. Bone Age Assessment Based on Two-Stage Deep Neural Networks. In *Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, Australia, 10–13 December 2018; pp. 1–6.

34. Zhao, C.; Han, J.; Jia, Y.; Fan, L.; Gou, F. Versatile Framework for Medical Image Processing and Analysis with Application to Automatic Bone Age Assessment. *J. Electr. Comput. Eng.* 2018, 2018, 2187247. [CrossRef]
35. 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. Digit. Imaging* **2017**, *30*, 427–441. [CrossRef]

36. 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 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1158–1161.

37. Son, S.J.; Song, Y.; Kim, N.; Do, Y.; Kwak, N.; Lee, M.S.; Lee, B. TW3-Based Fully Automated Bone Age Assessment System Using Deep Neural Networks. *IEEE Access* **2019**, *7*, 33346–33358. [CrossRef]

38. Tohka, J. Rigid-Body Registration. In *Brain Mapping*; Toga, A.W., Ed.; Academic Press: Waltham, MA, USA, 2015; pp. 301–305. [CrossRef]

39. Abdani, S.R.; Zulkifley, M.A.; Moubark, A.M. Pterygium Tissues Segmentation using Densely Connected DeepLab. In Proceedings of the 2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE), Penang, Malaysia, 18–19 April 2020; pp. 229–232.

40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.

41. Zulkifley, M.A.; Abdani, S.R.; Zulkifley, N.H. COVID-19 Screening using a Lightweight Convolutional Neural Networks with Generative Adversarial Network Data Augmentation. *Symmetry* **2020**, *12*, 1530. [CrossRef]

42. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]

43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.

44. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 1314–1324.

45. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.

46. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.

47. Simonyan, K.; Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*; Technical Report; University of Oxford: Oxford, UK, 2014.

48. 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.

49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.