This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2876916, IEEE Journal of Biomedical and Health Informatics

Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment from Hand Radiograph

Xuhua Ren, *Student Member, IEEE,* Tingting Li, Xiujun Yang*, Shuai Wang, Sahar Ahmad, Lei Xiang, Shaun Richard Stone, Lihong Li, Yiqiang Zhan, *Member, IEEE,* Dinggang Shen*, *Fellow, IEEE* and Qian Wang*, *Member, IEEE*

***Abstract*—Skeletal bone age assessment is a common clinical practice to investigate endocrinology, genetic and growth disorders of children. However, clinical interpretation and bone age analyses are time-consuming, labor intensive and often subject to inter-observer variability. This advocates the need of fully automated method for bone age assessment. We propose a regression convolutional neural network (CNN) to automatically assess the pediatric bone age from hand radiograph. Our network is specifically trained to place more attention to those bone age related regions in the X-ray images. Specifically, we first adopt the attention module to process all images and generate the coarse/fine attention maps as inputs for the regression network. Then, the regression CNN follows the supervision of the dynamic attention loss during training, thus it can estimate the bone age of the hard (or "outlier") images more accurately. The experimental results show that our method achieves an average discrepancy of 5.2-5.3 months between clinical and automatic bone age evaluations on two large datasets. In conclusion, we propose a fully automated deep learning solution to process X-ray images of the hand for bone age assessment, with the accuracy comparable to human experts but with much better efficiency.**

***Index Terms*—Bone age assessment, deep learning, regression convolutional neural network, hand radiograph**

I. Introduction

BONE age assessment is a common procedure used in pediatric radiology for both diagnostic and therapeutic investigations of endocrinology abnormalities [1]. Based on the discrepancy between the reading of the bone age and the chronological age, physicians can make more accurate diagnoses of abnormal development in children.

Currently, the left-hand X-ray image is widely used for assessing the bone age as it can clearly render the subtle bone/cartilage development pattern with minimum radiation exposure [2]. Although X-ray radiograph is widely available in many clinical sites, the reading of the bone age is non-trivial in radiology practice. There are currently two popular methods to determine the bone age, i.e., the Greulich and Pyle (G&P) method [3] and the Tanner-Whitehouse (TW) method [4]. The G&P method focuses on a set of specific regions of interest (ROIs) of the hand and wrist joints (c.f. Fig. 1(a)). The radiologist then compares the image appearance within the ROIs to the atlases for reference. The TW method (also the TW2 and TW3 for the second and third editions, respectively) analyzes 20 ROIs and assigns a staging score to each of them.

An overall bone age reading is then derived from all ROIs and their scores. Both methods are time consuming – an experienced radiologist may spend 1.4 min on average to assess a patient with the G&P method and 7.9 min for the TW2 [5]. Furthermore, both methods suffer from high intra- and inter-observer variability. The average spreads of the reading are 0.96 year (11.5 months) for the G&P and 0.74 year (8.9 months) for the TW2 [6]. In some stages of child development (e.g., 14-18 years old children as in Fig. 1(b)), archetypal changes can often be so subtle and lack the sensitivity to be captured by human eyes using radiological examination.

A. Related Work

Many efforts in the literature push forward computer-assisted assessment to the bone age. There are mainly two types of methods. The first type relies on image processing. Giordano et al. [7] proposed a bone extraction method which was carried out by integrating anatomical knowledge of the hand and trigonometric concepts, whereas a TW2 staging was achieved by combining the bone segmentation and the enhancement of Gaussian filtering. Pietka et al. [8] proposed to identify the epiphyseal/metaphyseal ROIs, from which feature extraction was conducted by the graphical model. Gertych et al. [9] developed an automated method to assess the bone age, whereas the digital hand atlas was derived from a collection of 1,400 X-ray images to account for evenly distributed normal children. Cao et al. [10] presented a web-based system for bone age assessment, with an atlas constructed from a large set of clinically normal images of diverse ethnic groups.

* Corresponding authors: Qian Wang (wang.qian@sjtu.edu.cn), Dinggang Shen (dgshen@med.unc.edu), Xiujun Yang (yangxj01@schchildren.com.cn)

Xuhua Ren, Xiang Lei, Yiqiang Zhan and Qian Wang are in the Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, 200030, Shanghai, China. Tingting Li, Xiujun Yang and Lihong Li are in the Department of Radiology, Shanghai Children's Hospital, Shanghai Jiao Tong University, 200062, Shanghai, China. Shuai Wang, Sahar Ahmad, Dinggang Shen are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. Dinggang Shen is also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea. Shaun Richard Stone is in the Aberdeen Biomedical Imaging Centre (ABIC), Lilian Sutton Building, Foresterhill, University of Aberdeen, Aberdeen, UK.

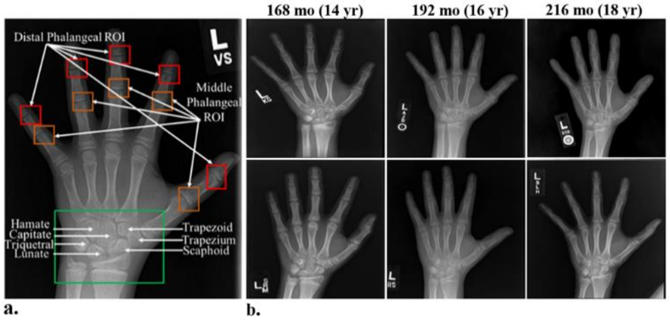

Fig. 1. (a) ROIs used in the G&P method for bone age assessment: distal phalangeal ROI, middle phalangeal ROI, hamate, capitate, triquetral, lunate, trapezoid, trapezium, and scaphoid [23]. (b) Subjects of different ages (i.e., 14-18 years) may have similar appearance in hand radiograph, which causes confusion to bone age assessment and also increases intra-/inter-observer variability.

The second type is knowledge-based, which primarily relies on decision rules or machine learning technologies. Mahmoodi et al. [11] presented a system to segment bones in a child's hand radiograph image, and determined the growth progress using decision theoretic approaches. Aja-Fernández et al. [12] proposed a fuzzy method to translate the natural language description of the TW3 method into an automatic classifier for bone age assessment. Thodberg et al. [13] proposed BoneXpert to reconstruct the borders of 15 bones automatically and then read the bone age from 13 bones (radius, ulna, and 11 short bones). Mahmoodi et al. [14] proposed a framework where phalanx bones were automatically localized; then several shape descriptors were obtained from the segmented bones.

Most existing computer-assisted bone age assessment methods try to extract features from the bones following clinical procedures such as TW and G&P. It imposes constraints to high-level understanding (i.e., machine learning) upon low-level (i.e., coming directly from image) visual descriptors, and limits the generalization capabilities of the devised solutions [15]. The deep learning technique [16], instead, aims at encoding visual features directly. For example, Spampinato et al. [17] proposed and tested several deep learning approaches to assess skeletal bone age automatically. The results showed an average discrepancy of 9.6 months with respect to the expert reading. Iglovikov et al. [18] utilized several deep learning architectures (i.e., U-Net [19], ResNet-50 [20], and the custom VGG-style neural networks [21]), and outperformed other common methods. Although these studies open a new paradigm for automated bone age assessment, they are not sufficient since:

(1) These methods are only evaluated upon the subjects of selected gender or within a narrow range of ages;

(2) These methods are vulnerable to "hard" subjects (e.g., the outliers that are hard to assess by computers);

(3) It is confusing to visualize and then interpret the results of deep learning by clinicians.

In this work, we propose a novel regression convolutional neural network [22] (CNN) method for automated pediatric bone age assessment and evaluate its performance upon two large-scale datasets. Different from other solutions, we input the coarse-to-fine attention maps of the hand radiograph to the regression CNN, such that the network became aware of the bone age related regions and visual appearance. Moreover, we propose a novel dynamic attention loss to supervise the training of the network, which can better handle the "hard" subjects in bone age assessment. With the experiments on two large datasets (consisting of 12,480 and 12,390 subjects, respectively), we conclude that our method could yield a low average discrepancy (5.2-5.3 months) with respect to the expert reading. The results imply that the accuracy of the deep learning-based bone age assessment system could be comparable to human expert, while the speed of reading a single image is much faster.

II. Method

A. Overview

We propose a regression CNN to complete the bone age assessment. The pipeline of our method is shown in Fig. 2. Particularly, we generate the coarse and fine attention maps from each hand X-ray image, and input to the regression CNN. The network then extracts feature maps and estimates the bone age after integrating the gender information of the subject.

B. Coarse and Fine Attention Maps

The X-ray radiographs of children hands vary considerably in intensity and appearance, which prevents automated method to learn the salient features effectively. Therefore, a pre-processing pipeline that standardizes all images is essential. Moreover, in this work, bones are the most important ROIs as their patterns are widely referred to in clinical bone age assessment. Therefore, we aim to (1) exclude the bias and irrelevant background and then focus on the foreground of hands; (2) focus on the bone-age-related image appearance information. The two steps above (i.e., in the attention module of Fig. 2) result in the coarse-to-fine attention maps, which help the subsequent regression CNN to better learn the salient features and estimate the bone age more accurately.

Coarse attention: We aim to localize and crop the foreground, where the hand is zoomed-in. In this way, the irrelevant background in the X-ray image can be removed. Before that, an input image must pass histogram matching [24] to normalize its intensity appearance. All images thus share similar histogram distributions. To generate the coarse attention map, we use Faster-RCNN [25] to place a bounding box for the hand foreground. The training of Faster-RCNN does not require accurate yet time-consuming delineation. Instead, we draw rectangular bounding boxes for 1,000 images, which are shown to be enough in our experiment. An example of the cropped coarse attention map can be found in Fig. 2.

Fine attention: In bone age assessment, bones are the most important anatomical structures to be referred to. Thereby, enhancing the bone regions in the X-ray image is essential to learn more salient features and benefit the regression task. We employ the Hessian filtering [26] to filter the coarse attention map and then obtain the fine attention map. It is found that the Hessian filtering can enhance the boundaries of the bones and also the areas of the bone joints effectively [27], as shown in Fig. 2.

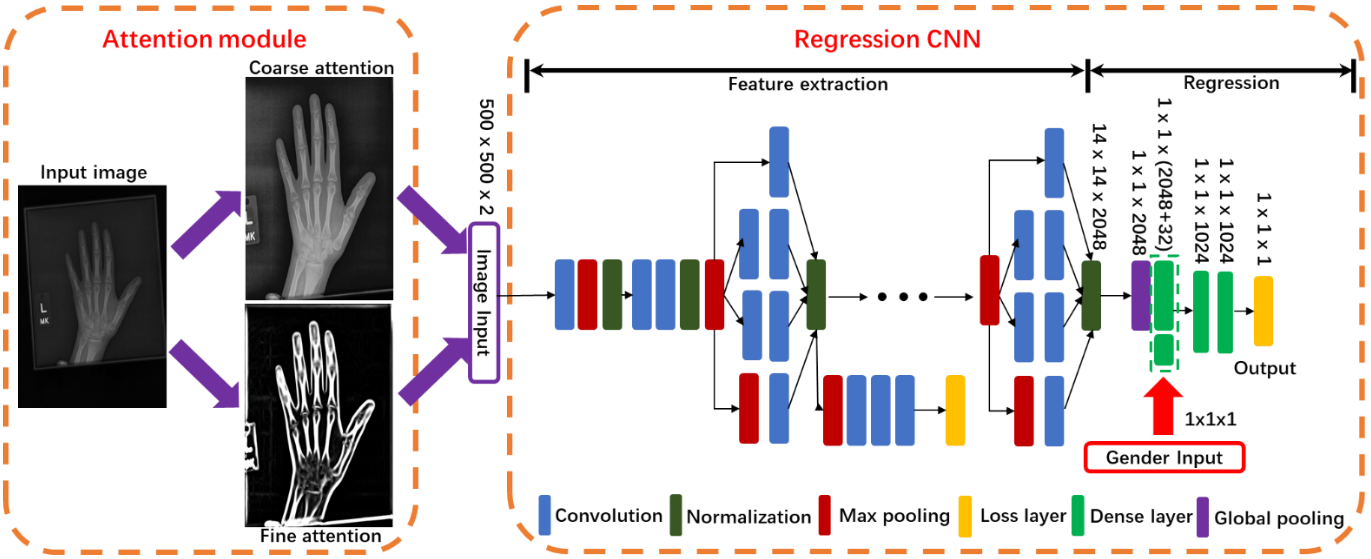

Fig. 2. Illustration of our proposed method: (1) The attention module completes image pre-processing and generates the coarse/fine attention maps; (2) The regression CNN adopts the Inception-V3 network for feature extraction and then integrates gender information for bone age assessment.

C. Regression CNN

The regression CNN to estimate the bone age as in Fig. 2. The input data (i.e., the coarse and fine attention maps) flow into an Inception-V3 [28] network for feature extraction, followed by a global pooling layer [29]. Then, the attention maps are encoded by a feature vector (length = 2048), and also the patient gender information is integrated into the network (i.e., 1 for male; 0 for female). Finally, we utilize two fully-connected layers [30] of 1,024 neurons and further regress out the bone age in the output layer. Batch normalization (BN) [31] and Rectified Linear Unit (RELU) [32] are applied after each convolutional layer in the network. The dropout strategy is adopted in the convolutional and fully-connected layers as it could greatly improve the modeling capability of the network and reduce the risk of overfitting [33].

Attention loss: We propose a novel attention loss to the network, in order to tackle the hard subjects that are often associated with abnormalities and diseases in clinical practice and may thus confuse automated bone age assessment. Denoting the expert reading of the bone age of patient i as g_i ("ground-truth") and the automated prediction result as p_i, one could define the mean absolute error (MAE) of all n patients as the loss to train the network:

$$\begin{aligned} \mathrm{MAE} &= \frac{\sum_{i=1}^{n}|g_i - p_i|}{n} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\frac{p_i}{g_i} - 1\right||g_i| \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\alpha_i|g_i|, \\ \alpha_i &= \left|\frac{p_i}{g_i} - 1\right|. \end{aligned} \tag{1}$$

As shown in the above, α_i calculates the relative difference between the ground-truth and the predicted bone age. We then define the attention loss as:

$$\mathrm{AMAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}{\alpha_i}^{k}|g_i|. \tag{2}$$

k is a hyper-parameter to highlight α, such that more attention is devoted to the "hard" subjects (i.e., with large difference between the predicted result p_i and the ground truth g_i). In this way, our network could automatically down-weight the majority of easy subjects and prevent them from dominating the training process [34].

Dynamic training configuration: Concerning the attention loss, we further adopt the dynamic training strategy to start with a high k and gradually reduce it to 1 alongside the training process. In this way, we derive the dynamic attention loss that could encourage the training to converge faster and help the network to better handle "hard" subjects. Note that all weights in the networks are initialized randomly with a Gaussian distribution (mean: 0; standard deviation: 0.001). During the training process, the weights are updated by Adam [35] with a learning rate of 0.01 initially. Data augmentation [36] includes image rotation within -20° to 20°, image translation within the ratio 0-0.2, image scaling within the ratio 0-0.2, image horizontal flip, and image brightness shifting in the whole image within the ratio 0.8-1.2. The data augmentation operation is performed by Keras Application Programming Interface [37].

D. Region Activation for Bone Age Assessment

A major challenge for automated bone age assessment is the interpretation of the outcome. Different from many other solutions, our method works in an end-to-end way and regresses out the bone age from the X-ray image directly. Although we generate the coarse and fine attention maps that help the network focus on the bone age related areas, our method indeed borrows little from existing clinical schemes (e.g., G&P and TW3). To this end, it is necessary to verify the contributions for individual regions in the hand X-ray image. Whereas the pattern of the activated regions by deep learning might be potentially helpful to understand the connection between bone age and visual appearance in radiograph.

In order to visualize the region activation for bone age assessment, we have proposed the Regression Activation Mapping (RAM) tool in this work. The RAM tool is generally inspired by the Class Activation Mapping (CAM), which exposes the implicit attention of CNN within the image space and highlights the most informative spatial regions relevant to the learning task [38]. Given the trained regression CNN for bone age assessment, we apply global pooling (GP) [29] to the feature maps after "feature extraction" yet before entering "regression" (c.f. Fig. 2). The features produced by GP are then adopted to regress out the bone age directly, i.e., through a fully-connected layer. While the regression assigns different coefficients to individual features, we use the coefficients to weight the corresponding feature maps before GP. Therefore, a final RAM response can be generated by averaging all feature maps in the image space according to their individual weights. The RAM result provides the spatial average of the feature maps and thus implies the importance of different regions in estimating the bone age.

III. Experimental Results

A. Datasets and Evaluation Settings

There are two datasets used in this work. The first public dataset ("RSNA") comes from the RSNA Pediatric Bone Age Machine Learning Challenge [1], which consists of hand radiographs provided by a consortium of US based research institutions. All images have associated bone ages provided by multiple expert observers. There are 12,480 X-ray images that are publicly available, while detailed demographic information is shown in Table I. In our experiment, we randomly split the dataset into training (9,984 images), validation (1,248 images), and test (1,248 images) sets. There is no patient overlap across different sets. The second private dataset ("SCH") comes from Shanghai Children's Hospital. All images also have associated bone age assessment in G&P provided by multiple expert observers. There are 12,390 subjects, with the detailed information also provided in Table I. In our work, we randomly split the dataset into training (9,912 images), validation (1,239 images), and test (1,239 images) sets, without patient overlap. Given a dataset, we use the training set to generate the deep learning model. The validation set is used to tune the hyper-parameters, while the performance of bone age assessment is evaluated upon the test set. Since this work targets on pediatric bone age assessment, we exclude all subjects older than 18 years old. In the SCH dataset which has been collected in clinical routines, there are no subjects over the age of 18 years. In the RSNA dataset, there are 21 (0.1%) subjects older than 18 years. We exclude those subjects since (1) the impact is minimal concerning the subject number and (2) the comparisons upon the two datasets will be consistent and fair for all pediatric age groups.

To evaluate the accuracy of the proposed method in reading the bone age, we compute MAE and its standard deviation between expert reading and automatic reading upon the test subjects. Our method in this work integrates two novel strategies:

(1) The coarse-to-fine attention maps, which could help the network to focus on bone age related regions;

(2) The dynamic attention loss, which could implement online mining of hard subjects in the bone age regression task.

We thus investigate the effectiveness of the two strategies by designing several experimental configurations for comparison, including the champion of the RSNA challenge (S1), S1 + course-to-fine attention maps (S2), S1 + attention loss (S3), S1 + dynamic attention loss (S4), and our full method (S5). The champion of the recent RSNA challenge (S1) also extracted the features from the Inception-V3 network, while excluding the use of course-to-fine attention maps or the dynamic attention loss. They only utilized the original image as the input (with similar data augmentation strategy as ours) and the MAE loss. We also conduct paired *t*-tests for the errors of the test subjects between individual experimental configurations. The statistical significance is attained at $p = 0.05$. All statistical analyses are performed by using Python.

TABLE I
Demographic Information of The Two Datasets Used for Bone Age Assessment in This Work

Dataset	Subject Number				Age Range	F/M Ratio
	Total	Train	Valid	Test		
RSNA	12,480	9,984	1,248	1,248	0-228	45.8/54.2
SCH	12,390	9,912	1,239	1,239	0-191	57.2/42.8

B. Choice of CNN Architecture

It is necessary to verify the optimality of the CNN architecture to the task of bone age assessment. We have thus evaluated the accuracy regarding several popular network backbones. There are two main architectures of CNN, i.e., Inception-like and Residual-like CNN. According to the RSNA challenge result, Inception-like network obtains better performance in BAA task. Based on this, we selected the Inception-V3 network. So, the networks compared here include Inception-V3 [28] (adopted in our method), Inception-V4 [39], Residual-50 [20], Residual-101 [20], Densenet-121 [40], and Densenet-161 [40]. We first replace the feature extraction part (Inception-V3, c.f. Fig. 2) in our method with a certain CNN architecture under comparison, and then conduct the evaluation upon the RSNA dataset. Note that both the attention module and the dynamic attention loss are disabled for fair comparison here. In particular, the Inception-V3 network has achieved a MAE of 6.0 ± 6.1 months (c.f. Table II), which is better than all other networks. Moreover, paired *t*-tests show that the outcomes of Inception-V3 are significantly better than other competing networks except for Densenet-161. It is also worth noting that the number of the network parameters is reasonable in Inception-V3, which is adopted as the backbone in our method finally.

C. Accuracy on the RSNA Dataset

The accuracy of each configuration upon the RSNA dataset is reported in Table III. Our method (S5) has achieved an

[1] https://www.kaggle.com/kmader/rsna-bone-age

average MAE of 5.2 ± 5.1 months (0.43 ± 0.43 year) upon 1,248 test subjects, compared with 6.0 ± 6.1 months (0.50 ± 0.51 year) for S1 ($p = 1.43\times10^{-6}$). We thus conclude that our method could outperform the existing champion of the RSNA challenge in terms of the accuracy to estimate the bone age automatically. Moreover, the outcomes of S2, S3, and S4 are better than S1, which implies the effectiveness of the strategies of attention module and dynamic attention loss. We further decompose the accuracy analysis to different age/gender groups in Table III. In particular, our method (S5) performs better (i.e., with lower MAE) than S1 in all age/gender groups, while the improvement is statistically significant.

D. Accuracy on the SCH Dataset

Similar to the RSNA dataset, we have compared all experimental configurations (S1-S5) upon the test subjects in the SCH dataset. The results are summarized in Table IV. Our method (S5) has achieved an average MAE of 5.3 ± 5.5 month (0.44 ± 0.46 years), which is significantly lower than S1 (6.2 ± 6.3 months, 0.52 ± 0.53 years, $p = 2.19\times10^{-7}$ in paired *t*-test) and all other configurations.

E. Impact of Transfer Learning

It is known that the number of the training subjects is critical to the performance of deep learning. Therefore, given a certain dataset, one may be able to train the network with more data transferred from another dataset. To this end, we investigate the role of transfer learning in automated bone age assessment. Specifically, we train the regression CNN with the RSNA dataset in the same way as in Section III.C. Then, the network parameters are fine-tuned by a limited quantity of the training subjects in the SCH dataset. Finally, we evaluate the bone age assessment accuracy by referring to the test set of the SCH dataset.

We aim to verify whether the network parameters transferred from the RSNA dataset would be potentially beneficial to reading the bone ages of the SCH dataset. And we have designed two cases for comparison.

(1) Random initialization. We have used random weights to initialize the network, which is trained, validated and tested with the SCH dataset only. Note that, the validation and test sets of the SCH dataset here are the same with Section III.D, while the numbers of the training subjects vary according to different configurations in Table V.

(2) Transferred initialization. We have trained the network with the RSNA dataset as in Section III.C. Then we use some training subjects from the SCH dataset to fine-tune the network parameters [41]. The numbers of the SCH training subjects are also shown in Table V.

The results in Table V show that the accuracy in bone age assessment is consistently higher in the case of transferred learning than the case of random initialization. Particularly, the MAE with 9,984 RSNA subjects for transferring and 496 SCH subjects for fine-tuning reaches 6.2 ± 8.3 months, compared to 6.9 ± 9.3 months without transferred network parameters from RSNA. The margin of the MAEs becomes narrow when more SCH subjects are available for training (i.e., 5.3 ± 5.5 months with 9,984 RSNA and 9,912 SCH subjects for training vs. 5.4 ± 5.6 months with 0 RSNA and 9,912 SCH subjects for training, no statistical significance found). The above results imply that more training subjects would be beneficial toward a lower MAE. Meanwhile, with more training subjects from the same dataset, the gain from transfer learning might decrease.

F. Region Activation Result

In Fig. 3 and Fig. 4, we provide the typical RAM responses upon the RSNA and SCH test sets, respectively. The examples are corresponding to individual age/gender groups. It is interesting to notice that the regression CNN focuses on a few distal metacarpal ossification centers (especially for subjects of 37-180 months old). At the two ends of the age range under consideration (e.g., below 36 months and above 181 months), the network extends its attention to more areas in hand radiograph. The above findings are consistent across the two datasets. A possible reason that may explain the region activation pattern is linked with the connection between the image appearance and the bone age. Although the activated regions (especially between 37 and 180 months) are quite different from the clinical schemes (e.g., TW3), follow-up studies are necessary to investigate the roles of different hand regions in diagnosing the bone age. We are unable to rule out the possibility that a careful reading of only the activated region could be potentially helpful to improve precision of bone age assessment.

Meanwhile, it is worth noting that the two ends of the age range are minorities in the whole datasets. As shown by the distributions in blue in Table III and Table IV, most subjects that need bone age assessment are between 37 and 180 months old. The impact of the relatively limited amount of data below 36 months old or above 181 months old is still unknown, concerning the different patterns of the region activation. It is possible that the regression CNN must borrow information from more activated regions to complete bone age assessment for the two ends of the age range, in order to compensate for the shortage of the training data.

IV. Discussion

In this study, we use supervised CNN to regress out the bone age from hand radiograph automatically. While it is highly desired by clinicians to read the pediatric bone age automatically, the accuracy and the efficiency of the reading should always be scrutinized. In our experiments over two large-scale datasets, our method has achieved impressive MAEs (i.e., 5.2 months in average for the RSNA dataset and 5.3 months for the SCH dataset). Meanwhile, it takes approximately 1.5s for the entire system to complete fetching the X-ray image from Picture Archiving and Communication System (PACS) in Shanghai Children's Hospital, processing the image, and then reading the bone age automatically. While the accuracy of the regression CNN for bone age assessment is mostly comparable to (or slightly better than) human expert, the speed performance turns out to be much better (e.g., 1.4 to 7.9 min for radiologist to read the bone age as reported in the literature). More sophisticated design of the network could potentially improve the bone age assessment further, which will be part of our future work. The deep learning tool will also be able to help suppress the inter-rater variation regarding the bone age estimation, which will be validated in future.

TABLE II
COMPARISON OF DIFFERENT CNN ARCHITECTURES FOR BONE AGE ASSESSMENT

Architecture	Inception-V3	Inception-V4	Residual-50	Residual-101	Densenet-121	Densenet-161
Parameter No.	24,850,841	43,713,305	26,611,385	45,682,297	9,090,297	29,925,593
MAE (month)	**6.0±6.1**	6.6±6.5	6.4±6.6	6.6±6.6	6.5±6.4	6.2±6.1

The evaluation of the accuracy is conducted on the RSNA dataset, while both the attention module and the dynamic attention loss are disabled.

TABLE III
OVERALL PERFORMANCE UPON THE RSNA DATASET FOR BONE AGE ASSESSMENT

		Overall performance (evaluated upon 1,248 test subjects between 0-18 years old)					
		MAE (month)			*p*-value (paired *t*-test vs. S5)		
Method	S1	6.0±6.1			1.43×10^{-6}		
	S2	5.8±5.9			8.22×10^{-5}		
	S3	5.5±5.8			9.59×10^{-3}		
	S4	5.3±5.3			4.77×10^{-2}		
	S5	**5.2±5.1**			--		
		Gender: Male (574 test subjects)					
Age: month		[0, 36)	[36, 72)	[72, 108)	[108, 144)	[144, 180)	[180, 216]
Age: year		[0, 3)	[3, 6)	[6, 9)	[9, 12)	[12, 15)	[15, 18]
Test Subject No.		11	58	161	173	138	33
Method	S1	6.4±3.2	8.6±6.5	5.7±4.0	5.3±5.6	5.4±6.9	8.4±4.8
	S2	5.6±5.4	8.2±6.6	**5.5±4.0**	5.6±4.6	5.7±5.1	5.4±4.3
	S3	5.0±3.7	7.6±5.8	6.2±5.3	4.9±5.8	4.7±4.2	5.1±3.7
	S4	4.8±3.3	**6.6±5.6**	5.6±5.5	5.7±5.8	5.2±4.6	4.8±3.8
	S5	**4.4±3.7**	6.7±5.9	5.6±4.0	**4.8±5.0**	**4.2±3.9**	**4.7±3.7**
		Gender: Female (674 test subjects)					
Age: month		[0, 36)	[36, 72)	[72, 108)	[108, 144)	[144, 180)	[180, 216]
Age: year		[0, 3)	[3, 6)	[6, 9)	[9, 12)	[12, 15)	[15, 18]
Test Subject No.		13	36	83	179	300	63
Method	S1	7.2±9.5	8.4±7.5	7.5±7.8	5.4±4.9	5.2±6.7	8.0±8.6
	S2	7.0±8.0	7.1±5.2	7.4±6.4	5.4±6.7	4.9±6.8	7.5±8.9
	S3	7.1±7.5	8.0±7.0	7.1±6.0	5.2±7.2	4.5±5.7	6.9±7.4
	S4	6.7±6.8	7.1±6.1	7.0±5.8	**5.0±5.6**	**4.1±4.8**	**5.9±6.9**
	S5	**6.5±7.7**	**7.0±6.5**	**7.0±5.4**	5.1±6.0	4.4±5.1	6.0±6.9

The experimental configurations under comparison include the champion of the RSNA challenge (S1), S1 + course-to-fine attention maps (S2), S1 + attention loss (S3), S1 + dynamic attention loss (S4), and our full method (S5). The blue shadowed curves in the table indicate the age/gender-specific distribution of all subjects in the test dataset(s).

TABLE IV
OVERALL PERFORMANCE UPON THE SCH DATASET FOR BONE AGE ASSESSMENT

		Overall performance (evaluated upon 1,239 test subjects between 0-18 years old)					
		MAE			*p*-value (paired *t*-test vs. S5)		
Method	S1	6.2±6.3			2.19×10^{-7}		
	S2	6.1±5.8			7.59×10^{-6}		
	S3	5.7±5.5			3.65×10^{-3}		
	S4	5.5±5.3			5.32×10^{-2}		
	S5	**5.3±5.5**			--		
		Gender: Male (523 test subjects)					
Age: month		[0, 36)	[36, 72)	[72, 108)	[108, 144)	[144, 180)	[180, 216]
Age: year		[0, 3)	[3, 6)	[6, 9)	[9, 12)	[12, 15)	[15, 18]
Test Subject No.		157	140	17	44	132	33
Method	S1	3.2±2.7	5.3±5.5	7.2±6.1	9.3±6.3	5.6±5.8	7.9±6.3
	S2	2.8±1.9	5.6±6.2	7.1±5.7	9.8±7.3	5.6±5.6	7.7±5.5
	S3	2.4±2.2	5.2±4.7	8.6±5.9	10.4±6.0	4.8±4.2	6.6±5.0
	S4	**2.2±1.8**	4.8±4.4	8.6±6.1	11.6±6.7	**4.6±5.6**	7.3±5.4
	S5	2.7±2.0	**4.7±4.3**	**6.9±6.5**	**8.7±6.4**	5.4±6.4	**6.5±5.7**
		Gender: Female (716 test subjects)					
Age: month		[0, 36)	[36, 72)	[72, 108)	[108, 144)	[144, 180)	[180, 216]
Age: year		[0, 3)	[3, 6)	[6, 9)	[9, 12)	[12, 15)	[15, 18]
Test Subject No.		79	161	134	232	67	43
Method	S1	5.9±5.9	6.7±5.2	5.6±7.0	6.9±7.9	10.1±9.5	8.0±8.6
	S2	5.6±4.0	8.2±7.3	5.6±6.5	6.3±5.7	7.7±8.7	7.5±10.9
	S3	7.3±3.6	6.4±7.4	4.8±5.7	6.1±7.7	8.3±5.7	6.9±7.4
	S4	5.4±3.6	7.2±6.7	4.6±5.6	5.6±6.8	7.7±5.5	**5.9±6.9**
	S5	**5.1±4.0**	**6.2±6.7**	**4.4±6.4**	**5.5±6.4**	**7.2±5.6**	6.0±6.9

The experimental configurations under comparison include the champion of the RSNA challenge (S1), S1 + course-to-fine attention maps (S2), S1 + attention loss (S3), S1 + dynamic attention loss (S4), and our full method (S5). The blue shadowed curves in the table indicate the age/gender-specific distribution of all subjects in the test dataset(s).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2876916, IEEE Journal of Biomedical and Health Informatics

TABLE V
BONE AGE ASSESSMENT ACCURACY IN THE SCH DATASET WITH DIFFERENT CONFIGURATIONS

Random	Training Subject No.	RSNA	0							
		SCH	496	991	1,487	1,982	2,974	4,956	7,930	9,912
	MAE (month)	Mean	6.9	6.4	5.9	6.1	5.7	5.7	5.6	5.4
		STD	9.3	9.1	8.7	7.8	6.9	6.1	5.8	5.6
Transfer	Training Subject No.	RSNA	9,984							
		SCH	496	991	1,487	1,982	2,974	4,956	7,930	9,912
	MAE (month)	Mean	6.2	5.9	5.7	5.8	5.7	5.6	5.4	**5.3**
		STD	8.3	7.9	7.5	7.4	6.5	5.9	5.8	**5.5**
p-value (paired *t*-test between two initialization cases)			<0.001	<0.001	0.007	0.044	0.37	0.53	0.22	0.78

Note that our regression CNN cannot simply replace human radiologist. In this study, there are in total 2,487 test images of high subject diversity, which are used to evaluate the performance of our method only. We conclude that our method could be comparable to human experts. However, we are unable to derive any conclusion regarding the quality of reading bone ages from individual raters and hospitals. It is also worth noting that there are multiple ways to read the bone ages in clinical practice. But it is not an aim of this work to compare different schemes of bone age reading.

A. Analysis of "Hard" Subjects

The "hard" subjects are often critical in clinical diagnosis and thus draw attention. In our proposed method, we have introduced the novel attention module and the dynamic attention loss to improve the accuracy of automated bone age assessment. Our experimental results show that the MAE of our method becomes lower than the champion of the RSNA challenge (c.f. S5 vs. S1 in Sections III.C -Section III.D). In Fig. 5 and Fig. 6, we provide several "hard" subjects for the RSNA and SCH datasets, respectively. Note that the bone age readings produced by S1 (the champion of the RSNA challenge) deviate a lot from the expert reading for the subjects in the figures. With the two strategies proposed in this work, our method (S5) has produced much more superior performance in estimating the respective bone ages.

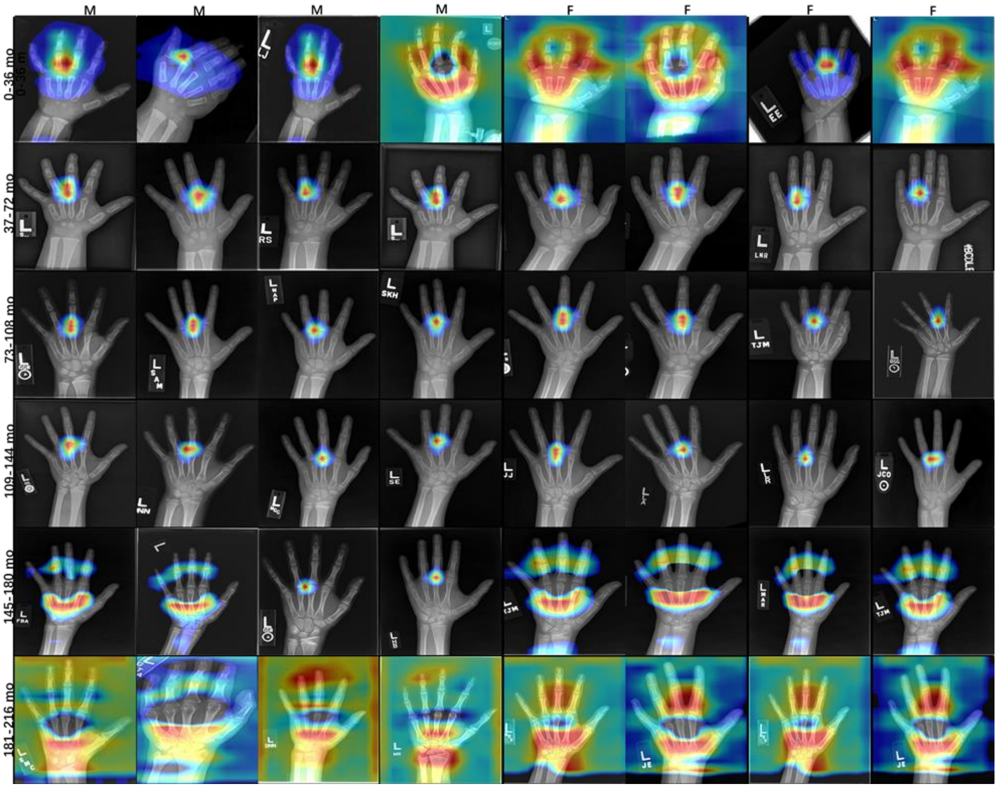

Fig. 3. Typical RAM results in the RSNA test set, corresponding to different age/gender groups. The colors indicate the strength for a location to be activated for bone age assessment (blue: low; red: strong).

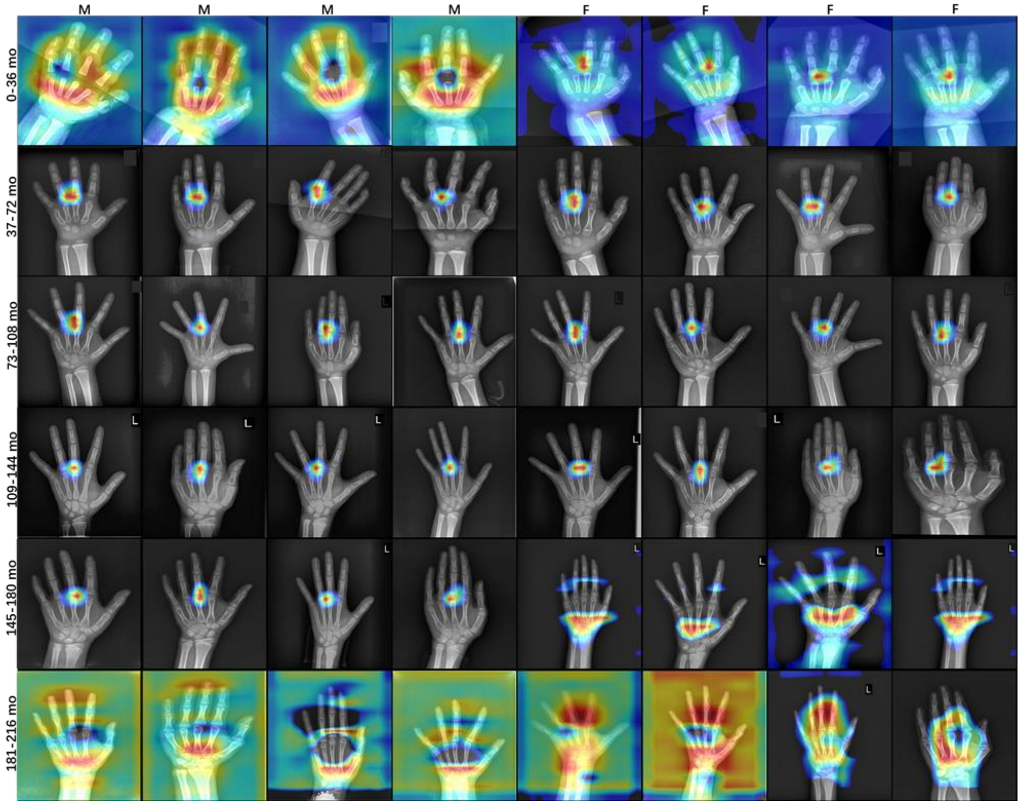

Fig. 4. Typical RAM results in the SCH test set, corresponding to different age/gender groups. The colors indicate the strength for a location to be activated for bone age assessment (blue: low; red: strong).

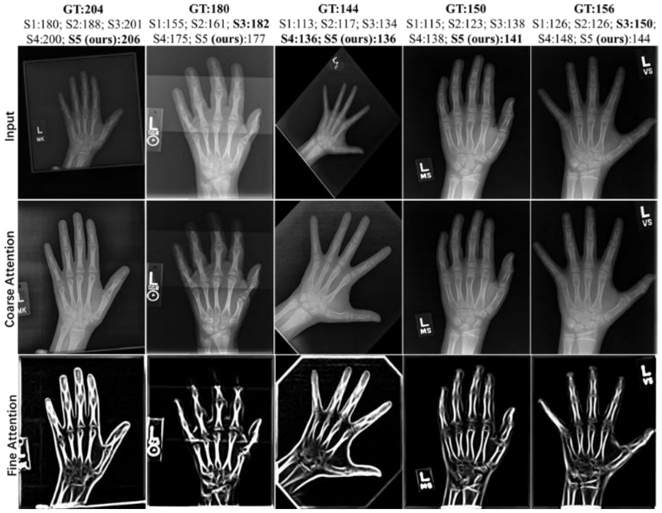

Fig. 5. Five "hard" subjects from the RSNA dataset. The ground-truth bone age reading from radiologist, as well as the automated readings from different methods, are marked on the top of the figure (unit: month). The coarse and fine attention maps are provided for better illustration. The methods under comparison include the champion of the RSNA challenge (S1), S1 + course-to-fine attention maps (S2), S1 + attention loss (S3), S1 + dynamic attention loss (S4), and our full method (S5). "GT" is short for ground-truth.

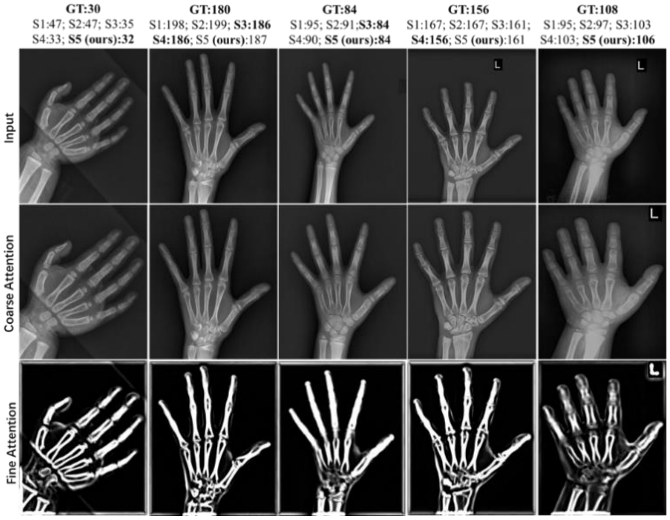

Fig. 6. Five "hard" subjects from the SCH dataset. The ground-truth bone age reading from radiologist, as well as the automated readings from different methods, are marked on the top of the figure (unit: month). The coarse and fine attention maps are provided for better illustration. The methods under comparison include the champion of the RSNA challenge (S1), S1 + course-to-fine attention maps (S2), S1 + attention loss (S3), S1 + dynamic attention loss (S4), and our full method (S5). "GT" is short for ground-truth.

V. Conclusion

We propose a fully automated deep learning solution to process X-ray images of children hands for bone age assessment in this work. Our work integrates attention module and dynamic attention loss, which are shown to be effective in improving the accuracy in reading the bone age. Our proposed method achieves 5.2 months in RSNA dataset and 5.3 months in the SCH dataset. Meanwhile, the entire pipeline can process a subject within ~1.5s, which is highly efficient for clinical usage.

Acknowledgements

This research was supported by the grants from Medical Engineering Cross Research Foundation of Shanghai Jiao Tong University (YG2017ZD08), National Natural Science Foundation of China (81471733, 61471390), National Key R&D Program of China (2017YFC0107602), Science and Technology Commission of Shanghai Municipality (16410722400).

References

[1] A. Albanese and R. Stanhope, "Investigation of delayed puberty," Clinical endocrinology, vol. 43, pp. 105-110, 1995.

[2] V. De Sanctis, S. Di Maio, A. T. Soliman, G. Raiola, R. Elalaily, and G. Millimaggi, "Hand X-ray in pediatric endocrinology: Skeletal age assessment and beyond," Indian journal of endocrinology and metabolism, vol. 18, p. S63, 2014.

[3] S. M. Garn, "Radiographic atlas of skeletal development of the hand and wrist," American journal of human genetics, vol. 11, p. 282, 1959.

[4] L. L. Morris, "Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method)," Journal of Medical Imaging and Radiation Oncology, vol. 47, pp. 340-341, 2003.

[5] A. Christoforidis, M. Badouraki, G. Katzos, and M. Athanassiou-Metaxa, "Bone age estimation and prediction of final height in patients with β-thalassaemia major: a comparison between the two most common methods," Pediatric radiology, vol. 37, pp. 1241-1246, 2007.

[6] D. King, D. Steventon, M. O'sullivan, A. Cook, V. Hornsby, I. Jefferson, et al., "Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods," The British journal of radiology, vol. 67, pp. 848-851, 1994.

[7] D. Giordano, R. Leonardi, F. Maiorana, G. Scarciofalo, and C. Spampinato, "Epiphysis and metaphysis extraction and classification by adaptive thresholding and DoG filtering for automated skeletal bone age analysis," in Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007, pp. 6551-6556.

[8] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. Huang, and V. Gilsanz, "Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction," IEEE transactions on medical imaging, vol. 20, pp. 715-729, 2001.

[9] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. Huang, "Bone age assessment of children using a digital hand atlas," Computerized Medical Imaging and Graphics, vol. 31, pp. 322-331, 2007.

[10] F. Cao, H. Huang, E. Pietka, and V. Gilsanz, "Digital hand atlas and web-based bone age assessment: system design and implementation," Computerized medical imaging and graphics, vol. 24, pp. 297-307, 2000.

[11] S. MAHMOODI, B. SHARIF, E. CHESTER, J. OWEN, and R. LEE, "Automated vision system for skeletal age assessment using knowledge based techniques," in IEE conference publication, 1997, pp. 809-813.

[12] S. Aja-Fernández, R. de Luis-Garcıa, M. A. Martın-Fernandez, and C. Alberola-López, "A computational TW3 classifier for skeletal maturity assessment. A computing with words approach," Journal of Biomedical Informatics, vol. 37, pp. 99-107, 2004.

[13] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," IEEE transactions on medical imaging, vol. 28, pp. 52-66, 2009.

[14] S. Mahmoodi, B. S. Sharif, E. G. Chester, J. P. Owen, and R. Lee, "Skeletal growth estimation using radiographic image processing and analysis," IEEE Transactions on Information Technology in Biomedicine, vol. 4, pp. 292-297, 2000.

[15] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," IEEE Access, vol. 6, pp. 9375-9389, 2018.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, p. 436, 2015.

[17] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," Medical image analysis, vol. 36, pp. 41-51, 2017.

[18] V. Iglovikov, A. Rakhlin, A. Kalinin, and A. Shvets, "Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks," arXiv preprint arXiv:1712.05053, 2017.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234-241.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[23] J. Liu, J. Qi, Z. Liu, Q. Ning, and X. Luo, "Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method," Computerized Medical Imaging and Graphics, vol. 32, pp. 678-684, 2008.

[24] D. Shapira, S. Avidan, and Y. Hel-Or, "Multiple histogram matching," in Image Processing (ICIP), 2013 20th IEEE International Conference on, 2013, pp. 2269-2273.

[25] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, et al., "Deformable Convolutional Networks," arXiv preprint arXiv:1703.06211, 2017.

[26] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 1998, pp. 130-137.

[27] M. Krčah, G. Székely, and R. Blanc, "Fully automatic and fast segmentation of the femur bone from 3D-CT images with no shape prior," in Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, 2011, pp. 2087-2090.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818-2826.

[29] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," 2015.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807-814.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," arXiv preprint arXiv:1708.02002, 2017.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[36] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," Technical report2017.

[37] F. Chollet, "Keras," ed, 2015.

[38] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?," arXiv preprint arXiv:1611.07450, 2016.

[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in AAAI, 2017, p. 12.

[40] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, p. 3.

[41] G. Montone, J. K. O'Regan, and A. V. Terekhov, "Gradual Tuning: a better way of Fine Tuning the parameters of a Deep Neural Network," arXiv preprint arXiv:1711.10177, 2017.