IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 9, 2019, accepted August 16, 2019, date of publication August 26, 2019, date of current version September 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937341

Bone Age Assessment Based on Rank-Monotonicity Enhanced Ranking CNN

BO LIU[1,2], YU ZHANG[1], MEICHENG CHU[1], XIANGZHI BAI[1,2], AND FUGEN ZHOU[1,2]
[1]Image Processing Center, Beihang University, Beijing 100191, China
[2]Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100083, China

Corresponding author: Xiangzhi Bai (jackybxz@buaa.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601012, and in part by the National Key Research and Development Program of China under Grant 2017YFC0113100.

ABSTRACT Skeletal bone age assessment based on hand x-ray is widely used in many fields. There is an urgent need for automated method to alleviate manual labor and address the problem of intra- and inter-observer variability. Most existing methods modeled the task as regression or multiclass classification problems. However, the regression method over-simplifies the relation between image features and bone age as linear while the bone development follows a nonlinear pattern; multiclass classification undesirably ignores the ordinal information of the age labels. In this work, we pioneered the applying of ranking learning to the problem of bone age assessment and proposed a two-stage bone age assessment network. The first stage is a CNN network VGG-U-Net used to segment the hand/wrist from the X-ray image. Then, a conditional GAN network was constructed to assess bone age. The generator was a ranking CNN which consisted of multiple binary classification outputs. We also proposed to use a rank-monotonicity loss to improve the performance. We validated the proposed method using the RSNA2017 Pediatric Bone Age dataset and the usefulness of different components was also investigated. The proposed method achieved an averaged mean absolute error (MAE) of 6.05 month (6.01 month for the male cohort and 6.09 month for the female cohort). Its performance was comparable with other state-of-the-art CNN based methods. Through ablation study, we found all proposed components (including the rank-monotonicity loss, adversary training strategy etc.) took effect and contributed to the final performance. In conclusion, we illustrated that rank-monotonicity enhanced ranking learning is more suitable for the task of bone age estimation. The proposed method is a valuable alternative for automatic bone age estimation.

INDEX TERMS Bone age assessment, ranking learning, convolutional neural network (CNN), computer-aided diagnosis (CAD).

I. INTRODUCTION

Bone age assessment of children and adolescents is widely used in clinical, legal and sports medicine fields [1]. It is usually performed by radiological examination of the left hand. Currently, bone age assessment is primarily conducted by trained radiologists, who manually assess hand bones in X-ray images in accordance with Greulich-Pyle (GP) [2] or Tanner-Whitehouse (TW) [3] methods. GP assesses the age by comparing the hand X-ray image with the atlas consisting of reference images from subjects of different ages. TW method considers a specific set of regions of interest (ROI), the development of each region of interest is divided into different stages and each stage is given a letter corresponding to a numerical score that varies by race and gender. Overall skeletal maturity score could be calculated by adding the scores for all ROIs. Manual evaluation is time-consuming and suffers from the problem of intra- and inter-observer variability. There is an urgent need for automated bone age assessment system.

A host of classical computer vision and image processing techniques have been applied to the automatic assessment of bone age [4]. Typically, they all assess bone age by extracting low-level features of bones commonly used in TW or GP clinical methods. Without the ability to extract advanced visual descriptors or semantic features, these methods usefully performed poor for relative matured bones which overlap with each other. One method needs to be noted is the BoneXpert

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

proposed by Thodberg et al. [5] proposed in 2009, which has been approved for the clinical use in Europe. It performs automatic bone age assessment through a unified model of TW and GP methods, but requires high quality X-ray scanning to get reliable results and does not utilize the carpal bones which is considered to be important for age assessment in infants and toddlers [4].

As a new technology in machine learning research, deep learning can characterize input data and realize complex function approximation by learning a deep nonlinear network structure. Researchers have explored this method for medical data analysis because of the success of deep learning techniques in general image classification. In most of these applications, convolutional neural networks have achieved results that went beyond the prior arts. Recently, deep learning methods were also been used to handle bone age assessment and achieved better performance than conventional methods. However, almost all these methods modeled the bone age estimation as regression or multiclass classification problems, which suffered from some drawbacks. The regression methods assume a linear relation between the image features and the bone age. However, the hand bone matures follow a different pattern depending on the person's age. For example, the bone maturation process is separated into six phases (infancy, toddlers, pre-puberty, early and mid-puberty, late puberty, post-puberty) and the bone develops in different ways in each phase [6]. This property requires the regression method to learn non-stationary kernels which is very difficult and will easily cause over-fitting problem in training [7], [8]. Besides, the regression model is also susceptible to the impact of outliers. As for the multiclass classification, a class probability vector was predicted with each class presenting the bone age in month and the class labels were assumed to be independent to one another. However, the age labels form a well-ordered set and thus have an ordinal relationship, which is completely ignored in multi-class classification methods. Exploiting the relative order among the age labels along with the exact age would help to achieve better results [9], [10].

In the field of face age estimation, ordinal regression (a.k.a. ranking learning) methods were investigated to address the aforementioned problems of the regressing and the multiclass classification methods [7], [11]–[13]. Ordinal regression can be considered as intermediate problem between regression and classification, and was more suitable for prediction of ordinal variables. For age estimation, ordinal regression could preserve the ordinal information of the age labels and also reflect the non-stationary nature of aging process. Better performance was reported compared with traditional multiclass classification and regression-based approaches. More recently, several works proposed to solve the ordinal regression problems by converting it into a series of simpler binary classification subproblems [8]–[10]. In this way, the well-studied binary classification algorithms could be directly used. Besides, it also had the theoretical benefit that new generalization bounds for ordinal regression could be derived from known bounds for binary classification [10].

Despite the successful use of the ranking learning in face age estimation, there are few published works investigating its performance in the problem of bone age assessment. To fill this knowledge gap, we proposed a new bone age assessment method in this work based on ranking learning methods. Specially, a ranking CNN was developed to solve the ranking learning problem which utilized the VGG16 network to extract the image features and had multiple binary classification outputs. We also proposed to use a novel rank-monotonicity loss and adversary training techniques to improve the performance of the ranking CNN. Besides, we also developed a VGG-U-Net which could accurately segment the hand/wrist from the X-ray images and effectively improve the age assessment accuracy. Part of this work was reported in the International Conference on Digital Image Computing: Techniques and Applications (DICTA) [14].

The rest of the paper is organized as follows. In Section 2, we described the ranking CNN based approach for bone age assessment. Then, in Section 3, we conducted a series of experiments using the dataset of RSNA2017 Pediatric Bone Age Challenge to validate our novel strategies and compare our method with the state-of-the-art methods. Finally, discussions are presented in Sections 4 with our conclusions in Section 5.

II. MATERIALS AND METHODS

Figure 1 showed the general architecture of the proposed method for bone age assessment. It is a two-stage network and contains (1) VGG-U-NET, a preprocessing stage to segment the input X-ray image. Post-segmentation processing was performed to generate three-channel images used in the next stage, and (2) age assessment stage which exploit a conditional GAN (cGAN) network to extract visual features and assess bone age. The generator is a ranking CNN with multiple outputs to solve the ranking learning problem.

A. PREPROCESSING WITH VGG-U-NET

As shown in Figure 2, the appearance of the X-ray images differ greatly with varying background, contrast and hand position, which may degenerated the performance of age estimation method. To address this problem, we would like to extract hand/wrist mask from the images and remove other irrelevant objects. Based on the segmentation mask, a three-channel image with enhanced gradient information will be finally constructed which was used in age assessment stage.

1) NETWORK ARCHITECTURE

Due to the varying appearance of the X-ray image, it is not an easy task to accurate segment the hand/wrist from the X-ray images [15]. We proposed a VGG-U-Net network to perform this task which combined the VGG16 encoding path with the decoding path of the U-Net.

U-Net is a successful segmentation method and was widely used in medical image segmentation. It has an encoder-decoder structure. The encoding path follows the typical architecture of repeating application of convolutions,

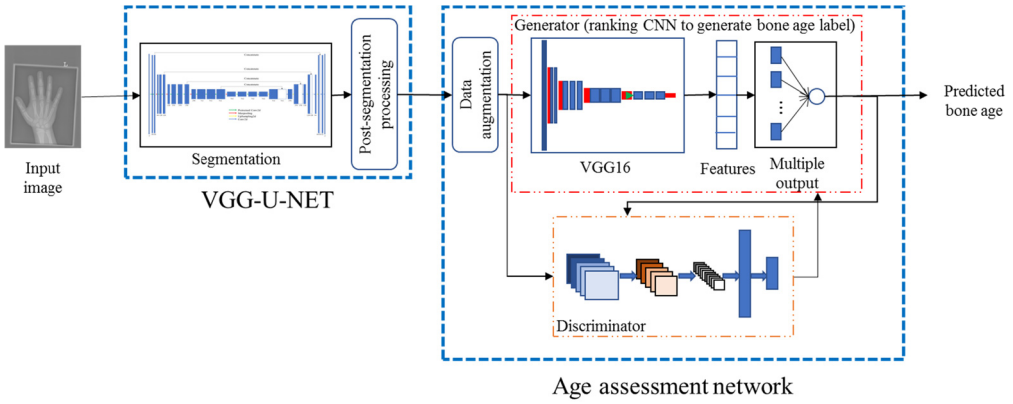


FIGURE 1. The general architecture of the proposed method.

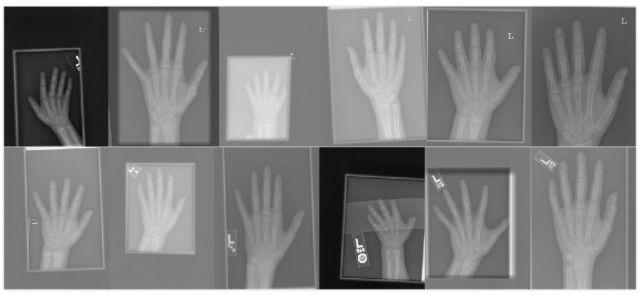


FIGURE 2. X-ray image samples.

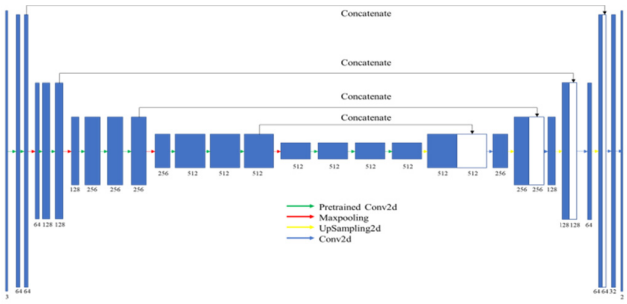


FIGURE 3. The architecture of the VGG-U-Net.

rectified linear unit and pooling to construct hierarchical image features. Every step in the decoding path consists of an upsampling of the feature map followed by a convolution which halves the number of feature channel. The encoder and decoder are symmetric structures, and there is a skip connection between the encoder and the decoder which combines low-level features with advanced features. It helps the decoder well fix target details and achieve pixel-level positioning.

Commonly, an adequate number of training samples were needed for the U-Net to obtain good segmentation results. However, it is laborious to manually label the image. Transfer learning is an effective technique to improve performance and training efficiency when the training data is scarce. Therefore, in this work, we proposed to replace the encoder of the U-Net with the pre-trained VGG16 which was pretrained on ImageNet. To connect VGG16 with U-Net, we substituted the 1x1 convolution with 512 channels for the original fully connection layer of VGG16 which acted as the bottleneck layer. And skip connections were added between the encoding and decoding paths at suitable places. The architecture of the VGG-U-Net was shown in Figure 3.

We manually labeled 100 images of which 80 images were used to train the VGG-U-Net network and 20 images were used for validation. The rest X-ray images were automatically segmented used the trained VGG-U-Net. As the result of the VGG-U-Net may have some isolated false segmentation and holes, we further processed the results by first filling holes and then keeping the largest connected components.

2) POST-SEGMENTATION PROCESSING

Using the segmentation mask, all other parts of the image except for the hand/wrist was discarded by assigning the intensity to zero. Then the contrast of the generated X-ray image was enhanced using the histogram equalization method. As the following age assessment network also utilized VGG network to extract image features which requires inputs with three channels, we need to construct a RGB image from the grayscale image. Instead of duplicating the grayscale image, we proposed to enhance the input with edge information extracted using the wavelet transform modulus maxima and the Canny edge detector methods. As we shown in the ablation study, add these two edge channels improved the performance of the proposed method.

B. AGE ASSESSMENT NETWORK

The structure of the bone age assessment network is shown in Figure 1. It is a conditional GAN network consisted of a generator to generate bone age label (based on which a four channel image was generated), and a discriminator to determine whether the generated four channel image was from real data distribution or not. The generator network is

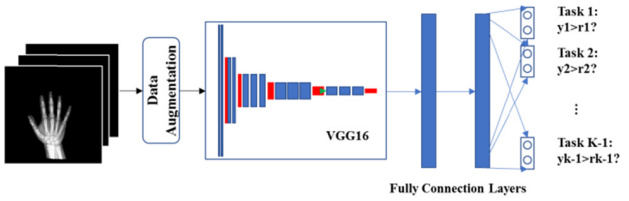


FIGURE 4. The structure of the bone age assessment network.

a multiple output convolution network which modeled the age estimation as a ordinal regression or ranking learning problem. Its loss function contained a multiple out loss (H_{mo}) and a ranking-consistency loss (H_{rc}). As we discussed in the Abstract, ranking learning could utilize the ordinal information of the age labels and also reflect the non-stationary nature of aging process. It was supposed to perform better than conventional regression or multiclass classification techniques.

As the adversarial training strategy can be considered as adding another loss function (supervision) to the generator network, the proposed method could be viewed as a CNN ranking network with three loss functions, i.e. H_{mo},H_{rc} andH_{gan}(the loss function introduced by adversarial training). This type of notation will be used in the ablation study.

1) NETWORK ARCHITECTURE

a: THE RANKING CNN NETWORK AS LABEL GENERATOR

As shown in Figure 4, the ranking CNN employed the first 13 layers of VGG16 to extract the features of the input images, followed by two fully connected layers with 1024 neurons for each layer. At the end of the network, $K-1$ output layers are connected with the fully connection layer, each containing 2 neurons, where K is the number of all possible ages in month. Each output layer corresponded to a binary classification problem. For a given input image I_k, the i_{th} binary classifier output $y_i^k = \{0, 1\}$ predicted whether the bone age of the image x_i is larger than age k, with k ranging from 1 to the largest possible age in month:

$$y_i^k = \begin{cases} 1, & if\,(age(x_i) > k) \\ 0, & otherwise, \end{cases}$$

Assuming consistent and correct answers were obtained for the $K-1$ binary classification problems, for an input image with true bone age r, the answers for the classification problem would be $y_i^k = 1$ for k =1 until r-1, and $y_i^k = 1 = 0$ afterwards. Then, a reasonable rule for bone age assessment based on the result of the binary classifications would be [10]:

$$Age_k = 1 + \sum_{i=1}^{K-1} y_k^i \tag{1}$$

which represent one plus the number of positive prediction among the $K-1$ classification problems. As discussed in the work of Li, *et al*[10], although consistent y_i^k is usually desired, the rule in (1) can be applied even when y_i^k is not consistent.

As for the feature extraction part, we had tried other networks (including Inception-V3, ResNet50, Inception-ResNetV2, DenseNet121) in developing this method and found that the VGG16 performed the best among these networks.

b: THE DISCRIMINATOR NETWORK

As shown in Figure 5, the ground-truth age and the predicted age were resized into a image of the same dimension with input, and concatenated to the input image as the fourth channel. The resulting image with ground-truth age channel (I_g, can be viewed as the output of the generator) or the image with predicted age (I_r) were input into the discriminator network to discriminate whether the image is real (concatenated with true label) or generated (concatenated with the predicted label). The discrimination network consisted of five Conv+ BatchNormalization+ MaxPool blocks and three fully connection layers, as shown in Figure 5.

2) TRAINING LOSSES

The ranking CNN had $K-1$ binary outputs $y_i^k = \{0, 1\}$ and it is reasonable to employ cross-entropy loss function for each binary classifier. For multiple binary classifiers, we combine the loss function for all outputs as follows

$$H_{mo} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K-1}\lambda_k 1\{\sigma_i^k = y_i^k\}\log(p(y_i^k|x_i))$$

where N indicates the number of input samples, $K-1$ indicates the number of binary classifiers. If inside is true, $1\{\cdot\}$ is 1, else 0. σ_i^k is the true label indicating whether the ground-true age of x_i is older than k. y_i^k is the prediction of the k_{th} binary classification problem. λ_k is weighting factor used to indicate the relative importance among different tasks, which was chosen according to the distribution of sample number with different ages [8]

$$\lambda_k = \frac{\sqrt{N_k}}{\sum_{k=1}^{K-1} N_k}$$

where N_k is the number of samples with age k.

The consistence of y_k is a desired feature for the ranking methods. Actually, the ordinal information would naturally require y_k to be rank-monotonic, i.e. $p(y_i^{k-1}|x_i) \geq p(y_i^k|x_i)$. For example, the probability of the age being older than 6 month should be no smaller than the probability of being older than 7 month. In this work, we proposed to explicitly enforce the rank-monotonicity by introducing a consistence penalty into the loss function

$$H_{rm} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{y_i}\max(0, (p(y_i^k|x_i) - p(y_i^{k-1}|x_i)))$$

which penalized the violation of the rank-monotonicity.

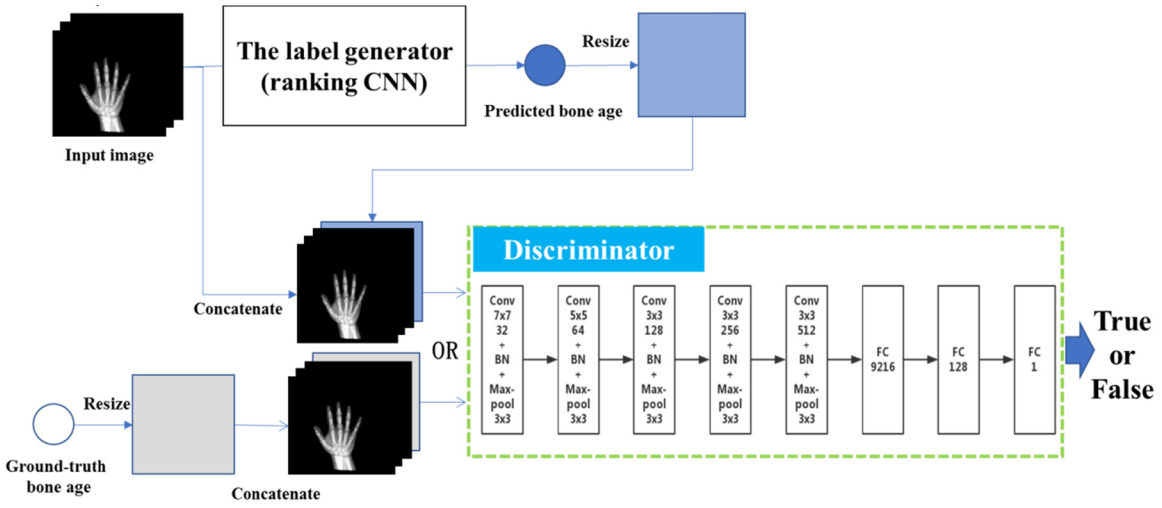


FIGURE 5. The architecture of the discriminator.

As for the discriminator D, we use the binary cross-loss entropy as the its loss function

$$H_{gan}(\tilde{x}_i) = \sum_{i=1}^{n} 1\{\tilde{x}_i \in I_r\} \log D(\tilde{x}_i) + 1\{\tilde{x}_i \in I_g\} \log(1 - D(\tilde{x}_i))$$

where $I_{r,i}$ and $I_{g,i}$ are the four channel inputs generated by concatenating the i-th image with the ground-truth and generated labels respectively. If the predicted age was close with the true age, the output was close to 1. Or else, the network output was close to zero.

The whole age assessment networks was trained by alternately minimizing

$$H_G = H_{mo} + \alpha H_{rm} + H_{gan}$$

and maximizing H_{gan}.

3) EXPERIMENTAL SETTING

The training dataset of RSNA2017 Pediatric Bone Age Challenge was used to develop and validate the method. It was provided by the North American Radiological Association and contains 12611 X-ray images of children hands (aging from 0 to 19 years old). Bone age in month for each image was manually labeled by experts.

To evaluate the performance of the proposed method, we randomly picked 1000 images for validation and 2000 images for testing with the same sex ratio as in the original data. The remaining images were used for training. As previous works shown that training separate networks for male and female cohorts may help achieve better accuracy, we also trained and validated three networks for the male, female and whole cohorts in this work for comparison.

The experiment was performed twice and the averaged results were reported in the following. The experiments were conducted using Keras with tensorflow [16] backend on a workstation equipped with two Intel Xeon E5 2630V4 CPU and a NVIDIA 1080Ti GPU. We took Adam as the optimizing algorithm and fixed the learning rate to 1e-5. A batch size of 32 images was used for each step.

TABLE 1. Parameters for data augmentation.

Transformation types	Parameter range
Horizontal translation	[0, image width×0.1]
Vertical translation	[0, image height×0.1]
Rotation	[-10°, 10°]
Scaling	[0.85, 1.15]
Horizontal flipping	0/1 (flip or not)

To avoid overfitting problem, we performed random geometric transformations on the input image to augment the training data. The applied geometric transformations and their parameter ranges are listed in Table 1.

As for the evaluation metrics, the most commonly used metric mean absolute error (MAE) was used for quantitative evaluation, which was calculated as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - x_i|$$

where x_i and y_i were the ground-truth and estimated bone age in month respectively

III. RESULTS

A. HAND/WRIST SEGMENTATION

Figure 6 shows the loss and accuracy curves for the training and validation dataset w/o using pretrained weights. As we can see, using the pretrained weights sped the convergence of the training and also improved the segmentation accuracy. The segmentation after post-segmentation processing achieved a satisfying accuracy of 99.9%. Figure 7 shows the result generated after applying histogram equalization.

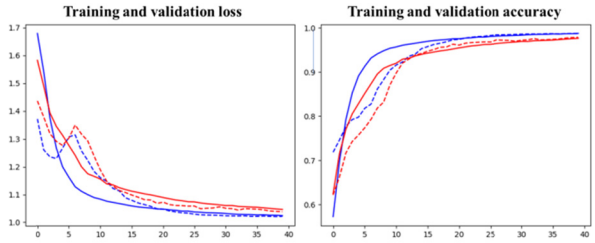


FIGURE 6. The preprocessing results (a) Input images; (b) Segmentation mask of VGG-U-Net; (c) Post-processed segmentation mask; (d) The results after performing applying histogram equalization.

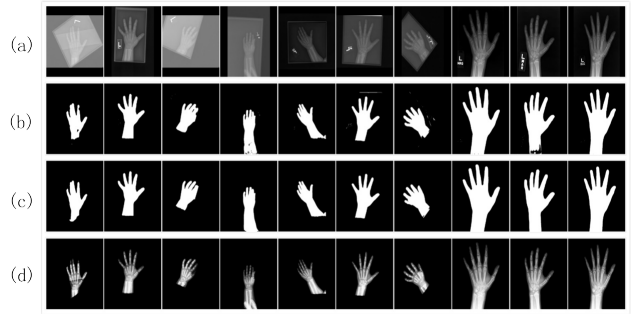
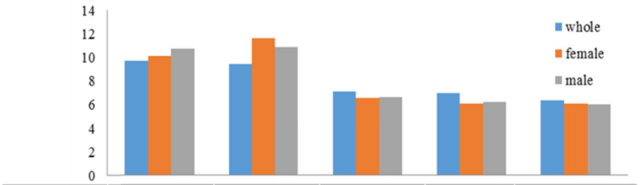


FIGURE 7. The preprocessing results (a) Input images; (b) Segmentation mask of VGG-U-Net; (c) Post-processed segmentation mask; (d) The results after performing applying histogram equalization.



	Variant1	Variant2	Variant3	Variant4	The proposed
Loss function	Multiclass Classification	Regression	H_{mo}	$H_{mo}+H_{rm}$	$H_{mo}+H_{rm}+H_{gan}$
Whole cohort	9.77	9.48	7.15	6.98	6.41
Female cohort	10.17	11.69	6.57	6.11	6.09
Male cohort	10.77	10.9	6.65	6.21	6.01

FIGURE 8. The preprocessing results (a) Input images; (b) Segmentation mask of VGG-U-Net; (c) Post-processed segmentation mask; (d) The results after performing applying histogram equalization.

B. ACCURACY OF AGE ESTIMATION

Figure 8 shows the performance of the proposed methods along with some variants. Different loss functions were applied for these variants with all other modules kept identical with the proposed method. As we can see in Figure 8, the proposed method achieved an accuracy of 6.41, 6.09 and 6.01 for the whole, female, and male cohorts. The result was comparable with other state-of-the-arts CNN-based methods which were evaluated on the same dataset, as shown in Table 2. Besides, Mutasa et al. achieved an mean MAE of 7.64 on a private dataset with similar number of samples [17].

Some studies reported that the standard error between observers is between 0.45 and 0.83 years, which is equivalent to the standard deviation of 0.64-1.17 years [18].

TABLE 2. Comparison of different methods on the RNSA competition dataset.

Methods	Accuracy (MAE)		
	Male cohort	Female cohort	Mean
Iglovikov et al.[15]	6.16	6.39	6.28
Larson et al. [18]	-	-	6.24
The proposed method	6.01	6.09	6.05

FIGURE 9. Selected examples of attention maps for female (upper rows) and male (lower rows) in the four major skeletal maturity stages.

The standard deviation of the difference between the predicted age of the proposed method and the evaluator's decision is 0.54 years, which is lower than this range, indicating that the proposed method performed as well with trained physicians.

Comparing the performance of different variants, the use of the multiple classifier output (ranking learning) greatly improved the age estimation accuracy compared with the conventional multiclass classification and regression loss functions. Both the H_{rc} and H_{gan} loss took effect and further improve the estimation accuracy for all the three cohorts. The regression loss performed slightly better than the multiclass classification loss for the whole cohort, but worse for the female cohort and male cohort. For the variant1 and variant2, better accuracy was obtained on the whole cohort. The methods using ranking learning achieved better performance on the female and male cohorts than the whole cohort.

In this paper, gradient-based class attention maps were generated by applying logit operator to the last convolution layer before the global pooling. The attention maps showed the importance of each pixel in the network decision process. Figure 9 shows which part of the image plays the most important role in bone age assessment for different age stages. As we can see, for the younger cohort, the network paid more attention on the carpal bones. As the ages increase, the network focused less importance on carpal bones but more on phalanx bones. When children approach their adult stage, importance returned to the carpal. The shift of attention conformed with the prior knowledge [6].

C. ABLATION STUDY

In this section, we conducted a set of ablation experiments on our method to justify the choices made in our design.

TABLE 3. Analysis of the usefulness of different components. Only the multiple classification output was used for these experiments.

Different settings		1	2	3	4
Optional components	Hand/wrist segmentation		√	√	√
	Adding edge channels		√		√
	Data augmentation	√		√	√
Results					
MAE	Whole	9.55	12.27	7.45	7.15
	Female	7.8	12.4	6.48	6.57
	Male	7.64	11.34	6.57	6.65

As we can see in Table 3, the data augmentation played a very important role for the performance. Without the data augmentation, the MAE increased from (7.15, 6.57 and 6.65) to (12.27, 12.4, and 11.34) for there cohorts. Comparing with setting #1 with setting #3, the hand/wrist segmentation and adding edge channels also had a large effect on the result and improve the estimation accuracy. The effect of adding edge channels was relatively small but also improved the results. As adding edge channels without segmenting hand/wrist would introduce many false feature, we did not evaluate the performance of this setting.

IV. DISCUSSION

For bone age assessment, almost all existing methods exploited multiclass classification or regression loss functions. Ranking learning technique which focuses on learning the rank between samples may be more suitable for the problem of age estimation and has been investigated in the field of face age estimation. In this work, we pioneered the applying of ranking learning to the problem of bone age assessment and compared its performance with other methods. As the experimental results shown, the ranking learning technique could achieve much better results, which was consistent with the findings in the field of face age estimation.

We used the reduction framework proposed by Li and Lin [10] to convert the ordinal regression or ranking learning into a set of binary classification problems. The final age was obtained by a simple summing of the output of the binary classification problems under the assumption of the consistence of binary outputs. The consistency issue was commonly not considered in the training because of the complexity [8], [19] or out of theoretical analysis [9]. In this works, we proposed to explicitly enforce the consistency by introducing a novel rank-monotonicity loss to the ranking CNN. The results shown that the rank-monotonicity loss could further improve the age estimation accuracy.

cGAN has been widely investigated for many medical imaging problems, including simulation and segmentation. The use of cGAN in this work was similar to its usage in segmentation where the discriminator tries to tell the input image is combined with generated mask or the ground truth, with the difference that the discriminator used for segmentation commonly has a patchwise or pixelwise loss. Its usefulness was validated by the experiments in this work, though the improvement was incremental. In the next works, we would like to further investigate the use of cGAN for bone age estimation.

The RSNA pediatric bone age machine learning challenge also provided another set of 200 images for testing. While the label for the training images were extracted from the radiology reports [18], the age label for these test images were determined as the weighted mean of the corrected reviewer estimates [20]. This caused inconsistency between the training and testing sets. Better results could be obtained for the test set than that achieved on validation sets. For example, the first place method achieved a MAE of 4.27 on the test set but 5.99 on the validation set [21]. As such, considering the main purpose of this work is to evaluate the performance of rank-monotonicity enhanced ranking CNN for bone age estimation, we chose to validate the proposed method on a portion of 2000 images randomly picked from the training images.

Though the proposed methods achieved a remarkable performance, there are still some rooms to improve. Firstly, as some works shown that ensembling the results from different key parts of the hand (such as the ossification centers of epiphysis and carpal bones) could improve the estimation accuracy [20], our method may also benefit from the ensemble technique. Secondly, preprocessing seems to be an important component for generalizability of the algorithm. In this way, we applied the random geometric transformations to handle the different pose of the hand in the original dataset. It worthies to investigate whether using the deformation layer [4] could improve the performance. Finally, the generalizability of the proposed method needs to be validated on other dataset before it could be reliably used in clinic.

V. CONCLUSION

In this work, we illustrated that ranking learning is more suitable for the task of bone age estimation than classification and regression. The proposed method based on rank-monotonicity enhanced ranking CNN could accurately estimate bone age with a MAE of around 6 month, which was comparable with trained physicians and other CNN-based automatic methods. Though further analysis is needed, it has the potential to be used in clinic for automatic bone age assessment.

REFERENCES

[1] D. D. Martin, J. M. Wit, Z. Hochberg, L. Sävendahl, R. R. Van Rijn, O. Fricke, N. Cameron, J. Caliebe, T. Hertel, D. Kiepe, K. Albertsson-Wikland, H. H. Thodberg, G. Binder, and M. B. Ranke, "The use of bone age in clinical practice—Part 1," *Hormone Res. Paediatrics*, vol. 76, no. 1, pp. 1–9, 2011.

[2] W. W. Greulich and S. I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist," *Amer. J. Med. Sci.*, vol. 238, no. 3, p. 393, 1959.

[3] G. Beunen, J. Lefevre, M. Ostyn, R. Renson, J. Simons, and D. Van Gerven, "Skeletal maturity in Belgian youths assessed by the Tanner-Whitehouse method (TW2)," *Ann. Hum. Biol.*, vol. 17, no. 5, pp. 355–376, Jan. 1990.

[4] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-Ray images," *Med. Image Anal.*, vol. 36, pp. 41–51, 2017.

[5] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 52–66, Jan. 2009.

[6] V. Gilsanz and O. Ratib, *Hand Bone Age: A Digital Atlas of Skeletal Maturity*. Berlin, Germany: Springer, 2005.

[7] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 585–592.

[8] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4920–4928.

[9] S. Chen, C. Zhang, and M. Dong, "Deep age estimation: From classification to ranking," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2209–2222, Aug. 2018.

[10] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 865–872.

[11] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3396–3399.

[12] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2570–2577.

[13] D. Cao, Z. Lei, Z. Zhang, J. Feng, and S. Z. Li, "Human age estimation using ranking SVM," in *Proc. Chin. Conf. Biometric Recognit.*, 2012, pp. 324–331.

[14] M. Chu, B. Liu, F. Zhou, X. Bai, and B. Guo, "Bone age assessment based on two-stage deep neural networks," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–6.

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

[16] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI)*, 2016, pp. 265–283.

[17] S. Mutasa, P. D. Chang, C. Ruzal-Shapiro, and R. Ayyala, "MABAL: A novel deep-learning architecture for machine-assisted bone age labeling," *J. Digit. Imag.*, vol. 31, no. 4, pp. 513–519, Aug. 2018.

[18] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, Nov. 2018.

[19] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.

[20] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, and M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, "The RSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, Nov. 2018.

[21] *Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge*. Accessed: Aug. 2, 2019. [Online]. Available: https://www.16bit.ai/blog/ml-and-future-of-radiology

BO LIU received the Ph.D. degree from Beihang University, Beijing, in 2015, where he is currently an Assistant Professor with the Image Processing Center. His research interests include biomedical image processing, pattern recognition, and medicalphysics.

YU ZHANG received the bachelor's degree in control engineering from Beihang University, Beijing, in 2017, where he is currently pursuing the master's degree in pattern recognition. His research interests include medical image processing and deep learning.

MEICHENG CHU received the M.S. degree in control engineering from Beihang University, Beijing, in 2019. Her research interests include biomedical image processing, pattern recognition, and deep learning.

XIANGZHI BAI received the B.S. and Ph.D. degrees from Beihang University, Beijing, in 2003 and 2009, respectively, where he is currently a Professor with the Image Processing Center. He has published more than 100 international journals and conference papers in the fields of mathematical morphology, fuzzy clustering, image analysis, pattern recognition, and bioinformatics. He holds 20 national invention patents. He also acts as an active Reviewer of around 40 international journals and conferences.

FUGEN ZHOU received the Ph.D. degree from Beihang University, Beijing, China, in 2006, where he is currently a Professor with the Image Processing Center. His research interests include the medical image processing techniques, image registration and fusion techniques, digital video processing techniques, and deep learning techniques.