CircleNet for Hip Landmark Detection

Hai Wu,[1] Hongtao Xie,[1]* Chuanbin Liu,[1] Zheng-Jun Zha,[1] Jun Sun,[2] Yongdong Zhang[1]
[1]School of Information Science and Technology, University of Science and Technology of China
[2]Anhui Province Children's Hospital of China
{wuh, lcb592}@mail.ustc.edu.cn, {htxie, zhazj, zhyd73}@ustc.edu.cn, sunjun500@aliyun.com

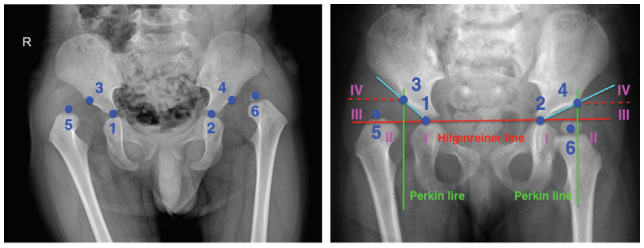
Abstract

Landmark detection plays a critical role in diagnosis of Developmental Dysplasia of the Hip (DDH). Heatmap and anchor-based object detection techniques could obtain reasonable results. However, they have limitations in both robustness and precision given the complexities and inhomogeneity of hip X-ray images. In this paper, we propose a much simpler and more efficient framework called CircleNet to improve the accuracy of landmark detection by predicting landmark and corresponding radius. Using the CircleNet, we not only constrain the relationship between landmarks but also integrate landmark detection and object detection into an end-to-end framework. In order to capture the effective information of the long-range dependency of landmarks in the DDH image, here we propose a new context modeling framework, named the Local Non-Local (LNL) block. The LNL block has the benefits of both non-local block and lightweight computation. We construct a professional DDH dataset for the first time and evaluate our CircleNet on it. The dataset has the largest number of DDH X-ray images in the world to our knowledge. Our results show that the CircleNet can achieve the state-of-the-art results for landmark detection on the dataset with a large margin of 1.8 average pixels compared to current methods. The dataset and source code will be publicly available.

Introduction

In medical image analysis, landmarks have significant clinical and scientific value. Clinical measurements, derived from the landmarks in X-ray images, are used for diagnosis and surgeries. Developmental Dysplasia of the Hip (DDH) is one of the most common diseases of skeletal system in infants and children. Current common method is proposed by Tonnis (Tönnis 1985). The key to the Tonnis's method is detecting six landmarks (see Figure 1(a)) to estimate the degree (see Figure 1(b)) of DDH.

The accuracy of detection directly affects the diagnosis results. Many children do not receive timely treatment based on two reasons. a) The characteristic of lower contrast in hip X-ray images and the diversities of bone morphology



(a) (b)

Figure 1: In image of DDH (a), six blue landmarks need to be detected. The figure (b) is a schematic diagram of the clinical diagnosis of the hip joint. We need detect four landmarks (1, 2, 3, 4) to draw Hilgenreiner line (Tönnis 1985) and Perkin line (Tönnis 1985) to divide areas shown as I, II, III, IV. When landmarks 5 and 6 are detected, the degree of DDH depends on which areas they are in.

(see Figure 2). b) The lack of medical facilities and professional doctors in remote rural areas hospitals. How to solve the shortage of medical resources and improve the accuracy of DDH diagnosis has become a significant problem in the health field of many countries.

Recent years have witnessed the progress of deep learning in object detection (Duan et al. 2019) (Zhou, Zhuo, and Krahenbuhl 2019) (Wang et al. 2019b) and landmark detection, especially in medical image analysis (Payer et al. 2016) (Xu et al. 2017) (Xie et al. 2019) (Liu et al. 2019). Because of noise, lower contrast, blurry boundaries and various shapes of bones in hip X-ray images, it is difficult to obtain precise landmarks. Meanwhile, segmentation (Ronneberger, Fischer, and Brox 2015) (Xu et al. 2017) is a common method in medical image processing. Due to the complex morphological structures of skeletons in hip X-ray images, it is difficult to mark the accurate bone contours for segmentation operation. In addition to landmark detection, we need to find 5- and 6-centered femoral head regions (red circles in Figure 3) to better judge whether the hip is dislocated. For example, two red circles in Figure 3 have big difference in size, which means the patient may have symptom of DDH. This suggests that we need cross-domain research

*Corresponding author

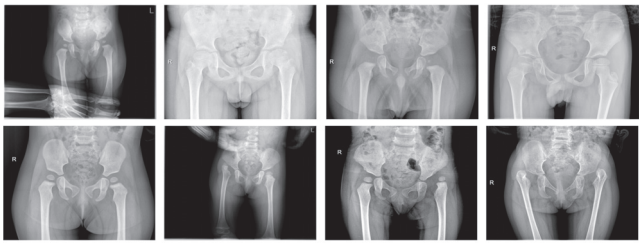


Figure 2: Lower contrast and the diversities of bone morphology in hip X-ray images. The upper four images from left to right represent four different characteristics (relatively clear, lower contrast, infant and child). The bottom four images represent four different morphology (normal, left hip dislocation, right hip dislocation, bilateral hip dislocation).

at landmark detection and object detection.

The recent approaches for object detection can be categorized into two classes. The first is to use anchors over an image and classify them directly. Anchor-based methods include two-stage detector (Ren et al. 2015) (Lu et al. 2019) and multi-stage methods (Cai and Vasconcelos 2018) (Chen et al. 2019). These methods need post-processing, namely Non-Maxima Suppression (NMS), then remove duplicated detections by computing IoU. The second is anchor-free methods (Lin et al. 2017) (Kong et al. 2019) (Duan et al. 2019), which usually need NMS or complex grouping of predicted landmarks.

In this paper, we provide a much simpler and more efficient framework called CircleNet which combines landmark detection and object detection. As shown in Figure 3, the CircleNet predicts a radius while detecting landmarks. For landmarks 1 and 3, these radii are the distance between them. For landmarks 2 and 4, these radii are also the distance between them. In this way, we constrain the relationship between landmarks instead of predicting them in isolation. For landmarks 5 and 6, we need these radii to get the size of their respective femoral head regions to provide a more accurate clinical diagnosis. Namely, landmark 5 and 6 are through self-constraint via respective radius to improve the accuracy of landmark detection. Our designed CircleNet combines landmark detection and object detection together to achieve end-to-end training through a unified framework.

At present, the mainstream method of landmark detection is based on heatmaps (Payer et al. 2016), but when this method extract features, it mainly focuses on local areas. To capture long-range dependency, repeating convolution operation is needed, which is computationally inefficient and hard to optimize (Wang et al. 2018). To address this issue, the non-local network (Wang et al. 2018) based on self-attention (Vaswani et al. 2017) is proposed to model the long-range dependency using only one layer. Because the non-local network computes the pairwise relations between the query position and all positions to form an attention map for each query position, this global modeling idea leads to higher computations. In order to better integrate the long-range dependency of images and simplify the computation,

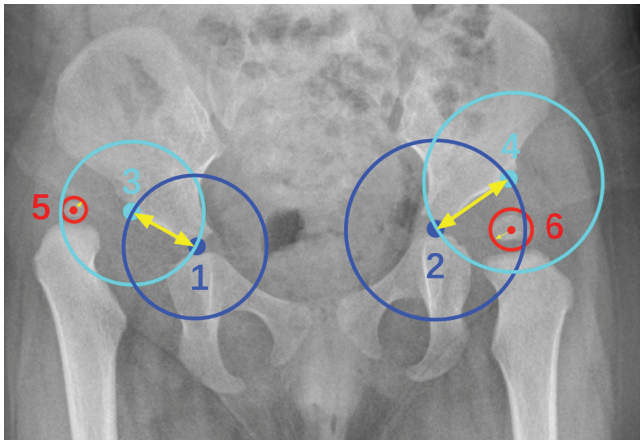


Figure 3: Six landmarks of different colors represent landmarks that need to be predicted. Circles of different colors are generated by the radii (yellow arrow) and corresponding color landmarks.

we design a block named Local Non-Local (LNL) which can be used in the backbone of the CircleNet.

In addition, we construct a professional DDH dataset for the first time, which has 9532 DDH X-ray images. According to the standards of professional doctors, the distribution (age and degree of DDH) of data is reasonable. The DDH dataset has significant clinical and scientific value. We evaluate the CircleNet on the dataset, and our results demonstrate that the CircleNet can achieve the state-of-the-art performance.

Related Work

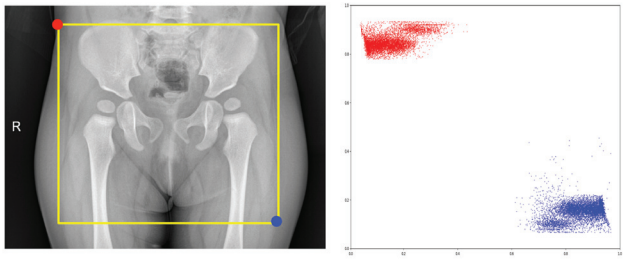
Landmark detection by heatmap or segmentation. Payer et al. (Payer et al. 2016) output a heatmap to detect landmarks. Ronneberger et al. (Ronneberger, Fischer, and Brox 2015) propose a fully convolution network (FCN) called U-net to segment objects. (Xu et al. 2017) adopt a supervised action map for image segmentation to extract landmarks.

Object detection with implicit anchors. Faster RCNN (Ren et al. 2015) generates region proposal and uses numerous anchors to detect objects. (Cai and Vasconcelos 2018) adopts cascade anchor-based detectors to balance positive and negative samples. Hybrid Task Cascade (Chen et al. 2019) uses a multi-stage network with multi-branch to detect objects and get segmentation masks. Guided Anchoring (Wang et al. 2019a) proposes a new anchoring scheme that predicts sparse and arbitrary-shaped anchors.

Object detection without anchors. CornerNet (Law and Deng 2018) detects two bounding box corners as landmarks. Duan et al. (Duan et al. 2019) propose CenterNet, which detects objects using a triple, including one center point and two corners. CornerNet-Lite (Law et al. 2019) is an improved version of CornerNet. However, these methods require a grouping stage after landmark detection, which significantly slows down each algorithm. FoveaBox (Kong et al. 2019) sets central area as landmarks to be predict. The problem of this method is that center areas of objects have

fewer features to identify objects.

Modeling long-range dependency. The main approaches for long-range dependency modeling is to model the pairwise relations. This operation has recently been successfully used in machine translation and visual recognition (Hu et al. 2018) (Wang et al. 2018) (Yuan and Wang 2018). Non-local (Wang et al. 2018) adopts self-attention mechanisms to model the pixel-level pairwise relations to capture long-range dependency between all positions. CCNet (Huang et al. 2018) improves non-local block via stacking two criss-cross blocks. GCNet (Cao et al. 2019) adopts a query-independent formulation to model global context. Looking closely at Figure 4, we can see that the skeleton is basically distributed in the central area of the image, named region of interest, and very little useful information is provided at the edge of the image. Based on this observation, we can find that calculating the edge of the image with non-local is a waste of computation to model pixel-level pairwise relations. The propsed LNL block can effectively model the effective context as non-local (Wang et al. 2018), with the lightweight computation and amount of parameters. At the same time, the LNL block can achieve better performance than the non-local and GC block on our task.



(a) (b)

Figure 4: Yellow box in (a) denotes region of interest in DDH image to efficiently capture long-range dependency with the LNL block. The red and blue points in (a) denote corners of region of interest, which are extracted by the function findContours in OpenCV in train dataset 7706 images. Distributions of these points are shown in (b). Best viewed in color.

Proposed Method

Figure 5 illustrates the overall CircleNet framework for landmark detection and radius prediction. The backbone of CircleNet is ResNet-50.

Landmark detection and radius prediction

Given an input image I with width W and height H, we need to produce a landmark heatmap $\hat{Y} \in [0,1]^{\frac{W}{S} \times \frac{H}{S} \times C}$, where S is the output stride and C is the number of landmarks in an image, here $C = 6$. Similarly to (Dai et al. 2017), we use the default output stride of $S = 4$. $\hat{Y}_{x,y,c} = 1$ means a detected landmark. We adopt ResNet-50 as backbone to predict $\hat{Y}$ from an image I. The CircleNet is trained following (Law and Deng 2018). For each ground truth landmark p of class c, we denote $\tilde{p} = \lfloor \frac{p}{S} \rfloor$. We use a Gaussian kernel $Y_{xyc} = \exp(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2})$ to generate ground truth of landmarks onto a heatmap $Y \in [0,1]^{\frac{W}{S} \times \frac{H}{S} \times C}$, where σ_p is a changeable standard deviation. If these Gaussian labels have overlaps, we take the element-wise maximum $M_{xyc} = \max_{c=1,2,\ldots,C} Y_{xyc}$. The training loss can be formulated as focal loss:

$$L_l = -\frac{1}{N} \sum_{xyc} \psi_{xyc} (1 - \hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}). \tag{1}$$

Where

$$\hat{Y}_{xyc} = \begin{cases} \hat{Y}_{xyc} & \text{if } Y_{xyc} = 1 \\ 1 - \hat{Y}_{xyc} & \text{otherwise} \end{cases} \tag{2}$$

and

$$\psi_{xyc} = \begin{cases} 1 & \text{if } Y_{xyc} = 1 \\ (1 - M_{xyc})^{\beta} & \text{otherwise.} \end{cases} \tag{3}$$

N is the number of landmarks in an image I, α and β are default parameters of the focal loss. We expect N to be 6. Similarly to (Law and Deng 2018), $\alpha = 2$ and $\beta = 4$ are default in our experiments.

To compensate for the error caused by downsampling, we additionally predict a local offset $\hat{O}$ for each landmark. We use L1 loss to train offset, and loss function is

$$L_o = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - (\frac{p}{S} - \tilde{p}) \right|. \tag{4}$$

We denote $(x^{(l)}, y^{(l)})$ as the landmark of image with category c_l. We use our final heatmaps to predict all landmarks. At the same time, we regress to the radius r_l for each class c_l. L1 loss is adopted at each landmark to regress radius

$$L_r = \frac{1}{N} \sum_{l=1}^{N} \left| \hat{R}_{p_l} - r_l \right|. \tag{5}$$

Here $\hat{R}_{p_l}$ represents the ground truth radius of each landmark. The whole training loss function consists of three basic parts:

$$L_{circle} = L_l + \lambda_r L_r + \lambda_o L_o. \tag{6}$$

In our experiments, we adopt $\lambda_r = 0.1$ and $\lambda_o = 1$ as default setting, and other values of λ_r are shown in experiments section. The CircleNet can predict different radii at different landmarks. As result shown in Figure 5, we treat landmark 1 and 3 (2 and 4) as a group, and radii of these two landmarks are the distance between them. For landmark 5 or 6 via self-restraint, we predict the circumcircle (red circle in Result) of the femoral head. Using the CircleNet, we not only constrain the relationship between landmarks but also integrate landmark detection and object detection into an end-to-end framework.

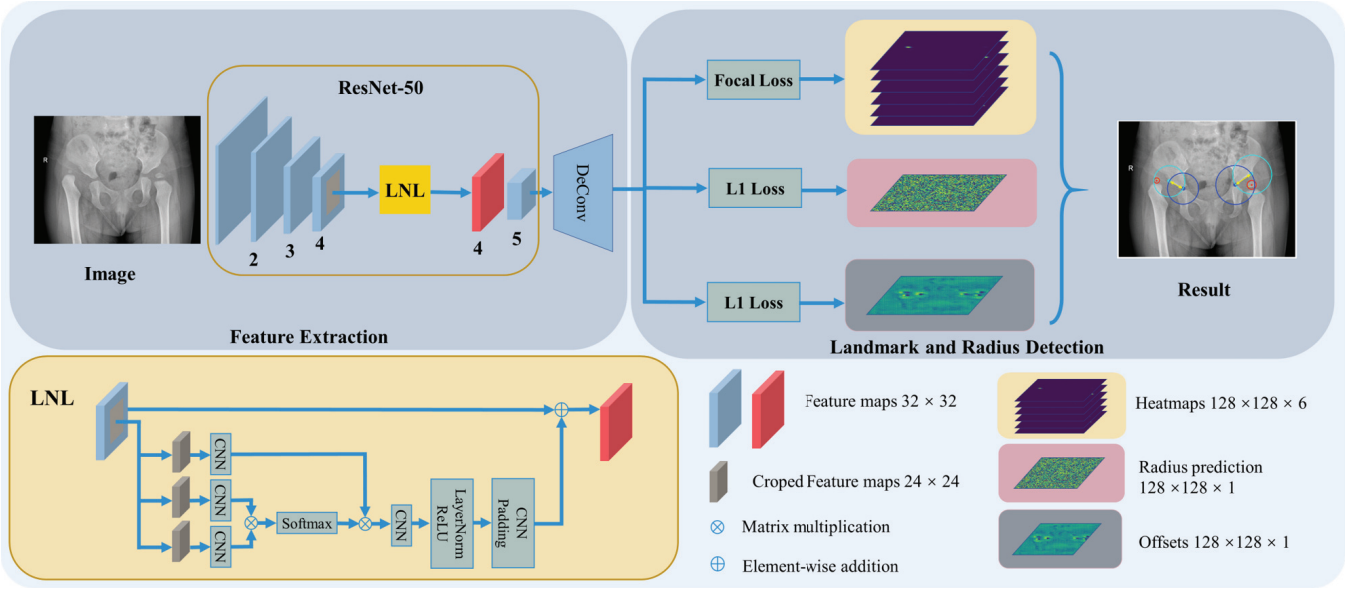


Figure 5: Illustration of the CircleNet for the DDH images landmark detection. The backbone is default ResNet-50 with four basic residual block named stage 2, 3, 4, 5. The DeConv in figure denotes transposed convolution. The overall architecture of CircleNet mainly comprises two components, i.e. the feature extraction section and landmark and radius detection section. We use an end-to-end network to predict the $\hat{Y}$, offset $\hat{O}$, and $\hat{R}$ of each landmark. As shown in figure, the LNL block is embedded after stage 4 of the backbone to efficiently capture long-range dependency of pixel-wise relationship. The detail of the LNL block is shown in figure. Best viewed in color.

Local non-local block

The classic non-local block can be used to improve the features between the query position and other positions. We denote $F = \{F_i\}_{i=1}^{N_p}$ as the feature map of an image, where $N_p = W \times H$. F is the input of the non-local block, and Z is output. F and Z have the same dimensions. We can express the non-local block as

$$Z_i = F_i + W_z \sum_{j=1}^{N_p} \frac{f(F_i, F_j)}{\phi(F)}(W_v, F_j). \tag{7}$$

In this formula, i is the query position, and j are other possible positions. We denote $f(F_i, F_j)$ as the relationship between position i and j. $\phi(F)$ is a normalization factor. W_z and W_v denote linear transform matrices. $\omega_{ij} = \frac{f(F_i,F_j)}{\phi(F)}$ denotes pariwise relationship between i and j. The most widely-used method, Embedded Gaussian, is illustrated in Figure 6(a). The ω_{ij} is defined as $\omega_{ij} = \frac{\exp(\langle W_q F_i, W_k F_j \rangle)}{\sum_m \exp(\langle W_q F_i, W_k F_m \rangle)}$.

In order to make full use of the effective information of the long-range dependency of landmarks in the image, and simplify the computation and amount of parameters, here we propose a novel Local Non-Local (LNL) block. The detailed architecture of the LNL block is illustrated in Figure 6(d), formulated as

$$Z_i = F_i + \eta(\sum_{j=1}^{\mu^2 N_p} \frac{f(F_i, F_j)}{\phi(F)}(W_v, F_j)), \tag{8}$$

where $\eta(\cdot)$ denotes $W_{z2}\mathrm{ReLU}(\mathrm{LN}(W_{z1}(\cdot))$.

Different from the traditional non-local block, our LNL block has three advantages. a) With less number of parameters compared to non-local. The LNL block is used between stage 4 and 5 of the backbone. We replace 1×1 convolution, namely W_z in Figure 6(a), with LayerNorm, ReLU and two 1×1 convolution, shown as W_{z1} and W_{z2} in Figure 6(d). The number of parameters drops from 1024×1024 to $2 \times 1024 \times 1024/\theta$, where θ is to reduce the number of channels. With default reduction ratio set to $\theta = 8$, the number of parameters can be reduced to 1/4. More results on different θ are shown in Table 3. b) Reduce computations. We crop the feature maps after stage 4, and the long-range dependency computing shrinks from $32 \times 32 \times 1024$ to $32 \times \mu \times 32 \times \mu \times 1024$, where μ is the area ratio of feature. We use function findContours in OpenCV to obtain regions of interest where main pelvises are in, as shown in Figure 4(a). The default area ratio is set to $\mu = 24/32$, and more parametric comparison results are shown in Table 5. c) Focus on effective region of interest to improve long-range dependency. We use padding $= 4$ in W_{z2} to recover the size to $32 \times 32 \times 1024$. Based on the region of interest, the LNL block can pay more attention to area where concentrate most of the relate information of all landmarks. At the same time, focusing on region of interest can suppress interference by unrelated information in edge region of image.

Inference

We obtain peaks in heatmap for each category of landmark in every channel to get six landmarks. We extract all peaks whose value is greater or equal to its neighbors points. Fi-

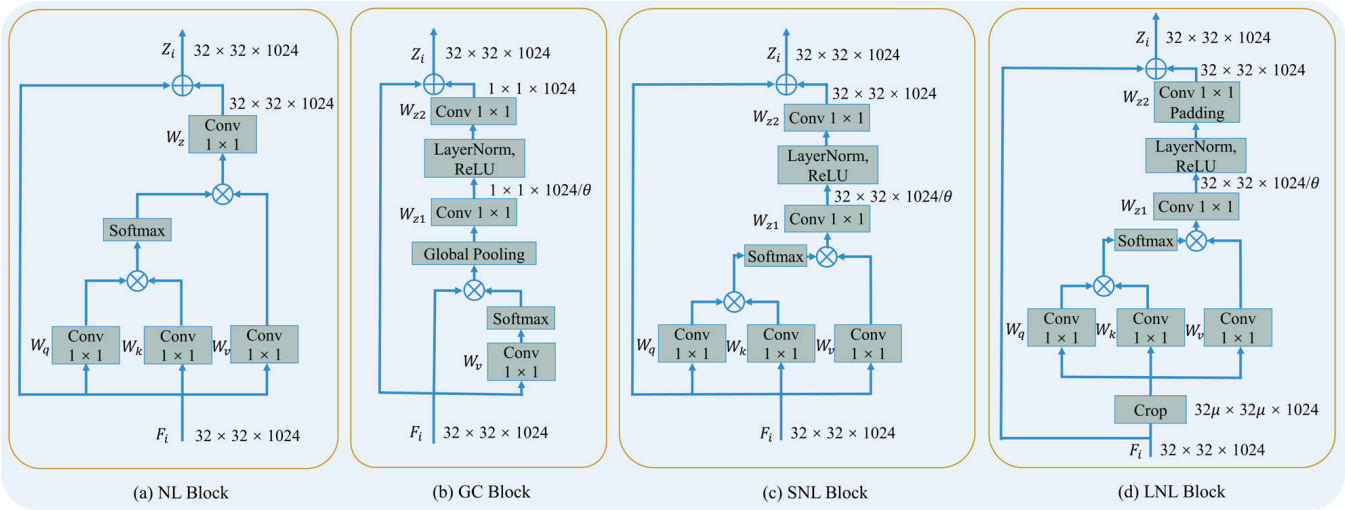


Figure 6: Illustration of several blocks. These blocks are used after stage 4 as default, and number of channels and size of feature maps are shown in blocks. (a) is basic non-local (NL) block. (b) is global context (Cao et al. 2019) (GC) block. (c) is simplified non-local (SNL) block to reduce the amount of parameters. (d) is the proposed LNL block. $\oplus$ denotes element-wise addition, and $\otimes$ denotes matrix multiplication.

nally, we extract the maximum peaks in every channel as six final outputs. At the same time, we use the offset to compensate for the error caused by downsampling. We adopt the predicted radius to obtain the areas of bone. All outputs are produced directly from the landmark estimation without the need of NMS or other post-processing. Finally, we find six maximum confidence landmarks of the six categories and the corresponding radii in each image as the final output.

Experiments

To evaluate the CircleNet, we carry out a series experiments on our DDH dataset. Experimental results demonstrate that the proposed CircleNet outperforms other methods. The proposed LNL block can bring further improvements in accuracy of landmark detection.

Dataset

The DDH dataset is collected in the process of clinical routine and contain all common conditions in clinical cases between 2013-2019. The DDH X-ray images are collected from Children's Hospital. We extracted the original DICOM format files from the hospital PACS system, and we converted these files into JPEG format images. All landmarks of dataset are labeled by fifteen professional doctors. Each image has a corresponding txt document which contains coordinates of landmarks and radii of femoral heads. Patients are between 0.1-12 years old. The total number of DDH images is 9532, in which 7706 images are used for training and the rest 1826 images are for testing. The dataset is ready for openness. Now, the dataset is available from authors upon reasonable request.

Experimental Setup

We apply the CircleNet to the DDH dataset for landmark detection. The CircleNet is trained using the Pytorch framework on a Ubuntu workstation equipped with an Intel i7-9700 CPU and two 11GB Nvidia GeForce 1080Ti GPUs. During training, the mini batch size is set to 12. Adagrad optimizer is used for updating with the learning rate of 1.25e-4. The default training epoch is 30. During training, we resize the input resolution to 512 $\times$ 512. At inference, we recover the output to original size to statistically analyze behaviors of different methods.

State-of-the-art comparison

We compare the CircleNet with other approaches in Table 1. Mean distance error of each landmark position in pixels is used for comparison, and we compare the Missed Detection (MD) and Frame-Per-Second (FPS). These approaches include mainstream segmentation networks such as Unet (Ronneberger, Fischer, and Brox 2015) and SAC (Xu et al. 2017) to detect landmarks. One-stage methods in object detection such as RetinaNet (Lin et al. 2017), FCOS (Tian et al. 2019), GHM (Li, Liu, and Wang 2019) are listed in the table. Other methods in object detection include Faster R-CNN (Ren et al. 2015), Fater R-CNN (Ren et al. 2015) with Dconv2 (Zhu et al. 2019b), Grid R-CNN (Lu et al. 2019), Cascade R-CNN (Cai and Vasconcelos 2018), Hybrid Task Cascade (Chen et al. 2019), GN (Wu and He 2018) with WS (Qiao et al. 2019), Libra R-CNN (Pang et al. 2019), and Generalized Attention (Zhu et al. 2019a). For most different methods, we use the same backbone ResNet-50 to test to ensure the comparability of results. Label of each landmark of these object detection methods in training is bounding box with a side length of $2r$. We can find in Tabel 1, thanks to its simple algorithm and no need for complex post-

Table 1: State-of-the-art comparison on the DDH test dataset which include 1826 images. Mean distance error of landmark detection is measured in pixels. The lmk in table denotes landmark. FPS is measured on the same computer with a Nvidia GeForce 1080Ti GPU. Miss detection (MD) denotes number of images on which at least one landmark is not found, and these images do not participate in the statistics of mean error in pixels. Average denotes mean error of six landmarks in an image.

	Backbone	FPS	MD	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
Unet	Unet	3.2	17	8.03	8.12	5.26	6.53	9.74	9.30	7.83
SAC	FCN	1.9	46	11.93	11.11	65.04	63.79	18.83	15.97	31.11
RetinaNet	ResNet-50	15.4	**0**	7.65	8.64	5.89	7.60	6.32	6.91	7.17
FCOS	ResNet-50	15.0	**0**	10.66	10.48	7.90	10.91	12.22	12.61	10.80
RetinaNet+GHM	ResNet-50	16.8	**0**	7.67	8.80	5.35	7.05	5.59	7.04	6.92
Faster R-CNN	ResNet-50	10.2	12	7.49	8.50	5.29	6.87	5.17	6.58	6.65
Faster R-CNN+Dconv2	ResNet-50	12.4	14	7.25	8.04	5.71	7.26	5.12	6.29	6.61
Grid R-CNN	ResNet-50	9.1	9	8.22	9.37	6.30	8.05	5.28	6.26	7.25
Cascade R-CNN	ResNet-50	7.4	15	7.74	8.51	5.74	7.46	5.14	6.16	6.79
Hybrid Task Cascade	ResNet-50	3.9	4	7.90	8.74	6.09	8.01	5.91	6.83	7.25
Faster R-CNN+GN+WS	ResNet-50	6.4	21	7.40	8.28	5.30	6.75	5.32	6.50	6.59
Libra R-CNN	ResNet-50	13	13	7.34	8.39	5.38	8.00	5.50	6.68	6.88
Generalized Attention	ResNet-50	9.8	**0**	7.52	8.30	6.05	7.75	5.38	6.79	6.97
CircleNet	ResNet-50	**25.6**	**0**	6.16	**6.13**	4.54	4.72	**4.22**	4.14	4.99
CircleNet+LNL	ResNet-50	22.7	**0**	**5.68**	6.21	**4.17**	**4.40**	4.26	**4.01**	**4.79**

Table 2: Different blocks in different stages of the backbone ResNet-50 to capture long-range dependency.

Block	Stage 4	Stage 5	FPS	MD	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
-	-	-	25.6	**0**	6.16	6.13	4.54	4.72	4.22	4.14	4.99
NL	✓		22.2	2	6.07	6.46	4.41	4.59	3.91	4.24	4.95
NL		✓	13.3	1	5.78	6.29	4.28	4.88	4.23	4.05	4.92
GC	✓		11.1	1	5.74	**6.08**	4.53	4.70	4.01	4.08	4.86
GC		✓	4.5	4	6.10	6.29	4.34	4.93	4.11	4.14	4.99
SNL	✓		21.7	1	5.88	5.97	4.19	4.93	4.10	4.44	4.92
SNL		✓	12.2	**0**	6.29	6.26	4.62	4.69	**3.95**	4.06	4.98
LNL	✓		**22.7**	**0**	**5.68**	6.21	**4.17**	**4.40**	4.26	**4.01**	**4.79**
LNL		✓	11.8	**0**	6.07	6.35	4.54	4.62	4.10	4.10	4.96

processing, the Circlenet achieves the highest speed of landmark detection (FPS=25.6) and zero MD. The CircleNet reduces 1.6 average pixels error compared to Faster R-CNN (Ren et al. 2015) with GN (Wu and He 2018) and WS (Qiao et al. 2019). Because of the LNL block can capture effective features in areas where pelvises are located, it can bring the precision of landmark detection. As shown in the tabel, the CircleNet with LNL block can improve 1.8 average pixels. Compared with the CircleNet, the LNL block contributes 0.2 average pixels.

Additional experiments

In order to explain the impact of other parameter settings in more detail, we make the following comparative tests on the DDH dataset.

Different blocks in different stages of the backbone ResNet-50 to capture long-range dependency. We compare four methods (NL (Wang et al. 2018) in Figure 6(a), GC (Cao et al. 2019) in Figure 6(b), simplified non-local (SNL) in Figure 6(c), LNL in Figure 6(d) after stage 4 and stage 5 of ResNet-50. The result is shown in Table 2. The GC block and SNL block can be respectively formulated as

$$Z_i = F_i + \eta(\sum\nolimits_{j=1}^{N_p} \frac{e^{W_v F_j}}{\sum_{m=1}^{N_p} e^{W_v F_m}} F_j) \tag{9}$$

and

$$Z_i = F_i + \eta(\sum\nolimits_{j=1}^{N_p} \frac{f(F_i, F_j)}{\phi(F)}(W_v, F_j)), \tag{10}$$

where $\eta(\cdot)$ is $W_{z2}\mathrm{ReLU}(\mathrm{LN}(W_{z1}(\cdot))$. We set $\lambda_r = 0.1$, $\mu = 24/32$, $\theta = 8$. As we can see, the LNL block after stage 4 can obtain lowest average pixels error with higher FPS.

Radius weight λ_r in the CircleNet. In order to illustrate the effect of introducing the radius loss on the detection accuracy of landmarks, we compare different values of radius weight λ_r, and results are shown in Table 3. We set $\mu = 24/32$, $\theta = 8$, and use the LNL after stage 4 of the backbone. Because of the values of MD and FPS are almost

same for different λ_r, these values are not shown in the table. We can find that $\lambda_r = 0.1$ gives a good result compared to other λ_r. The average error can reduce 0.11 average pixels with $\lambda_r = 0.1$ compared to $\lambda_r = 0.01$.

Table 3: Different radius weights λ_r in the CircleNet.

λ_r	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
0.01	5.88	**6.01**	4.31	4.97	4.17	4.05	4.90
0.1	**5.68**	6.21	**4.17**	**4.40**	4.26	4.01	**4.79**
0.2	5.96	6.26	4.36	4.96	**4.03**	3.93	4.92
0.5	5.90	6.03	4.36	4.71	4.06	**3.89**	4.82
0.7	5.87	6.10	4.26	4.61	4.05	4.11	4.83
1	5.92	6.39	4.37	4.60	4.00	4.13	4.90

Different μ in LNL block to pay attention to regions of interest. We analyze the sensitivity of the LNL block to the area ratio of regions of interest μ. Different values of μ affect computations. We use the function findCounter in OpenCV to get regions of interest where pelvises are located in 7706 images, and distribution of region corners are shown in Figure 4(b). After the stage 4 of backbone, the feature maps is 32×32. We can find in Figure 4(b) that the region of interest is probably located in the center of the image, which accounts for about 3/4 $\times$ 3/4 of the image. Three different values of μ (20/32, 24/32, 28/32) are for comparison experiments, results are shown in Table 4. $\mu = 24/32$ is the best with 0.05 average pixels error reducing compared to $\mu = 20/32$. Feature maps area with $\mu = 24/32$ can not only pay more attention to regions of interest but also suppress information of edge regions in images.

Table 4: Different μ in LNL block to pay attention to regions of interest.

μ	FPS	MD	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
20/32	**23.1**	3	**5.68**	**6.12**	4.21	4.99	**3.98**	4.08	4.84
24/32	22.7	**0**	**5.68**	6.21	**4.17**	**4.40**	4.26	**4.01**	**4.79**
28/32	22.2	1	5.80	6.36	4.45	4.71	4.20	4.00	4.92

Different methods in LNL block to capture long-range dependency. To meet various needs in practical applications, four methods of the non-local block (Wang et al. 2018) with different ω_{ij} are designed, namely Gaussian (Gau), Embedded Gaussian (E-Gau), Dot product (Dot pro), and Concat. Gaussian denotes ω_{ij} as the Gaussian function, which is defined as $\omega_{ij} = \frac{\exp(\langle F_i, F_j \rangle)}{\sum_m \exp(\langle F_i, F_m \rangle)}$. For Dot product, ω_{ij} is formulated as $\omega_{ij} = \frac{\langle W_q F_i, W_k F_j \rangle}{N_p}$. Concat is defined as $\omega_{ij} = \frac{\mathrm{ReLU}(\mathrm{W}_q[\mathrm{F}_i, \mathrm{F}_j])}{N_p}$. We compare these four methods in the LNL block, and results can be seen in Table 5. As can be seen from the table, the Embedding-Gaussian gives a lowest pixels average error compared to other methods. Gaussian has the highest FPS, which is 0.6 higher than Embedding-Gaussian.

Different θ in LNL block to reduce the amount of parameters. We alert the ratio θ to reduce redundancy in parameters and provide a tradeoff between performance and amount of parameters. Results are shown in Table 6. We can find that $\theta = 8$ can bring at least average 0.01 pixels improvement in landmarks accuracy with lowest MD.

Table 5: Different methods in LNL block to capture long-range dependency.

Method	FPS	MD	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
Gau	**23.3**	1	5.81	**6.14**	4.74	5.30	**3.99**	**3.95**	4.99
E-Gau	22.7	**0**	**5.68**	6.21	**4.17**	4.40	4.26	4.01	**4.79**
Dot pro	19.6	**0**	5.95	6.47	4.41	**4.35**	4.08	3.98	4.87
Concat	19.1	5	5.81	**6.14**	4.74	5.30	**3.99**	**3.95**	4.99

Table 6: Different θ in LNL block to reduce the amount of parameters.

θ	FPS	MD	lmk1	lmk2	lmk3	lmk4	lmk5	lmk6	Average
4	**22.7**	**0**	5.82	6.24	4.47	4.56	4.04	4.01	4.86
8	**22.7**	**0**	**5.68**	6.21	**4.17**	**4.40**	4.26	4.01	**4.79**
16	22.6	2	5.85	**5.90**	4.51	4.48	4.05	**3.99**	4.80
32	**22.7**	1	5.78	6.08	4.58	4.60	**4.00**	4.29	4.89

Conclusion

In this paper, we present a novel approach to address the problem of landmark detection in hip X-ray images. We first construct a professional DDH dataset which is of great significance to both clinical practice and scientific research. We propose the CircleNet by integrating landmark detection and object detection into an end-to-end framework. Based on this integration, the CircleNet can constrain the relationship between landmarks instead of predicting them in isolation. In addition, the LNL block is designed to effectively capture long-range dependency of regions of interest in DDH images. Using the CircleNet, we present superior performances against other methods. Further investigation on its clinical value will be performed.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (61525206, 61771468, 61976008), the Huawei-USTC Joint Innovation Project on Machine Vision Technology (FA2018111122), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209).

References

Cai, Z., and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q.

2019. Centernet: Object detection with keypoint triplets. *arXiv preprint arXiv:1904.08189*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on CVPR*, 3588–3597.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2018. Ccnet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*.

Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; and Shi, J. 2019. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*.

Law, H., and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

Law, H.; Teng, Y.; Russakovsky, O.; and Deng, J. 2019. Cornernet-lite: Efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*.

Li, B.; Liu, Y.; and Wang, X. 2019. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8577–8584.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE ICCV*, 2980–2988.

Liu, C.; Xie, H.; Zhang, S.; Xu, J.; Sun, J.; and Zhang, Y. 2019. Misshapen pelvis landmark detection by spatial local correlation mining for diagnosing developmental dysplasia of the hip. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 441–449. Springer.

Lu, X.; Li, B.; Yue, Y.; Li, Q.; and Yan, J. 2019. Grid r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7363–7372.

Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; and Lin, D. 2019. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 821–830.

Payer, C.; Štern, D.; Bischof, H.; and Urschler, M. 2016. Regressing heatmaps for multiple landmark localization using cnns. In *International Conference on MICCAI*, 230–238. Springer.

Qiao, S.; Wang, H.; Liu, C.; Shen, W.; and Yuille, A. 2019. Weight standardization. *arXiv preprint arXiv:1903.10520*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*.

Tönnis, D. 1985. Indications and time planning for operative interventions in hip dysplasia in child and adulthood. *Zeitschrift fur Orthopadie und ihre Grenzgebiete* 123(4):458–461.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wang, J.; Chen, K.; Yang, S.; Loy, C. C.; and Lin, D. 2019a. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2965–2974.

Wang, Y.; Xie, H.; Fu, Z.; and Zhang, Y. 2019b. Dsrn: a deep scale relationship network for scene text detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 947–953. AAAI Press.

Wu, Y., and He, K. 2018. Group normalization. In *Proceedings of the ECCV*, 3–19.

Xie, H.; Yang, D.; Sun, N.; Chen, Z.; and Zhang, Y. 2019. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition* 85:109–119.

Xu, Z.; Huang, Q.; Park, J.; Chen, M.; Xu, D.; Yang, D.; Liu, D.; and Zhou, S. K. 2017. Supervised action classifier: Approaching landmark detection as image partitioning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 338–346. Springer.

Yuan, Y., and Wang, J. 2018. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*.

Zhou, X.; Zhuo, J.; and Krahenbuhl, P. 2019. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 850–859.

Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; and Dai, J. 2019a. An empirical study of spatial attention mechanisms in deep networks. *arXiv preprint arXiv:1904.05873*.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019b. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9308–9316.