Medical Image Analysis 68 (2021) 101904

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Cascaded convolutional networks for automatic cephalometric landmark detection

Minmin Zeng[a,*], Zhenlei Yan[b], Shuai Liu[c], Yanheng Zhou[d], Lixin Qiu[a]

[a] Fourth Clinical Division, School and Hospital of Stomatology, Peking University, Beijing, China
[b] Ling. AI, Beijing, China
[c] Second Clinical Division, School and Hospital of Stomatology, Peking University, Beijing, China
[d] Department of orthodontics, School and Hospital of Stomatology, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 11 May 2019
Revised 15 June 2020
Accepted 11 November 2020
Available online 18 November 2020

Keywords:
Cephalometric landmark detection
Convolutional neural network
Computer vision
X-ray image applications

ABSTRACT

Cephalometric analysis is a fundamental examination which is widely used in orthodontic diagnosis and treatment planning. Its key step is to detect the anatomical landmarks in lateral cephalograms, which is time-consuming in traditional manual way. To solve this problem, we propose a novel approach with a cascaded three-stage convolutional neural networks to predict cephalometric landmarks automatically. In the first stage, high-level features of the craniofacial structures are extracted to locate the lateral face area which helps to overcome the appearance variations. Next, we process the aligned face area to estimate the locations of all landmarks simultaneously. At the last stage, each landmark is refined through a dedicated network using high-resolution image data around the initial position to achieve more accurate result. We evaluate the proposed method on several anatomical landmark datasets and the experimental results show that our method achieved competitive performance compared with the other methods.

1. Introduction

Cephalometric analysis is a fundamental examination which is routinely used in fields of orthodontics and orthognathics (Proffit et al., 2006). Annotating the landmarks of the dental, skeletal, and soft tissue structures from lateral cephalograms is the key procedure in cephalometric analysis, since they serve as the datum of the succeeding qualitative assessment of angles and distances which provide diagnosis information of the craniofacial condition of a patient and affect treatment planning decision.

Due to the X-ray imaging quality of the skull and the individual variations of anatomical types (Lindner et al., 2016), it is not easy to reliably locate the landmarks in lateral cephalograms within high precision (Baumrind and Frantz, 1971). Even for an experienced orthodontist, it's still time-consuming to manually identify the landmarks consistently (intra-observer variations) (Durão et al., 2015). Moreover, orthodontists with various training and experience backgrounds may result in inconsistent annotations (inter-observer variations). Therefore, it will be of great value to construct an automatic computerized system that identifies cephalometric landmarks accurately, consistently and rapidly.

During the last decades, as technologies develops in computer vision and machine learning, lots of approaches were proposed to address this issue (Zhou and Abdel-Mottaleb, 2005; Nikneshan et al., 2015; Wang et al., 2016). In 2014 and 2015, IEEE International Symposium on Biomedical Imaging (ISBI) Grand Challenges[1] that focused on this task were organized (Wang et al., 2015). The summarized performance in these challenges showed significant improvement. However, it is still far from the goal of actual clinical practice, since the best accuracy among reported results is only about 73% of the detections fall in the clinically accepted precision range of 2.0 mm (Payer et al., 2019).

The difficulties in developing a fully automatic cephalometric landmark detection system mainly come from two aspects. Firstly, the lateral cephalogram is acquired by projecting the skull object into a 2-dimensional gray image with overlapping structures (Lindner et al., 2016), therefore it is difficult to extract useful image features by hand-crafted approach. Secondly, the overall medical imaging dataset for training is usually small due to high cost of annotations, so the learned system is vulnerable to overfitting and leads to poor performance on test data (Domingos, 2012). Recently, convolutional neural network (CNN) technique has achieved great success in wide range of computer vision applications, including image classification (Krizhevsky et al., 2012), face recognition

* Corresponding author.
E-mail address: bdzengmw@163.com (M. Zeng).

[1] http://www-o.ntust.edu.tw/~cweiwang/ISBI2015/challenge1/index.html.

https://doi.org/10.1016/j.media.2020.101904
1361-8415/

(Parkhi et al., 2015), object detection (Ren et al., 2015) and image segmentation (Ronneberger et al., 2015), due to its excellent capability to learn useful features from images automatically (LeCun et al., 2015). Therefore, researchers begin to apply CNN in medical image analysis as a promising new tool (Litjens et al., 2017).

In this paper, we treat cephalometric landmark detection as a multi-level regression problem and propose a novel approach using cascaded convolutional neural networks to solve it. The automatic prediction pipeline is composed of three stages, following a coarse-to-fine detection strategy. The first stage is designed to locate the lateral face area in the raw X-ray image. It can be considered as an alignment processing to overcome the obvious appearance variations among images. Given the lateral face area, the next stage aims to predict the coordinates of all landmarks simultaneously, which implicitly encodes the geometric shape constraints among landmarks. Due to the high complexity of coordinates regression (Pfister et al., 2015), it's hard to predict all landmarks within high precision range directly. Therefore, in the last stage, each landmark is independently refined by a network by processing the high-resolution image patch around the initial position to achieve more accurate estimation. This three-stage structure could utilize more training data and help to prevent overfitting problem. To evaluate the performance of the proposed approach, we run it on a public dataset from IEEE 2015 ISBI Grand Challenge and compare its results with the other state-of-the-art approaches. Furthermore, we setup additional experiments to visually demonstrate the mechanism of our approach and evaluate its generalization ability to other cephalograms acquired by different equipment and software.

Generally speaking, the major contributions of this paper are summarized as follows: (1) We proposed a novel approach based on cascaded convolutional neural networks to detect cephalometric landmark automatically. (2) Extensive experiments were conducted on public datasets and the results showed that the proposed method is comparable to other recent methods in anatomical landmark detection. (3) We constructed a new cephalogram dataset to evaluate the proposed method and publish it to the research community.

The rest of this paper is organized as follows. Section 2 briefly outlines the most relevant works in anatomical landmark detection. Section 3 describes our approach in detail. Section 4 shows the extensive experiments on several anatomical datasets and the discussion of results. Finally, Section 5 concludes this paper.

2. Related work

In this section, we briefly describe the most representative methods for cephalometric landmark detection problem.

2.1. Traditional approach

In the last decades, a considerable amount of methods for cephalometric landmark detection have been studied. Grau et al. (2001) proposed a template matching approach which adopted the features computed by image edge detection and contour segmentation operators for automated identification of landmarks from cephalograms. Forsyth and Davis (1996) demonstrated a two-stage approach which firstly detected a candidate set of points around the landmarks using rough and fine image features, then exploited the spatial relationships between landmarks to find the optimal candidate points. El-Feghi et al. (2004) used machine learning methods such as k-means clustering for automated cephalometric analysis.

Since previous studies lacked a public benchmark, Wang et al. (2015) organized two challenges (IEEE 2014 and 2015 ISBI Grand Challenges) on this task and summarized the performance of the detection methods (Wang et al., 2016). Ibragimov et al. (2015a) used a random forest-based classifier with Haar-like features (Ibragimov et al., 2015b) to model the appearance of landmarks, then combined the statistical shape representation defined by Gaussian kernel estimation (Ibragimov et al., 2012) to achieve the optimal landmark positions by applying game-theoretic optimization framework Ibragimov et al. (2014a,b). Lindner and Cootes (2015) applied random forest regression-voting to predict the likely position of each landmark respectively (Lindner et al., 2013; 2015), then adopted a statistical shape model (Cootes et al., 1995) to optimize all landmark positions to ensure consistency across the whole set.

2.2. CNN-based approach

Recently, despite the limited amount of annotated training images in the medical imaging fields, many CNN-based approaches were still proposed to solve anatomical landmark detection problem successfully. Lee et al. (2017) treated cephalometric landmark detection as a regression problem and proposed a single convolutional neural network to directly learn the positions of all landmarks, but it's difficult to be optimized. Arik et al. (2017) proposed a framework that firstly used a convolutional neural network to learn the probability whether the input image patch's center is a landmark, for each landmark respectively, then combined with a statistical shape model to refine all landmarks' optimal positions. Zhang et al. (2017) proposed a two-stage task-oriented deep neural networks to address the limited availability of medical imaging data for network learning in anatomical landmark detection. Urschler et al. (2018) presented a unified framework that combined both image appearance information and geometric landmark configuration into a unified random forest framework which was optimized iteratively to refine joint landmark predictions by using the coordinate descent algorithm. Payer et al. (2019) proposed a fully convolutional SpatialConfiguration-Net (SCN) that dedicated one component to predict locally accurate but ambiguous candidate landmarks, while the other component improved robustness to ambiguities by incorporating the spatial configuration of landmarks.

Despite these previous studies, it is still a challenging task to detect cephalometric landmark automatically on small training dataset within such high precision that each landmark can be located in the clinically accepted 2.0 mm precision range.

3. Methods

In this section, we explain the approach how to detect cephalometric landmarks.

3.1. Formulation

The cephalometric landmarks used in this study are shown in Fig. 1, where 19 types of landmarks (Wang et al., 2015) are annotated to assist cephalometric analysis. The computerized system is designed to predict the positions of all 19 landmarks given the input X-ray dental images. To facilitate discussions, the cephalometric landmark detection problem is formally described as follows. Let $\mathcal{X} \in \mathbb{N}^{W \times H}$ denote the image set of cephalograms which is in gray scale, where W is the image width and H is the image height. Let $\mathcal{Y} \in \mathbb{R}^{2K}$ denote all landmarks' coordinates space, where K is the number of landmarks, i.e. $K = 19$. Given an input image $X \in \mathcal{X}$, predicting cephalometric landmarks can be considered as learning a nonlinear function Φ, which maps from $X \in \mathcal{X}$ to coordinate vector $Y \in \mathcal{Y}$, i.e.

$$\Phi : X \longrightarrow Y \tag{1}$$

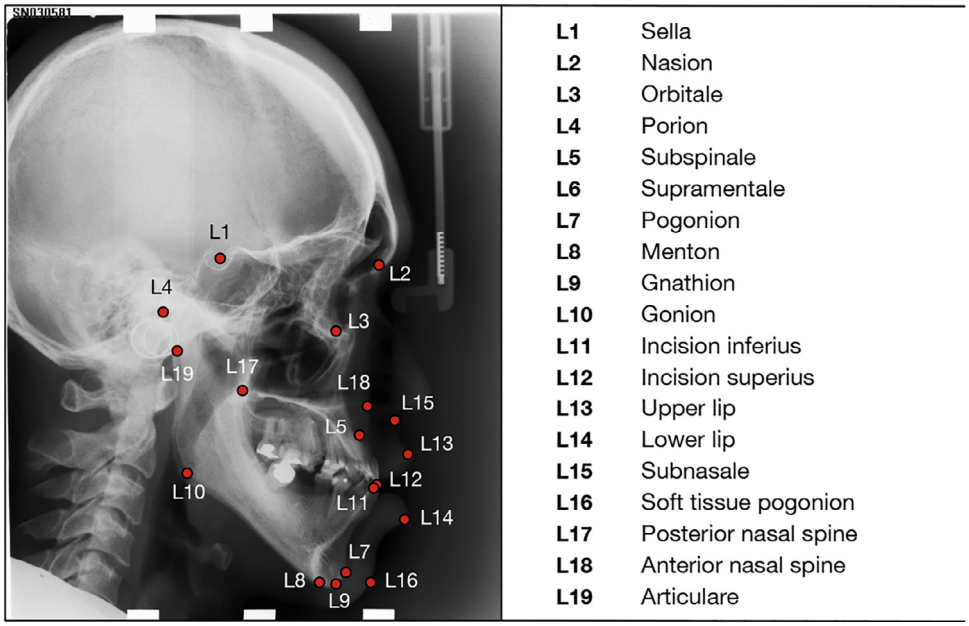


Fig. 1. The annotation example of the 19 cephalometric landmarks used in this study. The cephalogram is reproduced from image #001 of the IEEE 2015 ISBI Grand Challenge dataset.

3.2. Overall framework

Since regressing the coordinates directly involves a highly nonlinear mapping Φ (Pfister et al., 2015), it's difficult to optimize objective function especially on small training set. Motivated by the successful application of CNN-based models in facial keypoint detection task (Sun et al., 2013; Liu et al., 2015; Zhang et al., 2016), we treat cephalometric landmark detection as a multi-level regression problem by decomposing it into three subtasks, instead of solving Φ directly. These three subtasks perform a coarse-to-fine predicting procedure and are listed as below,

- How to align the lateral face area in cephalograms?
- How to estimate the initial coarse positions of all landmarks?
- How to refine the position of each landmark within desired precision?

For each subtask, we adopt convolutional neural networks to learn the corresponding objective function respectively. The input image of each subsequent network is extracted based on its preceding network's output. These learning stages constitute a casaded prediction pipeline. The overall framework is shown in Fig. 2.

Next, we demonstrate the three stages in details.

3.2.1. Alignment stage

The first stage is designed to locate the lateral face area given the input cephalogram and considered as a alignment procedure. The similar technique has widely been used in face recognition and facial keypoint prediction (Sun et al., 2013; Taigman et al., 2014). It's beneficial for the following landmark detection due to the existinghead position variations in cephalograms, and this process could discard the irrelevant image data. We treat this alignment task as a bounding regression problem and proposed a convolutional neural network called Align-Net to estimate the bounding box of lateral face area. In this paper, lateral face area is defined as the minimum enclosing rectangle of all landmarks with a specified margin (100 pixels in original image).

3.2.2. Proposal stage

The lateral face area located in Alignment stage, is extracted as the input data in this proposal stage. This stage is designed to yield the initial proposal of all landmarks' positions simultaneously. We treated it as a coordinates regression problem and employ a convolutional neural network called Proposal-Net to solve it. The joint learning procedure of all landmarks not only utilizes the local image features of lateral face area but also implicitly encodes the global geometric shape constraints among landmarks.

3.2.3. Refinement stage

Limited by the appearance variations in cephalograms and the small size of training data, it's hard to predict all landmarks within desired high precision by a single model. Furthermore, since the input image is downsampled to small size in previous stages, the loss of image details will lead to unwanted detection errors. In order to improve the predicting precision, for each landmark, we extract the image patch surrounding its proposal location in original image and adopt a convolutional neural network called Refine-Net to learn the optimal position. Since this image patch is in higher resolution than previous networks and retains more details of image intensity pattern, it is reasonable to achieve more accurate results.

3.3. Convolutional network architectures

In this section, we demonstrate the training procedures and structural designs of individual networks used in each stage in detail.

3.3.1. Align-Net

The first network is designed to locate the lateral face area given the raw X-ray dental image. This task is formulated as a bounding box regression problem and the learning objective is set to minimize the bounding box loss L^{box} which is defined as Euclidean loss of coordinates of target rectangle box as shown in below,

$$L^{box} = \left\| \hat{y}^{box} - y^{box} \right\|_2 \tag{2}$$

where $\hat{y}^{box} \in \mathbb{R}^4$ is the prediction results obtained from Align-Net and $y^{box} \in \mathbb{R}^4$ is the groundtruth coordinates. $\|\cdot\|_2$ is the Euclidean norm function. These 4-dimensional coordinates includes

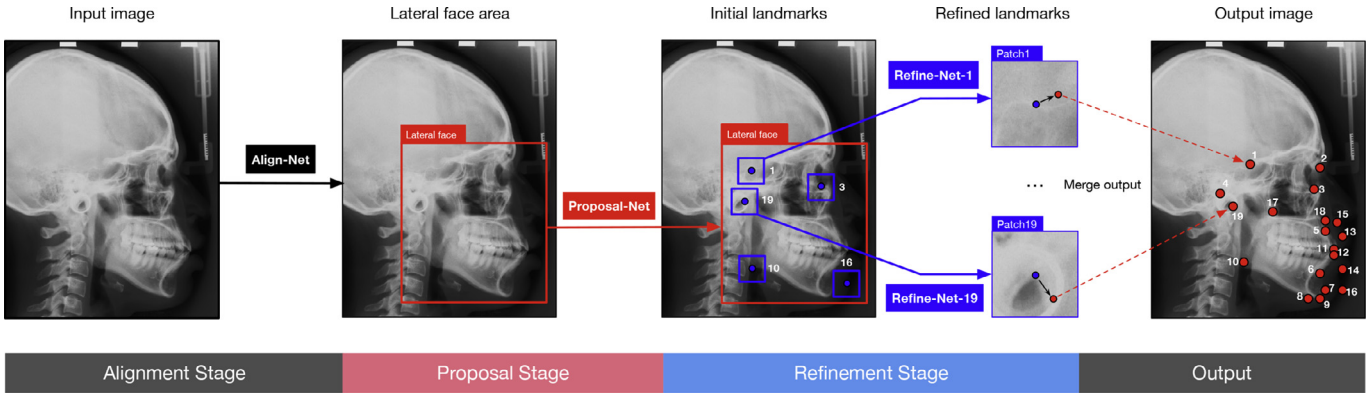


Fig. 2. The pipeline of our framework includes three-stage cascaded convolutional neural networks. The input is a raw X-ray dental image. In the first stage, lateral face area is located by using Align-Net, which is highlighted by a red rectangle box. In the next stage, initial positions of the 19 landmarks are predicted simultaneously through Proposal-Net, which are denoted by blue dots. In the last stage, each landmark is refined by its corresponding Refine-Net based on the image patch (highlighted by the blue rectangle box) which is extracted around its initial position respectively. The final landmark prediction results are annotated in the output image denoted by green dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

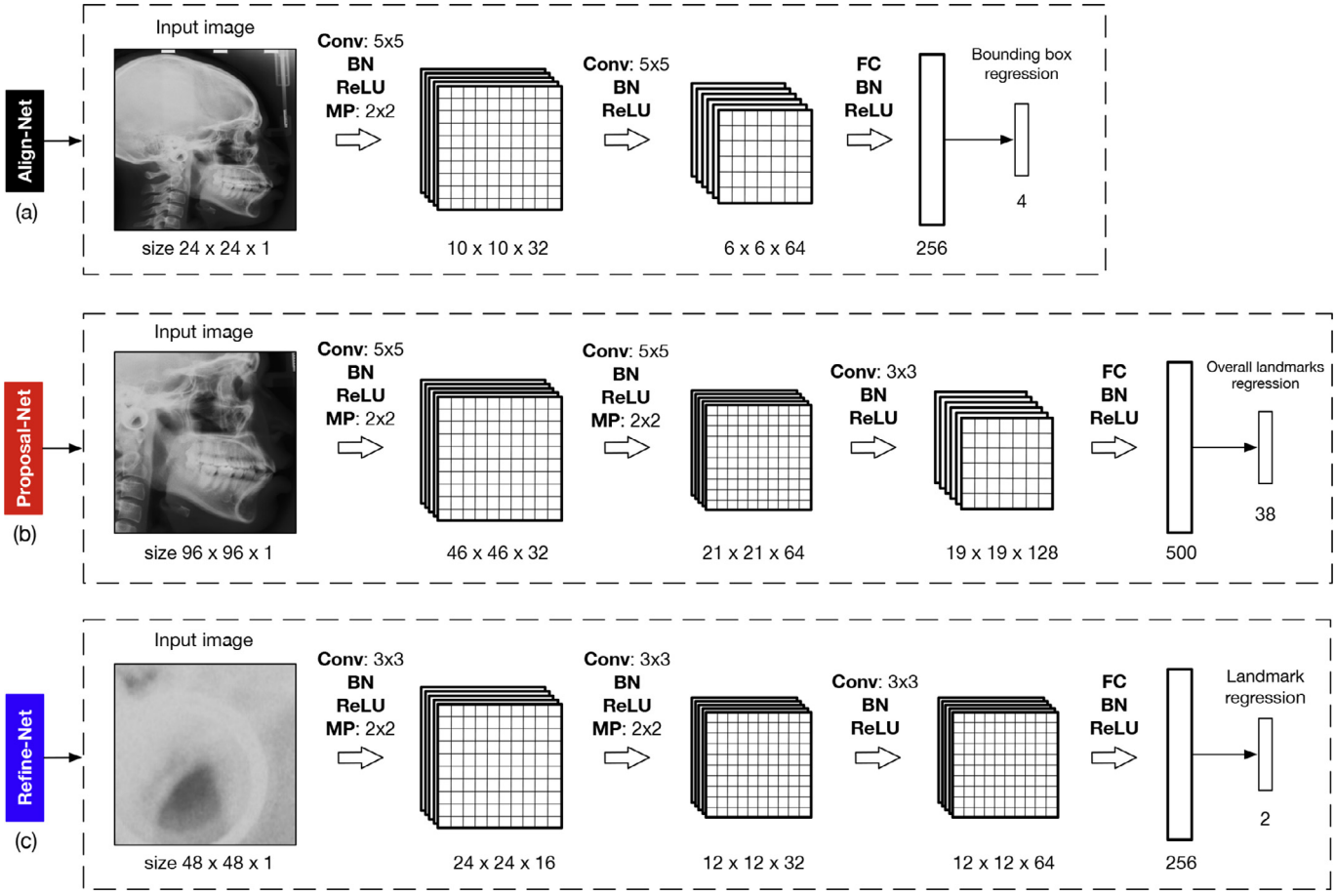


Fig. 3. The architectures of Align-Net, Coarse-Net and Refine-Net, where "Conv" means convolution, "BN" means batch normalization, "ReLU" means relu (rectified linear unit) activation, "MP" means max pooling and "FC" means full connection. The stride size in convolution and pooling is 1 and 2, respectively. The default padding size in convolution is 0, except Refine-Net where the padding size is 1.

left-top and right-bottom corners. The main structure of Align-Net is composed of two convolutional layers with 5×5 filters followed by two fully connected layers. The raw X-ray image is scaled to 24×24 size as the input data. The detailed structure is shown in Fig. 3(a).

3.3.2. Proposal-Net

Proposal-Net is the key component of our approach that is designed to predict the initial proposal of all landmarks' positions. This problem is formulated as a regression problem of all landmarks' coordinates. Similar to the bounding box loss, overall landmark loss $L^{proposal}$ is defined as the Euclidean loss of all landmarks' coordinates as below,

$$L^{proposal} = \left\| \hat{y}^{proposal} - y^{proposal} \right\|_2 \tag{3}$$

where $\hat{y}^{proposal} \in \mathbb{R}^{38}$ is the estimated coordinates of all landmarks and $y^{proposal} \in \mathbb{R}^{38}$ is the groundtruth coordinates. Vector $y^{proposal} =$

$(w_1, h_1, w_2, h_2, ..., w_{19}, h_{19})$ and $y_i^{proposal} = (w_i, h_i)$ is the coordinates of landmark i. Notice that the coordinates of $y^{proposal}$ is computed relative to the left-top corner of lateral face box predicted by Align-Net. In this network, we use three convolutional layers which filters are respectively 5×5, 5×5 and 3×3, followed by two fully connected layers. The input image is extracted in original image based on the predicted face area box and scaled to 96×96 size. The detailed structure is shown in Fig. 3(b).

3.3.3. Refine-Net

Following the previous networks, we have achieved the initial positions of all landmarks. For each landmark i, we extract a square image patch, the center of whichis set as the initial position $y_i^{proposal}$ with length l_{patch}, then employ the corresponding Refine-Net to refine the position. This refinement task can be formulated as a regression problem. The loss of Refine-Net is defined as the Euclidean loss over a single landmark as shown in Eq. (4),

$$L_i^{refine} = \left\| \hat{y}_i^{refine} - y_i^{refine} \right\|_2 \tag{4}$$

where $\hat{y}_i^{refine}$ is the ith landmark's coordinates obtained from Refine-Net-i and y_i^{refine} is the groundtruth. Since there is only one landmark, $y_i^{refine} = (w_i', h_i') \in \mathbb{R}^2$. Notice that the coordinates of y_i^{refine} is computed relative to the left-top corner of the input image patch, i.e. $w_i' = w_i - \frac{l_{patch}}{2}$ and $h_i' = h_i - \frac{l_{patch}}{2}$. In this network, we employ three convolutional layers with 3×3 filters followed by two fully connected layers to learn the Refine-Net. The input image patch is scaled to 48×48 size. The detailed structure is shown in Fig. 3(c).

At last, the final position $\hat{y}_i$ of landmark i can be calculated as Eq. (5),

$$\hat{y}_i = \hat{y}_{lt}^{box} + \hat{y}_i^{proposal} - \left(\frac{l_{patch}}{2}, \frac{l_{patch}}{2} \right) + \hat{y}_i^{refine} \tag{5}$$

where $\hat{y}_{lt}^{box}$ is the left-top corner coordinates of bounding box.

Additionally, several techniques which have been widely used in deep learning are also adopted in our CNN models. Batch normalization (Ioffe and Szegedy, 2015) is proposed to mitigate the internal covariate shift problem in deep neural networks. It normalizes the parameters of input layer in each mini-batch. Batch normalization is evidently proved to accelerate the training procedure and improve the performance of networks. Dropout (Srivastava et al., 2014) is a regularization technique that is aimed to prevent neural networks from overfitting problem. It randomly drops certain percentage of neurons of input layer in the training procedure and is able to prevent complex co-adaptations on training data. Dropout can be considered as a model averaging strategy of neural networks.

The details of network structure discussed above are shown in Fig. 3.

3.4. Data augmentation

As mentioned in Section 1, since it's resource-consuming to obtain the groundtruth labels of medical imaging data annotated by clinical experts, the amount of training data in medical image analysis is usually very small compared with other computer vision tasks. In order to prevent overfitting problem, the most common way is to artificially enlarge the dataset using label-preserving image transformations (Krizhevsky et al., 2012). In this paper, we employ three types of data augmentation in all the three training stages as shown in below,

- **Scale.** Change the size of input image by multiplying a scale factor f_S sampled from $[0.8, 1.1]$ uniformly.
- **Translation.** Translate the input image in both horizontal and vertical direction by applying a translation factor $f_T = (\Delta w, \Delta h)$. Δw is sampled from $[-w_{min}, W - w_{max}]$ in horizontal direction uniformly, where w_{min} is the minimum horizontal coordinate of all landmarks, w_{max} is the maximum horizontal coordinate and W is the image width. Δh is sampled from $[-h_{min}, H - h_{max}]$ in vertical direction uniformly, where h_{min} is the minimum vertical coordinate of all landmarks, h_{max} is the maximum vertical coordinate and H is the image height.
- **Brightness.** Sample a brightness coefficient f_B uniformly from $[0.7, 1.3]$ to simulate the variation of brightness.

Given an original training image X, these augmentation operations were applied step by step to generate a new training sample X' repeatably.

4. Experiments

4.1. Evaluation metrics

Three classical evaluation metrics in cephalometric radiography analysis are adopted in our experiments, same as in the previous studies (Wang et al., 2016). The definitions of these metrics are shown below,

- **Mean radial error (MRE).** Given the landmark i in image j, the radical error (RE) is defined as the Euclidean distance between estimated landmark coordinates $\hat{y}_i = (\hat{w}_i, \hat{h}_i)$ and the manual annotated landmark coordinates $y_i = (w_i, h_i)$, i.e. $RE_i^j = \left\| \hat{y}_i - y_i^{gt} \right\|_2$, where $\|\cdot\|_2$ is Euclidean norm function. The mean radial error (MRE) for landmark i is defined as shown in Eq. (6), where M is the the number of images.

$$MRE_i = \frac{\sum_{j=1}^{M} RE_i^n}{M} \tag{6}$$

  The associated standard deviation (SD) is defined as below,

$$SD_i = \sqrt{\frac{\sum_{j=1}^{M} (RE_i^j - MRE_i)^2}{M}} \tag{7}$$

- **Success detection rate (SDR).** For a detected landmark, if the radical error between it and the groundtruth is no greater than δmm, it's considered as a successful detection. The success detection rate for δ mm is defined as below,

$$SDR_\delta = \frac{\#(\{\hat{y}_i : \left\| \hat{y}_i - y_i \right\|_2 \le \delta\})}{\#(\Omega)} \tag{8}$$

  where $\#(\cdot)$ is the cardinal function and Ω is the set of predictions over all images.
- **Confusion matrix and success classification rate (SCR).** Confusion matrix usually describes the performance of a classification model on test data. Success classification rate (SCR) is defined as the average diagonal value of confusion matrix. In this paper, these two metrics are used to evaluate the classifications of anatomical types.

4.2. Dataset

We evaluated the proposed method on a public available cephalograms dataset[2] which is used in IEEE 2015 ISBI Grand Challenge #1: *Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image* (Wang et al., 2015). The dataset contains 400 cephalometric X-ray images, which were collected from

[2] https://figshare.com/s/37ec464af8e81ae6ebbf.

Table 1
Eight standard clinical measurement methods for classification of anatomical types.

Method	ANB[a]	SNB[b]	SNA[c]	ODI[d]	APDI[e]	FHI[f]	FHA[g]	MW[h]
Type 1	$3.2^\circ - 5.7^\circ$	$74.6^\circ - 78.7^\circ$	$79.4^\circ - 83.2^\circ$	$68.4^\circ - 80.5^\circ$	$77.6^\circ - 85.2^\circ$	$0.65 - 0.75$	$26.8^\circ - 31.4^\circ$	2–4.5 mm
Type 2	$> 5.7^\circ$	$< 74.6^\circ$	$> 83.2^\circ$	$> 80.5^\circ$	$< 77.6^\circ$	> 0.75	$> 31.4^\circ$	= 0 mm
Type 3	$< 3.2^\circ$	$> 78.7^\circ$	$< 79.4^\circ$	$< 68.4^\circ$	$> 85.2^\circ$	< 0.65	$< 26.8^\circ$	< 0 mm
Type 4	–	–	–	–	–	–	–	> 4.5 mm

[a] ANB: angle between point A (L5), nasion (L2) and point B (L6).
[b] SNB: angle between sella (L1), nasion (L2) and point B (L6).
[c] SNA: angle between sella (L1), nasion (L2) and point A (L5).
[d] ODI (Overbite depth indicator) : arithmetic sum of the angle between the lines L5-L6 and L8-L10, and the angle between the lines L3-L4 and L17-L18.
[e] APDI (Anteroposterior dysplasia indicator) : arithmetic sum of the angle between the lines L3-L4 and L2-L7, the angle between the lines L2-L7 and L5-L6, and the angle between the lines L3-L4 and L17-L18.
[f] FHI (Facial height index) : ratio of the posterior face height (distance from L1 to L10) to the anterior face height (distance from L2 to L8).
[g] FMA (Frankfurt mandibular angle) : angle between the lines from sella (L1) to nasion (L2) and from gonion (L10) to gnathion (L9).
[h] MW (Modified Wits): $((x_{L12} - x_{L11})/\|x_{L12} - x_{L11}\|)\|x_{L12} - x_{L11}\|$.

400 patients and acquired by Soredex CRANEX@Excel Ceph machine (Finland) and Soredex SorCom software (3.1.5, version 2.0) (Wang et al., 2015). The resolution of image is 2400×1935 pixels, while the pixel spacing is 0.1 mm/pixel in each dimension. For detection, 19 landmarks were annotated by two experienced medical doctors for each image. The groundtruth of landmarks are the average of two experts' annotations. For the classifications of anatomical types, 8 clinical measurements which is determined by the landmark positions are used. The definitions of 8 anatomical types are shown in Table 1. The groundtruth anatomical types are determined by the groundtruth landmark positions. Detailed illustrations of 19 landmarks and 8 anatomical types can be found in Wang et al. (2015).

Consistently with previous studies in evaluation, 400 images were split to three pieces: Train dataset (150 images), Test1 dataset (150 images) and Test2 dataset (100 images). We train on Train dataset and evaluate on both Test1 and Test2 datasets.

4.3. Training details

We trained the proposed models on a server machine with an Intel Xeon(R) E5-2678 CPU up to 2.5 GHz and a NVIDIA GTX Titan X GPU using cuDNN v8.0. For implementation, we use the Caffe framework (Jia et al., 2014).

4.3.1. Preprocessing

The train dataset includes 150 images. In the training stage, images are augmented using the operations mentioned in Section 3.4, each image yields 500 augmented samples. Subsequently, pixel values in each sample are converted from $\{p \in \mathbb{N} \mid 0 \le p \le 255\}$ to $\{p' \in \mathbb{R} \mid -1 \le p' \le 1\}$ as $p' = \frac{p - p_{avg}}{255}$, where $p_{avg} = 121.78$ is the average pixel value calculated over Train dataset.

4.3.2. Hyper-parameters

The batch size in the training procedure is chosen as 256. Initial network weights are independently sampled using xaiver policy (He et al., 2015). Weight regularization is applied with a weight decay coefficient of 0.001. The learning rate is initially chosen as 0.001, and step learning policy is used with gamma = 0.95. Back-propagation is applied with a momentum coefficient of 0.9. These hyper-parameters are set identically for each network in our framework. In refinement stage, the length l_{patch} of the square image patch which center is the position predicted by Proposal-Net is set to 200 pixels. The number of solver iterations in each learning stage is determined by cross validation that training the network on 90% of the training images and using the remaining 10% as a validation set, therefore the training iteration is 800,000 for Align-Net, 500,000 for Proposal-Net and 500,000 for all Refine-Nets.

4.4. Landmark detection results

We ran experiments on 250 test images collected from Test1 and Test2 datasets. For each test image, the locations of 19 landmarks are predicted automatically. Table 2 shows the landmark detection results on Test1 dataset and Test2 datasets. The results of MRE with SD and SCR for 2.0, 2.5, 3.0 and 4.0 mm ranges are listed for each individual landmark respectively in details. The average MRE of all landmarks on Test1 dataset is 1.34 ± 0.92 mm. This value significantly outperforms the previous result 1.67 ± 1.65 mm which was achieved by (Lindner and Cootes, 2015) above 19.8%, and the average SD is 0.92 mm which is also smaller than 1.65 mm. The average MRE and SD on Test2 dataset is 1.64 ± 0.91 mm, which is still better than previous result 1.92 ± 1.24 mm (Lindner and Cootes, 2015) about 14.6%. These results proved that our approach could locate cephalometric landmarks more accurately (smaller MRE) and consistently (smaller SD).

Besides this, for the clinically accepted precision range of 2.0 mm, i.e. SDR in 2.0 mm, our approach achieved 81.37% accuracy on Test1 dataset and 70.58% accuracy on Test2 dataset. These results are also higher than other published methods (Wang et al., 2016; Arik et al., 2017) as shown in Table 3. Additionally, our results are the best in 2.5-, 3.0- and 4.0 mm with significantly improved compared with previous benchmarks. The SDR in 4.0 mm even reached 97% and 93% on Test1 and Test2 dataset respectively.

Comparing the results between Test1 and Test2 datasets, we found that the performance on Test1 dataset is consistently better than Test2, this means that Test2 dataset is more challenge than Test1. In details, the drop of detection performance is mainly caused by L6, L13 and L16. The precision of the estimated positions of L6, L13 and L16 were much worse on Test2 dataset. Especially, the SDR in 2.0 mm of L16 only achieved poor 5% accuracy. Even in 4.0 mm group, the SDR of L16 is only 37%. The SDR in 2.0mm of L13 and L6 are only 13% and 30% respectively. It seems that the data distribution of train dataset is more closer to the distribution of Test1 dataset than Test2.

4.4.1. Impact of l_{patch}

l_{patch} is the only hyper parameter should be set manually of our approach in inference procedure. It is used to extract the image patch with size $l_{patch} \times l_{patch}$ given a proposal location of landmark in refinement stage. Thus, it's useful to evaluate the impact of the choice of l_{patch}. We ran experiments on Test dataset with $l_{patch} = 100, 200, 300, 400$ respectively. The corresponding cumulative distribution of IPE, i.e. $IPE^j = \frac{\sum_{i=1}^{K} RE_i^j}{K}$ where K is the number of landmarks, are shown in Fig. 4. The results show that the model with $l_{patch} = 200$ achieved the best performance on this dataset.

Table 2
Landmark detection results in terms of mean radial error (MRE) and successful detection rate (SDR) within 2.0, 2.5, 3.0 and 4.0 mm neighborhoods on IEEE 2015 ISBI Grand Challenge Test 1 Dataset.

	ISBI 2015 Challenge Test 1 Dataset					ISBI 2015 Challenge Test 2 Dataset				
	MRE (mm)	SDR (%)				MRE (mm)	SDR (%)			
		2.0 mm	2.5 mm	3.0 mm	4.0 mm		2.0 mm	2.5 mm	3.0 mm	4.0 mm
L1	1.04 ± 1.23	93.33	95.33	96.00	98.00	0.89 ± 0.73	95.00	99.00	99.00	99.00
L2	1.25 ± 1.02	82.67	90.67	92.67	97.33	1.04 ± 0.78	91.00	96.00	98.00	99.00
L3	1.30 ± 0.80	86.67	93.33	96.67	99.33	2.37 ± 0.89	35.00	61.00	74.00	97.00
L4	2.02 ± 1.25	54.67	68.67	80.67	94.00	2.04 ± 2.14	73.00	79.00	83.00	86.00
L5	1.74 ± 0.99	62.00	74.67	90.00	98.67	1.29 ± 0.66	85.00	92.00	99.00	100.0
L6	1.35 ± 0.76	80.67	91.33	96.67	100.0	2.81 ± 1.24	30.00	43.00	57.00	80.00
L7	1.33 ± 0.94	80.00	90.67	96.67	98.67	1.02 ± 1.03	91.00	94.00	96.00	97.00
L8	0.95 ± 0.87	91.33	95.33	97.33	98.67	1.03 ± 0.67	93.00	96.00	97.00	100.0
L9	1.01 ± 0.79	90.67	97.33	98.67	99.33	0.77 ± 0.72	97.00	98.00	98.00	99.00
L10	1.97 ± 1.10	57.33	72.00	81.33	93.33	1.59 ± 1.02	69.00	84.00	90.00	98.00
L11	1.07 ± 0.77	88.67	95.33	98.67	99.33	1.10 ± 0.76	90.00	94.00	97.00	98.00
L12	0.96 ± 0.61	95.33	96.00	98.00	100.0	0.98 ± 0.87	92.00	95.00	97.00	98.00
L13	1.63 ± 0.83	74.67	92.67	96.00	98.67	2.88 ± 0.76	13.00	26.00	59.00	93.00
L14	1.21 ± 0.72	95.33	98.00	98.00	98.00	2.30 ± 0.68	38.00	68.00	87.00	95.00
L15	0.95 ± 0.81	94.00	96.67	97.33	98.67	0.93 ± 0.65	91.00	96.00	99.00	100.0
L16	1.52 ± 0.99	77.33	88.67	94.00	97.33	4.49 ± 1.57	5.00	8.00	14.00	37.00
L17	1.01 ± 0.77	92.00	96.00	96.67	98.67	0.88 ± 0.55	95.00	99.00	100.0	100.0
L18	1.39 ± 1.09	86.00	87.33	93.33	96.67	1.46 ± 0.84	72.00	91.00	95.00	98.00
L19	1.83 ± 1.21	63.33	72.67	83.33	94.67	1.38 ± 0.74	86.00	92.00	96.00	99.00
Average	$\mathbf{1.34 \pm 0.92}$	81.37	89.09	93.79	97.86	$\mathbf{1.64 \pm 0.91}$	70.58	79.53	86.05	93.32

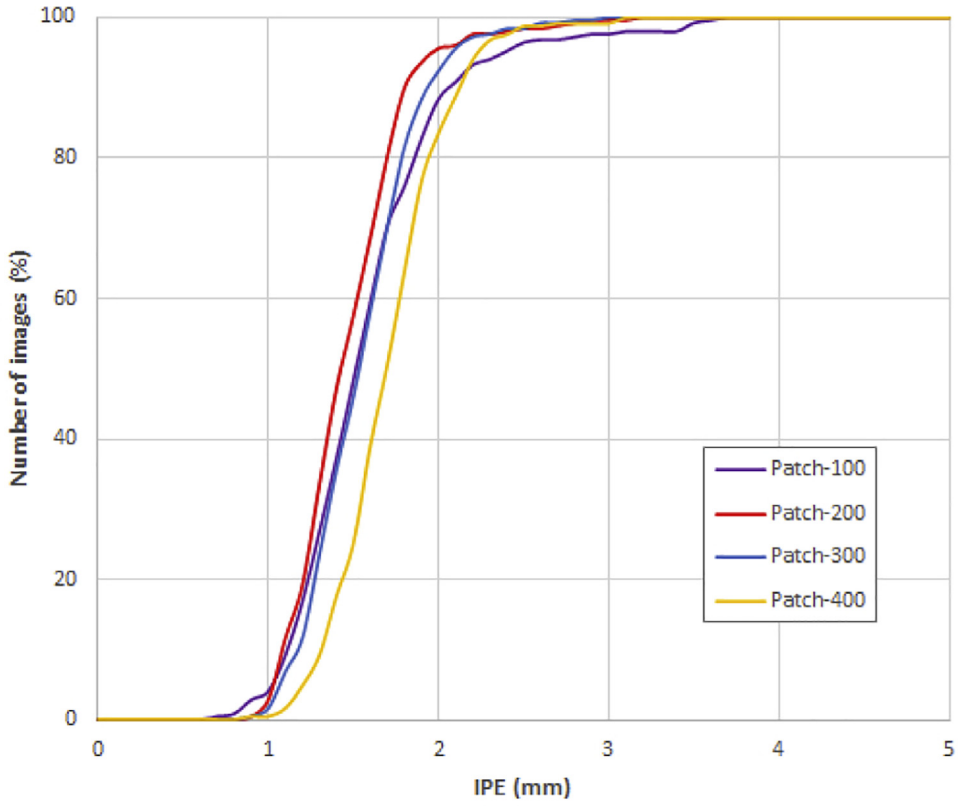


Fig. 4. Cumulative distribution of IPE with l_{patch} = 100, 200, 300 and 400 respectively on ISBI 2015 Test dataset.

4.5. Pathology classification results

Based on the classification schemes described in Section 4.2, pathological assessment is carried out using the estimated positions of landmarks by the evaluation program (Lindner et al., 2016). Table 4 shows the confusion matrix for classification of anatomical types on Test datasets. Table 5 shows the success classification rate (SCR), i.e. the diagonal average of confusion matrix on Test datasets. The results show that our method achieved the best classification accuracy for ANB, APDI, FMA both on Test1 and Test2 datasets. The average SCR over all anatomical types is 82.76% on Test1 dataset and is the best result compared with other methods. The average SCR over all anatomical types is 79.27% on Test2 dataset and is slightly lower than Lindner and Cootes (2015)'s work

Table 3
Comparison of mean results of success detection rates on ISBI 2015 Challenge Test dataset.

Method	SDR(%)			
	2.0 mm	2.5 mm	3.0 mm	4.0 mm
Ours	**76.82**	**84.97**	**90.00**	**95.58**
SCN Payer et al. (2019)	73.33	78.76	83.24	89.75
Localization U-Net (2015)	72.15	77.83	82.04	88.80
Arik et al. (2017)	72.29	78.21	82.24	86.80
Urschler et al. (2018))	70.21	76.95	82.08	89.01
Lindner and Cootes (2015)	70.65	76.93	82.17	89.85
Ibragimov et al. (2015a)	68.13	74.63	79.77	86.87

(80.99%). The drop of this performance is mainly caused by the less accuracy of SNA and FHI. Besides the success classification rate, we also exploited the average classification accuracy and achieved 83.41% and 81.25% on Test1 and Test2 datasets respectively.

4.6. Ablation study

In this section, we quantitatively analyze the impact of individual stages in this framework.

As mentioned in Section 3.2, Proposal-Net is the bridge connecting Align-Net and Refine-Net, thus it is kept while we evaluated the MRE performance without Align-Net and without Refine-Net respectively. The results are shown in Table 6. Align-Net could improve the MRE performance by about 7.5%, meanwhile Refine-Net could improve the performance by about 44.5%. This means that Refine-Net is the key component to achieve predictions within high precision compared with Align-Net.

Next, we show the cumulative distribution of IPE with Refine-Net and with Propasal-Net, i.e. without Refine-Net in Fig. 5. The figure shows that Refine-Net improves the cumulative distribution of IPE significantly compared to Proposal-Net.

In order to further investigate how Refine-Net improves the prediction precision, for each landmark, we plot a Proposal-Refinement scatter diagram of MRE. The result is shown in Fig. 6. For each landmark location L in the figure, the slope of line $(0, L)$ represents the improvement ability of Refine-Net. The lower the slope is, the stronger the improvement ability is. The abscissa of landmark L represents the proposal quality. We can find that the final prediction precision of a landmark depends on two factors: proposal quality and refinement ability. Based on the Proposal-Refine diagram, we can analyze the reasons of specific landmark's prediction performance. For example, we could find that L16, L13 and L4 are the top 3 landmarks with high errors according to the vertical axis value in Fig. 6. The main reason for L16 is that its proposal quality is really bad, although its refinement ability is rather good. For L13, the main reason is that the refinement improvement ability is poor, although its proposal is better than L4. For L4, the reason is that its proposal quality is bad.

4.7. Failure analysis

In this section, we analyze the failure cases, those with high errors, on ISBI 2015 Test dataset. We selected the top four cases (#208, #318, #389, #194) with the highest errors in IPE. The predicted landmark locations including proposal and refinement locations are presented on cephalogram in Fig. 7. The Fig. 7(a), i.e #208 has the highest error where the landmark L1 and L4 are estimated (in red color) far away from the groundtruth (in green color). We could find that the proposal quality of L1 and L4 (in blue color) are really bad, this leads to the low precision prediction as discussed in Section 4.6. The other cases are similar with #208.

4.8. Visual interpretation

Since the experimental results have shown the performance of our approach, it's important to reveal the underlying mechanism of the cascaded networks. In this section, we adopted Grad-CAM technique (Selvaraju et al., 2017) which is proposed to produce visual interpretations of CNN-based models. For each network, Grad-CAM uses gradients of target loss flowing into the last convolutional layer, to generate a localization map highlighting the

Table 4
Confusion matrix for classification of anatomical types on ISBI 2015 Challenge Test 1 and Test 2 datasets.

	ISBI 2015 Challenge Test 1 Dataset				ISBI 2015 Challenge Test 2 Dataset			
		Type 1 (%)	Type 2 (%)	Type 3 (%)		Type 1 (%)	Type 2 (%)	Type3 (%)
ANB	Type 1	68.09	8.51	23.40	Type 1	70.00	0.00	30.00
	Type 2	23.33	76.67	0.00	Type 2	21.43	78.57	0.00
	Type 3	8.22	0.00	91.78	Type 3	2.38	0.00	97.62
SNB	Type 1	78.57	7.14	14.29	Type 1	75.86	3.45	20.69
	Type 2	26.67	73.33	0.00	Type 2	0.00	100.00	0.00
	Type 3	7.53	0.00	92.47	Type 3	6.78	0.00	93.22
SNA	Type 1	66.67	14.81	18.52	Type 1	56.10	26.83	17.07
	Type 2	20.90	77.61	1.49	Type 2	25.00	75.00	0.00
	Type 3	20.69	10.34	68.97	Type 3	36.84	0.00	63.16
ODI	Type 1	83.33	1.52	15.15	Type 1	76.92	0.00	23.08
	Type 2	20.00	80.00	0.00	Type 2	62.50	37.50	0.00
	Type 3	8.70	0.00	91.30	Type 3	0.00	0.00	100.00
APDI	Type 1	82.98	8.51	8.51	Type 1	78.57	4.76	16.67
	Type 2	22.86	77.14	0.00	Type 2	9.09	90.91	0.00
	Type 3	1.47	0.00	98.53	Type 3	2.78	0.00	97.22
FHI	Type 1	93.85	0.00	6.15	Type 1	84.44	0.00	15.56
	Type 2	0.00	100.00	0.00	Type 2	50.00	50.00	0.00
	Type 3	22.89	0.00	77.11	Type 3	18.87	0.00	81.13
FMA	Type 1	73.33	3.33	23.33	Type 1	66.67	28.57	4.76
	Type 2	18.00	82.00	0.00	Type 2	16.18	83.82	0.00
	Type 3	0.00	0.00	100.00	Type 3	0.00	0.00	100.00
MW		Type 1	Type 3	Type 4		Type 1	Type 3	Type4
	Type 1	84.78	8.70	6.52	Type 1	83.33	9.52	7.14
	Type 3	12.70	87.30	1.59	Type 3	14.29	85.71	0.00
	Type 4	19.51	0.00	80.49	Type 4	23.33	0.00	76.67

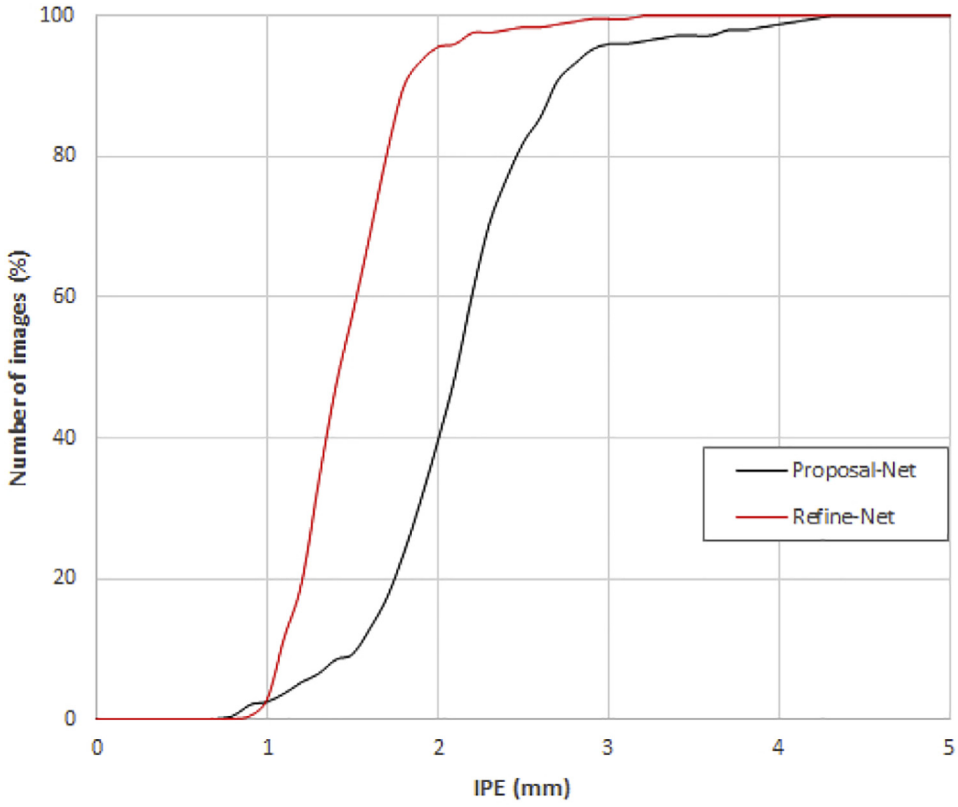


Fig. 5. Cumulative distribution of image-specific radical errors of Proposal-Net and Refine-Net on ISBI 2015 Test dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

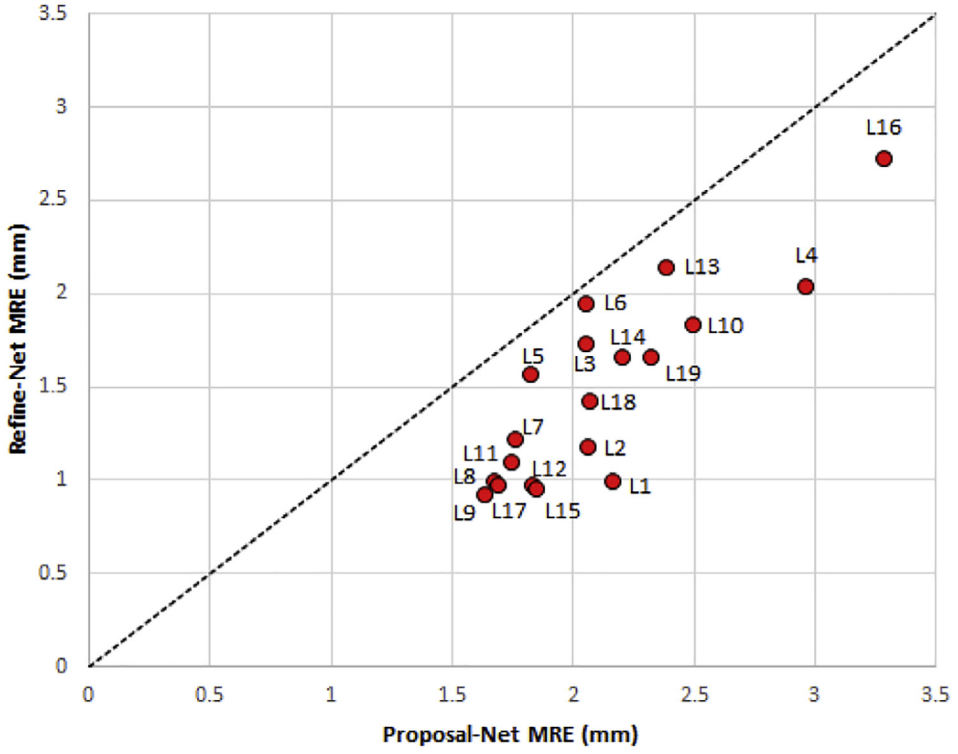


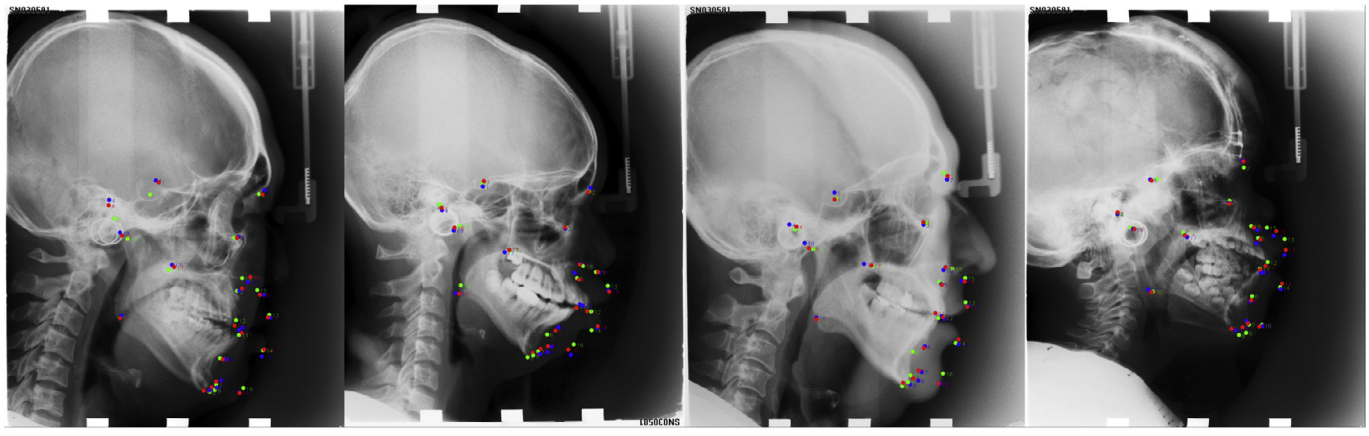
Fig. 6. Proposal-Refinement scatter diagram. The horizontal axis is the MRE in mm which is computed using Proposal-Net predictions. The vertical axis is the MRE in mm which is computed using Refine-Net predictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Comparison of success classification rate (%) for classification of anatomical types on IEEE 2015 ISBI Grand Challenge Test 1 and Test 2 Datasets.

	ISBI 2015 Challenge Test 1 Dataset				ISBI 2015 Challenge Test 2 Dataset			
	Ours	Arik et al. (2017)	Lindner and Cootes (2015)	Ibragimov et al. (2015a)	Ours	Arik et al. (2017)	Lindner and Cootes (2015)	Ibragimov et al. (2015a)
ANB	**78.84**	61.47	64.99	59.42	**82.06**	77.31	75.83	76.64
SNB	81.46	70.11	**84.52**	71.09	**89.69**	69.81	81.92	75.24
SNA	**71.08**	63.57	68.45	59.00	64.75	66.72	**77.97**	70.24
ODI	**84.88**	75.04	84.64	78.04	71.47	**72.28**	71.26	63.71
APDI	**86.22**	82.38	82.14	80.16	**88.90**	87.18	87.25	79.93
FHI	**90.32**	65.92	67.92	58.97	71.86	69.16	**90.90**	86.74
FMA	**85.11**	73.90	75.54	77.03	**83.50**	78.01	80.66	78.90
MW	**84.19**	81.31	82.19	83.94	81.90	77.45	**82.11**	77.53
Average	**82.76**	71.71	76.41	70.84	79.27	74.74	**80.99**	76.12

Table 6
Comparison of MRE with SD and SDR for ablation study of Align-Net and Refine-Net respectively.

Method	MRE ± SD	SDR(%)			
		2.0 mm	2.5 mm	3.0 mm	4.0 mm
Ours	**1.46 ± 0.92**	**76.82**	**84.97**	**90.00**	**95.58**
Ours (w/o Align-Net)	1.57 ± 1.14	75.07	83.03	88.88	94.68
Ours (w/o Refine-Net)	2.11 ± 1.30	55.70	68.82	78.46	90.34



(a) #208 (b) #318 (c) #389 (d) #194

Fig. 7. From left to right, failure cases with the top four IPE are selected from ISBI 2015 Test dataset. Green dots represent the groundtruth of landmarks, blue dots represent the landmark locations predicted by Proposal-Net and red dots represent the landmark locations predicted by Refine-Net. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

important regions in the image for the corresponding objective function. We randomly selected 4 test images (#190, #232, #302 and #314) from Test dataset to visualize Align-Net, Proposal-Net and Refine-Nets respectively as shown in Fig. 8. Visualization of the second convolutional layer in Align-Net are listed in the first row. It's interesting that the highlight area indicated by heatmap color space is mainly localized near the center of lateral face area. Visualization of the third convolutional layer of Proposal-Net is listed in the second row. The highlighted areas indicate the key supporting regions surrounding the lateral face area, including the forehead, nose, jaw, and neck spine regions. In refinement stage, we selected 4 representative Refine-Nets of landmarks L1, L2, L8 and L12 to show the visualization which used the third convolutional layer of Refine-Net. We can find that the highlighted areas almost cover the most important and obvious image patterns around the landmark. These visualization results of each network revealed which parts of the image are actually important when predicting landmarks using convolutional networks and could help us understand how CNNs are applied to cephalometric landmark detection.

4.9. Model complexity analysis

We use #params, i.e. the number of all parameters in a network, and FLOPs, i.e. the number of floating-point multiplication-adds to represent the model complexity. Our framework consists of an Align-Net, a Proposal-Net and 19 Refine-Nets, hence the total model complexity is the sum of these networks. The total complexity of this framework is about 606.34M flops and 69.11M parameters as shown in Table 7. It is rather lightweight compared with MobileNet (Howard et al., 2017). At inference stage, we ran the framework in a server machine in GPU and a normal PC server (2.5 GHz) in CPU respectively. The Runtime (seconds) results are shown in Table 7. Given a new image, the inference process can be completed in about 3 seconds which is fast enough compared with traditional manual way in clinical practice.

4.10. Validation on additional datasets

Although the experiments on ISBI 2015 Challenge dataset demonstrated the effectiveness of the proposed method, it's

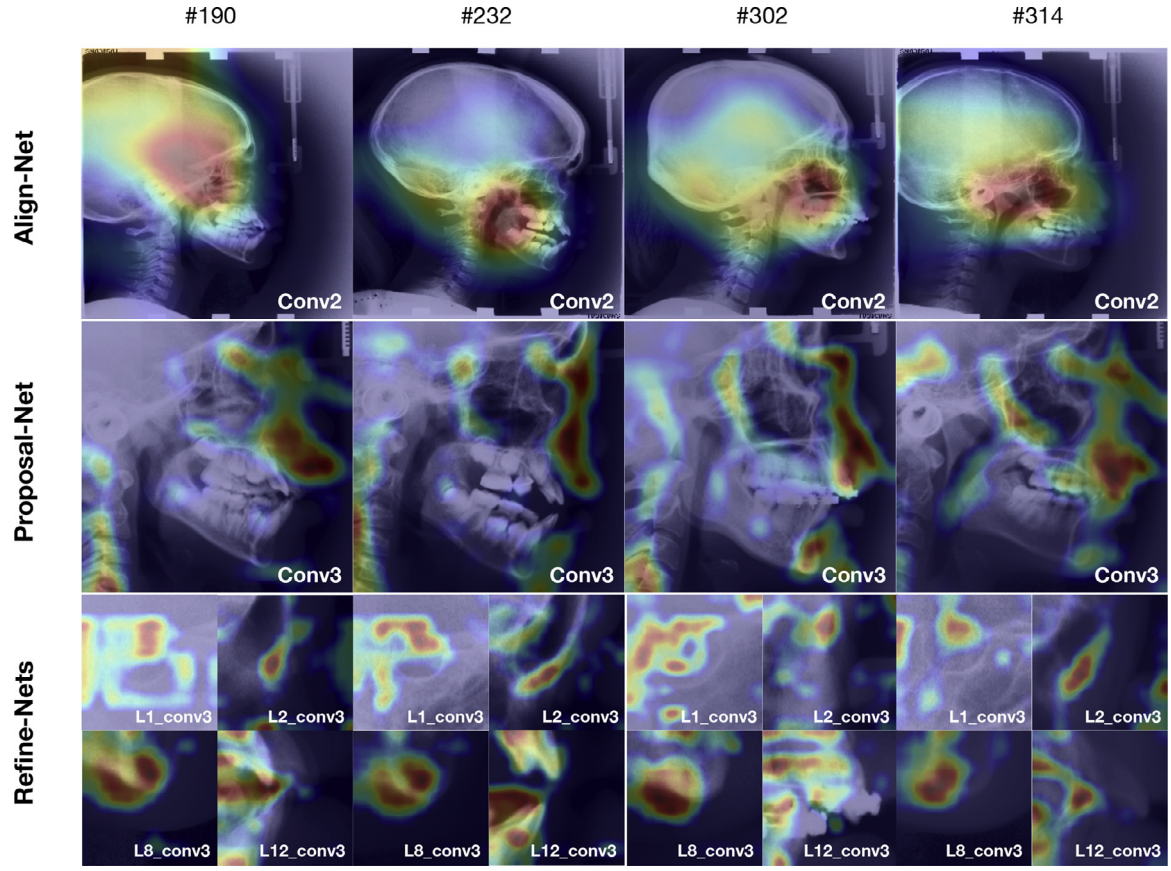


Fig. 8. Grad-CAM visualization demonstration of our cascaded framework, including Align-Net, Proposal-Net and 4 Refine-Nets (L1, L2, L8 and L12). Each sub figure is generated by blending raw image with the visualized gradients of the last convolutional layer in the corresponding network pixel by pixel. The gradients are visualized using heatmap color space. The red color represents the large gradient trend, while the blue color represents the small gradient trend. For convenience to investigate Refine-Nets, we highlight the groundtruth positions of landmarks by green dots in each sub figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Complexity analysis of our cascaded convolutional neural network framework. In the table, M represents million.

Network	# Params	FLOPs	Runtime (s)	
			GPU	CPU
Align-Net	0.64 M	5.54 M	0.56	1.04
Proposal-Net	23.25 M	294.52 M	0.33	0.52
Refine-Net	2.38 M	16.12 M	0.06	0.08
Total	69.11 M	606.34 M	2.03	3.08

Table 8
Comparison of mean results of Success Detection Rate (SDR) on PKU cephalometric landmark dataset.

Method	MRE ± SD	SDR (%)			
		2.0 mm	2.5 mm	3.0 mm	4.0 mm
Ours	2.02 ± 1.89	64.81	73.94	81.73	89.78

important to evaluate the method's effectiveness on other cephalometric or anatomical landmark datasets to validate its generalization ability.

4.10.1. PKU cephalogram dataset

To quantatively validate the generalization ability of our method, we constructed a new cephalogram dataset called PKU cephalogram dataset and published this dataset[3] to the research community. The patients' data were collected from Fourth Clinical Division, School and Hospital of Stomatology, Peking University. The dataset contains 102 patients' cephalograms, whose age are from 9 to 53 years. The average resolution size of these images is 2089×1937 pixels, while the pixel spacing is about 0.125 mm/pixel. These X-ray images were acquired by Planmeca ProMax 3D machine (Finland) and Planmeca Romexis software (3.7.0 R). Two senior doctors annotated the 19 cephalometric landmarks separately. In this experiment, we directly evaluated the previous trained model on the new dataset. The landmark prediction results are shown in Table 8. The results show that even without fine-tuning on the new dataset, the proposed method is still able to predict the landmarks within reasonable errors.

Additionally, we show the best, median, and worst case sorted by IPE in Fig. 9. For the worst case, Fig. 9(c), each predicted landmark which is in red color has a consistent offset compared to the groundtruth. The failure reason probably is that the head scale in this cephalogram is obviously different with other case which is not learned precisely by Align-Net. For the best case, Fig. 9(a), and median case, Fig. 9(b), the prediction results are rather close to the groundtruth.

4.10.2. Hand radiographs dataset

Although our method is proposed to detect cephalometric landmarks, it's also applicable to other anatomic landmark detection tasks. In this section, we evaluate the cascaded framework on a

[3] https://doi.org/10.6084/m9.figshare.13265471.v1.

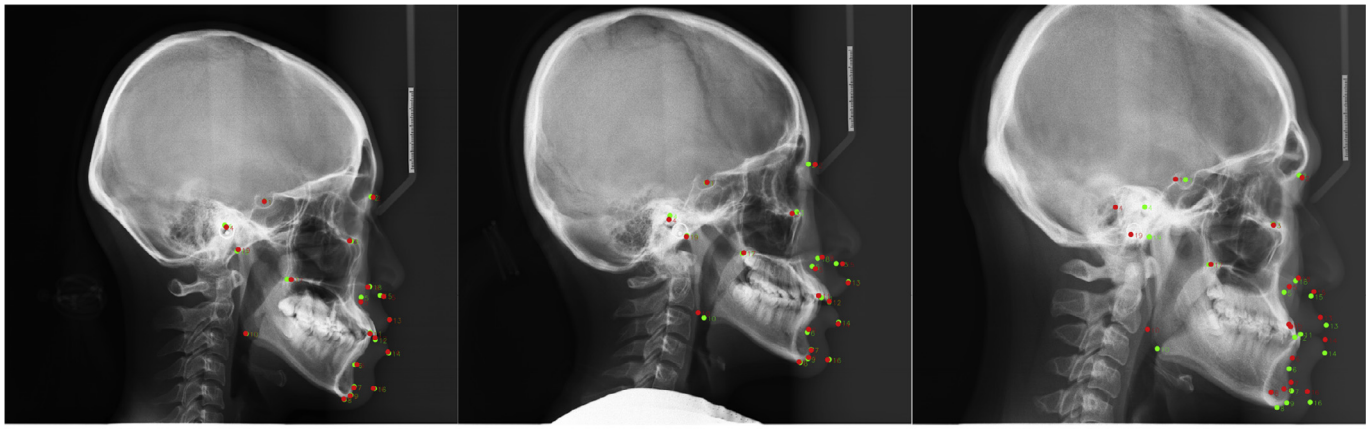


Fig. 9. Cephalometric landmark detection samples on PKU cephalogram dataset. The green dots represent the groundtruth landmarks and the red dots represent the predicted landmarks. #102 is the best case which has the least errors. #52 is the median case. #47 is the worst case which has the most errors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Comparison of Median RE, MRE ± SD and Success Detection Rates on Hand radiographs dataset.

Method	Median RE (mm)	MRE ± SD (mm)	O_r (%)		
			$r = 2$ mm	$r = 4$ mm	$r = 10$ mm
Ours	0.45	0.77 ± 1.14	1921 (5.80%)	271 (0.82%)	42 (0.12%)
SCN Payer et al. (2019)	**0.43**	$\mathbf{0.66 \pm 0.74}$	**1659 (5.01%)**	**241 (0.73%)**	**3 (0.01%)**
Localization U-Net (2015)	0.44	0.70 ± 2.18	1703 (5.14%)	270 (0.82%)	22 (0.07%)
Lindner and Cootes (2015))	0.64	0.85 ± 1.01	2094 (6.32%)	347 (1.05%)	20 (0.06%)
Urschler et al. (2018))	0.51	0.80 ± 0.93	2586 (7.81%)	510 (1.54%)	12 (0.04%)
Štern et al. (2016))	0.51	0.80 ± 0.91	2582 (7.80%)	512 (1.55%)	15 (0.05%)
Ebner et al. (2014))	0.51	0.97 ± 2.45	2781 (8.40%)	716 (2.16%)	228 (0.69%)
Payer et al. (2016))	0.91	1.13 ± 0.98	4109 (12.4%)	444 (1.34%)	12 (0.04%)

public available dataset of hand radiographs.[4] This dataset contains 895 radiographs of left hands. Different from ISBI 2015 cephalometric dataset, the radiographs were acquired by different X-ray scanners which results in varying image size and photo quality. We use the 37 characteristic landmarks on finger tips and bone joints annotated by Payer et al. (2019). As the images lack information about physical pixel resolution, We adopted the image-specific normalization factor setting proposed in Payer et al. (2019) and used the three folds with equal number of images, resulting in 597 training and 298 testing images per fold in our experiments.

We trained our model on Hand radiographs dataset as described in Section 4.3. The preprocessing of the input images is the same with Payer et al. (2016). The results are shown in Table 9. Although our method is not the best one (the 3rd place in MRE ± SD metric), the gap is not very big. Even in O_r metric, the performance is comparable to SCN Payer et al. (2019) which is the best method on this dataset. This proved that our approach is able to handle other anatomical landmark detection task within high precision.

5. Discussion and conclusion

In this paper, we proposed a novel approach which is able to detect cephalometric landmarks. We evaluated the approach on a public dataset published by IEEE 2015 ISBI Grand Challenge. The proposed approach achieved the least mean radical error (MRE) and the highest success detection rate (SDR) for 2.0 mm precision range, which is considered as the clinically accepted, and also for 2.5-, 3.0- and 4.0 mm ranges.

Additionally, our approach also achieved significant improvement in pathology assessment of 8 anatomical types. On Test1 dataset, types of ANB, SNA, ODI, APDI, FHI, FMA and MW were predicted better than the other methods, while only worse for SNB type. On Test2 dataset, types of ANB, SNB, APDI and FMA were predicted better than other methods. The results revealed the different data distributions between Test1 and Test2 datasets. Analogously, the performance of the published methods on Test1 dataset are all better than Test2. It seems that the data distribution of Test1 dataset is more consistent with Train dataset, which means that Test2 dataset is more difficult in this competition.

Different from the previous methods which usually adopted random forests to vote for positions of each individual landmark and combined a statistical shape model to refine all landmarks' positions, our approach use three-stage CNN models to constitute a cascaded pipeline with no more domain-specific priors. The experimental results evidently proved that this cascaded framework could predict cephalometric landmarks better than traditional ways with small training data. The cascaded structure could learn the shape constraints among landmarks implicitly efficiently which comparable to traditional methods which usually employ a statistical shape model explicitly. In addition, comparing with other CNN-based approaches, we purely use CNN model to compose prediction framework. Secondly, we use cascaded CNN models to predict landmarks from coarse to fine. This three-stage cascaded approach could utilize more training image data which helps to prevent overfitting problem and extract more useful multi-scale features of cephalograms than other methods.

Furthermore, we constructed a new cephalograms dataset which contains 102 patients. The prediction results on this dataset show that our approach could be considered as a practical CNN-based approach for cephalometric landmark detection task. Beside this, we performed the proposed method on Hand radiographs dataset to validate the generalization ability. The

[4] Digital Hand Atlas Database System, www.ipilab.org/BAAweb.

experimental results show that this approach is comparable to recent anatomical landmark detection methods.

This work is a good attempt to apply CNN technique to solve cephalometric landmark detection. Although it has achieved significant performance, there is stillroom for improvement in the future work. For instance, it is not an end-to-end learning framework. The framework consists of 21 individual CNN models which are trained respectively. This is inefficient in both training and testing stage. It's attractive to propose an end-to-end convolutional neural network to address this detection problem more efficiently.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Minmin Zeng: Conceptualization, Methodology, Writing - original draft. **Zhenlei Yan:** Software, Writing - review & editing. **Shuai Liu:** Data curation, Writing - review & editing. **Yanheng Zhou:** Supervision. **Lixin Qiu:** Supervision.

Acknowledgments

This work was financed by Grant-in-aid for scientific research from the National Natural Science Foundation for the Youth of China (No. 81701022). The authors would like to thank Christian Payer for the kind help in gaining access to the Hand radiographs dataset.

References

Arik, S.Ö., Ibragimov, B., Xing, L., 2017. Fully automated quantitative cephalometry using convolutional neural networks. J. Med. Imaging 4 (1), 014501.
Baumrind, S., Frantz, R.C., 1971. The reliability of head film measurements: 1. Landmark identification. Am. J. Orthod. 60 (2), 111–127.
Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59.
Domingos, P.M., 2012. A few useful things to know about machine learning.. Commun. ACM 55 (10), 78–87.
Durão, A.P.R., Morosolli, A., Pittayapat, P., Bolstad, N., Ferreira, A.P., Jacobs, R., 2015. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. Imaging Sci. Dent. 45 (4), 213–220.
Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M., 2014. Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 421–428.
El-Feghi, I., Sid-Ahmed, M.A., Ahmadi, M., 2004. Automatic localization of craniofacial landmarks for assisted cephalometry. Pattern Recognit. 37 (3), 609–621.
Forsyth, D., Davis, D., 1996. Assessment of an automated cephalometric analysis system. Eur. J. Orthod. 18 (5), 471–478.
Grau, V., Alcaniz, M., Juan, M., Monserrat, C., Knoll, C., 2001. Automatic localization of cephalometric landmarks. J. Biomed. Inform. 34 (3), 146–156.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.
Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
Ibragimov, B., Likar, B., Pernus, F., Vrtovec, T., 2014. Automatic cephalometric X-ray landmark detection by applying game theory and random forests. In: Proc. ISBI Int. Symp. on Biomedical Imaging.
Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2014. Shape representation for efficient landmark-based segmentation in 3-D. IEEE Trans. Med. Imaging 33 (4), 861–874.
Ibragimov, B., Likar, B., Pernus, F., Vrtovec, T., 2015. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. In: Proceedings of International Symposium on Biomedical imaging (ISBI).
Ibragimov, B., Likar, B., Pernus, F., et al., 2012. A game-theoretic framework for landmark-based image segmentation. IEEE Trans. Med. Imaging 31 (9), 1761–1776.
Ibragimov, B., Prince, J.L., Murano, E.Z., Woo, J., Stone, M., Likar, B., Pernuš, F., Vrtovec, T., 2015. Segmentation of tongue muscles from super-resolution magnetic resonance images. Med. Image Anal. 20 (1), 198–207.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp. 675–678.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.
Lee, H., Park, M., Kim, J., 2017. Cephalometric landmark detection in dental X-ray-images using convolutional neural networks. In: Medical Imaging 2017: Computer-Aided Diagnosis, 10134. International Society for Optics and Photonics, p. 101341W.
Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F., 2015. Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1862–1874.
Lindner, C., Cootes, T.F., 2015. Fully automatic cephalometric evaluation using random forest regression-voting. IEEE International Symposium on Biomedical Imaging. Citeseer.
Lindner, C., Thiagarajah, S., Wilkinson, J.M., Wallis, G.A., Cootes, T.F., arcOGEN Consortium, et al., 2013. Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE Trans. Med. Imaging 32 (8), 1462–1472.
Lindner, C., Wang, C.-W., Huang, C.-T., Li, C.-H., Chang, S.-W., Cootes, T.F., 2016. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci. Rep. 6, 33581.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer vision, pp. 3730–3738.
Nikneshan, S., Mohseni, S., Nouri, M., Hadian, H., Kharazifard, M.J., 2015. The effect of emboss enhancement on reliability of landmark identification in digital lateral cephalometric images. Iranian J. Radiol. 12 (2), e19302.
Parkhi, O.M., Vedaldi, A., Zisserman, A., et al., 2015. Deep face recognition.. In: BMVC, 1, p. 6.
Payer, C., Štern, D., Bischof, H., Urschler, M., 2016. Regressing heatmaps for multiple landmark localization using CNNs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 230–238.
Payer, C., Štern, D., Bischof, H., Urschler, M., 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal..
Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921.
Proffit, W.R., Fields Jr, H.W., Sarver, D.M., 2006. Contemporary Orthodontics. Elsevier Health Sciences.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Štern, D., Ebner, T., Urschler, M., 2016. From local to global random regression forests: exploring anatomical landmark localization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 221–229.
Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483.
Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708.
Urschler, M., Ebner, T., Štern, D., 2018. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. 43, 23–36.
Wang, C.-W., Huang, C.-T., Hsieh, M.-C., Li, C.-H., Chang, S.-W., Li, W.-C., Vandaele, R., Marée, R., Jodogne, S., Geurts, P., et al., 2015. Evaluation and comparison of anatomical landmark detection methods for cephalometric X-ray images: a grand challenge. IEEE Trans. Med. Imaging 34 (9), 1890–1900.
Wang, C.-W., Huang, C.-T., Lee, J.-H., Li, C.-H., Chang, S.-W., Siao, M.-J., Lai, T.-M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al., 2016. A benchmark for comparison of dental radiography analysis algorithms. Med. Image Anal. 31, 63–76.
Zhang, J., Liu, M., Shen, D., 2017. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans. Image Process. 26 (10), 4753–4764.
Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23 (10), 1499–1503.
Zhou, J., Abdel-Mottaleb, M., 2005. A content-based system for human identification based on bitewing dental X-ray images. Pattern Recognit. 38 (11), 2132–2142.