IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 19 July 2023, accepted 12 September 2023, date of publication 18 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316689

METHODS

An Automatic Assessment Model of Adenoid Hypertrophy in MRI Images Based on Deep Convolutional Neural Networks

ZILING HE[1], YUBIN XIAO[1], (Graduate Student Member, IEEE), XUAN WU[1], (Graduate Student Member, IEEE), YANCHUN LIANG[2], YOU ZHOU[1], AND GUANGHUI AN[3]

[1] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
[2] School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, Guangdong 519041, China
[3] Shanghai University of Traditional Chinese Medicine, Shanghai Municipal Hospital Traditional Chinese Medicine, Shanghai 201203, China

Corresponding authors: You Zhou (zyou@jlu.edu.cn) and Guanghui An (changhexiaoyu@126.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF1201200, in part by the Jilin Provincial Department of Science and Technology Project under Grant 20230201083GX, Grant 20220201145GX, and Grant 20230201065GX, in part by the National Natural Science Foundation of China under Grant 62072212, Grant 61972174, Grant 61972175, and Grant 12205114, in part by the Guangdong Universities Innovation Team Project under Grant 2021KCXTD015, and in part by the Key Disciplines Projects under Grant 2021ZDJS138.

ABSTRACT Adenoid hypertrophy is a pathological condition characterized by the enlargement of the adenoids in children, which may lead to various problems. The conventional manual measurement is time-consuming and subject to subjective errors. Previous automated methods based on landmark detection in X-ray images have shown reliability. However, these methods neglect global features, making it difficult to locate landmarks accurately and thus affecting adenoid-to-nasopharyngeal (AN) ratio estimation when migrating to MRI images. In this paper, we first apply a deep-learning method to automatically assess adenoid hypertrophy in MRI images. We propose an adenoid network (ADNet) to capture local and global features near landmarks to achieve accurate landmark localization. Specifically, ADNet uses an encoder-decoder architecture where we employ a depthwise separable convolution-based encoder to extract local features and then employ an adaptive convolution-based decoder to capture global features. We collected a dataset of 500 cephalometric MRI images to train and evaluate the performance of the proposed model. Our experimental results demonstrate that our network achieves state-of-the-art performance with an average radial error of 4.85 pixels for landmark detection, an average successful detection rate of 96% within 15 pixels, and an error of 0.026 for the calculated AN ratio.

INDEX TERMS Adenoid hypertrophy, deep learning, medical image processing, landmark detection.

I. INTRODUCTION

The adenoid is a collection of lymphoid tissue situated in the posterior region of the nasopharyngeal airway, forming a part of Waldeyer's ring. Adenoids can become physiologically enlarged at the age of 2-10 years old, then degenerate around the age of 8-10 years, and usually atrophy completely by the age of 12-14 years [1]. However, adenoidal infection and inflammation may result in pathological adenoid hypertrophy (AH), obstructing the upper airway. This can lead to mouth breathing, recurrent sinusitis, otitis media, changes in facial growth in children, or more severe complications such as obstructive sleep apnea syndrome, impaired cognitive function, and abnormal intellectual development. These serious complications often result in the need for adenoidectomy to improve the overall quality of life [1], [2], [3], making timely diagnosis and treatment of adenoid hypertrophy crucial.

Currently, the main diagnostic methods for detecting adenoid hypertrophy include flexible fiber-optic nasal endoscopy and nasopharyngeal radiological examination (such as lateral cephalography) [1], [4]. Nasal endoscopy enables direct visualization of the adenoids and adjacent structures, allow-

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

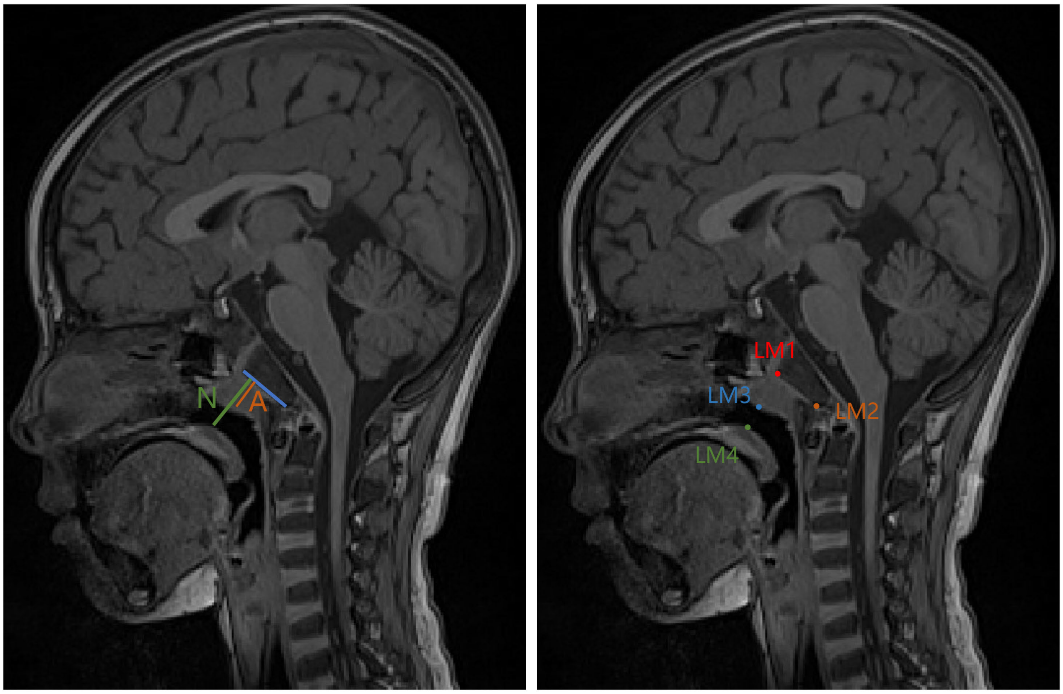


FIGURE 1. An example of the Fjioka assessment method based on cephalograms, the AN ratio is measured through four landmarks. LM1, LM2, LM3, and LM4 are the four landmarks that we need to detect.

ing for measurement of adenoid enlargement severity and proportionate airway blockage. This provides strong evidence for preoperative diagnosis and evaluation of postoperative outcomes. However, the invasive nature of nasal endoscopy makes it challenging for many children to cooperate with physicians during the examination, limiting its clinical use [5]. As a result, lateral cephalography has become the most commonly used tool for detecting adenoid hypertrophy, and numerous studies demonstrate its high reliability in identifying AH [6], [7].

Among all assessment methods based on cephalograms, the most notable is based on calculating the adenoid-to-nasopharyngeal (AN) ratio, described by Fujioka [8]. This method aims to determine the ratio between the measurement of the adenoid tissue (defined by the distance between the basiocciput region and the most convex part of the adenoid pad) and the nasopharyngeal aperture (defined by the distance between the sphenobasiocciput and the posterior edge of the hard palate) [8]. As shown in FIGURE 1, four relevant landmarks are manually marked on the cephalograms to measure the AN ratio. However, accurate identification of these landmarks is highly dependent on the examiner's clinical experience, leading to different results among different examiners. Furthermore, this task is time-consuming and involves repetitive work that may affect doctors' productivity. Therefore, it's meaningful to develop an accurate and efficient algorithm for the automatic measurement of AN on lateral cephalograms.

Benefiting from the rapid development of artificial intelligence, deep learning-based methods have made great progress in many tasks such as medical image classification [9], retinal vessel segmentation [10], brain tumor segmentation [11], [12] and so on [13], [14]. Despite these advances, research on the use of deep learning-based methods for radiographic adenoid hypertrophy (AH) assessment remains limited [15], [16]. Furthermore, existing methods are primarily based on X-ray images and do not perform well when applied to other image formats such as Magnetic Resonance Imaging (MRI) images. Given the potential radiation risks associated with X-rays, especially for lateral cephalometric X-rays, patients and their families may be uncomfortable with this, making taking regular check-ups during conservative treatment may be difficult. In contrast, MRI is a non-invasive, non-radiographic examination, with no radiological or biological damage to brain tissue, and MRI images have a relatively high resolution of soft tissue, thus it is more suitable for examining the adenoids [17]. However, up to date, no automated diagnosis method for adenoid hypertrophy based on MRI images has been proposed. Given the above, the automatic method for AH assessment in MRI images is extremely essential.

In this paper, we propose the first deep learning model for automatically measuring AN ratios in MRI images to detect AH and assess its severity in patients. Specifically, we propose a novel network called Adenoid-Net (ADNet) with encoder-decoder architecture for accurate landmark detection around adenoid sites and then calculate the AN ratio. Our model involves a sequence of convolutional neural networks as the feature extractor to extract features that guide landmark detection and a landmark detection head to export the position of predicted landmarks. In the feature extractor, we employ a depthwise separable convolution-based encoder to extract local features, making the network pay more attention to important regions. We then introduce adaptive convolution in the decoder to catch long-range and deformation features. The landmark detection head predicts the heatmaps for landmarks and finally outputs the coordinates of four landmarks. For efficient acquisition of numerical coordinates, we integrate a numerical coordinate regression layer [18] into the detection head after the heatmap-generating layer. The numerical coordinate regression layer enables the network to be fully differentiable, thereby overcoming the limitations of heatmap-based methods [18]. To evaluate the effectiveness of our model, we collected 500 lateral MRI images with professional annotations from Shuguang Hospital for this study. Our experimental results show that our model can automatically locate the four adenoid landmarks with an average point error of 4.85 pixels and then measure the AN ratio with an average error of 0.026. Thus, our model can be a powerful tool for physicians to access AH in MRI images, reducing the burden of repetitive work for physicians and improving their efficiency.

In summary, the key contributions of this work are as follows:

- We first apply the deep learning-based approach to the automatic assessment of AH in MRI images. We propose a novel model named ADNet to study the task of landmark detection around adenoids in MRI images and thereby automatically evaluate the presence of AH in patients, which can effectively mine local and global features to achieve fast and accurate landmark detection.
- We propose an adaptive convolution module that can make the convolution randomly sampled by learning the offset of sampling points to adapt to the deformation features. We propose a simple and effective encoder to extract features which significantly reduces the number of model parameters.
- Experimental results show that our network can effectively detect four landmarks and assess adenoid hypertrophy with high accuracy, indicating it can be an effective method to assist physicians in analysis and diagnosis.

The rest of this paper is organized as follows. Section II outlines the work related to medical image landmark detection. Section III describes the details of the proposed model. Section IV presents the experimental setup and discusses the result. Finally, Section V draws the conclusions and discusses further work of this paper.

II. RELATED WORK

In this section, we present recent work that is most relevant to our work, including human pose estimation, medical landmark detection methods and deformable convolution.

A. HUMAN POSE ESTIMATION

The main goal of Human Pose Estimation (HPE) is to detect joint key points of the human body, which is a key point localization task in computer vision and is widely used as a fundamental task for semantic segmentation, pedestrian re-identification, action recognition, etc.

The original HPE methods [19] depend on hand-crafted features to learn the relations between different body parts to describe the human body. However, they are limited in accuracy especially under severe occlusions and complex conditions. Toshev et al. [20] applied Convolutional Neural Networks to the human pose estimation task for the first time, which can directly get the specific coordinates of the body parts. However, this method directly regression fits the image coordinates, which in turn is highly nonlinear, making it difficult for the network to learn such a mapping. To cope with this problem, Thompson et al. [21] proposed a method to obtain key point coordinates by generating a Gaussian heatmap, where each pixel in the heatmap represents the probability of the existence of an articulation point. Since then, most HPE methods employ Gaussian heatmap based methods to detect joint key points. To deal with body parts with different scales, such as face, hands, and feet, Newell et al. [22] proposed a Stacked Hourglass Network (SHG) that captures features at each scale by stacking several hourglass modules with pooling and upsampling. To cope with the challenge of occluded joints, invisible joints, and complex backgrounds in the wild for HPE, Chen et al. [23] proposed a two-stage network called Cascaded Pyramid Network (CPN) with a GlobalNet and a RefineNet, and the GlobalNet is responsible for the easy samples and the RefineNet aims at handling those challenging samples. Most existing architectures use the high-to-low and low-to-high processes to learn multi-scale features. However, recovering from low-resolution representations does not cover the loss of information of downsampling. To maintain high-resolution representations through the whole process, Sun et al. [24] presented a novel architecture named HighResolution Net (HRNet).

Medical image landmark detection and HPE are both key points detection tasks, and there are some correlations in methodology. Some papers apply HPE methods to medical landmark detection tasks without adjustment. However, compared with natural images, medical images are not such rich in information and have very little difference between pixels. Therefore, the human pose approach cannot be directly transferred to medical landmark detection tasks. In the next section, we describe the work related to medical landmark detection.

B. MEDICAL LANDMARK DETECTION

In the field of medical image analysis, accurate localization of anatomical landmarks is a critical step in treatment planning. Recent advances in deep learning have lead to the development of several effective methods for landmark detection. Zhong et al. [25] used a two-stage U-Net heatmap regressing method for skull landmark detection. They embed attention mechanisms with global stage heatmaps to guide the local stage heatmap patches. To address the challenges of detecting landmarks with different levels of resolutions and semantics, Chen et al. [26] proposed an attentional feature pyramid fusion module to fuse high resolution and semantic enhanced features to achieve higher accuracy for cephalometric landmark detection. To achieve large-scale landmark detection, Liu et al. [27] proposed a two-stage model to segment two bones and detect 175 key points in CT images, in which the first stage produces coarse segmentation and landmark and then the second stage crop the regions of interest from the original image for further segmentation refinement and landmark detection. Khanal et al. [28] proposed a method combining with Densenet to detect the landmarks of spinal bone corners. Noothout et al. [29] employ a global-to-local localization approach in which the global convolution network performs regression and classification to simultaneously get displacement vectors and the presence of landmarks of interest separately, and then landmarks are refined by analyzing local sub-images. To allow a model to learn anatomical context rather than depending on handcrafted graphical models, Oh et al. [30] proposed a novel framework consisting of the Local Feature Perturbator and the Anatomical Context loss, forcing the network to gaze relevant features more globally and learn the anatomical context based on spatial relationships between the landmarks.

In conclusion, most existing landmark detection methods are based on Gaussian heatmap regression to learn the positions of landmarks of medical images, suggesting the great applicability of heatmap regression. In view of this, in this paper, we consider utilizing heatmap regression to detect landmarks.

C. DEFORMABLE CONVOLUTION

In the field of visual recognition, a key challenge is accommodating geometric variations in object scale, pose, and viewpoint. Typically, researchers address this challenge by either training models on a large dataset with sufficient variations or using transformation-invariant features and algorithms like SIFT (scale invariant feature transform) [31]. However, the design of such algorithms relies heavily on human expertise and empirical intuition. To overcome these limitations, deformable convolution [32] was introduced to learn spatial offsets that enable the network to adjust its sampling position based on the previous feature map and adapt to the geometric changes of the object. This method adds offsets to the regular grid sampling locations in standard convolution, thereby achieving free deformation of the sampling network. As a result, the network can adjust the convolution receptive field and sampling locations according to the object scale and shape, significantly improving its ability to model deformed features.

In our task, the morphology of adenoids is very complex and adaptive, which makes the localization of key points more difficult. Therefore, we introduce deformable convolution to help our network extract the features of adenoids better.

III. METHOD

A. NETWORK STRUCTURE

We propose a novel model named ADNet for adenoid landmark detection in MRI images. FIGURE 2(a) illustrates the overall framework of ADNet. Our proposed model consists of an encoder, a decoder, and a landmark detection head. The encoder is used for multi-level feature extraction of the image, and then the decoder upsamples to recover the image resolution. Skip connections are used between each encoder layer and its corresponding decoder layer to compensate for the loss of information due to downsampling. Finally, the detection head predicts the landmarks coordinates based on the extracted features. We introduce each component of the proposed network in the following sections.

1) ENCODER

For the overall structure of the encoder, we first use a standard Visual Geometry Group (VGG) block [33], i.e., two standard 3×3 convolution layers (Conv) each followed by a batch normalization layer (BN) and a rectified linear unit (ReLU) to extract the image features initially. Then, to extract information at each resolution for multi-scale features, we stack four blocks of our depthwise-inverse-bottleneck (DI) block, between which 2×2 maximum pooling is adopted for downsampling to reduce the resolution.

The details of the DI block are shown in FIGURE 2(b). Each DI block consists of a depthwise convolution, two vanilla convolution layers, and residual connections. Depthwise convolution applies a different convolution kernel to each input channel to extract spatial features with the kernel size of 7×7 and padding of 3. And then, the vanilla convolution layers use a 1×1 convolution kernel to combine features of different channels to perform cross-channel feature selection. By introducing these modules, the network can learn the importance distribution of the input terms to focus more on features in important regions. This DI block has an inverse bottleneck layer structure, with the middle layer having the largest number of channels. It extracts, expands, and compresses features at each layer and then adds the final features to the original input features through residual connections. The inverse bottleneck layer structure can effectively reduce information loss. The effectiveness of the encoder is demonstrated using experimental results in Section IV-F.

2) DECODER

Similar to the way the encoder is organized, we stack a VGG block and three convolution-adaptive (Conv-adapt) blocks, where each Conv-adapt block consists of a standard 3×3 convolution block (Conv + BN + ReLU layer) and an adaptive convolution. To upsample the low-resolution feature maps to the high-resolution feature maps, we use transpose convolution operation between adjacent Conv-adapt blocks. To achieve interaction between deep and shallow information, we use skip connections to connect the encoder with the

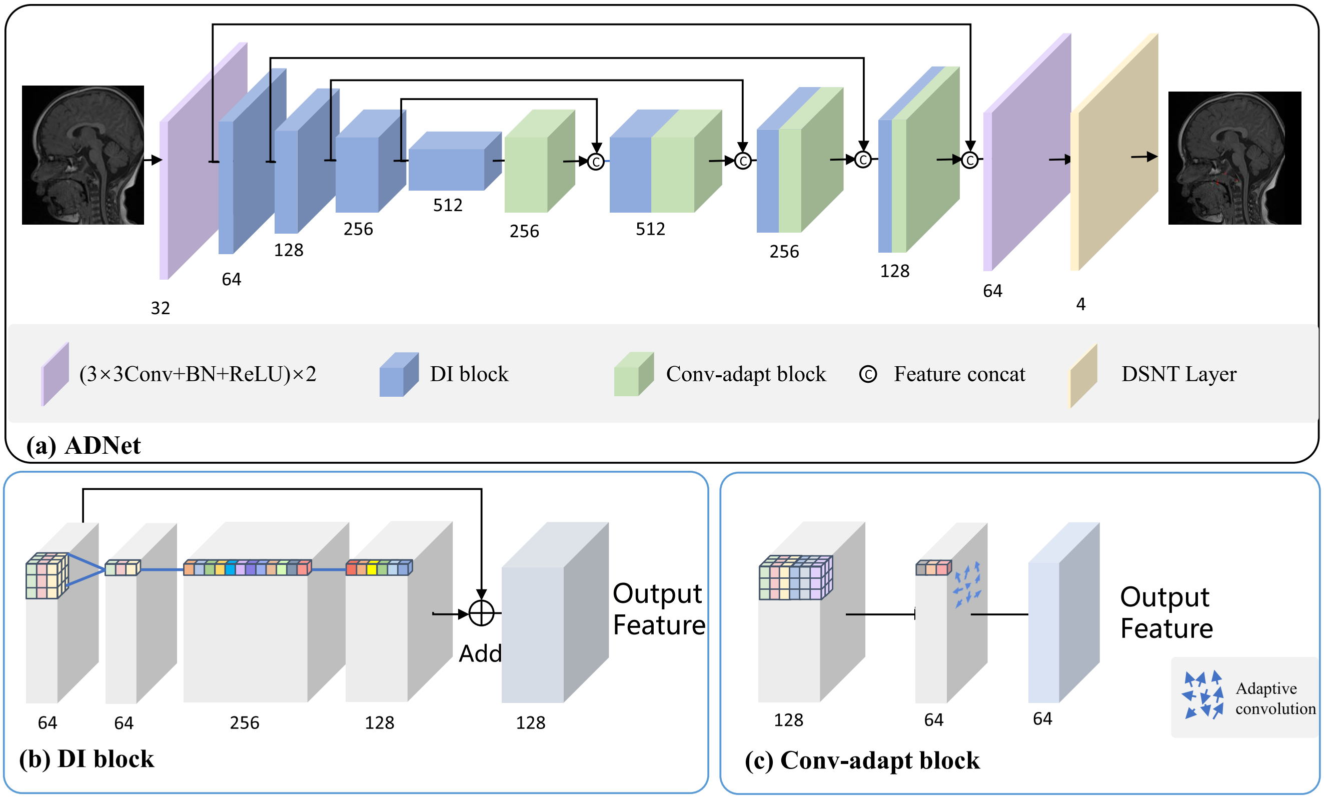


FIGURE 2. Framework of ADNet. Firstly it extracts features through an encoder-decoder type network, with skip-connection to retain spatial information. Secondly, it outputs four heatmaps to predict landmarks. Finally, it employs a numerical coordinate regression layer to infer the position of landmarks based on heatmaps.

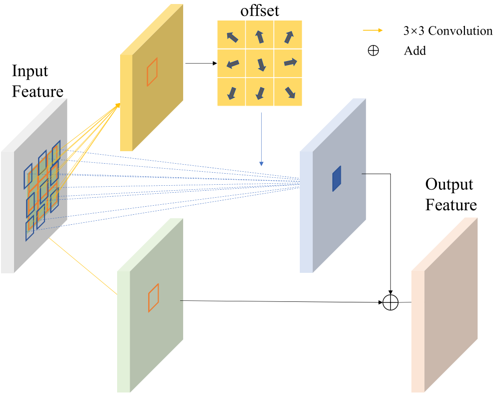


FIGURE 3. Structure of adaptive convolution module.

decoder. This approach allows for the efficient extraction of features from multiple resolutions and the integration of both deep and shallow information for more accurate predictions.

The adaptive convolution module is implemented using a two-branch structure and the its details are shown in the FIGURE 3.

One of the branches employs deformable convolution to learn the offset of each sampling point position in the convolution kernel. The deformable convolution enables the kernel to be randomly sampled within a certain range of the current position rather than being restricted to the regular grid points of conventional convolution. The deformable convolution consists of two steps: first, it gets the offsets and deformation parameters via 3 × 3 convolution; second, it calculates the offset sampling points using bilinear interpolation. Formally, we calculate the output of each input point as follows:

$$Y_{\text{Deform}}(p) = \sum_{k} w_k x(p + s_k p_k + \Delta p_k), \tag{1}$$

where w_k denotes the weight of convolution kernel, $x(p)$ represents the feature vector of the input feature map x at position p, and $y(p)$ represents the feature vector of the output feature map Y at position p. s_k is the adaptive dilation factor, p_k denotes the handpicked offset, and Δp_k is the spatial offset. Note that usually the offsets Δp_k and s_k are fractional, resulting in non-integer coordinates that cannot be localized in the image data. To address this, we apply the bilinear interpolation method to determine the location of the sampled points, which is defined as follows:

$$x(p) = \Sigma_p \mathrm{G}(q, p) x(q), \tag{2}$$

where $\mathrm{G}(\cdot)$ denotes the bilinear interpolation kernel and q represents all integer positions in the feature map x.

The other branch employs a vanilla 3 × 3 convolution, whose output is calculated as follows:

$$Y_{\text{Normal}}(p) = \Sigma_k w_k * x(p + p_k). \tag{3}$$

Then the final output of adaptive convolution is obtained by combining the output of the two branches, which is defined as follows:

$$Y_{\text{Adap}} = Y_{\text{Deform}} + Y_{\text{Normal}}. \tag{4}$$

The adaptive convolution module can dynamically adjust the shape and size of the convolution kernel according to the input, which enables the module to adapt to various perceptual field sizes and shapes. This ability to modify the kernel size and shape facilitates more precise feature extraction of the target object, resulting in improved performance. Furthermore, the module can learn the importance of each pixel point in the input feature map, which allows the network to focus more on the features of crucial regions and enhance the model's perceptual and expressive power.

3) LANDMARK DETECTION HEAD

Landmarks in the vicinity of adenoids maintain local morphological features and structural stability despite local disorder. Therefore, the local neighborhood centered on the landmarks is a remarkable marker for dependent mining. Intuitively, the features closer to the landmarks can provide more accurate guidance for detection. With this in mind, we define the heatmap using Gaussian functions to highlight the location of landmarks.

It is worth noting that there are two defects in the process of generating coordinates from the heatmap: 1) obtaining the final coordinates from the heatmap requires the use of *argmax*, which is non-differentiable and cannot be learned directly; 2) the coordinates obtained are restricted to integers, leading to inaccuracies that depend on the resolution of the heatmap. While the supervisory signal is directly derived from the heatmap, these two defects lead to the separation of the loss function from our target coordinates, thus impacting the prediction results. To address these problems, we adopt the method proposed in [18]. Specifically, we add the differentiable spatial to numerical transform (DSNT) layer to our landmark detection head at the last layer so that we can get numerical landmarks from the heatmap and make the network a fully differentiated end-to-end network, which leads to better performance.

B. LOSS FUNCTION

Considering that the final output of the model is numerical coordinates, we define the core term of loss function as computing the two-dimensional Euclidean distance between the prediction μ and ground truth q, which can be expressed as follows:

$$L_{euc}(\mu, q) = ||q - \mu||_2. \tag{5}$$

The Euclidean loss function has been shown to be effective in optimizing the distance between predicted and actual position according to the previous studies [34].

However, relying solely on this metric may lead to confusion with the problem of different heatmaps producing the same coordinate output. To address this issue, we incorporate pixel-level supervision of the heatmap during training by introducing regularization. Regularization has also been shown to play a critical role in improving model generalization [35]. Therefore, the final loss function is a combination of the Euclidean loss and a regularization term defined as:

$$L = L_{euc}(DSNT(P), q) + L_{reg}(P, Q), \tag{6}$$

where P and Q denote the predicted heatmap and heatmap generated from ground truth, respectively. And the regularization is Jensen-Shannon divergence that is defined as follows:

$$\begin{aligned} JS(P||Q) = &\frac{1}{2}\Sigma p(x) log(\frac{2p(x)}{p(x)+q(x)}) \\ &+ \frac{1}{2}\Sigma q(x) log(\frac{2q(x)}{p(x)+q(x)}). \end{aligned} \tag{7}$$

TABLE 1. Summary of the dataset used in this study.

AN ratio	Train set	Valid set	Test set	Total
≥0.6	263	37	75	375
<0.6	87	13	25	125
Total	350	50	100	500

IV. EXPERIMENT

A. DATASET

To our knowledge, there is no publicly available landmark dataset of adenoid MRI images. In this study, we collected lateral cephalometric MRI images from 331 patients in Shuguang Hospital. Then we select three images for each patient, including the nasal septum image, the left one and the right one next to it. After removing images of poor quality, we obtained a total of 500 usable images. These images are converted from DCM format to PNG, then uniformly resized and cropped to 540 × 640. The dataset is annotated with four marker points on MRI images by two specialized physicians and in consent. For each image, we obtain the positions of four landmarks and the AN ratio. If the AN ratio is greater than 0.6, the child is suspected of having AH. We divide the dataset into a training set, validation set, and test set in the ratio of 7:1:2, respectively, for our experiments. And the details of the dataset are shown in TABLE 1.

B. IMPLEMENT DETAILS

In this study, we conduct experiments on a Tesla V100 GPU using the PyTorch deep learning framework. We resize the images to 256 × 256 pixels in the training process. We use random resize, horizontal flip and random rotation (±15°) as data augmentation techniques to reduce overfitting. During training, we use the Adam [36] optimizer with the learning

TABLE 2. Confusion matrix.

		Ground-truth	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

rate initialized as 1e-4 and multiplied by 0.5 every 50 epochs. The number of epochs and batch size are 200 and 16, respectively.

C. EVALUATION METRICS

To verify the performance for the detection of landmarks of different models, we calculated the mean radial error (MRE) between the predicted coordinates and ground truth, which is defined as:

$$MRE = \frac{1}{n}\sum_{i}^{n} R_i, \tag{8}$$

where n denotes the number of landmarks and R_i stands for the Euclidean distance between the predicted landmark coordinates and the ground-truth coordinates. We also calculate the successful detection rate (SDR) to evaluate the percentage of predicted landmarks located within a certain threshold distance from the ground truth. Besides, for the determination of adenoid hypertrophy, we calculated the average AN ratio error (ANE), average accuracy (ACC), precision (PRE), recall (Recall), and macro F1 score to test the performance of the model as follows:

$$ANE = \frac{1}{n}\sum_{i} |AN_{pre}^{i} - AN_{gt}^{i}|, \tag{9}$$

where AN_{pre}^{i} and AN_{gt}^{i} denote the predicted AN ratio and the ground-truth AN ratio of the ith sample respectively.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

$$PRE = \frac{TP}{TP + FP}, \tag{11}$$

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

where TP, TN, FP, and FN denote the amount of true positive, true negative, false positive, and false negative samples respectively, the details are shown in TABLE 2.

$$F1 = 2\frac{PRE \times Recall}{PRE + Recall}. \tag{13}$$

D. BASELINE COMPARISON

To evaluate the performance of our model, we compare ADNet with other baseline models. These models include three well-known key point detection networks: SHG [22], CPN [23] and HRNet [24], and UNet-a baseline network for medical image processing. These key point detection methods are all based on Gaussian heatmaps, where they model keypoint locations as heatmaps and train convolutional networks to predict the heatmaps. In this work, we utilize the SHG with 4 stages for comparison, where stage denotes the number of hourglass modules. CPN [23] is another classical baseline for keypoint detection that uses a cascaded two-stage network and we follow its official implementation. HRNet [24] is a strong and popular baseline that maintains high resolution at each stage of the network and fuses features at different scales. It has been shown to perform well in several areas such as human pose estimation, face feature point detection, and object detection. It is worth noting that there are two main variants of HRNet, including a small network and a large network: HRNet-W32 and HRNet-W48, where 32 and 48 denote the widths (channels) of the latter three stages of the high-resolution sub-networks, respectively. In this task we use HRNet-32 and adopt the official implementation of open source. For UNet, we extended the official implementation by incorporating a landmark detection head, making it able to predict landmarks based on the heatmap.

E. MODEL PERFORMANCE

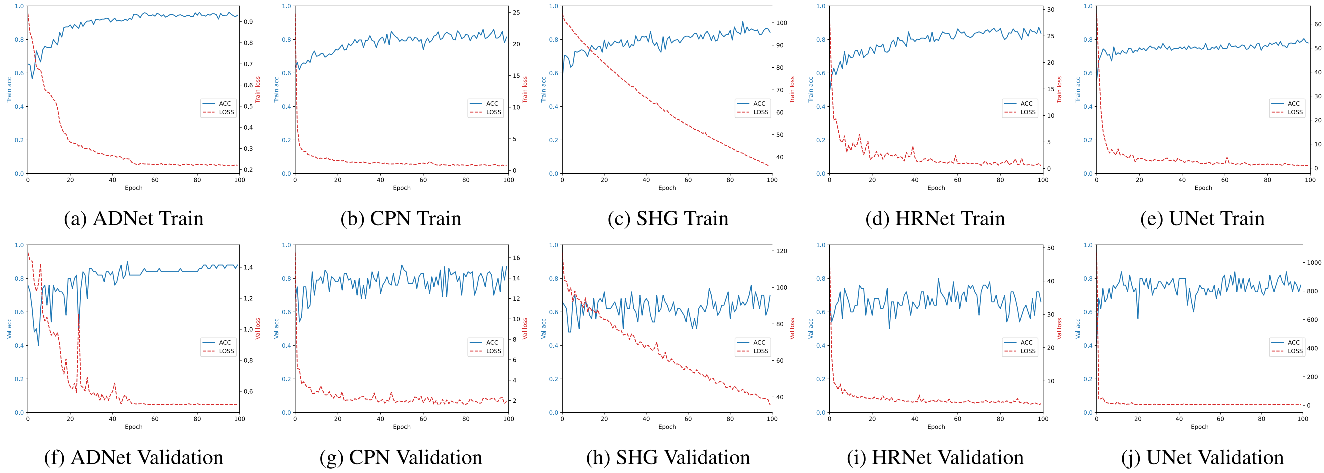
The accuracy and loss curves over epochs of each model on the training and validation sets are shown in FIGURE 4. And we present the comparison curves of the accuracy, ANE, and MRE for each model on the training and validation sets in different stages in FIGURE 5. In FIGURE 5a, after roughly 20 epochs, our model consistently outperforms other models in terms of accuracy. Specifically, our model achieves an accuracy of over 90% after just 40 epochs and maintains this level with a relatively smooth curve. In contrast, other models require around 160 epochs to reach 90% accuracy and experience greater fluctuations throughout training. Similarly, in FIGURE 5d, our model outperforms other models after 50 epochs, and the accuracy stays between 84% and 86%. Due to the small size of the validation set (only 50 images), the accuracy of all models fluctuates more significantly. However, our model demonstrates smoother fluctuations than other models, indicating its more stable performance and highlighting the effectiveness of our model. FIGURE 5b, 5e, 5c, and 5f demonstrate that our model has lower ANE and MRE values than the other models after 35 epochs, suggesting that our model enables more precise landmark localization and assessment of AH.

The results of our model and the baseline models and the manual method on the test set we collected are presented in TABLE 3. From the results, it's obvious that the SHG, CPN, and HRNet don't perform well in this task with the ANE value above 0.07, and HRNet has the worst result with the ANE value of 0.095. The accuracy of AN (evaluated by ANE) is determined by the results of landmark localization, which also indicates their poor performance for landmark detection in this task. Whereas UNet is relatively better with the ANE value of 0.063, possibly because of the concentrated feature regions and smaller inter-pixel differences in medical images compared with natural images and the limited amount of data available. It's clear that our method performs much

TABLE 3. Experimental results of our ADNet and baseline models.

Model	Params (M)	ANE↓	ACC (%)↑	PRE (%)↑	F1 (%)↑	Recall (%)↑
Manual	N.A.	0.033	95.00	94.05	96.93	100.00
SHG [22]	711.31	0.075	85.00	94.59	90.32	86.42
CPN [23]	26.54	0.075	89.00	96.05	92.99	90.12
HRNet [24]	28.54	0.095	82.00	83.16	89.77	97.53
UNet [37]	22.95	0.063	86.00	85.26	92.04	**100.00**
ADNet (ours)	28.86	**0.026**	**94.00**	**96.30**	**96.30**	96.30

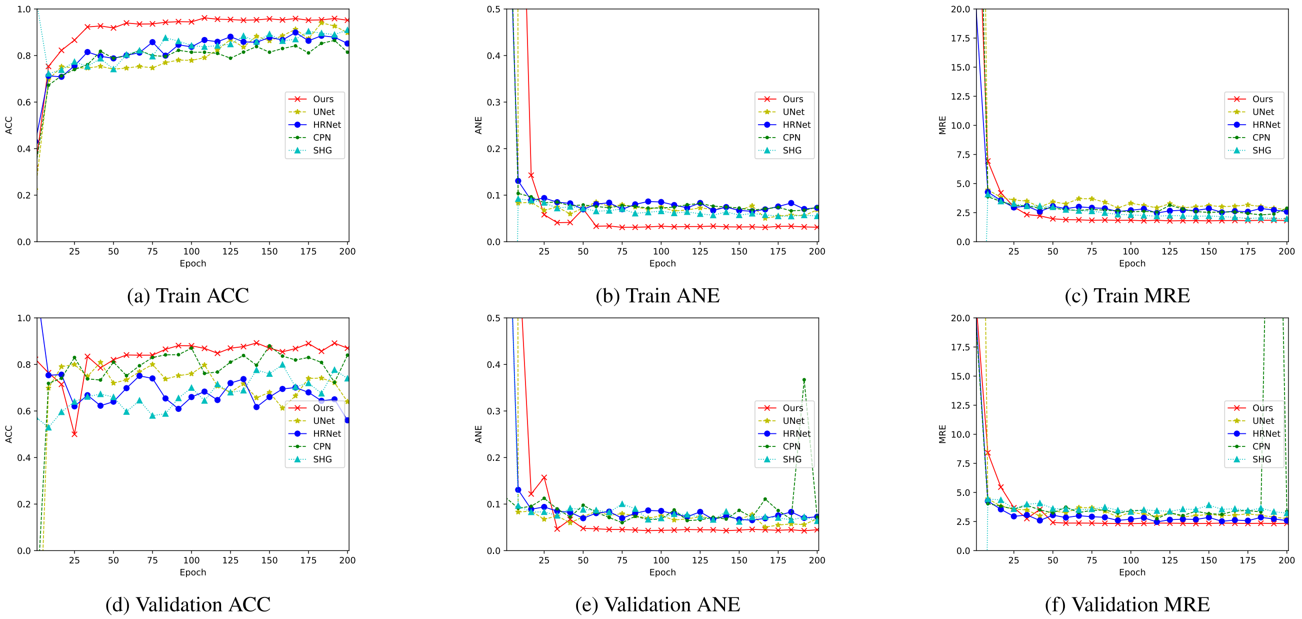
The manual results were obtained by a medical professional with extensive experience in AH diagnosis measuring on the test set.



(a) ADNet Train (b) CPN Train (c) SHG Train (d) HRNet Train (e) UNet Train

(f) ADNet Validation (g) CPN Validation (h) SHG Validation (i) HRNet Validation (j) UNet Validation

FIGURE 4. The accuracy and loss curves over epochs of each model on the training and validation sets.



(a) Train ACC (b) Train ANE (c) Train MRE

(d) Validation ACC (e) Validation ANE (f) Validation MRE

FIGURE 5. Comparison curves of accuracy, ANE, and MRE over epochs of each model on the training and validation sets.

better than all of them, with the ANE value of 0.026 and the ACC value of 94%, indicating that our network can determine whether the child has adenoid hypertrophy or the severity of the adenoid hypertrophy from the pictures more accurately. In addition, the ANE of the medical professional's manual method is 0.033 and the classification accuracy reaches 95%. These results indicate that our model reaches the level of medical experts in diagnosing of AH. Besides, the manual

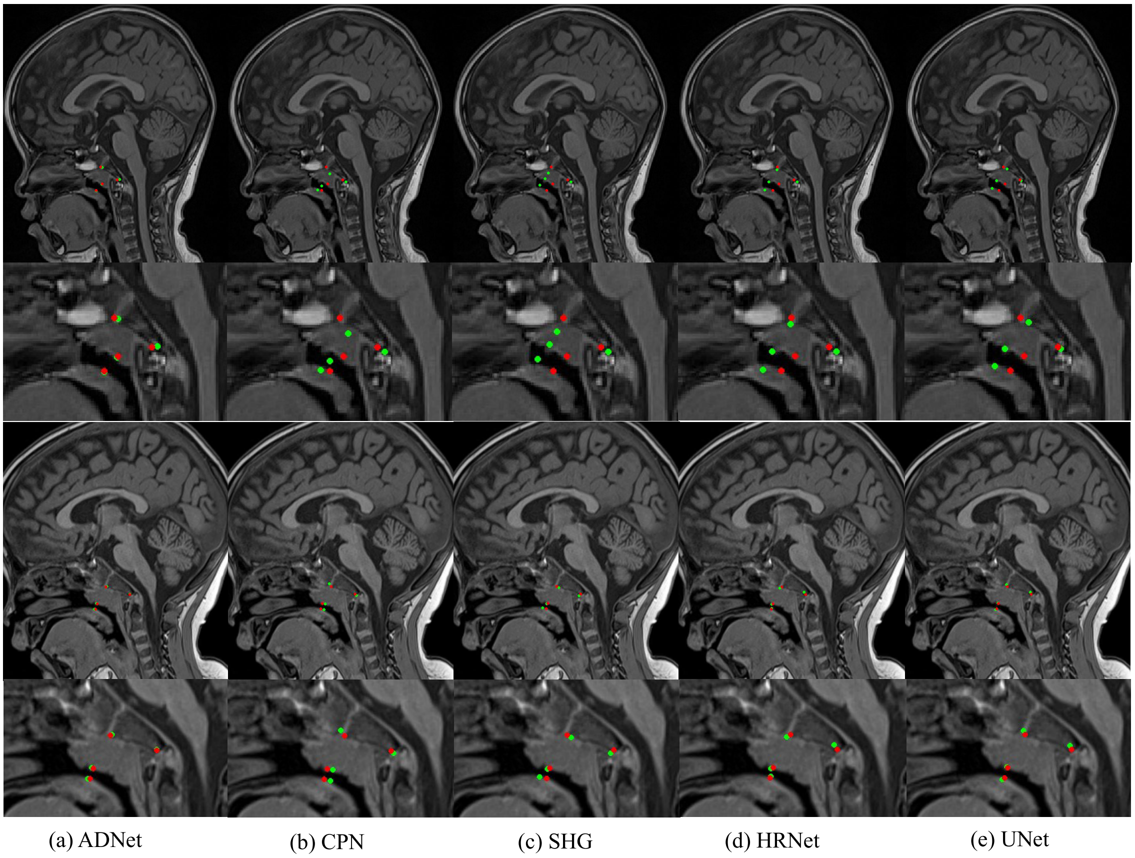


FIGURE 6. Examples of predictions from different models. Automatically located (red) and manually annotated landmarks (green) in MRI images.

TABLE 4. MRE results of our proposed ADNet and baseline models in MRI images.

Model	MRE (pixel)				
	LM1	LM2	LM3	LM4	Average
SHG [22]	10.45	8.35	10.78	10.05	9.91
CPN [23]	8.91	7.67	11.01	11.01	9.65
HRNet [24]	13.31	16.31	16.76	17.43	15.95
UNet [37]	7.95	5.87	9.90	9.47	8.30
ADNet (ours)	**4.28**	**3.76**	**5.45**	**5.92**	**4.85**

method is very time-consuming, taking 3-5 seconds per image processing on average. In comparison, our model is faster, with the inference process completed in about 0.6 seconds per image. As a result, ADNet can be a reliable and efficient tool to assist doctors in AH diagnosis. While our model has achieved a classification accuracy of 94%, it is necessary to improve its accuracy further to ensure the reliability in medical applications.

Then we compare our model with baseline models on the performance of landmark detection evaluated by MRE shown in TABLE 4. Our proposed method achieves the smallest error for all landmarks and the average MRE value is 4.85, which is 3.55 less than the suboptimal method UNet of 8.30. And SHG, CPN, and HRNet perform much worse in this task with the MRE value exceeding 9, and HRNet reaching 15.95. This result may be due to the fact that other models share a common limitation in that they focus mainly on local features, whereas lacking a global representation of the structure or shape, which may lead to errors in landmark detection. Given that our task involves tiny detection area with significant interference information in the head, it is crucial to consider global contextual information. Our proposed encoder and adaptive convolution can effectively integrate contextual information, allowing the network to learn global structure information and improve performance. Besides, those baseline models predict landmark positions directly from the heatmap, and the accuracy of positions is heavily dependent on the resolution of heatmaps. In contrast, our proposed method with the DSNT layer enables the network to obtain numerical coordinates so that it can achieve more accurate positioning regardless of the resolution of heatmaps. Given the critical importance of accurate landmark detection in ensuring precise AN ratio calculation and AH assessment, the excellent performance of our model in landmark detection strongly suggests its excellence in diagnosing AH.

We compare the ground truth with the prediction of our proposed ADNet and baseline models in FIGURE 6, where the green and red points denote the prediction and the ground truth of the landmarks, respectively. As is shown in the figure,

TABLE 5. The performance of our proposed ADNet and baseline models in MRI images for SDR.

Model	SDR (%)↑														
	1			2			3			4			Average		
	< 10	< 15	< 20	< 10	< 15	< 20	< 10	< 15	< 20	< 10	< 15	< 20	< 10	< 15	< 20
SHG [22]	42	85	88	68	91	97	62	78	87	67	83	88	59.75	84.25	90
CPN [23]	66	87	96	81	92	99	54	74	85	53	73	90	63.5	81.5	92.5
HRNet [24]	48	72	92	39	62	88	33	67	83	40	68	83	40	67.25	86.5
UNet [37]	76	91	**99**	90	98	**99**	70	85	87	66	85	89	75.5	89.75	93.5
ADNet (ours)	**92**	**96**	96	**98**	**98**	98	**93**	**95**	**95**	**90**	**95**	**95**	**93.25**	**96**	**96**

TABLE 6. Ablation studies for each module of our proposed ADNet.

Model	Components			MRE (pixel)					ANE	ACC (%)	Params (M)
	DIBlock	Conv-adap Block	DSNT	LM1	LM2	LM3	LM4	Average			
AD-1		✓	✓	4.23	3.68	6.9	7.07	5.47	0.044	91	42.74
AD-2	✓		✓	4.39	4.48	6.10	7.45	5.60	0.036	94	25.52
AD-3	✓	✓		6.37	6.37	7.85	8.79	7.35	0.048	90	28.86
ADNet (ours)	✓	✓	✓	**4.28**	**3.76**	**5.45**	**5.92**	**4.85**	**0.026**	**94**	28.86

our proposed ADNet achieves more accurate localization for four landmarks than other models.

We present the SDR results of our model and baseline models in TABLE 5. The results demonstrate that our proposed method outperforms the baseline models for different value ranges of SDR. The ADNet can achieve landmark detection errors within 10 pixels with 93.25% probability, which is about 20% higher than the second-best performing method UNet, and within 15 pixels with 96 probability, which is about 6% higher than UNet. These results powerfully demonstrate the high accuracy of our model and its potential as a clinical aid for diagnosis. And although UNet outperforms ADNet for detection rates within 20 pixels of error-a clinically acceptable error at landmark LM1 and landmark LM2, our method performs better by 8% and 6% at landmark LM3 and landmark LM4, respectively. Landmark LM1 and landmark LM2, which possess distinct features, are relatively easier identified. Whereas landmark LM3 and landmark LM4 rely more on the network's ability to capture adenoidal region features, indicating our approach of modeling long-range reliance and capturing global information can improve the CNN's performance for landmark detection.

F. ABLATION STUDY

In this section, we conduct ablation studies to analyze the impact of each module on the model's performance. Specifically, we evaluate three variants of our proposed model: AD-1, AD-2, and AD-3. AD-1 uses the standard VGG blocks in the encoder stage. AD-2 removes the adaptive convolution module in the decoder, making it also the standard VGG blocks. AD-3 only supervises the heatmap like other methods without DSNT layer [22], [23], [24].

The results of ablation studies are shown in TABLE 6. The result of AD-1 shows that our encoder is more suitable for this task, with the MRE values increasing by 0.45, 1.15, and 0.62 for landmark LM3, landmark LM4, and average, respectively, and the ANE value increasing by 0.018. This result suggests that our proposed encoder can extract features more effectively, with an attention-like function allowing the network to focus more on crucial features and the inverse bottleneck structure avoiding information loss, leading to more precise localization. Besides, our encoder significantly reduces the number of parameters added by the adaptive module. Then the result of AD-2 indicates that the model without the adaptive convolution module shows poor performance on the location of landmark LM2, landmark LM3, and landmark LM4, with the average MRE value 0.75 higher than ADNet. This result demonstrates the importance of capturing global dependencies for landmark detection. At last, the result of AD-3 suggests that removing the DSNT layer has the most significant impact on the performance of the model, with an average MRE value of 7.35 and the value of ANE of 0.048. Compared to our model ADNet, AD-3 has a 2.50 higher MRE value and 0.022 higher ANE value, indicating that the DSNT layer is very effective in ensuring accurate landmark detection.

V. CONCLUSION

In this paper, we propose an efficient model for adenoid hypertrophy diagnosis with limited data. We consider AH measurement as a landmark detection task, and propose an end-to-end network named ADNet to detect the landmarks. We improve the feature extractor to capture deformation features and long-range dependency, and use a new landmark detection method to obtain landmarks coordinates. We conduct substantial experiments on our collected dataset. The results show that our network can effectively measure the AN ratio based on MRI images to predict and assess adenoid hypertrophy, eliminating possible error caused by human operation and greatly reducing time consumption. Therefore, this automated assessment method can be applied to relevant clinical studies as well as community level health screening.

The current version of ADNet still has limitations: in the field of medical imaging, models trained by one imaging protocol are often not applicable to data collected by another imaging protocol. To further improve the performance of the model, we will try to collect more datasets on different imaging protocols to train our model. Making the network better adapted to data from other domains is one of our future research directions. Besides, there is potential for improvement of the model to achieve more precise landmark detection and accurate assessment of AH. To this end, we will continue to collect the head MRI images and explore more accurate methods for landmark detection and AH diagnosis.

REFERENCES

[1] M. P. Major, C. Flores-Mir, and P. W. Major, "Assessment of lateral cephalometric diagnosis of adenoid hypertrophy and posterior upper airway obstruction: A systematic review," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 130, no. 6, pp. 700–708, Dec. 2006.

[2] I. Brambilla, A. Pusateri, F. Pagella, D. Caimmi, S. Caimmi, A. Licari, S. Barberi, A. M. Castellazzi, and G. L. Marseglia, "Adenoids in children: Advances in immunology, diagnosis, and surgery," *Clin. Anatomy*, vol. 27, no. 3, pp. 346–352, Apr. 2014.

[3] L. Pereira, J. Monyror, F. T. Almeida, F. R. Almeida, E. Guerra, C. Flores-Mir, and C. Pachêco-Pereira, "Prevalence of adenoid hypertrophy: A systematic review and meta-analysis," *Sleep Med. Rev.*, vol. 38, pp. 101–112, Apr. 2018.

[4] C.-Y. Chien, A.-M. Chen, C.-F. Hwang, and C.-Y. Su, "The clinical significance of adenoid–choanae area ratio in children with adenoid hypertrophy," *Int. J. Pediatric Otorhinolaryngol.*, vol. 69, no. 2, pp. 235–239, Feb. 2005.

[5] D. I. Filho, D. B. Raveli, R. B. Raveli, L. D. C. L. Monteiro, and L. G. Gandin, "A comparison of nasopharyngeal endoscopy and lateral cephalometric radiography in the diagnosis of nasopharyngeal airway obstruction," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 120, no. 4, pp. 348–352, Oct. 2001.

[6] R. A. Wood, "Assessment of adenoidal obstruction in children: Clinical SignsVersus roentgenographic findings," *Pediatrics*, vol. 104, no. 2, p. 370, Aug. 1999.

[7] L. Soldatova, H. J. Otero, D. A. Saul, C. A. Barrera, and L. Elden, "Lateral neck radiography in preoperative evaluation of adenoid hypertrophy," *Ann. Otol., Rhinol. Laryngol.*, vol. 129, no. 5, pp. 482–488, May 2020.

[8] M. Fujioka, L. Young, and B. Girdany, "Radiographic evaluation of adenoidal size in children: Adenoidal-nasopharyngeal ratio," *Amer. J. Roentgenol.*, vol. 133, no. 3, pp. 401–404, Sep. 1979.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[10] W. Ma, S. Yu, K. Ma, J. Wang, X. Ding, and Y. Zheng, "Multi-task neural networks with spatial activation for retinal vessel segmentation and artery/vein classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 769–778.

[11] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 109–119.

[12] W. Ren, A. H. Bashkandi, J. A. Jahanshahi, A. Q. M. AlHamad, D. Javaheri, and M. Mohammadi, "Brain tumor diagnosis using a step-by-step methodology based on courtship learning-based water strider algorithm," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104614.

[13] L. Wang, Y. Xiao, J. Li, X. Feng, Q. Li, and J. Yang, "IIRWR: Internal inclined random walk with restart for LncRNA-disease association prediction," *IEEE Access*, vol. 7, pp. 54034–54041, 2019.

[14] Y. Xiao, Z. Xiao, X. Feng, Z. Chen, L. Kuang, and L. Wang, "A novel computational model for predicting potential LncRNA-disease associations based on both direct and indirect features of LncRNA-disease pairs," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–22, Dec. 2020.

[15] Y. Shen, X. Li, X. Liang, H. Xu, C. Li, Y. Yu, and B. Qiu, "A deep-learning-based approach for adenoid hypertrophy diagnosis," *Med. Phys.*, vol. 47, no. 5, pp. 2171–2181, May 2020.

[16] T. Zhao, J. Zhou, J. Yan, L. Cao, Y. Cao, F. Hua, and H. He, "Automated adenoid hypertrophy assessment with lateral cephalometry in children based on artificial intelligence," *Diagnostics*, vol. 11, no. 8, p. 1386, Jul. 2021.

[17] R. B. Snow, R. D. Zimmerman, S. E. Gandy, and M. D. F. Deck, "Comparison of magnetic resonance imaging and computed tomography in the evaluation of head injury," *Neurosurgery*, vol. 18, no. 1, pp. 45–52, Jan. 1986.

[18] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," 2018, *arXiv:1801.07372*.

[19] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*, Jun. 2011, pp. 1385–1392.

[20] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[21] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[22] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 483–499.

[23] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[24] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[25] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, and X. Gao, "An attention-guided deep regression model for landmark detection in cephalograms," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 540–548.

[26] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, "Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 873–881.

[27] Q. Liu, H. Deng, C. Lian, X. Chen, D. Xiao, L. Ma, X. Chen, T. Kuang, J. Gateno, and P. T. a. Yap, "Skullengine: A multi-stage (CNN) framework for collaborative CBCT image segmentation and landmark detection," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2021, pp. 606–614.

[28] B. Khanal, L. Dahal, P. Adhikari, and B. Khanal, "Automatic cobb angle detection using vertebra detector and vertebra corners regression," in *Proc. Int. Workshop Challenge Comput. Methods Clin. Appl. Spine Imag.*, 2019, pp. 81–87.

[29] J. M. H. Noothout, B. D. De Vos, J. M. Wolterink, E. M. Postma, P. A. M. Smeets, R. A. P. Takx, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning-based regression and classification for automatic landmark localization in medical images," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4011–4022, Dec. 2020.

[30] K. Oh, I.-S. Oh, V. N. T. Le, and D.-W. Lee, "Deep anatomical context feature learning for cephalometric landmark detection," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 806–817, Mar. 2021.

[31] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[32] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[34] P.-E. Danielsson, "Euclidean distance mapping," *Comput. Graph. image Process.*, vol. 14, no. 3, pp. 227–248, Nov. 1980.

[35] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," in *Readings in Computer Vision*. San Francisco, CA, USA: Morgan Kaufmann, 1987, pp. 638–643.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 15–29.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

ZILING HE received the B.Sc. degree from the Ocean University of China, Qingdao, China, in 2017. She is currently pursuing the M.Sc. degree with Jilin University. Her main research interest include deep learning, computer vision, and medical image processing.

YUBIN XIAO (Graduate Student Member, IEEE) received the M.Sc. degree in computer science from Xiangtan University, China, in 2021. He is currently pursuing the Ph.D. degree with Jilin University. His main research interests include neural combinatorial optimization, machine learning, and medical image processing.

XUAN WU (Graduate Student Member, IEEE) received the B.Sc. degree from Jilin University, Changchun, China, in 2020, where he is currently pursuing the Ph.D. degree. His current research interests include neural combinatorial optimization, swarm intelligence, neural architecture search, and medical image processing.

YANCHUN LIANG received the Ph.D. degree in applied mathematics from Jilin University, Changchun, China, in 1997. He was a Visiting Scholar with The University of Manchester, U.K., from 1990 to 1991, a Visiting Professor with the National University of Singapore, from 2000 to 2001, a Guest Professor with the Institute of High Performance Computing, Singapore, from 2002 to 2004, a Visiting Professor with the University of Trento, Italy, from 2006 to 2008, and a Visiting Professor with the University of Missouri, from 2011 to 2017. He is currently a Professor with the College of Computer Science and Technology, Jilin University, and the School of Computer Science, Zhuhai College of Science and Technology. He has published more than 400 papers. His research interests include computational intelligence, machine learning methods, text mining, and bioinformatics. He was a recipient of several grants from NSFC and EU. His research was featured in the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, the IEEE Transactions on Geoscience and Remote Sensing, Bioinformatics, the *Journal of Micromechanics and Microengineering*, *Physical Review E*, *Smart Materials and Structures*, *Applied Artificial Intelligence*, and *Medical Image Processing*.

YOU ZHOU received the B.Sc. and Ph.D. degrees from Jilin University, Changchun, China, in 2002 and 2008, respectively. He is currently a Professor with the College of Computer Science and Technology, Jilin University. He has published more than 70 journal and conference papers. His research interests include neural combinatorial optimization, pattern recognition, and bioinformatics.

GUANGHUI AN is currently a Doctor with the Shanghai Municipal Hospital Traditional Chinese Medicine. He specializes in the treatment of pediatric adenoid hypertrophy.

• • •