Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 15, No. 1, 2026

ECAAF-CenterNet: Efficient Channel Attention and Adaptive Fusion for Cephalometric Landmark Detection

Feng Xu [1], Guangping Zhuo [1, *], Guanghua Zhang [2], Xiuping Wu [3]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong Shanxi, 030619, China
[2] Department of Computer Science and Technology, Taiyuan College, Taiyuan Shanxi, 030032, China
[3] Stomatological Hospital, Shanxi Medical University, Taiyuan Shanxi, 030001, China

*** Corresponding author:** Guangping Zhuo

Abstract: In response to the challenges of complex local structures and significant differences in scale distribution in key point detection of lateral cranial films, this paper proposes an automatic key point detection method for lateral cranial films based on the improved CenterNet, ECAAF-CenterNet. Based on the CenterNet anchor-free detection framework, DLA34 is selected as the backbone network. Combined with the characteristics of dense distribution of key points and obvious scale differences in head images, the network structure is improved from the two aspects of feature enhancement and feature fusion. Introduce the Efficient Channel Attention Mechanism (ECA) in the backbone residual module to enhance the network's expression ability for key channel features. Meanwhile, in the multi-scale feature fusion stage, an adaptive weighted cross-scale feature fusion mechanism AF is designed to perform dynamic weighted fusion on features at different levels. Through experimental verification on the private dataset of 1603 lateral cranial films, the MRE reached 1.54mm, and the SDR could reach 94.51% at the 4.0mm threshold, which was superior to the baseline model.

Keywords: Lateral Cranial X-ray; Channel Attention; Adaptive Feature Fusion; Deep Learning.

1. Introduction

Dental X-rays play a significant role in the diagnosis and treatment of oral diseases. They can be used to detect hidden tooth structures, malignant or benign masses, bone loss, and dental caries, among other issues. They serve as an important basis for root canal treatment, dental caries diagnosis, and the formulation of treatment plans for orthodontic patients[1]. Cephalometric measurement technology helps doctors gain a deeper understanding of the structural features of patients' teeth and jaws as well as craniofacial bones by identifying specific anatomical key points on cranial images and measuring the distances and angles between these points.

In recent years, artificial intelligence technology based on deep learning has made breakthrough progress in the field of key point detection of lateral cranial radiography[3, 4, 17]. Especially technologies such as convolutional neural networks have demonstrated obvious advantages in the accuracy and stability of key point localization, gradually becoming a research hotspot [5]. Convolutional neural networks can automatically learn features from a large number of images and achieve end-to-end keypoint detection. For instance, Arik et al. [12] were the first to apply convolutional neural networks to fully automatic head shadow measurement, achieving detection results superior to traditional methods. Zeng et al. [7] adopted a cascading network structure. The average radial errors on two test datasets were 1.34mm and 1.64mm respectively, and the analysis time for a single image was only about 3 seconds. Lee et al. [8] proposed an automatic method for detecting cephalometric landmarks based on Bayesian convolutional neural networks. The method introduces uncertainty modeling while achieving landmark localization, providing confidence regions for each landmark, thereby improving the reliability of the detection results.

This study focuses on the task of automatic key point detection in lateral cranial images. This task has high requirements for model performance in terms of spatial positioning accuracy, multi-scale feature fusion, and key feature expression. In order to achieve stable and accurate key point localization while ensuring controllable model complexity, this paper selects the CenterNet framework as the basic detection framework [15] and adopts DLA34 as the backbone network [6]. CenterNet avoids anchor box design and complex post-processing by converting target detection into keypoint heat map prediction, while DLA34 can effectively integrate feature information of different scales through a hierarchical feature aggregation mechanism.

Nevertheless, there are still certain deficiencies in directly using the CenterNet-DLA34 network for cranial keypoint detection. On the one hand, the residual features in the traditional DLA network have relatively limited discrimination ability in the channel dimension, and some feature channels closely related to key point localization have not been fully emphasized. On the other hand, in the multi-scale feature fusion stage, the IDAUp module integrates features at different levels by means of equal-weight addition, without fully considering the differential contributions of shallow details and deep semantic information to key point detection.

In response to these problems, this paper focuses on the automatic detection task of key points in head shadow measurement. Based on the CenterNet-DLA34 network structure, improvement methods are proposed from two aspects, feature enhancement and feature fusion: (1) A lightweight and efficient channel attention mechanism (ECA) is incorporated into the residual modules of the backbone

network to enhance the perception of key channel features; (2) In the multi-scale feature fusion stage, design a cross-scale fusion strategy based on adaptive weights to achieve dynamic adjustment of feature contributions at different levels. Experiments prove that the above-mentioned method effectively improves the accuracy and robustness of key point detection without significantly increasing the complexity of the model.

2. Methods

2.1. CenterNet-DLA Network

CenterNet-DLA34 is a deep learning model for key point detection and target localization, consisting of the CenterNet anchor-free detection framework and the DLA34 backbone network. CenterNet models the object detection task as a problem of locating the target center point. By predicting the heat map of the target center point and regressing attributes such as the target size at the corresponding position, it avoids the complex anchor box design and post-processing procedures in traditional methods [15]. As the backbone network, DLA34 adopts a hierarchical feature aggregation structure to fuse the shallow spatial detail features with the deep semantic information, providing multi-scale feature representations for subsequent cephalometric landmark localization [6]. Its structure diagram is shown in Figure 1.

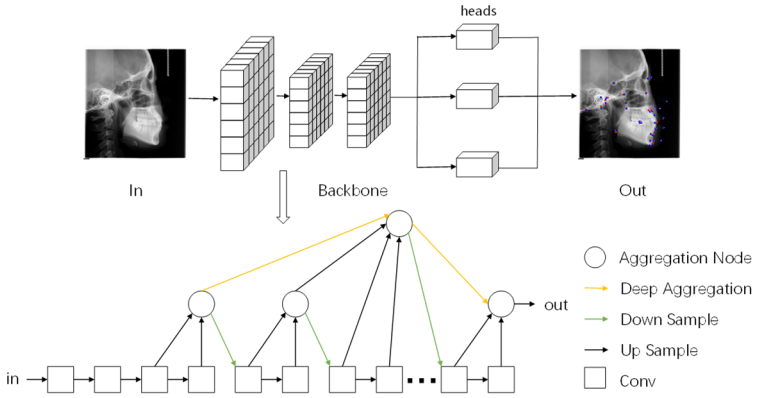

Fig 1. Network structure diagram of CenterNet-DLA34

2.2. Efficient Channel Attention Mechanism Module

ECA (Efficient Channel Attention) is a lightweight and efficient channel attention mechanism, which can enhance the feature representation ability of convolutional neural networks without significantly increasing the computational complexity. This method avoids the dimensionality reduction operation in the traditional channel attention mechanism, directly maintains the corresponding relationship of channel features, and achieves efficient channel modeling [16]. Its structure is shown in Figure 2.

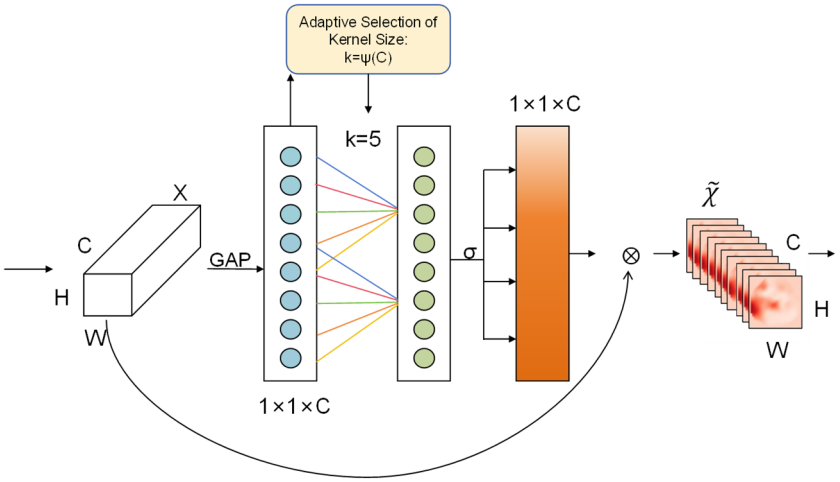

Fig 2. Schematic diagram of the ECA module

Specifically, the ECA module first performs the Global Average Pooling (GAP) operation on the input feature map, compressing the spatial information into channel-level global vectors with a size of $1 \times 1 \times C$, where C represents the number of channels. This process not only retains the channel information but also provides global information for the subsequent channel modeling. Then, ECA uses one-dimensional convolution in the channel dimension to model the local dependencies between adjacent channels and generates the corresponding channel weights through the Sigmoid function. The generated channel weights are further multiplied element-by-element with the original feature map to achieve adaptive adjustment of the response intensity of different channels, enhance the channel features related to key point localization, and suppress irrelevant redundant information.

In the ECA module, the kernel size of one-dimensional convolution is not fixed but adaptively determined based on the number of channels C, and its calculation method is as follows:

$$k = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \quad (1)$$

Here, k represents the kernel size of the one-dimensional convolution; C represents the number of channels; γ and b are hyperparameters, set to 2 and 1 respectively; "$|t|_{odd}$" indicates taking the odd number closest to "t".

2.3. Improvement of the CenterNet-DLA34 algorithm

2.3.1. Lightweight Channel Attention Residual Enhancement Mechanism

Aiming at the problem of insufficient discrimination ability of residual features in the Channel dimension in the traditional DLA network, this paper introduces the Efficient Channel Attention (ECA) mechanism in the basic residual module of the backbone network. This mechanism utilizes local cross-channel interactions to adaptively adjust feature responses. The network's ability to model key channel features can be enhanced without significantly increasing the number of parameters and computational complexity. By embedding the ECA attention mechanism into the residual module, the network can more effectively highlight information related to cephalometric landmark localization during feature extraction, thereby improving the model's ability to express detailed structures in complex head images. Its structure is shown in Figure 3.

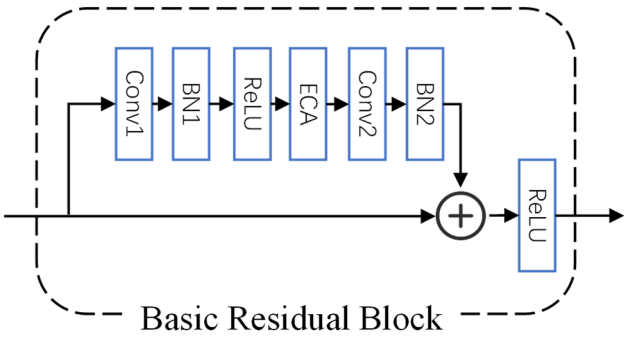

Fig 3. Diagram of the lightweight channel attention residual enhancement module

2.3.2. Adaptive Weighted Cross-scale Feature Fusion Mechanism

Aiming at the problem that the equal-weight Fusion of features at different levels in the IDAUp module of the traditional DLA34 network may lead to insufficient coupling of semantic information and positioning information, this paper designs an adaptive weighted cross-scale feature fusion mechanism (Adaptive Fusion, AF). This strategy achieves a dynamic balance of cross-scale information by learning the fusion ratio of features of different scales, enabling the model to adaptively adjust the contribution degree of shallow and deep features according to the feature hierarchy, thereby improving the overall accuracy and stability of cephalometric landmark localization. Its structure diagram is shown in Figure 4.

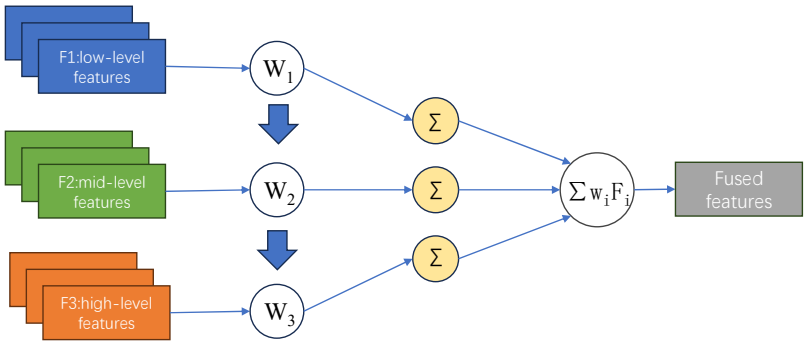

Fig 4. Diagram of the adaptive weighted cross-scale feature fusion module

F_1, F_2, F_3 respectively represent shallow features, middle features and deep features. Among them, the shallow features contain rich spatial detail information; Middle-level features possess both structural and semantic information. Deep features contain rich category, context and global semantic information, but have relatively low positional accuracy. Before fusion, the deep features will restore resolution through upsampling and adjust the channel dimension using convolution, so that the features of each layer can be fused at the same scale.

W_1, W_2, W_3 are learnable scale weight parameters used to measure the contribution degree of features of different scales in the fusion process. The features of each scale are first weighted:

$$\hat{F}_i = \omega_i \cdot F_i \quad (2)$$

Subsequently, the weighted features are summed to obtain the final fused features:

$$F_{fusion} = \sum_{i=1}^{3} \omega_i \cdot F_i \quad (3)$$

The fusion features are used as the input for the subsequent key point heat map prediction and regression tasks.

3. Experiment

3.1. Data Collection

The data is derived from a private dataset of 1603 lateral cranial films of a certain stomatological hospital.

3.2. Data Annotation

The accurate calibration of key points in head shadow measurement is of vital importance for doctors to formulate treatment plans. The ideal key points are generally selected as anatomical key points that are relatively stable and easy to locate during the growth and development process. However, not all commonly used key points can meet this requirement. The determination of many key points is based on different measurement methods proposed by various scholars [2]. This study investigated 36 key points of soft and hard tissues calibrated by doctors in a certain dental hospital on lateral head images, and the annotations were reviewed by two experienced doctors. The final Ground Truth is the average of the positions marked by the two doctors. As shown in Fig. 5.

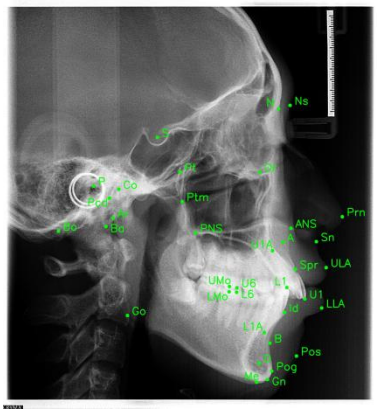

Fig 5. Diagram of 36 key points

These 36 key points are respectively: S (Sella),N (Nasion),P (Porion),Ba (Basion),Bo (Bolton Point),Or (Orbitale), Ptm (Pterygomaxillary fissure point),Pt (Pterygoid point),PNS (Posterior Nasal Spine),ANS (Anterior Nasal Spine),A (Point A / Subspinale),Spr (Superior alveolar ridge point),U1 (Upper central incisor point),U1A (Upper central incisor apex),L1 (Lower central incisor point),L1A (Lower central incisor apex),Id (Infradentale),B (Point B / Supramentale), Pog (Pogonion),Me (Menton),Gn (Gnathion),D (Symphysis center point),Go (Gonion),Ar (Articulare), Pcd (Posterior condylion),Co (Condylion),U6 (Upper first molar mesiobuccal cusp),L6 (Lower first molar mesiobuccal cusp),UMo (Upper molar point),LMo (Lower molar point), Ns (Soft tissue nasion),Prn (Pronasale),Sn (Subnasale), ULA (Upper lip anterior point),LLA (Lower lip anterior point),Pos (Soft tissue pogonion).

3.3. Experimental Environment and Training Settings

This experiment is implemented based on the PyTorch1.11.0 framework and Python3.8. During the training process, 8 NVIDIA GeForce RTX 3080 Ti Gpus are used for parallel acceleration. The resolution of the input image is set to 1024×1024 pixels. The model training involves a total of 200 training rounds, with a batch size of 16. The initial learning rate is 0.000125, and the cosine annealing strategy is used to dynamically adjust the learning rate in the 100th training round.

3.4. Evaluation Indicators

The performance evaluation in this field mainly adopts the mean radial error (MRE) and the success detection rate (SDR) within a specific accuracy range.

3.4.1. Mean Radial Error (MRE)

MRE measures the mean of the physical distance error between the predicted point and the true point. The calculation formula is as follows:

$$\mathrm{MRE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sqrt{(\mathrm{X_i} - \bar{\mathrm{X}}_\mathrm{i})^2 + (\mathrm{Y_i} - \bar{\mathrm{Y}}_\mathrm{i})^2} \qquad (4)$$

Here,$(\mathrm{X_i},\mathrm{Y_i})$ represents the true coordinates,$(\bar{\mathrm{X}}_\mathrm{i},\bar{\mathrm{Y}}_\mathrm{i})$ is the predicted coordinates, and N is the sample size.

3.4.2. Success Detection Rate (SDR)

The significance of SDR lies in how many target key points can be successfully located when considering a certain error range. The calculation formula is as follows:

$$P_b = \frac{\#\{j : || L_d(j) - L_r(j) || < b\}}{M} \times 100\% \qquad (5)$$

Among them, L_d and L_r are respectively the coordinates of the detected key points and their corresponding ground-truth coordinates, M is the number of all detected key points, $\#\{\Omega\}$ indicates the number of detected key points that meet the condition Ω, and the error range is usually 2mm, 2.5mm, 3mm, and 4mm.

3.5. Ablation Experiment

To verify the effectiveness of the model proposed in this paper in the task of head keypoint detection and analyze the specific impact of each component module on the model performance, an ablation experiment was designed in this paper under the same dataset division and training Settings.

This paper adopts the transfer learning strategy to train the model. All experiments were loaded with the same pre-trained model as the initial weights and fine-tuned under the same training configuration. CenterNet-DLA34 serves as the Baseline model. Baseline does not introduce any structural improvements on this basis, while the other models merely add corresponding modules to Baseline to verify the effectiveness of each improvement strategy. The experimental results are shown in Table 1.

Table 1. Ablation Experiment Data Table

Model	MRE (mm)	SDR (%)			
		2.0mm	2.5mm	3.0mm	4.0mm
CenterNet_DLA34	1.57	73.11	81.76	87.56	94.15
CenterNet_DLA34+ECA	1.57	73.09	81.82	87.71	94.42
CenterNet_DLA34+AF	1.56	73.67	82.84	88.23	94.47
CenterNet_DLA34+ECA+AF	**1.54**	**73.96**	**82.63**	**88.35**	94.**51**

It can be seen from Table 1 that after integrating the Efficient Channel Attention Mechanism (ECA) separately on the basis of the baseline model, while maintaining the stability of MRE, the SDR of the model has slightly improved under the thresholds of 2.5mm, 3.0mm, and 4.0mm. This indicates that ECA enhances the discriminability of the feature map through adaptive channel recalibration.

After introducing the adaptive weighted cross-scale feature fusion strategy AF separately based on the baseline model, the model achieved stable performance improvements in both MRE and SDR metrics, with MRE reduced to 1.56mm. It is indicated that this strategy helps alleviate the problem of uneven contribution of features of different scales during the fusion process, enabling the shallow spatial detail information and deep semantic information to be integrated more reasonably, thereby improving the positioning accuracy of the model.

When the ECA attention mechanism and the AF adaptive weighted cross-scale feature fusion strategy are simultaneously introduced, the model achieves the optimal comprehensive results under MRE and most SDR thresholds. Although there is a slight fluctuation in the 2.5mm metric of SDR compared to only introducing the AF adaptive fusion strategy, the overall performance is superior to that of the single-module and baseline models. This indicates that the two are not simply superimposed but have formed an effective collaborative optimization mechanism: The ECA module highlights the features related to key point localization in the channel dimension through adaptive weighting; The AF module dynamically weighted fuses shallow, middle and deep features in the scale dimension, effectively balancing the contribution of spatial details and semantic information. The two work together to achieve the unified optimization of local key features and global scale

information, thereby enhancing the model's overall performance in locating key points of the skull.

3.6. Comparative Experiment

Table 2. Comparison Test Table with Other Methods

Method	SDR (%)			
	2.0mm	2.5mm	3.0mm	4.0mm
SCN Payer et al.[13]	73.33	78.76	83.24	89.75
Localization U-Net[13][14]	72.15	77.83	82.04	88.80
Arik et al.[12]	72.29	78.21	82.24	86.80
Urschler et al. [11]	70.21	76.95	82.08	89.01
Lindner and Cootes [10]	70.65	76.93	82.17	89.85
Ibragimov et al. [9]	68.13	74.63	79.77	86.87
Ours	**73.96**	**82.63**	**88.35**	**94.51**

Most of the traditional methods are based on the ISBI 2015 challenge dataset, and the number of key points is 19. The dataset of this study labeled 36 key points for head shadow measurement. On the basis of covering all the key points of ISBI 2015, more key points with clinical significance were introduced, making the detection task have a higher scale in terms of quantity and structural complexity. Given the differences in datasets and annotation specifications among various methods, the relevant results are mainly used to reflect the overall performance level of the method proposed in this paper in complex key point detection scenarios. Despite this, the comparison of experimental results in Table 2 still reflects the key point positioning advantage of the model proposed in this paper in the key point detection accuracy of head shadow measurement.

4. Conclusion

The key point detection method ECAAF-CenterNet based on the improved CenterNet proposed in this paper achieves the collaborative optimization of channel features and scale information by introducing a lightweight ECA channel attention mechanism in the residual module and adopting an adaptive weighting strategy in the multi-scale feature fusion stage. It has effectively improved the accuracy of key point detection. However, this study still has limitations, such as the limited scale of the dataset and the influence of image noise and annotation bias on some key points. Future work will focus on expanding the dataset, multimodal image fusion, and optimizing the lightweight network structure to enhance the applicability and stability of the model in clinical environments.

Acknowledgments

Project of Shanxi Provincial Clinical Research Center for Oral Diseases (Maxillofacial Deformities) (Project No.: LC ZX0004)

References

[1] Wang C W, Huang C T, Lee J H, et al. A benchmark for comparison of dental radiography analysis algorithms[J]. Medical image analysis, 2016, 31: 63-76.

[2] Hlongwa P. Cephalometric analysis: manual tracing of a lateral cephalogram[J]. South African Dental Journal, 2019, 74(6): 318-322.

[3] Hwang H W, Moon J H, Kim M G, et al. Evaluation of automated cephalometric analysis based on the latest deep learning method[J]. The Angle Orthodontist, 2021, 91(3): 329-335.

[4] Nishimoto S, Sotsuka Y, Kawai K, et al. Personal computer-based cephalometric landmark detection with deep learning, using cephalograms on the internet[J]. Journal of Craniofacial Surgery, 2019, 30(1): 91-95.

[5] Hwang H W, Park J H, Moon J H, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? [J]. The Angle Orthodontist, 2020, 90(1): 69-76.

[6] Yu F, Wang D, Shelhamer E, et al. Deep layer aggregation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2403-2412.

[7] Zeng M, Yan Z, Liu S, et al. Cascaded convolutional networks for automatic cephalometric landmark detection[J]. Medical Image Analysis, 2021, 68: 101904.

[8] Lee J H, Yu H J, Kim M, et al. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks[J]. BMC oral health, 2020, 20(1): 270.

[9] Ibragimov B, Likar B, Pernus F, et al. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement [C]//Proceedings of International Symposium on Biomedical imaging (ISBI). 2015.

[10] Lindner C, Bromiley P A, Ionita M C, et al. Robust and accurate shape model matching using random forest regression-voting[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(9): 1862-1874.

[11] Urschler M, Ebner T, Štern D. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization[J]. Medical image analysis, 2018, 43: 23-36.

[12] Arık S Ö, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks[J]. Journal of Medical Imaging, 2017, 4(1): 014501-014501.

[13] Payer C, Štern D, Bischof H, et al. Integrating spatial configuration into heatmap regression based CNNs for landmark localization[J]. Medical image analysis, 2019, 54: 207-219.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234-241.

[15] ZHOU XY, WANG DQ, KRÄHENBÜHL P. Objects as Points [EB/OL]. (2019-04-25) [2026-01-15]. arXiv:1904. 07850. https: // doi.org/10.48550/arXiv.1904.07850.

[16] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[17] Kunz F, Stellzig-Eisenhauer A, Zeman F, et al. Künstliche Intelligenz in der Kieferorthopädie: Evaluierung einer vollständig automatisierten Fernröntgenseitenanalyse unter Anwendung eines individualisierten „convolutional neural network “[J]. Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie, 2020, 81: 52-68.