CMF-Net: craniomaxillofacial landmark localization on CBCT images using geometric constraint and transformer

Gang Lu, Huazhong Shu, Han Bao, Youyong Kong, Chen Zhang, Bin Yan, Yuanxiu Zhang, Jean-Louis Coatrieux

IOP Publishing Journal Title
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

CMF-Net: Craniomaxillofacial Landmark Localization on CBCT Images Using Geometric Constraint and Transformer

Gang Lu[1,2,3], Huazhong Shu[1,2,3], Han Bao[4,5,6], Youyong Kong[1,2,3], Chen Zhang[1,2,3], Bin Yan[4,5,6], Yuanxiu Zhang[1,2,3,4,5,6] and Jean-Louis Coatrieux[2,7,8]

[1] Laboratory of Image Science and Technology, Southeast University, Nanjing 210096, China
[2] Centre de Recherche en Information Biomédicale Sino-Français, 35000 Rennes, France
[3] Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China
[4] Department of Orthodontics, the Affiliated Stomatological Hospital of Nanjing Medical University, Nanjing 210029, China
[5] Jiangsu Province Key Laboratory of Oral Diseases, Nanjing Medical University, Nanjing 210029, China
[6] Jiangsu Province Engineering Research Center of Stomatological Translational Medicine, Nanjing 210029, China
[7] National Institute for Health and Medical Research, F-35000 Rennes, France
[8] Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, F-35000 Rennes, France
E-mail: shu.list@seu.edu.cn and byan@njmu.edu.cn

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Accurate and robust anatomical landmark localization is a mandatory and crucial step in deformation diagnosis and treatment planning for patients with craniomaxillofacial (CMF) malformations. In this paper, we propose a trainable end-to-end cephalometric landmark localization framework on CBCT scans, referred to as CMF-Net, which combines the appearance with transformers, geometric constraint, and adaptive wing (AWing) loss. More precisely: 1) We decompose the localization task into two branches: the appearance branch integrates transformers for identifying the exact positions of candidates, while the geometric constraint branch at low resolution allows the implicit spatial relationships to be effectively learned on the reduced training data. 2) We use the AWing loss to leverage the difference between the pixel values of the target heatmaps and the automatic prediction heatmaps. We verify our CMF-Net by identifying the 24 most relevant clinical landmarks on 150 dental CBCT scans with complicated scenarios collected from real-world clinics. Comprehensive experiments show that it performs better than the state-of-the-art deep learning methods, with an average localization error of 1.108 mm (the clinically acceptable precision range being 1.5 mm) and a correct landmark detection rate equal to 79.28%. Our CMF-Net is time-efficient and able to locate skull landmarks with high accuracy and significant robustness. This approach could be applied in 3D cephalometric measurement, analysis, and surgical planning.

Keywords: 3D cephalometric analysis, CBCT, craniomaxillofacial, landmark localization

1. Introduction

Congenital and acquired deformities are the major causes of craniomaxillofacial (CMF) malformations. Cone-beam computed tomography (CBCT) is often used in treatment due to its low radiation dosage compared to spiral multi-slice computed tomography. In addition, it has short imaging time and low examination cost. The goal of landmark localization is to accurately detect the location of each predefined meaning key point on the bonny boundary annotated by an orthodontist, which can assist clinicians to determine the degree of the deformity and make further a surgical plan more precise.

Presently, most of the advanced automatic or semi-automatic locating methods focus on 2D lateral cephalograms. Linear and angular measurements are commonly performed to evaluate the relationships among teeth, facial skeleton, and soft tissue profile. However, the 2D radiographs are not fully satisfying to analyze the extent of the CMF deformities (Troulis *et al.*, 2002) since 34% of patients with dentofacial varieties have asymmetric conditions (Severt and Proffit, 1997). This motivates us to make more precise cephalometric measurements and analyses on volumetric CBCT images, especially for patients with CMF malformations.

Whereas, going from 2D to 3D is not so trivial. Manual landmark annotation is labor-intensive and requires domain-specific expertise. In addition, some significant discrepancies in annotations are exhibited among experts. Therefore, developing a fully automatic and highly accurate localization system that can robustly identify cephalometric landmarks could help to circumvent the aforementioned shortcomings.

A considerable number of works have been devoted to automatic landmark detection fulfilling the clinically acceptable precision requirements (Zhang *et al.*, 2015; Zhang *et al.*, 2020b; Lang *et al.*, 2020; Lian *et al.*, 2020; Chen *et al.*, 2022; Torosdagli *et al.*, 2019a; Chen *et al.*, 2021b). Nevertheless, an efficient, accurate, robust, and fast automatic localization method is still expected and the reasons for such a situation can be summarized as follows: 1) *Severe aliasing artifacts*. Braces, metal alloy implants, and dental fillings will introduce severe streaking and shading aliasing artifacts, which often reduce the contrast of bone boundaries. As shown in Fig. 1(a), metal artifacts appear in coronal and sagittal views due to the presence of amalgam dental fillings. 2) *Large morphological variations*. Anatomical structures are diverse across individuals, especially for patients who are subject to congenital or acquired CMF deformities. Therefore, some singular anatomical structures are seldom to be covered due to the inherently limited datasets leading to a poor generalization potential of the model. For example, mandibular retrognathia and maxillary prognathia as exhibited in Fig. 1(b) are categorized into skeletal class II malocclusions. While in Fig. 1(c), the subject shows an apparent maxillary retrognathia that refers to skeletal class III malocclusion. Irregularities from metal alloy artifacts or underlying diseases are marked with red rectangle boxes. It is thus extremely hard to detect the landmarks of incisors accurately and robustly from maxilla

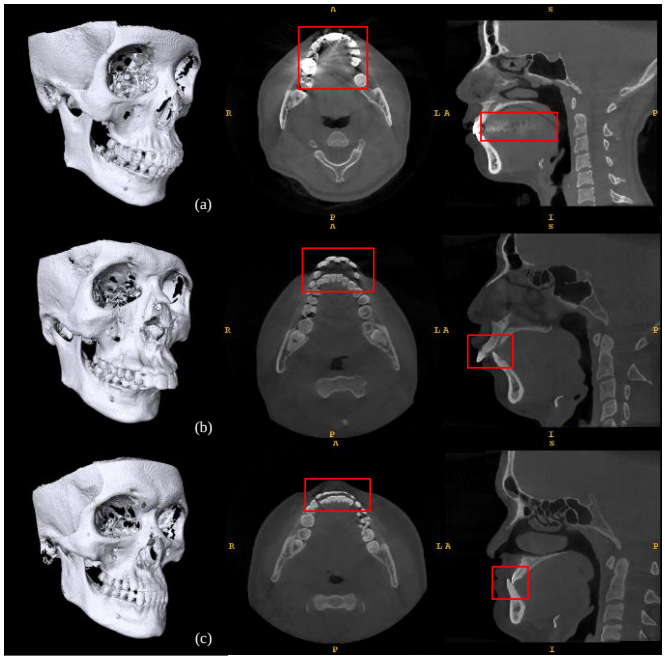


Figure 1. Three challenging cases were caused by metal artifacts and complex pathologies as labeled in the red rectangle boxes from randomly selected subjects in CBCT scans. The left column shows the 3D volumetric renderings of the skulls, the middle and right columns display coronal and sagittal views for the corresponding subject, respectively. (a) Severe streaking and shading metal artifacts arise from dental alloy fillings. (b) The subject is classified as skeletal Class II. (c) The subject belongs to skeletal Class III.

and mandible with serious anatomical abnormities. 3) *Local similarities*. Landmarks are hard to be recognized solely on local patterns because of high local similarities. Their detections and locations can only be obtained by exploiting global context information within the CBCT scans. 4) *Large volume data*. The majority of our collected CBCT volumes have a size of 610×610×610. Using such data as inputs to any deep learning (DL) framework would be computationally expensive and require large memory resources. 5) *Imaging issues*. The problems related to the imaging devices such as beam hardening, inhomogeneity, truncation effects, noise, and low tissue contrast increase the challenge of detecting anatomical landmarks in a fully automatic way.

Among the recent contributions on the localization task, most efforts concentrate on using U-Net or its variants (Payer *et al.*, 2019; Chen *et al.*, 2022; Zhang *et al.*, 2020b; Lang *et al.*, 2020; Payer *et al.*, 2016; Torosdagli *et al.*, 2019a; Lian *et al.*, 2020). However, the intrinsic locality of convolution operators leads to a limited receptive field and deteriorates the capability of the model to correctly regress the activations of the output heatmaps. As shown later, the combination of the local appearance branch with long-range dependency based on transformer (Carion *et al.*, 2020) allows capturing global context information and provides an enhanced accuracy in 3D anatomical landmark localization.

Detecting the anatomical landmarks using only local neighborhoods but regardless of geometric information will lead to numerous false positives. Methods like point distribution model (PDM) (Lindner *et al.*, 2015; Li *et al.*, 2018), Markov random field (MRF) (Glocker *et al.*, 2012;

Donner *et al.*, 2013; Donner *et al.*, 2010) and handcrafted energy function (Chen *et al.*, 2015; Wang *et al.*, 2021) have been reported to deal with this problem. They all decompose the identification task into successive processing steps including detection, feature extraction, filtering out false-positive landmarks and so on, with sometimes additional and time-consuming user selection. Here, the implicit prior knowledge is embedded within the geometric constraint component which enables the network to effectively learn the spatial hidden distribution of landmarks and eliminates the false-positive candidates generated in the appearance component.

Heatmap-based regression methods have been widely employed in localization tasks. In practice, it is essential to accurately estimate the values of foreground pixels compared to the ones on background, and adaptive wing (AWing) loss (Wang *et al.*, 2019) has obtained excellent results and outperforms other loss functions like mean square error (MSE), L1 norm, wing loss (Feng *et al.*, 2018) in facial landmark localization benchmarks. Therefore, AWing loss is utilized to predict against pseudo-probable Gaussian heatmaps generated at each annotated landmark.

In this paper, we propose an accurate and efficient CMF landmark localization method for CBCT scans based on 3D convolutional neural networks with heatmap regression. In summary, our main contributions include:

1) Our CMF-Net consists of two branches, the appearance branch aims to identify predefined landmarks with high accuracy by integrating transformer modules, and the geometric constraint branch focuses on eliminating false-positive candidates. Therefore, the prior knowledge is embedded into the geometric constraint that allows the model to effectively learn the implicit spatial relationships between the landmarks from a limited number of CBCT scans.

2) AWing loss is used to penalize the difference between the ground truth and the prediction, which enables the detection framework to pay much attention on pixel values near the mode of the regressed volumetric heatmap and results in lower localization error.

3) Experimental results conducted on two test datasets have shown that our model can locate skull landmarks with high accuracy and significant robustness. It reveals that the approach could assist clinicians in analyzing 3D cephalometric CBCT data and this way in planning surgical correction, especially for patients who suffer from severe CMF malformations.

The rest of the paper is organized as follows. Section 2 briefly introduces related work. In Section 3, we present the architecture of the proposed CMF-Net, the objective function, and the coordinate decoding during inference. Then, evaluation dataset, implementation details, and evaluation criteria are described in Section 4. Section 5 compares our approach with other advanced landmark localization methods. The effectiveness of various components has been studied, and the landmarking performance against the different amounts of training data is analyzed. Some conclusions and perspectives are drawn in Section 6.

2. Related work

2.1 Craniomaxillofacial landmark localization

CMF landmark localization is a fundamental examination in deformation analysis and surgical planning. A significant number of efforts have been made on this issue using 3D dental CBCT scans. Registration-based approaches usually transfer landmarks from template images to test images via registration techniques (Codari *et al.*, 2017; Shahidi *et al.*, 2014). Although both affine and non-rigid transforms are used for image registration, perfect alignment between the reference and moving images in the same spatial coordinate system remains difficult due to large variations in anatomy and pathological patterns. In general, segmentation is performed before registration and its outcomes determine the performance of the overall process. Besides, the registration procedure may have a high computational complexity which makes challenging for real-time clinical applications. The knowledge-based approach proposed in (Gupta *et al.*, 2015) involves seed point selection and contour detection, two sensitive steps required to be handled due to complex anatomical structures, and potential missing structures like teeth and image artifacts. The interest point detection method (Donner *et al.*, 2010) often limits to identify the anatomical landmarks with distinctive geometry (*e.g.*, salient corners or boundaries).

While conventional methods mainly rely on intensity information, machine leaning-based approaches can capture local context information and overall shape information. They can be roughly divided into two categories: classification and regression-based methods. Classification-based methods (Cheng *et al.*, 2011; Criminisi *et al.*, 2011) use classifiers to discriminate the positive patch surrounding certain landmark's position from the negative ones and often lead to poor localization performance if complex and similar anatomical structures exist across subjects. In contrast to classification-based methods, the ultimate goal of regression models is to estimate offsets from any voxel to the target landmark (Lindner *et al.*, 2015; Criminisi *et al.*, 2013; Ebner *et al.*, 2014; Gao and Shen, 2015; Urschler *et al.*, 2018; Zhang *et al.*, 2015). As shown in (Ebner *et al.*, 2014), the method (Criminisi *et al.*, 2013) using multivariate regression forests with context-rich visual features for detecting the bounding boxes of multiple anatomical structures can be modified to identify multiple landmarks as well. Nevertheless, the local appearances of faraway voxels being less informative, a spherical sampling strategy (Gao and Shen, 2015) was used to draw training voxels within a certain distance of a voxel to the target, which achieved higher localization accuracy on dental landmarks. Urschler *et al.* (2018) integrated the geometric configuration and image appearance into a unified random forest scheme, then they iteratively refined the multiple landmarks using a coordinate descent optimization scheme. In (Zhang *et al.*, 2015), before aggregating evidence for each landmark localization, an extra prior step as bone segmentation must be carried out. Although promising results

have been achieved, the limited capability of handcrafted feature extraction results in a sub-optimal outcome.

DL methods are now widely applied in all image processing tasks and whatever the application target. Heatmap regression methods become mainstream for accurate localization due to their image-to-image dense prediction and visual intuition. Zhang *et al.* (2017; 2020b) explored a multi-task-oriented approach using context-guided fully convolutional networks for concurrent bone segmentation and landmark digitalization from CBCT data. In the first stage, they introduced three volumetric displacements at *x*, *y*, and *z*-axes to bridge context information for every landmark, which is extensively parameter-heavy and computationally inefficient. Zhong *et al.* (2019) adopted a two-stage U-Net for cephalometric landmark localization, where the global stage identifies the coarse locations further refined in position through individual local stages. Lang *et al.* (2020) reported a cascaded network built on a U-Net for predicting CMF landmarks and a graph network for deciding whether a given landmark exists. Lian *et al.* (2020) presented a multi-task dynamic transformer network that learns task-oriented feature embedding in a "learning-to-learn" fashion for joint mandible segmentation and localization on dental CBCT data. Palazzo *et al.* (2021) designed a coarse-to-fine three-stage localization architecture for sequentially processing a CT scan. Instead of heatmap regression, coordinate regression methods (Gilmour and Ray, 2020; Li *et al.*, 2020; Zeng *et al.*, 2021) are rarely reported because of their high-dimensional nonlinear mapping from the input space to the coordinate representation which may introduce visual ambiguity. Due to the curse of dimensionality in CT scans, Lee *et al.*, (2019) first generated 2D shadowed images from 3D skull surface data, then VGG-19 (Simonyan and Zisserman, 2015) serves as a regressor for obtaining a set of 2D coordinates. Zeng *et al.* (2021) regressed the 2D coordinates through three sequential stages, where the last stage operates over a local image patch with high resolution to enhance the localization precision. Chen *et al.* (2022) progressively refined the relative displacements of landmarks from the cropping high-resolution patches with a structure-aware long short-term memory (SA-LSTM) network and obtained encouraging results on 3D cephalograms. Instead of using the Euclidean metric, Torosdagli *et al.* (2019a) argued that regression based on the bone manifold is more reliable for landmarking. They segmented the mandible bones and then generated the learning-based geodesic map to detect the landmarks with an LSTM. Nevertheless, their approach strongly relies on previous segmentation results and an underperformed outcome at this level would inevitably lead to inferior localization performance. To get rid of segmenting the regions of interest, a relational reasoning network is designed to infer other landmarks given locations of several representative ones (Torosdagli *et al.*, 2019b). Although excellent localization results are obtained, the initial landmarks must be deliberately selected which limits the application in the real clinical setting.

Recently, in (Chen *et al.*, 2021b; Lang *et al.*, 2022), they formulated landmark localization as an object detection problem. It is reported in (Chen *et al.*, 2021b) that a Faster R-CNN (Ren *et al.*, 2015) can be used to estimate the landmark centers. The hybrid method proposed by (Noothout *et al.*, 2020) makes use of a global-to-local localization approach to locate the landmarks by performing coordinate regression and patch-based classification simultaneously. The major limitation imposed by such a strategy of cascaded coarse-to-fine is that it increases the computational complexity of the localization task. Building upon the work of (Alansary *et al.*, 2019; Ghesu *et al.*, 2019), reinforcement learning methods have been introduced for CMF landmark detection (Kang *et al.*, 2021), but they require a large number of training data and have an expensive computational cost. Additionally, only a few landmarks can be detected in general because each agent is solely responsible for identifying the distinct and optimal trajectory to the specific meaning point.

2.2 Transformer

The attention mechanism has since a few years a great impact on computer vision (Vaswani *et al.*, 2017; Veličković *et al.*, 2018). Transformer (Carion *et al.*, 2020) introduced a shift in object detection by formulating it as a direct set prediction problem that gets rid of proposal, anchor, and non-maximal suppression. The recent attempts in segmentation (Chen *et al.*, 2021a) and recognition (Dosovitskiy *et al.*, 2021) have demonstrated its outstanding performance over other CNN techniques. Here, we integrate transformers into the appearance component, which allows the network to capture global context and long-range dependency by using self-attention mechanisms. To the best of our knowledge, we are the first to explore its potential for automatic landmark localization on medical images.

2.3 Geometric constraint

Despite of complex deformities or potential pathologies, the predefined fiducial points on a 3D hard skull are relatively stable and constrained by other landmarks when compared with ones on soft tissue. General localization approach exploits image appearance information with handcrafted graphical models encoding the geometric landmark configuration. Once the geometrical structure constraints serving as the prior knowledge are properly applied, these landmarks would be precisely identified from limited training data. Early methods usually identify the exact locations of landmarks through local appearance features that bring false positives due to similar anatomical structures, then a PDM (Lindner *et al.*, 2015; Li *et al.*, 2018) and graphical MRF (Glocker *et al.*, 2012; Donner *et al.*, 2013; Donner *et al.*, 2010) are applied to disambiguate the candidate predictions and to provide a better accuracy.

However, the exact positions are estimated by employing an extra postprocessing step based on a parametric or a graphic model. Payer *et al.* (2016) implicitly encoded structural knowledge in a single-stage CNN. To mitigate the need for large amounts of medical data, their later contribution (Payer *et al.*, 2019) decomposed the localization problem into local

appearance and spatial configuration components, providing thus highly accurate results by training the network in an end-to-end manner. Inspired by their work, we integrate the appearance and the geometric constraint branches into a trainable end-to-end framework to detect CMF landmarks accurately and robustly, even in the presence of anatomical malformations, dental aliasing artifacts and low contrasts.

3. Method

The overall architecture of our CMF-Net is illustrated in Fig. 2. It consists of two branches, the appearance branch aims to identify predefined landmarks with high accuracy by integrating two transformer modules, while the geometric constraint branch focuses on eliminating false-positive candidates to locate robustly. In the following, they are fused by the Hadamard product (element-wise product) operator to produce the final regressed volumetric heatmaps capturing both global and local context information. The responses at each heatmap are activated only if a landmark is identified by appearance feature and lies in a feasible region according to implicit structural constraints between these landmarks. Furthermore, we use the AWing loss to estimate the responses on the regressed volumetric Gaussian heatmaps, permitting the network to effectively detect the target landmarks being located. Finally, multiple skull landmarks in sub-pixel are all exactly determined from the regressed heatmaps with low resolution by using a quadratic curve fitting method during inference.

3.1 CMF-Net

3.1.1 Appearance branch with transformers

Although now classical CNN-based techniques have shown impressive performance in computer vision, convolution operations cannot fully perceive large receptive fields and suffer from limitations in our localization task. To surmount the drawbacks of capturing global context and long-range dependency, we integrate two transformers into a multi-level U-shape structure to accurately estimate the location of each landmark as shown in Fig. 2.

A given CBCT image being fed into the appearance branch, K levels are used to generate the CNN feature maps $\{\boldsymbol{f}_k\}_1^K \in \mathbb{R}^{C\times D_k\times H_k\times W_k}$ in the analysis path, where D_k, H_k, W_k denote depth, height, and width of the kth level, respectively. Considering the computational complexity of 3D volume, we retain the number of channels C unchanged in the multi-level structure and set its value to 96. All levels consist of two consecutive convolution blocks except for the highest level, each of them including a 3×3×3 convolution, a leaky ReLU (LReLU) with a slope of -0.2 (Maas *et al.*, 2013), and a dropout rate of 0.25. A 2×2×2 average pooling with stride 2 is applied to half the resolution at the last convolution layer in the analysis path to form the input to the next higher level. In the synthesis path, the high-level features are up-sampled with a 3D linear interpolation to recover the resolution quadratically. They are concatenated to the low-level outputs

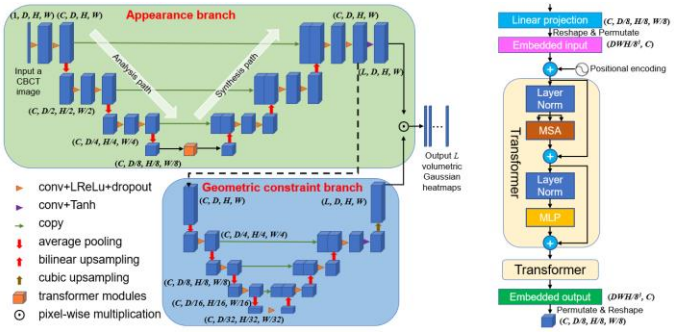


Figure 2. An overview of the CMF-Net with the schematic diagram of transformer modules is shown on the right side.

in the synthesis path and formulated as higher-resolution outputs that are equivalent to skip connection. To speed up the process of training and to avoid the issue of vanishing/exploding gradients, we rely on the idea exposed in (He *et al.*, 2016). Finally, a 1×1×1 convolution with a Tanh activation function restricts the output values of the appearance branch with L channels in the range (-1, 1) following the two convolution layers.

DL on 3D will introduce a much larger input vector when compared to 2D, which is prohibitively expensive for graphics hardware. Moreover, deploying transformer architectures over all levels in multi-level structures will exacerbate this problem. To mitigate the computational cost and memory requirements, we add two transformers onto the bottom in the appearance branch to build long-range dependency and exploit global context information. The highest level representative feature maps $\boldsymbol{f}_K$ are transformed to embedded space vectors $\mathbf{z}_0 \in \mathbb{R}^{D_K H_K W_K \times C}$ using a learnable linear projection with a convolution kernel size of 1×1×1 and C output channels, where the values of D_K, H_K and W_K are changed to $D/8$, $H/8$ and $W/8$ after three down-sampling operators, respectively. Moreover, we preserve the spatial information by adding a trainable position embedding initialized with zeros. The architecture of the transformer is shown on the right side of Fig. 2. The output of the mth transformer is expressed by (Chen *et al.*, 2021a)

$$\begin{cases} \mathbf{z}'_m = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{m-1})) + \mathbf{z}_{m-1} \\ \mathbf{z}_m = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_m)) + \mathbf{z}'_m \end{cases} \quad (1)$$

where $\mathrm{MSA}(\cdot)$, $\mathrm{MLP}(\cdot)$ and $\mathrm{LN}(\cdot)$ represent the multiple self-attention, multi-layer normalization and layer normalization operators, respectively. Finally, the embedded output of the transformer is permuted and reshaped with the same spatial resolution as $\boldsymbol{f}_K$ and then passes through the synthesis path to produce the appearance outputs. We experimentally set the number of attention heads to 8 and transformers to 2, respectively.

3.1.2 Geometric constraint branch

Even if the appearance branch can detect these landmarks with high accuracy, it less considers the spatial relationships between them and may thus lead to false positives. Furthermore, the human skull structure has rather stable

geometric characteristics and so the anatomical landmarks placed on the surface of it are anatomically-constrained, despite of complex pathologies or severe deformities. The embedded prior knowledge can ensure these landmarks are in feasible regions and facilitate the network to learn the spatial relations between them, thus reducing the requirement for a large amount of training data.

In our work, the implicit spatial relationships between the landmarks are effectively learned by using a geometric constraint branch which is formulated with a U-shape network at low resolution. We process the down-sampled feature maps of the appearance branch by a factor of 4 as the input of the geometric constraint branch, which ultimately avoids landmark misidentification and makes the localization performance to be further improved.

In (Payer *et al.*, 2019), both appearance and spatial configuration components are combined to improve the localization results, the relationships between the landmarks have not been adequately explored due to the limited capability of consecutive convolution layers with large kernel sizes in their spatial configuration branch. This is done here with a specific multiscale U-shape network at low resolution where each level involves a single convolutional layer of 3×3×3 kernel size with an LReLU. Then, they convoluted with a 1×1×1 kernel to produce the geometric constraint outputs for modeling the spatial relationships with a stronger representative capability, as shown in Fig. 2. Finally, a heatmap regressor ended with the Hadamard product between the outputs of the appearance and the up-sampled geometric constraint with a 3D cubic interpolation in the same resolution as the prediction heatmaps. In this way, the implicit spatial coherence between the landmarks is enhanced and further boosts the locating robustness of the proposed detection framework. We will validate the effectiveness of these components in the ablation experiments reported later.

3.2 Loss function

MSE and L1 norm are commonly used to penalize the difference between the intensities of the ground truth heatmaps and the prediction heatmaps. However, it has been shown in (Wang *et al.*, 2019) that Wing loss (Feng *et al.*, 2018) provides a function allowing to better adaptation to coordinate-based regression. AWing loss, initially employed in face alignment, can not only exactly estimate the targets of a heatmap, but also tolerate slight errors appearing in the background. The AWing loss function is given by (Wang *et al.*, 2019)

$$L(g,\hat{g}) = \begin{cases} w\ln(1+|\frac{g-\hat{g}}{\varepsilon}|^{\alpha-g}) & if\ |g-\hat{g}|<\theta \\ A|g-\hat{g}|-B & \text{otherwise} \end{cases} \quad (2)$$

where g and $\hat{g}$ denote the pixel values in the ground truth heatmap and the prediction heatmap. The parameters w, α, ε, θ will be set respectively to 14, 2.1, 1, 0.5 as recommended in (Wang *et al.*, 2019). $A = w(1/(1+(\theta/\varepsilon)^{\alpha-g})(\alpha-g)((\theta/\varepsilon)^{\alpha-g-1})(1/\varepsilon)$ and $B = (\theta A - w\ln(1+(\theta/\varepsilon)^{\alpha-g}))$ make the objective function differentiable under the condition $|g-\hat{g}|=\theta$.

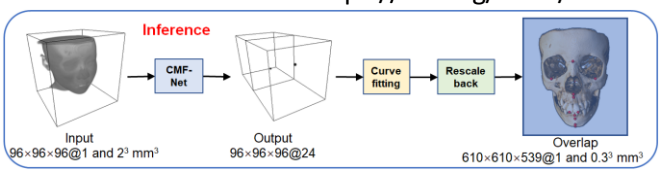


Figure 3. Illustration of the inference process when inputting a CBCT.

Finally, the network parameters $\mathbf{w},\mathbf{b}$ are updated to better estimate the L Gaussian volumetric heatmaps by minimizing the AWing loss between the ground truth heatmaps $g_l(\mathbf{x};\sigma)$ and the regressed prediction heatmaps $\hat{g}_l(I;\mathbf{w},\mathbf{b})$ over all landmarks for a given CBCT image I by the following formula

$$\min_{\mathbf{w},\mathbf{b}} \sum_{l=1}^{L}\sum_{\mathbf{x}} L\ (\hat{g}_l(I;\mathbf{w},\mathbf{b}), g_l(\mathbf{x};\sigma)) \quad (3)$$

where σ denotes the Gaussian standard deviation and is set as a constant value of 1.5. Note that $\mathbf{x}$ and $\mathbf{x}_l^*$ respectively refer to the coordinate in the heatmap domain and the target position of ground truth of *l*th landmark, and it is expressed as follows

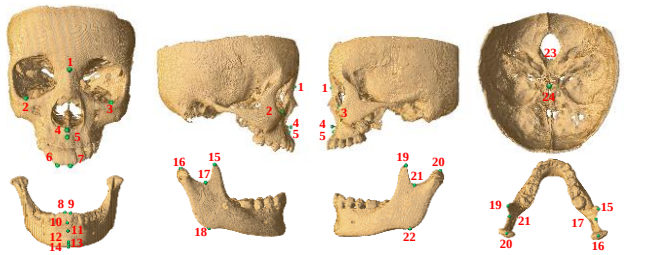
$$g_l(\mathbf{x};\sigma) = \exp(-\frac{1}{2\sigma^2}\left\|\mathbf{x}-\mathbf{x}_l^*\right\|_2^2) \quad (4)$$

3.3 Inference procedure

Fig. 3 illustrates the inference process for a given 3D dental CBCT image. Note that ground truth heatmaps are determined by the positions of landmarks in decimal format through the known down-sampling factors, rather than rounding them up to the nearest integer value. A drawback of heatmap-based regression is that the coordinates cannot be estimated directly. The location of the sub-pixel including the maximum value from the regressed heatmap at low resolution must be recognized using advanced techniques. Ideally, the regressed heatmap would follow a Gaussian distribution as well as the target heatmap generated around each landmark. However, the response signal from the predicted heatmap does not strictly obey a Gaussian distribution due to the influence of complex anatomical and pathological structures. Thus, the distribution-aware landmark regression method (*i.e.* DARK (Zhang *et al.*, 2020a)) may output final coordinates that are far away from the targets. We estimate the location of true extreme point around the maximum in the predicted heatmaps by a quadratic curve fitting method (Bailey, 2003). Then, the coordinates in sub-pixels are rescaled back to the original space to obtain the final regressed coordinate representations. Our algorithm can thus accurately detect multiple 3D cephalometric landmarks simultaneously when inputting a down-sampled volumetric CBCT image with a variable size.

4. Experimental configurations

4.1 Evaluation dataset



1. Nasion (Na) 2. Orbitale Right (OrR) 3. Orbitale Left (OrL)
4. Anterior Nasal Spine (ANS) 5. A Point (A) 6. Maxillary incisor-right (UR1)
7. Maxillary incisor-left (UL1) 8. Mandibular incisor-left (LL1) 9.Mandibular incisor-right (LR1)
10. Infradentale (Id) 11. B Point (B) 12. Pogonion (Pg)
13. Gnathion (Gn) 14. Menton (Me) 15. Coronoid Right (CorR)
16. Condylar Right (CoR) 17. Sigmoid Notch Right (SNR) 18. Gonion Right (GoR)
19. Coronoid Left (CorL) 20. Condylar Left (CoL) 21. Sigmoid Notch Left (SNL)
22. Gonion Left (GoL) 23. Basion (Ba) 24. Sella (S)

Figure 4. All landmarks annotated on a 3D volume rendering of a subject from frontal, right, left, and top views with 9 landmarks on midface and 15 on mandible. Their names and the corresponding abbreviations are given.

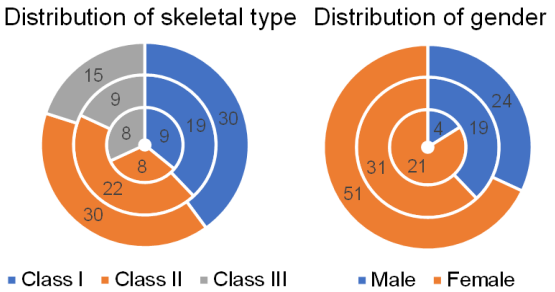


Figure 5. The distribution of skeletal classification, and gender information in our dataset. The outer, middle, and inner rings represent the training set, Test1 data and Test2 data, respectively.

3D CBCT data from 150 subjects (from 7 to 41 years old, average age = 17.97 years, standard deviation = 7.05, including 44 males and 81 females) were acquired from the Affiliated Hospital of Stomatology, Nanjing Medical University. Among them, 58, 60, 32 subjects were respectively identified as Class I, Class II, and Class III malocclusion according to skeletal classification types. The CBCT device was NewTom (Quantitative Radiology, Verona, Italy) and the parameter setting was as follows: tube voltage of 110 kV and current of 10 mAs, image matrix size of 610×610×(509~610) with a voxel of 0.25×0.25×0.25 mm^3 or 0.30×0.30×0.30 mm^3. It is worth noting that if a subject had a missing incisor, then the imaging data were excluded because there are four landmarks placed on incisors of maxilla and mandible. The research associated with the CMF landmark localization was approved by the ethics committee and has no implication on patient treatment.

Manual marking on volumetric CBCT images at original resolution was carried out on multiplanar reconstruction (MPR, *i.e.*, axial, sagittal, and coronal cross-sections) views and then confirmed the accuracy of the outcome on a segmented volume-rendered hard skull. The ground truths of all datasets were obtained by a domain-specific and experienced expert who served as reference. A second one, with less experience, was asked to annotate the datasets again to estimate the inter-observer variability and the performance of our approach when compared with the competing methods.

In total, 24 skeletal landmarks were identified, including 9 on midface and 15 on the mandible. All landmarks lay on a hard skull as shown in Fig. 4 and the objective is to locate these predefined significant points accurately and robustly.

As shown in Fig. 5, we randomly selected 50% of the 150 CBCT scans for training, 33% for Test1 data, and the remaining 17% CBCT images used as Test2 data to further validate the generalization capability of the proposed architecture. Among them, Classes I, II, and III represent respectively 40%, 40%, 20% in the training set, 38%, 44%, 18% for Test1 and 32%, 32%, 36% for Test2. The distributions of gender as male and female are 32%, 68% for training and 38%, 62% for Test1 and 16%, 84% for Test2.

4.2 Implementation details

Every image is smoothed using a 3D Gaussian filtering kernel with a standard deviation of 0.75 before entering the network. The CBCT images are down-sampled to get isotropic volumes with size of 96×96×96 and voxel spacing of 2×2×2 mm^3 by linear interpolation. We truncate the pixels' values in each dataset to the range of [-1024, 2048] and they are then normalized to [-0.5, 1] by dividing by 2048. To get rid of overfitting issues, a series of data augmentations are carried out to enrich the diversity of training data. Intensity perturbation operations are performed including shift ([-0.25, 0.25] pixel value) and scale ([0.75, 1.25] times). Spatial augmentation is also conducted to enrich the diversity of the training dataset using off-the-shelf SimpleITK (Beare *et al.*, 2018). It includes translation ([-30, 30] mm), rotation ([-0.25, 0.25] rad), scale ([0.95, 1.15] times) and flip (randomly flip along the *z*-axis with the possibility of 0.5). All data augmentation operations follow a uniform distribution within the predefined intervals.

Our CMF-Net is fully implemented using the TensorFlow platform on a GPU (*i.e.*, NVIDIA GeForce RTX 2080 Ti, 11 GB), and Ubuntu 18.04 equipped with 64 GB RAM. The Adam optimizer with an initial learning rate of 0.0001 (Kingma and Ba, 2014) is employed to minimize the objective function. It follows an exponential decay function as the training step increases. The number of iterations and batch size are set to 50000 and 1, respectively.

4.3 Evaluation metrics

The point-to-point error is computed by the Euclidean distance between the location of the ground truth $\mathbf{x}_{n,l}$ and the predicted position $\hat{\mathbf{x}}_{n,l}$. For L landmarks in N images, the mean radial error (MRE) and associated standard deviation (SD) are computed by

$$MRE = \frac{1}{NL}\sum_{n=1}^{N}\sum_{l=1}^{L}\left\|\mathbf{x}_{n,l} - \hat{\mathbf{x}}_{n,l}\right\| \quad (5)$$

$$SD = \frac{1}{\sqrt{NL-1}}\sum_{n=1}^{N}\sum_{l=1}^{L}\left\|\hat{\mathbf{x}}_{n,l} - MRE\right\|_2 \quad (6)$$

The successful detection rate (SDR) measures the rate of correctly detected anatomical landmarks: a landmark is considered as successfully detected if the distance between the ground truth and the automatically detected landmark of the

network is within various ranges of a certain radius *r*, such as 1.5 mm, 2 mm, 3 mm, 4 mm. It is worth noting that if a landmark is not been detected, then we regard it as missing, which will not take part in calculating the detection precision. Additionally, the non-parametric Wilcoxon signed rank test (*p*-value) is also used to evaluate the agreement for paired data.

5. Results and analysis

5.1 Methods for comparison

We compare our method to state-of-the-art methods on the training and two test datasets. The approach described in (Alansary *et al.*, 2019) is based on learning the optimal search strategy that maximizes the reward function to localize the target anatomical landmark with multiscale deep Q-network (DQN) (Mnih *et al.*, 2015) in a coarse-to-fine fashion. We repeatedly train an individual detection model for each landmark using an officially released source code[1]. To take advantage of anatomical interdependence, a collaborative multi agent landmark detection framework (Collab DQN) (Vlontzos *et al.*, 2019) is used to increase accuracy by sharing the agents' experience. We set the number of agents to 2 and train it using the official implementation[2]. The U-Net approach (Ronneberger *et al.*, 2015) is first developed for medical image segmentation and its architecture has been adopted here to fit for CMF landmark localization purpose. SCN (Payer *et al.*, 2019) suggests a spatial configuration network to encode the spatial relationships between the landmarks and optimize objective function with learnable standard deviations. It achieved promising results on both 2D and 3D medical images and the fact that its code is publicly available[3] and is used for the comparison with our method. In contrast to the conventional inference rule of finding the extreme value (Payer *et al.*, 2019) or the center of mass (Sun *et al.*, 2018) in the predicted heatmaps, DARK (Zhang *et al.*, 2020a) exploits unbiased sub-pixel centered coordinate encoding and Taylor expansion for coordinate decoding, a model-agnostic method used to refine positions on human pose estimation. We extend its original 2D solution to a 3D version for predicting the coordinates in the original image space. SA-LSTM (Chen *et al.*, 2022) first localizes landmarks via heatmap regression and then progressively updates the relative offsets based on the cropped high-resolution patches, and is used for the comparison with its released code[4]. 3D Faster R-CNN (Zhang *et al.*, 2020a) is designed for detecting a varying number of landmarks. We drop the random erasing operation to boost the discrimination of the model since all subjects have the same number of landmarks. Furthermore, we train an individual model using a single-scale U-Net for each landmark because no refinement code is freely available[5]. Specifically, the local refinement architecture is analogous to the appearance branch of our solution but without transformers. We crop a high-resolution volume patch with size 96×96×96 with an isotropic spacing of 0.4 mm around the initial position estimated from the first stage as input of the corresponding framework to achieve a better prediction. To make a fair comparison, we perform the same refinement strategy in our localization framework dubbed CMF-Net+Refinement.

5.2 Experimental results

5.2.1 Quantitative results

Quantitative evaluations of different methods are presented in this subsection. We report runtime, MRE±SD, the number of missing landmarks, and *p*-value as shown in Table 1. Furthermore, SDR is also used to measure the reliability of each method when given a certain error range (*i.e.*, 1.5 mm, 2 mm, 3 mm, 4 mm) as tabulated in Table 2. The best results are marked in bold.

Benefitting from the implicit communication between agents, Collab DQN achieves a lower MRE±SD and a higher SDR within the four precision ranges in contrast with DQN on test sets as summarized in Tables 1 and 2. We can see that U-Net achieves a lower localization error, as it regresses according to local patches and lacks global context information, which still results in false-positive predictions as demonstrated in a lower SDR within 3 mm and 4 mm when compared to DQN. There is a large performance gain of over 0.5 mm in terms of MRE and also a significant improvement in SDRs with SCN in comparison to localization U-Net on test sets. It proves that the spatial configuration component can reveal spatial dependency between neighboring landmarks and greatly disambiguate similar looking candidates and further enhance the localization precision. When we replaced the coordinate decoding method in SCN by using the scheme proposed in (Zhang *et al.*, 2020a), a severe deficit of up to 0.3 mm on MRE and a drop in SDRs as well are observed. This indicates that a quadratic curve fitting method predicts coordinates more accurately than by analyzing the distribution of responses from the regressed heatmap, since the response itself may not necessarily resemble a Gaussian distribution because of large morphological variations and pathological structures. Conversely, the fitting method is regardless of these drawbacks. SA-LSTM (Chen *et al.*, 2022) brings competitive results on both MRE±SD and SDR owing to a cascaded coarse-to-fine scheme. The disadvantages of intrinsic visual ambiguity and high-dimensional nonlinear mapping involved by subsequent displacement regression for coordinate refinement still make it underperform our method even if we use a down-sampled image as an input. The CMF-Net consistently performs better than other competing

[1] https://github.com/amiralansary/rl-medical
[2] https://github.com/thanosvlo/MARL-for-Anatomical-Landmark-Detection
[3] https://github.com/christianpayer/MedicalDataAugmentationTool-HeatmapRegression
[4] https://github.com/runnanchen/SA-LSTM-3D-Landmark-Detection
[5] https://github.com/xychen2022/3DFasterRCNN

Table 1. Results for inter-observer variability and nine different methods on training (75 images), Test1 (50 images), Test2 (25 images), and both test sets (75 images) (*i.e.*, runtime (s), MRE±SD (mm), number of missing landmarks and *p*-value compared to the latter of the proposed method)

Method	Runtime (s)	Training		Test1		Test2		Test1+Test2	
		MRE±SD (mm)	*p*-value	MRE±SD (mm)	#Miss	MRE±SD (mm)	#Miss	MRE±SD (mm)	*p*-value
Inter-observer	-	1.053±1.031	0.0007	0.963±0.999	0	0.984±1.012	0	0.970±1.003	0.0061
DQN	123.04	1.511±2.913	0.0050	2.026±3.085	0	2.365±7.513	0	2.139±5.016	<0.0001
Collab DQN	35.20	1.291±1.690	<0.0001	1.666±2.090	0	1.691±2.639	0	1.675±2.287	0.0019
Localization U-Net	4.27	1.697±3.219	0.5036	1.819±2.280	0	1.964±3.990	1	1.867±2.961	0.0583
SCN	8.60	0.912±0.561	<0.0001	1.257±0.895	0	1.240±0.874	0	1.268±0.880	0.0003
SCN+DARK	8.70	1.257±0.618	0.0005	1.565±0.925	0	1.592±0.876	0	1.574±0.909	0.0347
SA-LSTM	**3.08**	0.754±0.610	0.0009	1.122±0.913	0	1.187±0.959	0	1.144±0.929	0.3659
3D Faster R-CNN	85.36	0.104±**0.049**	0.1840	0.925±0.901	39	0.979±0.893	28	0.943±0.898	0.0944
CMF-Net	4.68	0.701±0.515	<0.0001	1.100±**0.839**	0	1.123±**0.773**	0	1.108±**0.817**	0.5179
CMF-Net+Refinement	62.94	**0.102**±0.050	-	**0.923**±0.894	0	**0.968**±0.876	0	**0.938**±0.888	-

Table 2. Results for inter-observer variability and nine different methods on training (75 images), Test1 (50 images), Test2 (25 images), and both test sets (75 images) (*i.e.*, SDR (%) within four precision ranges)

Method	Training				Test1				Test2				Test1+Test2			
	1.5 mm	2 mm	3 mm	4 mm	1.5 mm	2 mm	3 mm	4 mm	1.5 mm	2 mm	3 mm	4 mm	1.5 mm	2 mm	3 mm	4 mm
Inter-observer	78.72	88.28	95.00	97.39	82.08	88.50	94.92	97.75	80.67	88.83	95.17	97.67	81.61	88.61	95.00	97.72
DQN	71.67	86.06	95.33	97.22	51.17	69.50	86.00	92.83	49.33	70.50	87.00	93.00	50.56	69.83	86.33	92.89
Collab DQN	85.00	91.72	95.50	96.72	65.17	77.58	91.58	94.75	65.50	79.67	90.33	94.83	65.28	78.28	91.17	94.78
Localization U-Net	66.61	78.89	88.11	93.61	61.25	73.92	85.83	92.67	61.77	73.62	85.81	91.15	61.42	73.82	85.83	92.16
SCN	88.61	95.61	99.17	99.72	72.58	85.58	95.50	98.33	71.50	84.00	95.00	98.67	71.78	84.89	95.33	98.44
SCN+DARK	71.33	89.78	98.56	99.72	55.50	77.42	92.83	97.83	54.67	75.17	92.50	97.67	55.22	76.67	92.72,	97.78
SA-LSTM	86.67	92.08	97.25	98.25	75.75	86.58	95.50	98.42	73.67	84.17	93.50	98.00	75.06	85.78	94.83	98.28
3D Faster R-CNN	**100.00**	**100.00**	**100.00**	**100.00**	85.96	91.99	97.16	98.19	84.79	90.21	95.98	98.08	85.57	91.40	96.77	98.15
CMF-Net	95.17	97.94	99.39	99.89	80.25	90.08	96.83	**98.92**	77.33	87.33	97.00	**99.00**	79.28	89.17	96.89	**98.94**
CMF-Net+Refinement	**100.00**	**100.00**	**100.00**	**100.00**	**86.67**	**92.08**	**97.25**	98.25	**85.00**	**90.67**	**96.33**	98.33	**86.11**	**91.61**	**96.94**	98.28

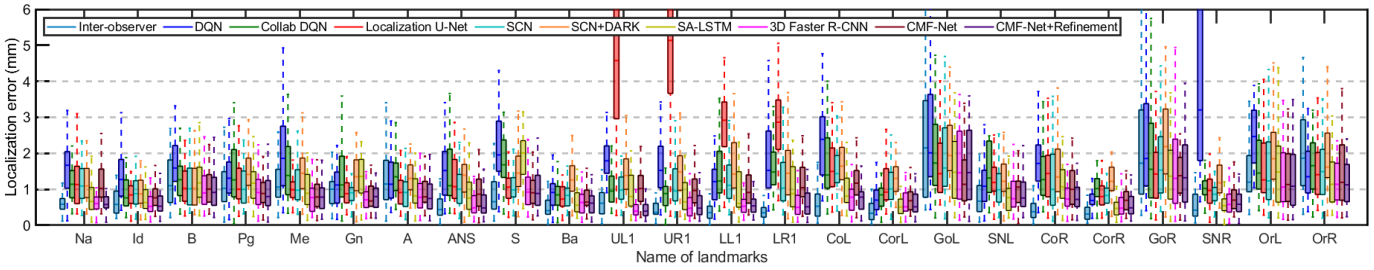


Figure 6. Localization errors (mm) for each of the 24 landmarks with inter-observer variability and nine different methods.

methods except for 3D Faster R-CNN for both MRE±SD and SDR with respect to all precision ranges, with an average error limited to 1.108 mm and an SDR of 79.28% within 1.5 mm over all test sets. This demonstrates that the model consisting of the appearance branch and the geometric constraint branch has the capability of achieving sub-pixel position accuracy according to a quadratic curve fitting method applied in predicted volumetric heatmaps, providing an enhanced accuracy in 3D anatomical landmark localization. It is worth recalling that the error range of 1.5 mm is clinically acceptable as reported in the literature (Zhang *et al.*, 2020b) and CMF-Net clearly fulfills this condition. By comparing the CMF-Net with its variant on all test sets, it can be seen that the difference ($p > 0.05$) between them for landmark detection is not statistically significant.

Even though 3D Faster R-CNN offers encouraging performance with less than 1 mm on MRE and SD owing to a cascaded global-to-local pipeline, the detection-based method often fails to identify all landmarks with the number of missing ones of 67 on both test sets as presented in Table I. Finally, our CMF-Net with refinement exhibits the lowest MRE and the highest successful detection accuracy except for SDR within 4 mm and has no missing landmarks whilst it takes less computational cost during inference compared to 3D Faster R-CNN. It also particularly outperforms the second observer in SDRs with significant gaps, *e.g.*, 4% and 3% higher within 1.5 mm and 2 mm precision ranges on all test sets. This indicates that our refinement version has a performance capable of landmarking these clinically relevant points with high accuracy and reaching the level of an experienced orthodontist.

Fig. 6 reports the error statistic on each of the 24 landmarks with a boxplot for comparing distributions between predictions of competing methods and our CMF-Net solution and its refinement version. We can see that U-Net provides a better localization accuracy than DQN, but it often fails to detect upper and lower teeth since the teeth possess local similar anatomical appearance features as opposed to other landmarks. With its spatial configuration encoding structural constraints between the landmarks, SCN leads to a large gain in both MRE and SD. The performance begins to decrease when the coordinate decoding is changed to DARK. Despite our method yielding the best results among all tested methods, too high localization errors remain for teeth on the mandible,

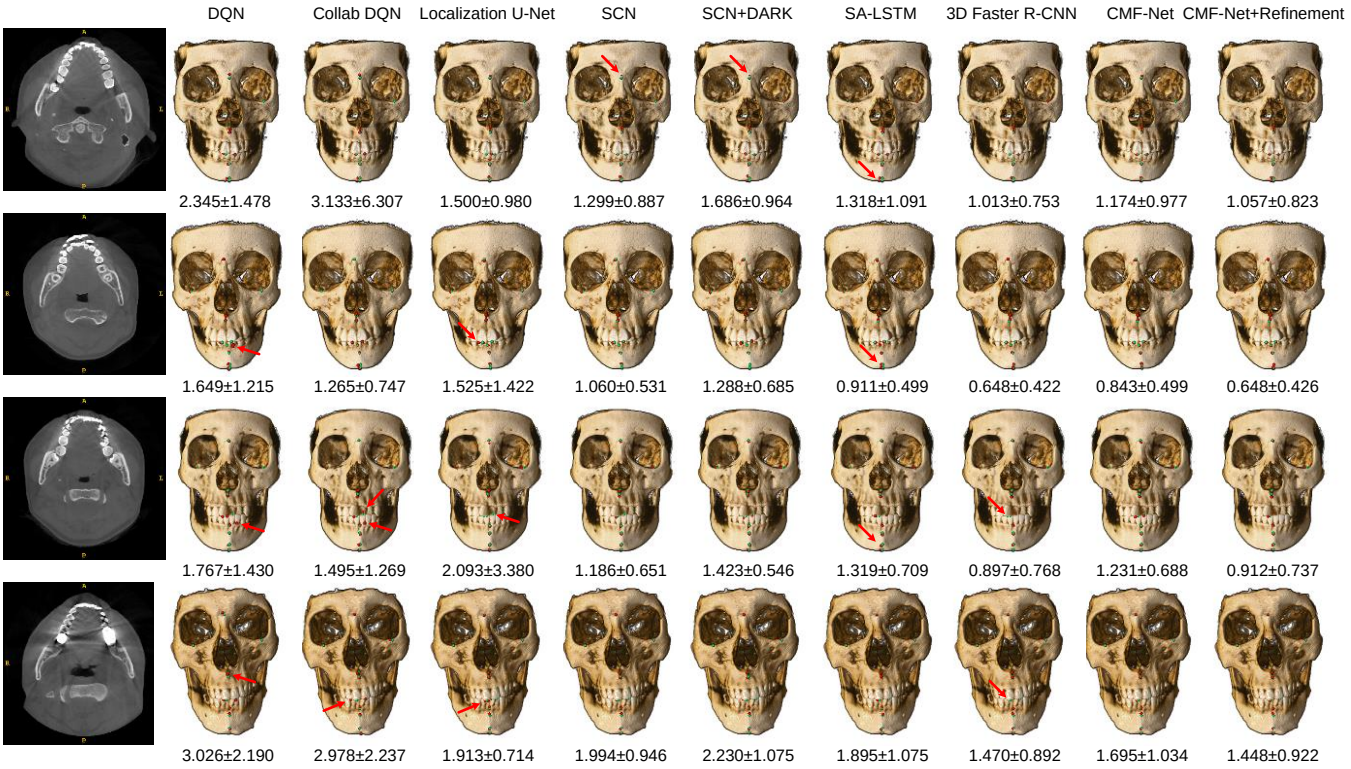


Figure 7. Visualization of localization results on 4 representative cases. The columns specify various methods, the rows different subjects, and the first columns display coronal views. MRE±SD is also shown below the corresponding case for each landmarking method. Red and green points respectively denote the landmarks detected by each method and the ground-truth annotations. Red arrows indicate large errors or misdetections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

Gonia, and Orbitale. Consider that providing sufficient receptive field can help the model to better localize these meaning points lay on the bony structure. In the future, we will work on building long-range dependency at multiple scales in a computationally-efficient way. The refinement variant of CMF-Net achieves considerably superior performance capable of landmarking comparable with the level of the second observer. Whereas, the coarse-to-fine manner still falls behind CMF-Net owing to not effectively capturing global-local dependence, especially for Sella and Gonion.

As we know, images acquired from patients with metallic implants undergoing CBCT examinations suffer from metal artifacts, and the following undesirable detection results will lead to an unacceptable diagnosis. Therefore, we selected 9 images from Test1 and Test2 data to evaluate the impact of metal artifacts on the localization performance of all techniques. The results reported in Table 3 clearly show that our proposed model performs better than other competing methods in terms of MRE±SD and SDR within various error ranges. Especially given the 1.5 mm error range, it achieves a remarkable improvement of 8% over SA-LSTM, enabling anatomical landmark detection with significant robustness.

5.2.2 Qualitative results

To complement these quantitative features, we propose to visually illustrate some representative cases of landmark localization. Four randomly selected subjects are displayed in Fig. 7. It can be seen that DQN fails to identify the correct locations of teeth when considering a subject who suffers from severe CMF deformities as pointed out by case 3. The reason is that the design of patch selection usually exposes a limitation in that it moves towards to the target according to the current local environment being located but neglects the global representation of anatomical structure. Collab DQN less considers the spatial relationships between landmarks, leading to false-positive predictions. In addition to high sensitivity to position initialization, both reinforce learning methods contribute to significant performance degradation, especially for limited annotated medical images. The results obtained by using localization U-Net tend to exhibit large errors as marked in the red arrow in the fourth column, especially for the second subject where the predicted tooth is far away from the target and is false-positively detected. U-Net focusing more on appearance learning is sensitive to their variations and leads to landmark misidentifications. SCN injects weak structural knowledge and the postprocessing step provided by DARK also brings imprecise coordinate representative owing to inaccurate response estimation on the prediction heatmaps. Large distance errors in locating challenging landmarks are marked by the red arrow in the first subjects in Fig. 7. SA-LSTM(Chen *et al.*, 2022) behaves worse in detecting

Table 3. Results for nine different methods on selected images with metal artifacts from test data (*i.e.*, MRE±SD, number of missing landmarks, SDR)

Method	MRE±SD (mm)	#Miss	SDR (%)			
			1.5 mm	2 mm	3 mm	4 mm
DQN	2.688±8.609	0	47.22	66.20	85.19	92.59
Collab DQN	1.836±2.242	0	62.04	73.15	88.89	93.52
Localization U-Net	1.867±1.688	0	55.09	73.15	84.26	92.59
SCN	1.423±0.937	0	62.04	80.09	95.37	98.61
SCN+DARK	1.730±1.006	0	43.06	71.76	89.81	98.15
SA-LSTM	1.267±1.054	0	66.67	80.56	93.98	97.69
3D Faster R-CNN	0.972±0.929	12	83.33	89.71	**98.04**	98.53
CMF-Net	1.240±0.957	0	74.07	87.04	97.22	**98.61**
CMF-Net+Refinement	**0.958±0.916**	0	**84.72**	**90.74**	97.69	**98.61**

landmarks located on the mandible in the first three cases, which justifies that the coordinate-based method would compromise the detection accuracy. Although 3D Faster R-CNN obtains a favorable detection precision, it fails to detect all landmarks perfectly. In these four representative cases, the locations of the predicted landmarks obtained by CMF-Net are much closer to the targets. We observe that the variant of CMF-Net achieves a much more accurate detection when applying a local refinement, *e.g.*, the positions of Nasion are nearly aligned across the four cases.

Case 4 shows severe aliasing artifacts due to metallic implants, and the performance of all methods begins to decrease when compared to that in the first three cases. Ours yields the best results with a mean radial error of 1.448 mm of all. This indicates that poor image quality will affect the downstream localization task, while it consistently surpasses other state-of-the-art methods on the metric of MRE.

5.3 Ablation studies

A set of ablation experiments on Test1 data were conducted to validate the effectiveness of the proposed model.

1) Evaluation on Different Components: We analyzed the influence of different components (i.e., transformer (Trans), geometric constraint (GC), and AWing loss) and considered the localization U-Net as the baseline (Base). The results are reported in Table 4. Adding transformers enhances the discriminative capability of the model (Base+AWing vs. Base+Trans+AWing, Base+GC+Awing vs. Base+Trans+GC +AWing). This result confirms the interest to utilize long-range information in feature maps to suppress false-positive candidates while achieving accuracy gains. In addition, adding GC dramatically decreases the localization error by more than 0.5 mm in MRE and improves the successful detection rate by over 15% in SDR (Base+Awing vs. Base+GC+Awing, Base+Trans+Awing vs. Base+Trans+GC+Awing). It boosts the performance by a large margin and such a result reveals that the prior knowledge on spatial constraints of landmarks has been effectively learned. When replacing the traditional MSE loss with the AWing loss, the best performance is achieved by using the transformer together with the geometric constraint branch (Base+Trans+GC vs. Base+Trans+GC+Awing). 3.8%

Table 4. Influence on Trans, GC, and AWing loss

Method	MRE±SD (mm)	SDR (%)			
		1.5 mm	2 mm	3 mm	4 mm
Base	1.819±2.280	61.25	73.92	85.83	92.67
Base+Awing	1.702±1.854	62.17	76.33	87.42	93.25
Base+GC+AWing	1.142±0.895	78.17	88.58	96.17	98.75
Base+Trans+AWing	1.607±1.517	63.00	77.58	89.08	94.75
Base+Trans+GC	1.143±0.875	78.33	88.93	96.58	98.58
Base+Trans+GC+AWing	**1.100±0.839**	**80.25**	**90.08**	**96.83**	**98.92**
Base+Trans(w/o multi-head)+GC+AWing	1.132±0.877	78.67	89.50	96.67	98.83

Table 5. Results achieved by selecting different parameter combinations for CMF-Net (*i.e.*, MRE±SD and SDR for 24 landmarks)

heads	transformers	MRE±SD (mm)	SDR (%)			
			1.5 mm	2 mm	3 mm	4 mm
8	1	1.105±0.864	79.58	89.58	97.17	98.67
8	2	1.100±0.839	80.25	90.08	96.83	98.92
8	4	1.107±0.834	79.50	90.08	97.00	99.08
16	2	1.127±0.853	78.58	89.50	97.17	99.00

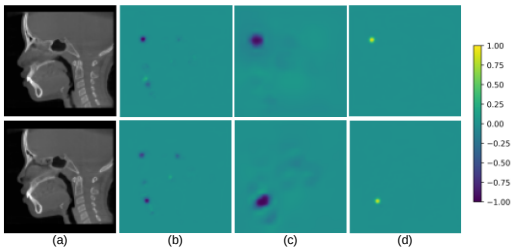


Figure 8. Visualizations of outputs of the appearance (b) and the geometric constraint (c) and the final prediction heatmap (d) on the associated sagittal slice (a) of the identified landmark when inputting a down-sampled volumetric CBCT. Examples of Nasal and Infradentale for the same case are shown in the top and bottom rows.

and 4.1% relative improvements are obtained on MRE and SD evaluation criteria respectively. A similar observation is shown in SDR metrics as well. This demonstrates that an accurate estimation of the pixel values near the mode of the regressed heatmap can be achieved. The effect of combining transformers, GC, and Awing loss is therefore positive and with a slight loss in time computation (refer to Table 1). As shown in Table 4, transformer with multi-head attention performs better than that with single-head attention. Besides, we find that the strategy of increasing the depth of the network does not bring a significant improvement.

2) Evaluation on Scaling: We perform different combinations to choose the optimal parameter settings (number of heads and transformers) for CMF-Net as summarized in Table 5. We set the number of heads to 8, the performance begins to saturate using 2 transformers. Scaling the size of heads brings no performance enhancement in our implementation.

5.4 Visualization

To explain how the proposed CMF-Net uses appearance information and prior knowledge to localize the predefined landmarks, we visualized the outputs of both branches during inference. We randomly choose an image from Test1 and its visualization results are presented in Fig. 8.

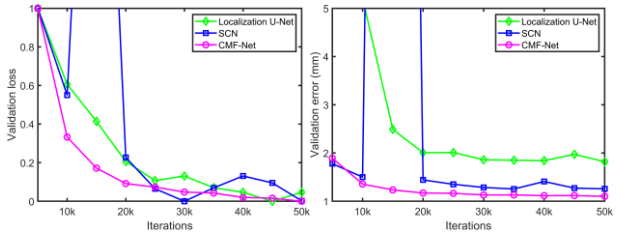


Figure 9. Comparisons of localization U-Net, SCN, and CMF-Net in terms of convergence speed and landmarking accuracy over iterations.

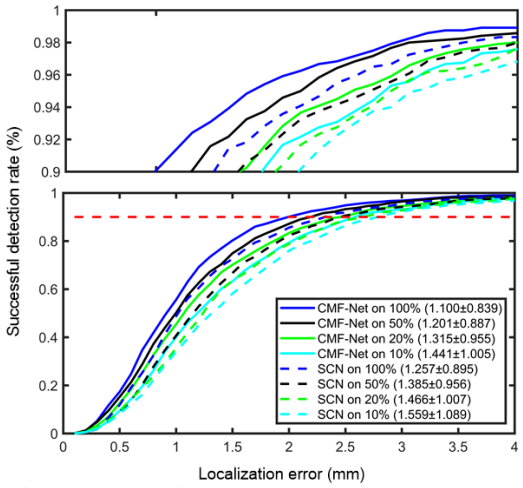


Figure 10. Comparisons of CMF-Net and SCN methods on SDR (%) and MRE±SD (mm) when trained on different percentages of training images, *i.e.*, 10%, 20%, 50%, and 100%. An enlarged picture of the curves above the red line corresponding to 90% is also displayed.

We can see that the appearance branch generates an ambiguous response with multiple peaks, especially in the bottom row, ignoring the latent distribution between the landmarks. While the geometric constraint ensures the detected landmark is contained within a large feasible region as shown in the third column. By combining with it, a precise heatmap being generated facilitate the model to robustly locate the corresponding landmark. This demonstrates that the geometric constraint branch uses prior knowledge to learn the implicit spatial relationships and screen out unlikely predictions.

5.5 Effectiveness of training optimization

The validation loss and the localization error of CMF-Net, SCN, and U-Net are shown in Fig. 9. We normalized the training loss to stay in the range [0,1] for comparison purpose because different loss functions were applied (their values at the first and last iterations are used to do that). CMF-Net optimizes the training process by accelerating the training speed and improving localization precision. Benefitting from the long-range dependency of the appearance branch with two transformers and the class-balanced AWing loss, CMF-Net provides a faster convergence speed and a lower localization error than SCN and U-Net. SCN presents a higher landmarking accuracy in comparison with U-Net because its spatial configuration eliminates locally similar anatomical structures. However, the objective function is difficult to converge as the validation loss and error oscillate at 15k, likely owing to a lack of long-range contextual understanding and spatial dependencies.

5.6 Influence of the amount of training data

The availability of large image datasets in medicine is an important issue due to strict ethical rules applied to patient privacy and the cost of expert-specific annotations. It is therefore important to understand how the performance evolves according to the amount of data used. We investigate this issue by considering 10%, 20%, 50%, and 100% of our training dataset with a comparison between the best two methods, SCN and CMF-Net. Fig. 10 shows the cumulative successful detection distribution versus discrete precision thresholds (the MRE±SD is also specified in each case). The CMF-Net consistently performs better than SCN at all different ratios of training images with an average error of 1.44 mm using only 10% of the training images. It proves that our CMF-Net can effectively learn representative features and generalize well on unseen data using limited sets of medical images, even in the presence of patients who suffer from severe malformations. Looking at the zoom displayed in the upper part of Fig. 10, we can see that the localization accuracy trained on 50% medical images has little difference from that trained on 100%, demonstrating that the embedded prior knowledge permits the model to effectively learn the implicit spatial relationships between the landmarks from very few annotated images.

5.7 Discussion

It is of major importance to fulfill the minimal medical requirements to see any algorithmic application potentially transferable to clinical settings. They cover both the method accuracy and the robustness but also other considerations like time computations, easiness to use, adaptability to future technology and so on. Only two methods fit the accuracy clinical constraint, SCN, and CMF-Net with an advantage over the latter. Regarding the time overheads, the results displayed in Table 1 clearly show that multiple landmarks up to 24 can be inferred simultaneously using CMF-Net in 4.68 s for a given CBCT dental volume, almost half of the time needed by SCN. The encouraging detection performance is ascribed to: 1) the appearance branch can accurately identify candidates due to the global context and long-range dependency incorporated with the transformers, and the geometric constraint branch can robustly determine the locations of these landmarks using implicit spatial relationships; 2) the exact estimation of the pixel values near the mode of the prediction heatmap with the AWing loss function.

6. Conclusion

In this paper, we have developed a trainable end-to-end localization approach referred to as CMF-Net by detecting up to 24 anatomical landmarks in 3D dental CBCT volumes. Despite low image quality, aliasing artifacts, and severe morphological variations, the proposed approach fulfills the clinical requirements in terms of both accuracy and robustness using a limited number of scans. The transformer was first introduced to the appearance branch to capture global context information and identify candidates with high accuracy, while the geometric constraint branch implicitly encoded prior knowledge for filtering out unlikely predictions. Then, the incorporation of the AWing loss allows estimating the pixel values of the prediction heatmaps precisely, especially for those near the mode of the Gaussian distribution. The experiments conducted so far show that this approach can save a significant amount of time for orthodontists in their landmark pointing. A local refinement strategy can also be adopted to further improve the localization performance at the expense of the computation time. Additionally, because our model does not rely on pretrained backbone networks to avoid overfitting, it can be easily extended to localization tasks on other medical data dealing with different modalities and organs.

Although the present method has achieved encouraging results relying on a small set of labeled images compared to state-of-the-arts, the work under progress concerns the use of this method in clinical environments in order to get a user feedback and identify the potential solutions for the most difficult landmarks to be detected. Furthermore, a CBCT volume must be down-sampled before passing it into our CMF-Net due to the limited computational resources of a graphics processing unit (GPU) which will result in local details loss. It may be solved by a cascaded coarse-to-fine style while preserving global-local dependence.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2022YFE0116700), the National Natural Science Foundation of China (62171125, 31800825, 31640028, 61876037, 82071143), the Natural Science Foundation of Jiangsu Province (BE2019748), the Fundamental Research Funds for the Central Universities (2242020K40039, 2242020K30012), the Short-term Recruitment Program of Foreign Experts (WQ20163200398), the Key Medical Research Projects of Jiangsu Health Commission (ZDA2020003). We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Alansary A, Oktay O, Li Y, Le Folgoc L, Hou B, Vaillant G, Kamnitsas K, Vlontzos A, Glocker B and Kainz B 2019 Evaluating reinforcement learning agents for anatomical landmark detection *Med. Image Anal.* **53** 156-64

Bailey D G *2003* Sub-pixel estimation of local extrema *Image Vis. Comput. New Zealand* pp 414-9

Beare R, Lowekamp B and Yaniv Z 2018 Image segmentation, registration and characterization in R with SimpleITK *J. Stat. Softw.* **86**

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S *2020* End-to-end object detection with transformers *Eur. Conf. Comput. Vis.* pp 213-29

Chen H, Shen C, Qin J, Ni D, Shi L, Cheng J C and Heng P-A *2015* Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 515-22

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y 2021a Transunet: Transformers make strong encoders for medical image segmentation *arXiv preprint arXiv:2102.04306*

Chen R, Ma Y, Chen N, Liu L, Cui Z, Lin Y and Wang W 2022 Structure-aware long short-term memory network for 3D cephalometric landmark detection *IEEE Trans. Med. Imaging* **41** 1791-801

Chen X, Lian C, Deng H H, Kuang T, Lin H-Y, Xiao D, Gateno J, Shen D, Xia J J and Yap P-T 2021b Fast and accurate craniomaxillofacial landmark detection via 3D Faster R-CNN *IEEE Trans. Med. Imaging* **40** 3867 - 78

Cheng E, Chen J, Yang J, Deng H, Wu Y, Megalooikonomou V, Gable B and Ling H *2011* Automatic dent-landmark detection in 3-D CBCT dental volumes *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* pp 6204-7

Codari M, Caffini M, Tartaglia G M, Sforza C and Baselli G 2017 Computer-aided cephalometric landmark annotation for CBCT data *Int. J. Comput. Assist. Radiol. Surg.* **12** 113-21

Criminisi A, Robertson D, Konukoglu E, Shotton J, Pathak S, White S and Siddiqui K 2013 Regression forests for efficient anatomy detection and localization in computed tomography scans *Med. Image Anal.* **17** 1293-303

Criminisi A, Shotton J, Robertson D and Konukoglu E *2011* Regression forests for efficient anatomy detection and localization in CT studies *International MICCAI Workshop on Medical Computer Vision* pp 106-17

Donner R, Langs G, Mičušik B and Bischof H 2010 Generalized sparse MRF appearance models *Image Vis. Comput.* **28** 1031-8

Donner R, Menze B H, Bischof H and Langs G 2013 Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization *Med. Image Anal.* **17** 1304-14

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G and Gelly S *2021* An image is worth 16x16 words: Transformers for image recognition at scale *Int. Conf. Learn. Represent.*

Ebner T, Stern D, Donner R, Bischof H and Urschler M 2014 *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 421-8

Feng Z-H, Kittler J, Awais M, Huber P and Wu X-J *2018* Wing loss for robust facial landmark localisation with convolutional neural networks *IEEE Conf. Comput. Vis. Pattern Recognit.* pp 2235-45

Gao Y and Shen D 2015 Collaborative regression-based anatomical landmark detection *Phys. Med. Biol.* **60** 9377-401

Ghesu F C, Georgescu B, Zheng Y, Grbic S, Maier A, Hornegger J and Comaniciu D 2019 Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 176-89

Gilmour L and Ray N *2020* Locating cephalometric X-Ray landmarks with foveated pyramid attention *Med. Imaging Deep Learn.* pp 262-76

Glocker B, Feulner J, Criminisi A, Haynor D R and Konukoglu E *2012* Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 590-8

Gupta A, Kharbanda O P, Sardana V, Balachandran R and Sardana H K 2015 A knowledge-based algorithm for automatic detection of cephalometric landmarks on CBCT images *Int. J. Comput. Assist. Radiol. Surg.* **10** 1737-52

He K, Zhang X, Ren S and Sun J *2016* Deep residual learning for image recognition *IEEE Conf. Comput. Vis. Pattern Recognit.* pp 770-8

Kang S H, Jeon K, Kang S-H and Lee S-H 2021 3D cephalometric landmark detection by multiple stage deep reinforcement learning *Sci. Rep.* **11** 1-13

Kingma D P and Ba J *2014* Adam: A method for stochastic optimization *Int. Conf. Lean. Representat.*

Lang Y, Lian C, Xiao D, Deng H, Thung K H, Yuan P, Gateno J, Kuang T, Alfi D M, Wang L, Shen D, Xia J J and Yap P T 2022 Localization of craniomaxillofacial landmarks on CBCT images using 3D mask R-CNN and local dependency learning *IEEE Trans. Med. Imaging* **41** 2856-66

Lang Y, Lian C, Xiao D, Deng H, Yuan P, Gateno J, Shen S G F, Alfi D M, Yap P-T, Xia J J and Shen D 2020 Automatic localization of landmarks in craniomaxillofacial CBCT images using a local attention-based graph convolution network *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 817-26

Lee S M, Kim H P, Jeon K, Lee S H and Seo J K 2019 Automatic 3D cephalometric annotation system using shadowed 2D image-based machine learning *Phys. Med. Biol.* **64** 055002

Li W, Lu Y, Zheng K, Liao H, Lin C, Luo J, Cheng C-T, Xiao J, Lu L, Kuo C-F and Miao S *2020* Structured landmark detection via topology-adapting deep graph learning *Eur. Conf. Comput. Vis.* pp 266-83

Li Y, Alansary A, Cerrolaza J J, Khanal B, Sinclair M, Matthew J, Gupta C, Knight C, Kainz B and Rueckert D *2018* Fast multiple landmark localisation using a patch-based iterative network *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 563-71

Lian C, Wang F, Deng H H, Wang L, Xiao D, Kuang T, Lin H-Y, Gateno J, Shen S G F, Yap P-T, Xia J J and Shen D 2020 Multi-task dynamic transformer network for concurrent bone segmentation and large-scale landmark localization with dental CBCT *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pp 807-16

Lindner C, Bromiley P A, Ionita M C and Cootes T F 2015 Robust and accurate shape model matching using random forest regression-voting *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 1862-74

Maas A L, Hannun A Y and Ng A Y *2013* Rectifier nonlinearities improve neural network acoustic models *Int. Conf. Mach. Learn.*

Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K and Ostrovski G 2015 Human-level control through deep reinforcement learning *Nature* **518** 529-33

Noothout J M H, De Vos B D, Wolterink J M, Postma E M, Smeets P A M, Takx R A P, Leiner T, Viergever M A and Isgum I 2020 Deep learning-based regression and classification for automatic landmark localization in medical images *IEEE Trans. Med. Imaging* **39** 4011-22

Palazzo S, Bellitto G, Prezzavento L, Rundo F, Bagci U, Giordano D, Leonardi R and Spampinato C *2021* Deep multi-stage model for automated landmarking of craniomaxillofacial CT scans *Int. Conf. Pattern Recognit.* pp 9982-7

Payer C, Stern D, Bischof H and Urschler M 2019 Integrating spatial configuration into heatmap regression based CNNs for landmark localization *Med. Image Anal.* **54** 207-19

Payer C, Štern D, Bischof H and Urschler M *2016* Regressing heatmaps for multiple landmark localization using CNNs *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 230-8

Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: Towards real-time object detection with region proposal networks *Adv. Neural Inf. Process. Syst.* **28**

Ronneberger O, Fischer P and Brox T *2015* U-Net: Convolutional networks for biomedical image segmentation *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 234-41

Severt T and Proffit W 1997 The prevalence of facial asymmetry in the dentofacial deformities population at the University of North Carolina *Int. J. Adult Orthodon. Orthognath. Surg.* **12** 171-6

Shahidi S, Bahrampour E, Soltanimehr E, Zamani A, Oshagh M, Moattari M and Mehdizadeh A 2014 The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images *BMC Med. Imaging* **14** 1-8

Simonyan K and Zisserman A *2015* Very deep convolutional networks for large-scale image recognition *Int. Conf. Learn. Represent.*

Sun X, Xiao B, Wei F, Liang S and Wei Y *2018* Integral human pose regression *Eur. Conf. Comput. Vis.* pp 529-45

Torosdagli N, Liberton D K, Verma P, Sincan M, Lee J S and Bagci U 2019a Deep geodesic learning for segmentation and anatomical landmarking *IEEE Trans. Med. Imaging* **38** 919-31

Torosdagli N, McIntosh M, Liberton D K, Verma P, Sincan M, Han W W, Lee J S and Bagci U 2019b Relational reasoning network (RRN) for anatomical landmarking *arXiv preprint arXiv:1904.04354*

Troulis M J, Everett P, Seldin E B, Kikinis R and Kaban L B 2002 Development of a three-dimensional treatment planning system based on computed tomographic data *Int. J. Oral Maxillofac. Surg.* **31** 349-57

Urschler M, Ebner T and Stern D 2018 Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization *Med. Image Anal.* **43** 23-36

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L u and Polosukhin I *2017* Attention is all you need *Adv. Neural Inf. Process. Syst.*

Veličković P, Cucurull G, Casanova A, Romero A, Lio P and Bengio Y *2018* Graph attention networks *Int. Conf. Learn. Represent.*

Vlontzos A, Alansary A, Kamnitsas K, Rueckert D and Kainz B *2019* Multiple landmark detection using multi-agent reinforcement learning *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 262-70

Wang F, Zheng K, Lu L, Xiao J, Wu M and Miao S *2021* Automatic vertebra localization and identification in CT by spine rectification and anatomically-constrained optimization *IEEE Conf. Comput. Vis. Pattern Recognit.* pp 5280-8

Wang X, Bo L and Fuxin L *2019* Adaptive wing loss for robust face alignment via heatmap regression *IEEE Int. Conf. Comput. Vis.* pp 6971-81

Zeng M, Yan Z, Liu S, Zhou Y and Qiu L 2021 Cascaded convolutional networks for automatic cephalometric landmark detection *Med. Image Anal.* **68** 101904

Zhang F, Zhu X, Dai H, Ye M and Zhu C *2020a* Distribution-aware coordinate representation for human pose estimation *IEEE Conf. Comput. Vis. Pattern Recognit.* pp 7093-102

Zhang J, Gao Y, Wang L, Tang Z, Xia J J and Shen D 2015 Automatic craniomaxillofacial landmark digitization via segmentation-guided partially-joint regression forest model and multiscale statistical features *IEEE Trans. Biomed. Eng.* **63** 1820-9

Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen S G-F, Tang Z, Chen K-C and Xia J J *2017* Joint craniomaxillofacial bone segmentation and landmark digitization by context-guided fully convolutional networks *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 720-8

Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen S G, Tang Z, Chen K C, Xia J J and Shen D 2020b Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization *Med. Image Anal.* **60** 101621

Zhong Z, Li J, Zhang Z, Jiao Z and Gao X *2019* An attention-guided deep regression model for landmark detection in cephalograms *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* pp 540-8