Computers in Biology and Medicine 192 (2025) 110294

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

THERE: Toward an easy and reliable tool for automatic cephalometric analysis

Riccardo Zese [a,*], Matteo De Maio [b], Francesca Cremonini [b]

[a] DOCPAS, University of Ferrara, Via Luigi Borsari 46, Ferrara, 44121, Italy
[b] DMTR, University of Ferrara, Via Luigi Borsari 46, Ferrara, 44121, Italy

ARTICLE INFO

Keywords:
Machine learning
Diagnostic methods and tools
Odontology
Artificial intelligence
Computer vision

ABSTRACT

Background: Cephalometric analysis in orthodontics is a meticulous process requiring high precision in identifying anatomical landmarks on lateral cephalometric radiographs. Manual analysis by clinicians remains the standard, as even slight deviations in landmark placement can lead to incorrect diagnoses. Despite advancements in AI, automatically localizing these landmarks remains challenging, with few systems meeting the exacting standards of European orthodontic practice.

Methods: We developed THERETransKey, a model designed for the automatic and accurate localization of anatomical landmarks in cephalometric analysis. To evaluate its practical utility, we integrated the model into THERE (auTomatic HElpeR for cEphalometry), an open-access, user-friendly AI-based tool for cephalometric tracing. Unique in its strict adherence to the guidelines of the European Board of Orthodontics, THERE provides immediate online access and continuously collects user data to improve its performance. The tool has been validated by users with different levels of experience in Orthodontics using a PSSUQ-based questionnaire.

Results: THERETransKey has shown enhanced accuracy with respect the previous models underlying THERE, as confirmed by its integration into daily clinical workflows at the University of Ferrara. Moreover, user feedback collected through the administered questionnaire confirms THERE's improved usability. Finally, its regular use enables the generation of a clinician-validated dataset during everyday practice.

Conclusions: By effectively supporting clinicians, THERE not only enhances cephalometric analysis but also contributes to building a robust dataset for future research. This initiative promotes a more generalized approach to automatic landmark detection across diverse radiographic equipment.

1. Introduction

In orthodontics, performing an accurate cephalometric analysis is a crucial diagnostic step, but it requires great precision to be used effectively. Cephalometric analysis is always performed on a lateral cephalometric radiograph with the dental arches in maximum intercuspation to identify a set of anatomical reference points on both soft and hard tissues, including teeth and skeletal contours. These landmarks are used in different combinations to calculate important distances, planes, and angles necessary for diagnosing various pathologies. The choice of these measures depends on the type of investigation to be conducted with the cephalometry. A key characteristic of this process is the need for extreme precision, as a deviation of even a few millimetres in the placement of these points can lead to an incorrect diagnosis. Given that the diagnosis is made by calculating the relative positions between different points, it becomes even more evident how critical high precision is.

For instance, the anatomical landmarks S (centre of the sella turcica), N (most anterior point of the fronto-nasal suture), and A (most posterior point of the maxillary anterior concavity) define an angle that represents the sagittal position of the upper jaw in relation to the cranial base, describing its normal or excessively anterior/posterior positioning. Similarly, the angle defined by S, N and B (most posterior point of the mandibular anterior concavity), represents the sagittal position of the mandible with respect to the cranial base. The difference ANB between angles SNA and SNB represents the sagittal intermaxillary relationship. Normal values range in the interval 2 ± 2 degrees, while a value outside this range represents anomalies, namely skeletal Class II if ANB > 4 degrees or skeletal Class III if ANB < 0 degrees.

Even though Yee et al. [1] concluded that Artificial Intelligence (AI) assistance can improve efficiency without compromising accuracy in cephalometric tracings for routine clinical practice and research, there is still room for future improvement. Errors in cephalometric analysis

* Corresponding author.
E-mail addresses: riccardo.zese@unife.it (R. Zese), matteo.demaio@edu.unife.it (M. De Maio), francesca.cremonini@unife.it (F. Cremonini).

https://doi.org/10.1016/j.compbiomed.2025.110294
Received 10 October 2024; Received in revised form 11 April 2025; Accepted 25 April 2025
Available online 16 May 2025

stem from tracing, landmark identification, and measurement inaccuracies [2]. Incorporating additional manual supervision and adjustments in automatic programs can enhance both accuracy and efficiency, as automated landmark identification continues to face challenges such as skeletal structure variations, blurry images due to device magnifications, and overlapping contralateral structures [3]. Even minor errors may result in misclassification and misdiagnosis. For three-dimensional cephalometric analysis, mean errors still range from <0.50 mm to >5 mm [4].

More recently, Polizzi et al. confirmed high variability in accurately identifying cephalometric landmarks. Many of the systematic reviews included in the umbrella review relied on an incorrect 2 mm cut-off threshold to compare AI-based landmark identification with human operators [5]. These findings align with the meta-analysis conducted by de Queiroz Tavares Borges Mesquita G et al. [6] which reported agreement rates of 79% and 90% between AI and manual detection of cephalometric landmarks, using error margins of 2 mm and 3 mm, respectively, and an average divergence of 2.05 mm. If we consider that even when two experts perform manual landmarking, divergences greater than 1 mm can occur, these results are indeed promising [7,8].

Certain hard and soft-tissue landmarks are particularly difficult to identify accurately due to image complexity caused by variations in X-ray projections between the left and right sides of the cranial structure. For instance, Gonion can be challenging to locate because it is often defined as the midpoint between two mandibular angle contours. Similarly, Basion and Orbitale are typically regarded as difficult to detect and unreliable in cephalometric analysis. The identification of upper and lower incisor landmarks is further complicated by open root apices and dental crowding, common in patients with malocclusion, which can reduce the precision of AI detection [9,10]. Considering the significant advancements in the application of AI in cephalometry, while remaining mindful of its limitations, THERE project was based on the idea of creating an automated AI-based tool for the production of cephalometric tracings. Although the offer of this type of technology is growing, until now there was no European programme dedicated to this application. The main competitor, WEBCEPH, is based in Korea. To emphasize the uniqueness and European character of our software, we decided to implement, among the various analyses available, an output that conforms to the principles of the EBO (European Board of Orthodontics), which represents the highest standard of excellence for European orthodontic clinicians.

Therefore, the primary objective of this paper is to develop a *general* and *accurate* model capable of handling data from heterogeneous sources. To this end, we implemented *THERETransKey*, a model that integrates the strengths of Convolutional Neural Networks (CNNs) with the Transformer architecture, drawing inspiration from the work of Yang et al. [11].

THERETransKey forms the core of the web-based application THERE (auTomatic HElpeR for cEphalometry), a fully open-access system requiring neither membership nor payment. It leverages THERETransKey to generate cephalometric tracings in accordance with the principles set by the European Board of Orthodontics. The platform is accessible online without the need to install software locally and is compatible with radiographs from any X-ray machine, regardless of brand or configuration. This aspect not only meets the approval of clinicians but is also a strategic element in the development of the system itself.

Moreover, each cephalometric analysis performed and, if necessary, corrected by a clinician contributes to improving the model's accuracy. Corrected tracings are collected and used to fine-tune THERETransKey, thereby progressively enhancing its reliability. For the tool to be effective in clinical practice, it must also be intuitive and user-friendly, ensuring it can genuinely assist clinicians in their workflows.

Another significant challenge in training THERETransKey is the scarcity of sufficiently large and diverse datasets. In 2024, Hendrickx et al. [12] conducted a systematic review with meta-analysis and found that training datasets in the field ranged from as few as 15 images to a maximum of 1983 images [13], while test sets ranged from just 4 to 400 images [14]. Through the deployment of THERE, we aim to contribute to the creation of a substantially larger and more diverse dataset, thereby improving the generalizability and performance of the model. We have committed to making this dataset available to the research community to foster the development of a large, open, and anonymized resource. This openness may encourage collaborations with external institutions and research teams and facilitate integration with models from other projects in future versions of THERE. By collecting data from a range of clinical environments using diverse radiographic equipment, we also aim to renew interest in the challenge of automatic anatomical landmark detection.

Building on preliminary results in landmark detection [15], this paper introduces the latest version of THERE. It outlines the architecture and training of the THERETransKey model and presents findings from a user experience questionnaire used to evaluate the system's usability and clinical utility.

To summarize, the main objective of the paper is the implementation of THERETransKey model which is able to accurately locate anatomical landmarks for the cephalometry analysis, achieving performance at least in line with the state of the art. Additionally, we also describe the creation of a dataset containing a significant number of images having the most heterogeneous characteristics, used to train THERETransKey, and the development of THERE, an easy-to-use and open tool for cephalometric analysis.

The paper is organized as follows. Section 2 introduces THERE, explaining how images are collected to build the dataset and how usability challenges are addressed. The next sections discuss how we handled these two aspects of THERE. In particular, Section 3 presents the dataset collected by THERE and Section 4 presents the methodology and results of the user experience questionnaire. Finally, the last sections focus on the main objective of the paper. Sections Section 5 describes the THERETransKey model and its training process, while Section 6 evaluates the model's performance. Finally, Section 7 concludes the paper with final remarks.

2. THERE Web application

The front-end of the application was previously presented in [15]. In this paper we briefly summarize the main application's features. THERE was developed using the Python Flask Framework, with two main objectives:

simplicity of use: the application workflow, depicted in Fig. 1, is linear and designed to be as user-friendly as possible allowing even users with limited technical expertise to operate it easily.

anonymity: the application ask users to upload teleradiographs from patients and saves them together with the coordinates defined by the user, so the application must ensure the highest privacy.

The first objective has been achieved by developing an application following three main steps, described below and depicted in Fig. 1, each one implemented in its main view:

- ① **Upload of the radiography:** managed by *Home* view, the landing page, reachable at https://there.ai.unife.it/;
- ② **Calibration:** the user must specify the scale of the radiography in the *Calibrate* view, to allow the computation of correct measurements;
- ③ **Visualization and correction of the cephalometry:** managed by the *Dashboard* view, where the anatomical landmarks are shown and the user can both modify the coordinates of the landmarks and read the measurements necessary for the diagnosis.

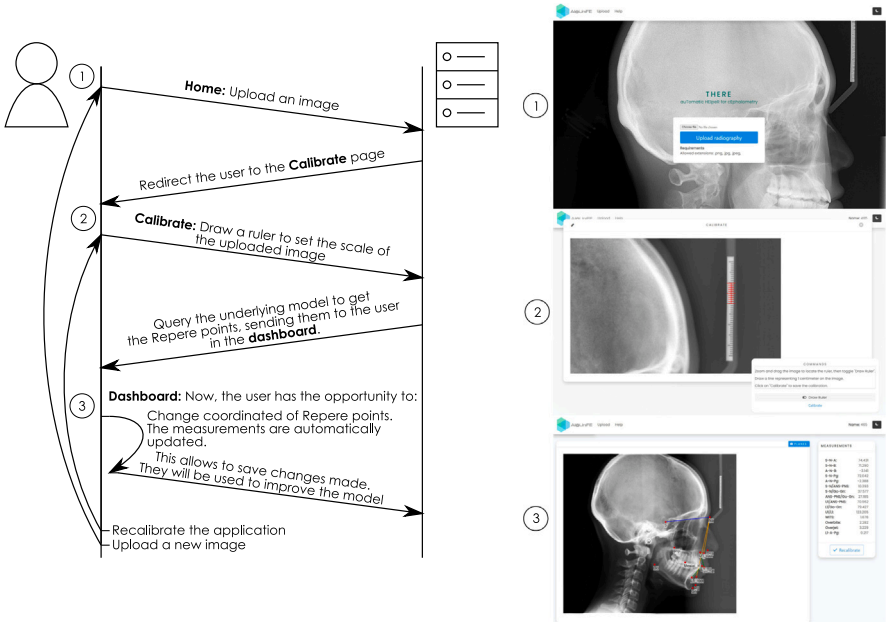

Fig. 1. Workflow of the application and screenshots for the three pages. The numbers refer to the different views, showing their interaction. A screenshot for each view is also reported.

For each view (①, ② and ③), Fig. 1 shows when the view is used, along with a corresponding screenshot. The application also presents a video tutorial showing the functionalities and the steps to perform to use THERE.

For the second objective, ensuring anonymity, the images are saved using a numeric sequential ID created by the DBMS at upload time, no sensible data of the patients is collected by the application. THERE only stores the anonymous image with the corresponding calibration information and the coordinates of the landmarks.

The back-end of the application allows administrators to access a fourth view not visible to users where it is possible to check the status of the application, download the uploaded images and launch the update of the model using the new radiographies collected. Currently, this fine tuning process is launched manually in order to define the best hyper-parameters, such as the starting learning rate and the number new radiographies necessary to trigger the training, to use during this phase. In the future, once defined the parameters, this process will be automated. In this view, the administrator can also collect statistics about the model performance in terms of how many radiographies have been corrected and the size of the corrections.

3. Dataset

One of the results of the project is the creation of a large, heterogeneous and open dataset, which can be used to help the future improvements of THERE and other systems trained on the here collected data. The dataset currently comprises 3153 images of anonymized lateral teleradiographs of the head of different patients. Images are taken using different devices, present different size, levels of saturation, colour and brightness, and refer to patients with heterogeneous characteristics. To enhance the generalizability and clinical applicability of AI models, it is recommended that datasets be expanded to include a more diverse range of data [12]. Following this approach, all acquisition devices, methods, and patients (varying in sex, age, and presence of orthodontic devices) were considered. In this way, the dataset should better reflects working conditions of clinicians and it should also force the networks trained on it to better generalize and hopefully being able to work also with possible new different devices not saw during training.

The dataset is composed of 183 PNGs and 2970 JPGs images of size spanning between 2685 × 2232 and 316 × 224 pixels, most of them in RGB format. Images' bit depth spans from 8 (there are 56 greyscale images) and 32. The age of the patients ranges from adolescence to old age, and they could have dental devices or implants.

Each image is labelled with the coordinates in pixels of the 14 anatomical landmarks considered in this project, namely:

1. **S:** Central point of the Sella Turcica.
2. **N:** Deepest antero-posterior point of the nose-frontal suture.
3. **PNS:** Radiological point determined by the perpendicular drawn from the apex of the pterygopalatine fossa to the bispinal plane.
4. **ANS:** Most anterior bony point of the anterior nasal spine.
5. **A:** Most posterior point of the anterior concavity of the maxillary alveolar process.
6. **B:** Most recessed point of the anterior concavity of the mandibular alveolar process.
7. **Pg:** Most anterior point of the chin symphysis contour.
8. **GN:** Lowest point of the chin symphysis contour.
9. **Go:** Geometric point constructed at the point where the tangent to the ascending branch of the mandible meets the plane of the mandible.
10. **U1 root:** Root apex edge of the upper central incisor.
11. **U1 tip:** Incisal edge of the upper central incisor.
12. **L1 root:** Root apex edge of the lower central incisor.
13. **L1 tip:** Incisal edge of the lower central incisor.
14. **Mesial:** apex of the mesiovestibular cusp of the upper first molar.

The position of the considered points on the skull is shown in Fig. 2. All teleradiographs are captured with the patient facing the right side of the image.

The dataset can be requested using the form available at https://there.ai.unife.it/contactus.

4. Validation of THERE Web application

To validate the effectiveness and practical impact of the application we prepared a questionnaire to be administered to clinicians and students of the Department of Orthodontics of the University of Ferrara. The sample consisted of 10 specialists with more than 3 years of experience after completing the Postgraduate School, 57

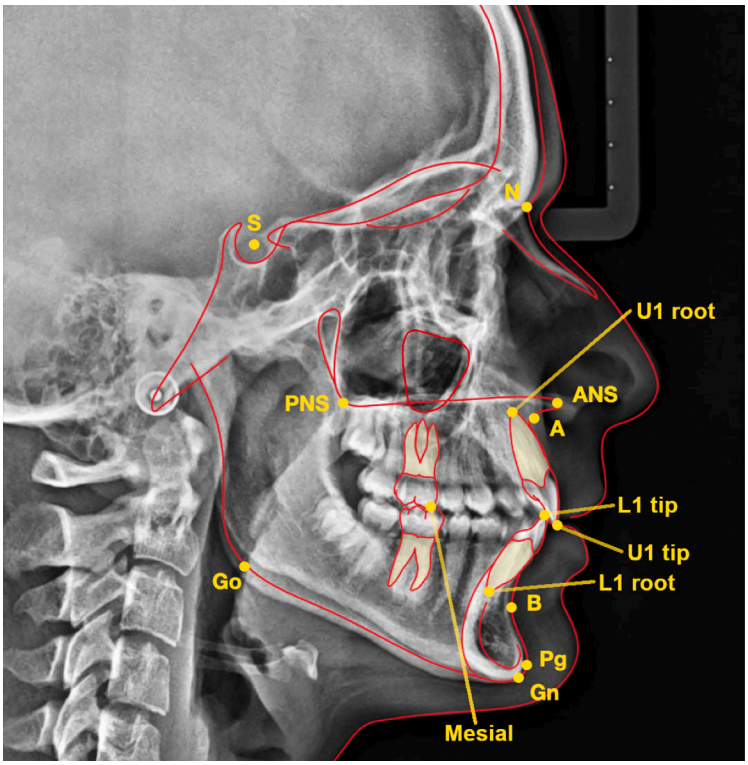

Fig. 2. Anatomical landmarks considered in the project, i.e., **S:** Central point of the Sella Turcica, **N:** Deepest antero-posterior point of the nose-frontal suture, **PNS:** Radiological point determined by the perpendicular drawn from the apex of the pterygopalatine fossa to the bispinal plane, **ANS:** Most anterior bony point of the anterior nasal spine, **A:** Most posterior point of the anterior concavity of the maxillary alveolar process, **B:** Most recessed point of the anterior concavity of the mandibular alveolar process, **Pg:** Most anterior point of the chin symphysis contour, **GN:** Lowest point of the chin symphysis contour, **Go:** Geometric point constructed at the point where the tangent to the ascending branch of the mandible meets the plane of the mandible, **U1 root:** Root apex edge of the upper central incisor, **U1 tip:** Incisal edge of the upper central incisor, **L1 root:** Root apex edge of the lower central incisor, **L1 tip:** Incisal edge of the lower central incisor, and **Mesial:** apex of the mesiovestibular cusp of the upper first molar.

residents, i.e., postgraduate students, from the Postgraduate School of Orthodontics, and 12 undergraduate students of the Dental School. In particular, residents are qualified doctors who are studying and practising to become specialized in orthodontics. The participants in the statistical sample were not involved in the design and implementation of the application; therefore, they were unfamiliar with its workflow and functionality when completing the questionnaire. They were only aware of the application's purpose.

The questionnaire is based the Post-Study System Usability Questionnaire (PSSUQ) Version 3 [16,17], also called Computer System Usability Questionnaire (CSUQ), and adds two more questions regarding the video tutorial available in the system. The choice of PSSUQ is due to the fact that PSSUQ has a high reliability [18], even with a small number of participants: for example, Tullis and Stetson [19] found that a sample size of 12 answers generated the same results as a larger sample size 90% of the time. As said above, in total, we collected 79 responses, ensuring a meaningful analysis. Focusing on the subsamples, residents and students have at least 12 responses, which is sufficient for the statistical analysis to be considered significant. Considering the subsample of specialists, we unfortunately only collected 10 answers, but the low standard deviation of the answers of each question (see Appendix) shows the unanimity of positive opinion.

PSSUQ Version 3 consists of 16 questions following a 7-point Likert Scale, i.e., every question has 7 options to choose from on a scale between Strongly Agree (score 1) to Strongly Disagree (score 7). We also added 2 extra questions following the same 7-point Likert Scale and asking about user-friendliness and usefulness of the tutorial included in the application. Table A.5 in the Appendix provides the list of the questions and reports the detailed distribution of the scores for each of them.

Table 1
Score obtained in each sub-scale averaged on the collected answers, the lower the better. Column "Overall" reports scores computed on all the 79 answers, while the following columns report the scores computed on respectively specialists with more than 3 years of experience, residents from the School of Specialization in Orthodontics, and undergraduate students.

Scale	Overall	Specialists	Residents	Students
System Usefulness	1.58	1.27	1.50	1.75
Information Quality	1.76	1.53	1.72	1.67
Interface Quality	1.70	1.53	1.65	1.56
PSSUQ score	1.67	1.41	1.62	1.67
System Tutorial score	1.66	1.40	1.57	1.92
THERE score	1.67	1.41	1.61	1.70

The PSSUQ provides an "overall score" by averaging the scores of all 16 items. Additionally, it has three sub-scales: System Usefulness, Information Quality, and Interface Quality. We also added 2 new scores: the first, called System Tutorial score, considering the effectiveness and usefulness of the tutorial, the second is the THERE Overall score, computed averaging the scores of all the question of the questionnaire. More detail can be found in Appendix. All the scores range between 1 and 7, the lower the better.

Table 1 reports the sub-scales and overall scores, calculated by averaging the collected responses, considering both the entire sample and sub-samples based on experience levels. As can be seen, all scores are significantly below 2 (with standard deviation around 1.00 or lower, see Appendix for more details), with specialists that assigned the best scores on average, following by residents and students. These scores demonstrate the system's ability to meet the requirements of user-friendliness, which is perceived more with increasing years of experience, but always remains very high. This result is also highlighted

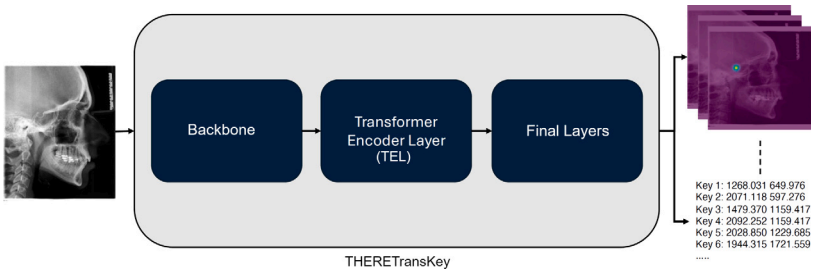

Fig. 3. Architecture of THERETransKey. The whole model, the grey box, consists of three sub-models (blue boxes): a Backbone, a Transformer Encoder Layer (TEL) and a set of Final Layers computing the prediction of the model.

by the System Tutorial score. From the collected answers, the users, although appreciating the tutorial included in the application, did not need to follow it in order to use THERE effectively. The simplicity of THERE is expected to facilitate widespread adoption, ensuring a continuous flow of newly labelled images. This will contribute to an ever-growing dataset and ongoing improvements to the underlying model.

Finally, in our questionnaire, every question was associated with a text field allowing the user to comment the score given to the answer. Moreover, a final free text field was added to give the users the possibility to write general suggestions or comments about THERE. We collected only three comments suggesting interesting features to add and one, which we report with some pride, congratulating us for the good work done. These results and comments stimulate us on to continue working to constantly improve the quality of THERE.

Therefore, we can conclude that the need of an implementation of an easy-to-use and open tool for cephalometric analysis has been fulfilled. This will help clinicians to use the tool, increasing even more rapidly the size of the dataset possibly allowing for a better model for the anatomical landmark detection.

5. Anatomical landmarks detection model

In the previous sections we described the front-end of the application and the dataset collected by the system. In this section we will focus on the main objective of this paper, i.e., on the model defined to perform the anatomical landmarks detection for the automatic cephalometry.

In the context of orthodontics, precise prediction of cephalometric points is extremely important for an accurate diagnosis. Numerous models based on Convolutional Neural Networks (CNNs) have been tested starting from the results of [15]. However, to address this challenge more effectively, we finally adopted a hybrid approach, combining the capabilities of CNNs with the Transformer architecture. This model was inspired by the work of Yang et al. [11] presented in the paper "TransPose: Keypoint Localization via Transformer", which proposes integrating a Transformer architecture with a CNN for human pose prediction. Specifically, the architecture we present in this paper, named *THERETransKey*, consists of three main components, as shown in Fig. 3: a backbone CNN model, a Transformer encoder, and final layers for generating heatmaps and landmarks. The CNN is used to extract significant features from the input image. The Transformer encoder captures global relationships between the features extracted from the preceding convolutional layer. Finally, the final layers increase the resolution of the feature maps and produce the model's output, namely the heatmaps and corresponding predicted landmarks.

The heatmaps are particularly important as they visually highlight where the network focuses most for keypoint prediction, indicating the image areas the network considers most relevant. This enhances interpretability and explainability of the model's decision. This factor is extremely important in the medical field because it can increase the user's trust in the results.

5.1. Transformer encoder

The output from the backbone is a set of feature maps that represent the features extracted from the image that must be analysed to extract information useful for the construction of the heatmaps. This task is demanded to the Transformer Encoder Layer (TEL), which implements the encoder part of a classical Transformer model. The connection between the backbone and TEL is implemented by:

1. a 1×1 convolution, which is used to adjust the number of channels in the feature maps returned by the backbone to match the TEL's dimension of the vectors encoding each input token;
2. a flattening and permutation step, where the spatial dimensions (height and width) are multiplied to obtain a single sequence dimension, and the initial tensor dimensions are permuted to fit the TEL's required format.

In TEL we use positional encodings, which are added to the embedding vectors before passing them through the encoder layers [20]. Positional encodings provide information about the spatial coordinates of the image, allowing the Transformer to understand the spatial arrangement of elements. In the THERETransKey model, TEL uses Sinusoidal positional encoding, which exploits sinusoidal functions to generate encoding values for each position. This encoding is used to provide information about the sequence elements' positions. TEL was implemented using the TransformerEncoder and TransformerEncoderLayer classes from the PyTorch library. The encoder layer is created with the following parameters:

- `d_model`: embedding dimension, set to 512.
- `nhead`: number of heads in the multi-head attention mechanism, set to 8.
- `dim_feedforward`: dimension of the feed-forward network, set to 1024.
- `activation`: activation function used, in this case, GELU.

TEL is built by stacking multiple encoder layers using PyTorch's TransformerEncoder class. We set this number to 4

5.2. Final layers

The final layers of the model produce the network's output, i.e., the heatmaps indicating the keypoints' predicted positions. Before reaching the final layers, the output tensor from the encoder is reordered and resized to be processed correctly by the following layers. After this transformation, the tensor passes through a deconvolution layer, which increases the spatial dimension of the feature map.

The deconvolution layer, which serves as the final layer, reduces the number of tensor channels to 14, corresponding to the keypoints to be predicted. This is achieved by sliding a 1×1 kernel with a stride of 1 and padding set to 0, allowing the modification of the number of channels without altering the spatial dimensions. The deconvolution layer's output is then passed through a sigmoid function, which restricts

Table 2
THERETransKey best performance for each combination of image size, loss function, and backbone model. In bold the best overall result, in italic the best result for each size.

Backbone	Loss	Loss value	PCK@0.5%	PCK@1%	PCK@5%
256					
HRNet based	*MSE*	*0.492*	*10.50*	*61.79*	*93.96*
HRNet based	Huber	0.385	11.73	61.35	93.16
HRNet based	Dice+MSE	0.246	12.14	61.08	92.68
HRNet based	Dice+Huber	0.201	10.73	59.78	91.72
ResNet based	Dice+Huber	0.041	10.01	40.93	97.32
ResNet based	Huber	0.096	9.12	38.56	97.18
ResNet based	MSE	0.078	8.34	37.39	96.53
ResNet based	Dice+MSE	0.068	7.89	31.53	95.08
EfficientNetV2-S	MSE	0.180	1.31	11.68	95.54
EfficientNetV2-S	Dice+MSE	0.024	1.01	11.13	96.68
EfficientNetV2-S	Huber	0.050	1.03	10.92	95.86
EfficientNetV2-S	Dice+Huber	0.036	1.12	10.83	96.06
512					
ResNet based	**Dice+Huber**	**0.137**	**49.73**	**73.98**	**93.26**
ResNet based	Dice+MSE	0.124	45.23	55.13	92.81
ResNet based	Huber	0.090	45.10	53.43	91.72
EfficientNetV2-S	Dice+Huber	0.004	11.37	28.71	99.49
EfficientNetV2-S	Huber	0.012	11.16	28.11	99.32
EfficientNetV2-S	Dice+MSE	0.034	10.89	27.01	98.84
EfficientNetV2-S	MSE	0.020	10.77	26.79	99.35
ResNet based	MSE	0.238	19.03	21.70	89.32
1024					
EfficientNetV2-S	*Huber*	*0.037*	*20.26*	*58.96*	99.08
EfficientNetV2-S	Dice+Huber	0.024	17.74	56.61	99.28
ResNet based	MSE	0.827	33.52	54.89	78.92
EfficientNetV2-S	MSE	0.050	17.34	54.84	99.17
EfficientNetV2-S	Dice+MSE	0.067	15.76	51.62	98.98
ResNet based	Dice+Huber	0.357	24.88	43.28	62.47
ResNet based	Dice+MSE	0.727	24.08	42.99	68.87
ResNet based	Huber	0.645	24.92	42.83	63.95

the values to the $[0, 1]$ range, transforming the heatmaps into representations of the corresponding keypoint's presence at each correct spatial location. Therefore, for each image, there are one heatmap for each keypoints to predict. To obtain the model's predicted landmark coordinates, the heatmaps pass through a function that uses the argmax operation to find the coordinates with the highest value (the hottest point) on each heatmap.

6. Experiments

We now present and discuss the results obtained during the model's training process, with particular attention to the graphs related to loss and evaluation metrics. Then, the model's accuracy will be analysed through a series of concrete prediction examples on radiographic images, highlighting the strengths and weaknesses of this approach. Finally, based on these observations, some potential improvements will be proposed to further optimize the model's performance, reducing the likelihood of errors during landmark prediction.

6.1. Dataset preprocessing

We conduct experiments on the current version of THERE dataset, described in Section 4. Before training, the images underwent a series of essential transformations. First, they were resized to a fixed dimension. To determine the optimal resize dimension, we tested three different settings: 256×256, 512×512, and 1024×1024. Then, the image were converted into greyscale and pixels were normalized to bring the pixel values into the range $[0, 1]$. Similarly, the coordinates of the keypoints were also normalized to the $[0, 1]$ range, ensuring consistency between the keypoint positions and image dimensions, eliminating discrepancies that could arise from different image sizes. The dataset was split into training, validation and test sets in a 0.70-0.15-0.15 ratio. Additionally, data augmentation was applied through random image rotation and random application of filters to adjust contrast or brightness.

6.2. Backbone model

THERETransKey exploits a backbone based on convolutional layers to extract relevant features from the radiographies. In our experiments we considered three different architectures for the backbone: (1) an architecture using residual blocks based on ResNet50 [21], (2) an architecture using High Resolution modules based on HRNet W32 [22,23], and (3) EfficientNetV2-S [24]. The objective of using the listed backbones is to create a small and lightweight model. In particular, to create a model as small as possible, we extracted from ResNet50 the first two residual blocks and from HRNet32 the first High Resolution block. Depending on the backbone and the image size, THERETransKey's parameters number ranges between 14.839M for backbone (1) and image size of 256×256, and 48.180M for backbone (2) and image size of 1024×1024. All the models used as backbone were pre-trained on ImageNet [25] and then fine-tuned on the THERE dataset during the training of THERETransKey model. Since we used a network pre-trained on the ImageNet dataset, which is composed of colour images with three channels (RGB), we modified the first convolutional layer to accept greyscale input, which consists of a single channel. This was done by adjusting the first convolutional layer and calculating the mean of the RGB channel weights, which was then applied to the new convolutional layer.

6.3. Training process

The training was guided by the Adam optimizer with Learning Rate Decay, that dynamically reduced the learning rate by a factor of 0.1 whenever the optimizer stops improving for two consecutive epochs. The initial learning rate was set to 0.0001. Finally, Early Stopping was used with patience set to 8 epochs. Number of epochs was set to 100. The training has been performed on a Linux machine equipped with 2 AMD EPYC 9124 16-Core and one GPU NVIDIA H100. During the training phase, real heatmaps are initially generated from the

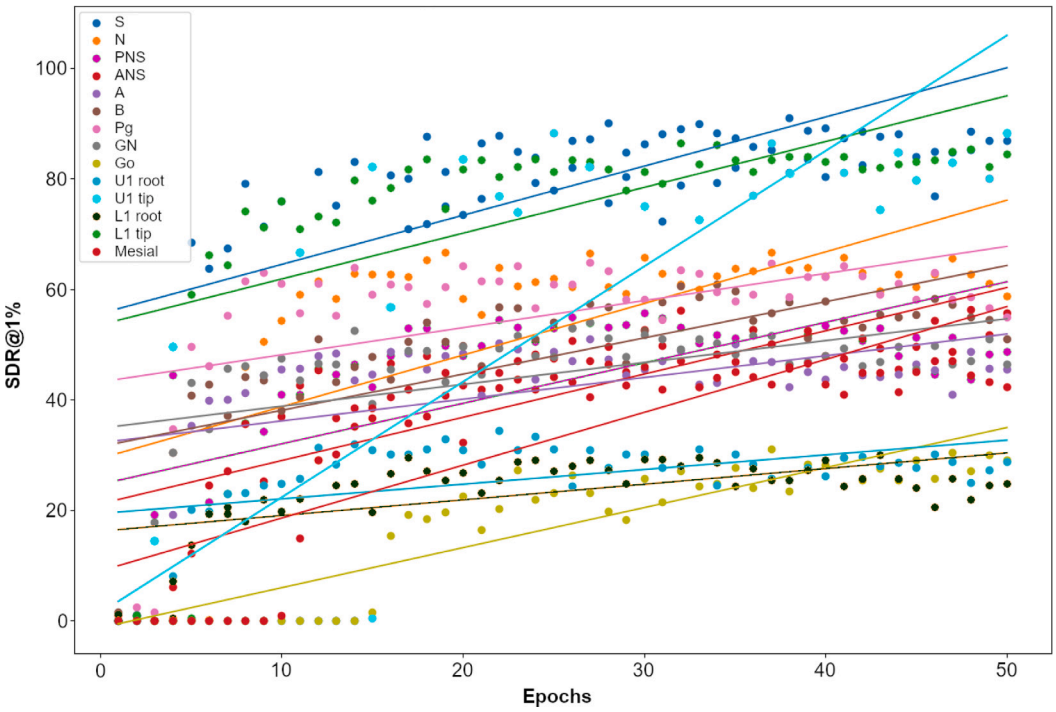

Fig. 4. Successful Detection Rate during training for each keypoint, namely, **S:** Central point of the Sella Turcica, **N:** Deepest antero-posterior point of the nose-frontal suture, **PNS:** Radiological point determined by the perpendicular drawn from the apex of the pterygopalatine fossa to the bispinal plane, **ANS:** Most anterior bony point of the anterior nasal spine, **A:** Most posterior point of the anterior concavity of the maxillary alveolar process, **B:** Most recessed point of the anterior concavity of the mandibular alveolar process, **Pg:** Most anterior point of the chin symphysis contour, **GN:** Lowest point of the chin symphysis contour, **Go:** Geometric point constructed at the point where the tangent to the ascending branch of the mandible meets the plane of the mandible, **U1 root:** Root apex edge of the upper central incisor, **U1 tip:** Incisal edge of the upper central incisor, **L1 root:** Root apex edge of the lower central incisor, **L1 tip:** Incisal edge of the lower central incisor, and **Mesial:** apex of the mesiovestibular cusp of the upper first molar.

keypoints coordinates in the ground truth, allowing alignment with the model-generated heatmaps.

6.4. Choice of the loss function

To choose the best loss function to guide the training process we considered Mean Squared Error (MSE) and Huber Loss alone and in combination with the Dice Loss. In the former case, the final loss is computed as $L = 0.5 \cdot L_B + 0.5 \cdot L_{Dice}$, where L_B can be MSE or Huber. In particular, the Dice Loss [26], based on the Dice coefficient, is used to measure the overlap of the heatmaps built by the model and those built using the ground truth. The Huber Loss [27] is a loss based on the MSE, which uses a squared term if the absolute error between the coordinates of the keypoints falls below a fixed parameter delta and a delta-scaled L1 term otherwise. The use of the L1 term is used to make the loss less sensitive to outliers than MSE. In our experiments we considered a delta value of 1.

6.5. Results

We use the Probability of Correct Keypoint (PCK) [28] to evaluate the quality of the model. In particular, PCK considers a keypoint as correct if the coordinates returned by the model falls below a fixed distance from the position specified by the ground truth. We consider three different distances, namely 0.5%, 1%, and 5% of the image size. Table 2 reports the results obtained from THERETransKeyconsidering all the combinations of image input size (256 × 256, 512 × 512, and 1024 × 1024), different backbones and loss functions.

HRNet is the best backbone, among the considered models, for input size of 256 × 256, followed by ResNet, while EfficientNet performs significantly worse. For larger images HRNet became impossible to train with batches of a significant size because it saturated the available memory in the machine on which it was trained. By reducing the batch size to 2, with images of 512 × 512 THERETransKey using HRNet quickly tended to overfitting, while with larger images it needed at least 10 h per epoch to be trained. ResNet achieved the best results considering images of 512 × 512, with the best model obtained in our experiments trained using Dice+Huber Loss. With images of 1024 × 1024 ResNet struggles more to reach good results, while EfficientNet obtained significantly good results, which however are far from the top results.

Table 3
SDR@1% for each keypoint achieved by the THERETransKey best model on the test set.

1 S	2 N	3 PNS	4 ANS	5 A	6 B	7 Pg
89.64	81.40	72.73	75.69	72.09	71.67	73.57
8 GN	9 Go	10 U1 root	11 U1 tip	12 L1 root	13 L1 tip	14 Mesial
74.00	62.79	62.58	91.12	52.01	89.01	67.44

In detail, the best model, i.e., considering backbone based on ResNet, with image size of 512 × 512 trained on Dice+Huber Loss, has 16.412M parameters and achieved a loss value of 0.13738 on the test set. Table 3 reports SDR@1% for each single keypoint.

Fig. 4 shows the SDR@1% trend on the validation set for each keypoint over the training epochs. In general, an increasing trend can be observed for the metric, with some points being predicted more accurately than others. For example, Point 11 (U1 tip, incisal edge of the upper central incisor) appears to be the easiest to locate, reaching an SDR of more than 91% in the final epochs, indicating that the model can detect this point with greater precision. Conversely, Points 10 (U1 root, root apex edge of the upper central incisor) and 12 (L1 root, root apex edge of the lower central incisor) are more difficult to predict, as shown by the less steep curves and lower SDR values (an

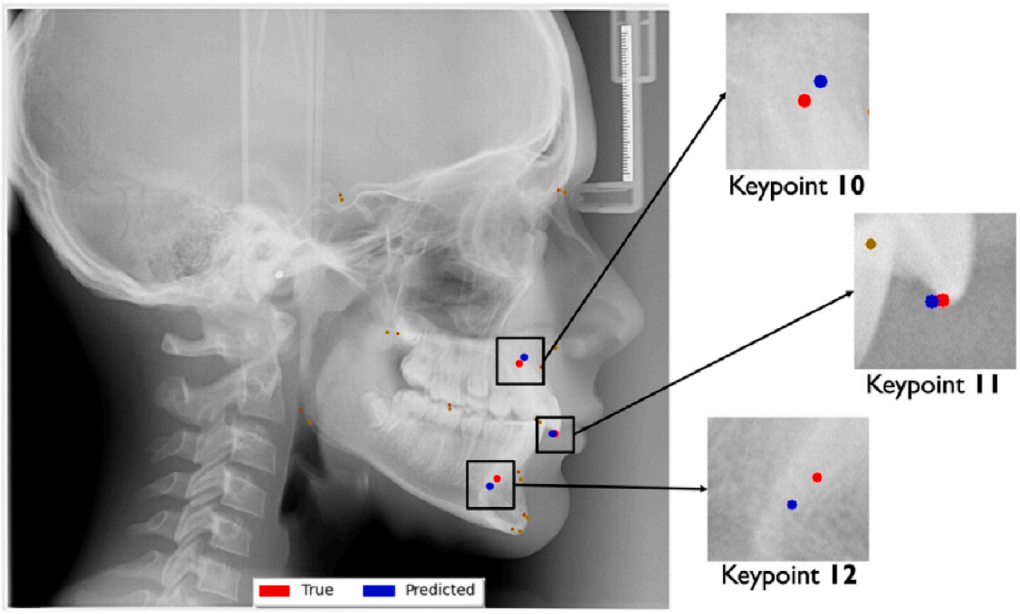

Fig. 5. Localization of keypoints 10 (L1 tip), 11 (Pg), and 12 (GN).

example of the three points can be seen in Fig. 5). This result may be due to Points 10 and 12 being located in anatomical regions with fewer visual details compared to others, such as those along the cranial edge. The lack of visual details complicates the model's ability to make precise predictions. Additionally, the position of these landmarks tends to vary slightly across images, making it more challenging for the network to learn a stable or consistent position. Fig. 4, therefore, allows for the clear identification of strengths and weaknesses, pinpointing which landmarks require further optimization to improve the network's overall performance.

An interesting aspect to observe is the evolution of the heatmaps generated by the model during the training epochs. Fig. 6 shows an example from a validation image, illustrating how the heatmaps corresponding to keypoint 1 (Sella Turcica) become increasingly precise as training progresses. In the initial epochs the network produces highly diffuse heatmaps, indicating limited accuracy in point identification. However, as training progresses, a noticeable increase in precision can be observed, even though some dispersion remains until clear localization is achieved in the final epochs. The visual example of heatmaps confirms the trend observed in Fig. 4.

6.6. Prediction analysis

As reported in the previous section, after the model's training, the test set was used to evaluate its performance numerically and by visualizing the results to understanding its final behaviour. The findings revealed generally good performance, although the network struggled with blurry, overly bright (overexposed) or dark images, which led to the loss of possible reference points, as shown in Fig. 7. In these cases, the error could reach values of 60 mm, indicating that image quality significantly impacts network performance; with low-quality images or those lacking clear details, the predictions become less accurate.

To further test the performance of the model, we analysed the results obtained considering the images uploaded in the THERE after the training of the network during the standard use of the application. Overall, the clinician had to correct at least one landmark position returned by the model in the 67.86% images. This number seems high, however it also considers cases where the clinician changed one single keypoint of less than 1 mm. Considering only the images with coordinates corrected by the clinician, the model presents an average error in millimetres on the x coordinates of 1.05 ± 1.43 and on the y coordinates of 1.06 ± 1.46 computed on all the keypoints, with an average distance of 1.49 ± 2.04 mm. Table 4 shows the average distance for each landmark computed on the images with corrected coordinates. If we compare the results with those presented in our preliminary work [15] we can see an improvement from 6.3 ± 4.6 mm to 1.49 ± 2.04 mm.

Table 4
Average distance between the positions of the landmarks in millimetres returned by THERE and the position corrected by the user.

1	2	3	4	5	6	7
S	N	PNS	ANS	A	B	Pg
0.64	2.76	1.11	1.19	2.15	1.95	1.30
8	9	10	11	12	13	14
GN	Go	U1 root	U1 tip	L1 root	L1 tip	Mesial
1.25	2.83	1.60	1.10	2.06	0.83	2.68

Finally, the data collected from the use of the application by the clinicians allows us to compare THERETransKey with the results reported in [12]. We computed:

Mean Relative Error (MRE): This metric measures how far, on average, the detected points are from the reference points. The lower the result, the higher the prediction accuracy.

Successful Detection Rate (SDR): This metric measures the percentage of landmarks correctly detected within a certain threshold. It is similar to PCK, but it considers Euclidean distance in millimetres from the predicted and the real coordinates of the keypoints. We set a threshold of 2 mm as in [12].

Our model achieved an average MRE of 1.06 ± 1.45 and an SDR@2 mm of 81.87%, which is in line with the results obtained by the systems considered within Test 1, i.e., using a public benchmark dataset from the IEEE International Symposium on Biomedical Imaging 2015 grand challenge [29], and Test 2, i.e., systems tested on their

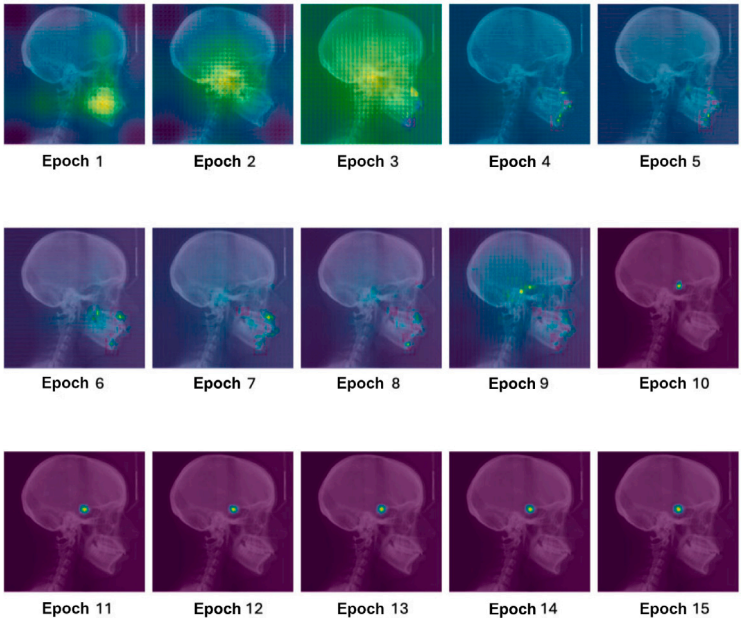

Fig. 6. Evolution of the heatmap for Point 1 (S, Sella Turcica).

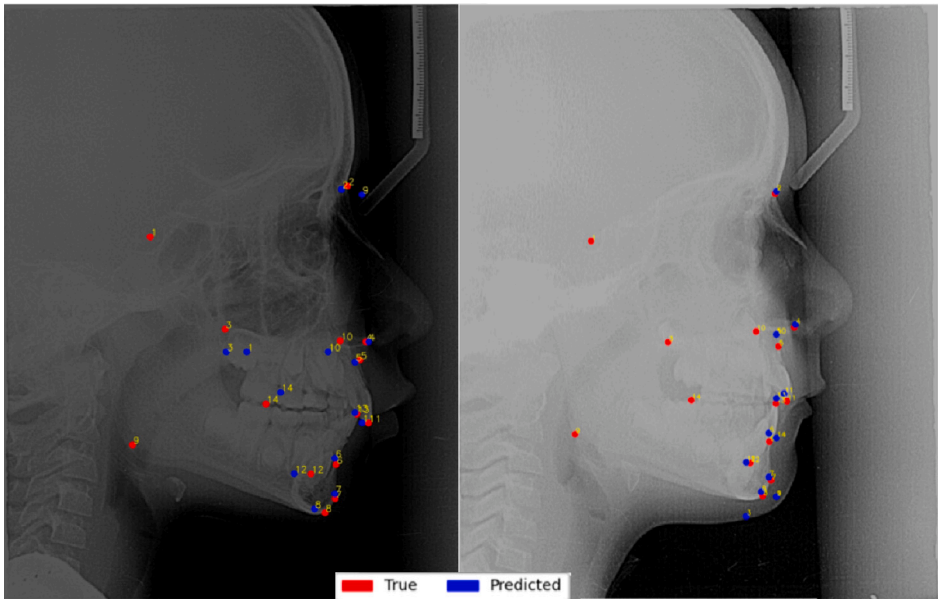

Fig. 7. Two examples of images of the test set with lowest precision. The example on the left is too dark, making difficult to follow the skin contours. The example on the right is overexposed, making difficult to analyse the bones edges. These conditions lead to significant errors in keypoint localization. For example, keypoint 9 should be positioned on the jaw, but due to difficulty in locating its edge, the model completely misplaces it in both figures.

own specific datasets which are characterized by different landmarks and different image conditions (size, quality, machine that took the radiography), that may significantly affect the performance of the network. For example, a network trained only on images taken from the same machine may work better on such images than a network trained on images taken by different machines, however it may struggle to generalize to different usage conditions. Moreover, the MRE results is also in line with the divergences that could occur considering manual landmarking performed by different experts [7,8], while SDR@2 mm is significantly higher than the rates of 79% reported in [6].

7. Conclusions and future work

In this paper we presented the evolution of THERE, firstly introduced in [15]. As proved by the validation of the system via the PSSUQ-based questionnaire, the THERE Web application allows clinicians to easily obtain the position of 14 anatomical landmarks necessary to perform a cephalometry by uploading a teleradiography which is analysed by an underlying neural network called THERETransKey. It also allows to easily correct these points and to visualize the measurements obtained from them. The assistance provided by THERE has proven

Table A.5
Distribution of the scores 1–7 over all the 79 answers assigned for each question of the questionnaire, the lower the better. For each score and question also the corresponding percentage is reported (rounded to the nearest integer — so values in rows may not sum exactly to 100%). The last column reports the average score for each question and its standard deviation.

Question #	Answer distribution – Count (%)							Avg ± std. dev.
	1	2	3	4	5	6	7	
1. Overall, I am satisfied with how easy it is to use this system.	49 (62%)	23 (29,1%)	4 (5,1%)	1 (1,3%)	1 (1,3%)	0 (0%)	1 (1,3%)	1,56 ± 0,98
2. It was simple to use this system.	51 (64,6%)	20 (25,3%)	5 (6,3%)	1 (1,3%)	0 (0%)	1 (1,3%)	1 (1,3%)	1,56 ± 1,05
3. I was able to complete the tasks quickly using this system.	50 (63,3%)	19 (24,1%)	6 (7,6%)	1 (1,3%)	0 (0%)	3 (3,8%)	0 (0%)	1,62 ± 1,11
4. I felt comfortable using this system.	46 (58,2%)	24 (30,4%)	6 (7,6%)	0 (0%)	0 (0%)	2 (2,5%)	1 (1,3%)	1,66 ± 1,13
5. It was easy to learn to use this system.	57 (72,2%)	12 (15,2%)	7 (8,9%)	0 (0%)	0 (0%)	1 (1,3%)	2 (2,5%)	1,54 ± 1,21
6. I believe I could become productive quickly using this system.	58 (73,4%)	11 (13,9%)	7 (8,9%)	0 (0%)	1 (1,3%)	0 (0%)	2 (2,5%)	1,52 ± 1,16
7. The system gave error messages that clearly told me how to fix problems.	27 (34,2%)	18 (22,8%)	26 (32,9%)	3 (3,8%)	0 (0%)	2 (2,5%)	3 (3,8%)	2,35 ± 1,44
8. Whenever I made a mistake using the system, I could recover easily and quickly.	38 (48,1%)	30 (38%)	9 (11,4%)	0 (0%)	0 (0%)	1 (1,3%)	1 (1,3%)	1,75 ± 1,03
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	47 (59,5%)	22 (27,8%)	6 (7,6%)	1 (1,3%)	1 (1,3%)	1 (1,3%)	1 (1,3%)	1,66 ± 1,12
10. It was easy to find the information I needed.	47 (59,5%)	22 (27,8%)	8 (10,1%)	0 (0%)	0 (0%)	1 (1,3%)	1 (1,3%)	1,62 ± 1,04
11. The information was effective in helping me complete the tasks.	47 (59,5%)	23 (29,1%)	7 (8,9%)	0 (0%)	1 (1,3%)	0 (0%)	1 (1,3%)	1,59 ± 0,98
12. The organization of information on the system screens was clear.	51 (64,6%)	18 (22,8%)	6 (7,6%)	2 (2,5%)	1 (1,3%)	0 (0%)	1 (1,3%)	1,58 ± 1,05
13. The interface of this system was pleasant.	47 (59,5%)	17 (21,5%)	12 (15,2%)	1 (1,3%)	0 (0%)	1 (1,3%)	1 (1,3%)	1,7 ± 1,11
14. I liked using the interface of this system.	48 (60,8%)	19 (24,1%)	9 (11,4%)	1 (1,3%)	0 (0%)	1 (1,3%)	1 (1,3%)	1,65 ± 1,09
15. This system has all the functions and capabilities I expect it to have.	46 (58,2%)	19 (24,1%)	8 (10,1%)	3 (3,8%)	1 (1,3%)	1 (1,3%)	1 (1,3%)	1,75 ± 1,19
16. Overall, I am satisfied with this system.	52 (65,8%)	17 (21,5%)	4 (5,1%)	3 (3,8%)	0 (0%)	2 (2,5%)	1 (1,3%)	1,63 ± 1,21
17. I was able to use the interface without the help of the tutorial	45 (57%)	20 (25,3%)	10 (12,7%)	1 (1,3%)	0 (0%)	3 (3,8%)	0 (0%)	1,73 ± 1,14
18. The tutorial is complete and sufficient to use the interface	52 (65,8%)	16 (20,3%)	8 (10,1%)	0 (0%)	1 (1,3%)	1 (1,3%)	1 (1,3%)	1,59 ± 1,12

effective as the model presents a good precision in the localization of the anatomical landmarks. Moreover, the use of the system in the workplace enables the automatic and transparent collection of correctly labelled images, validated by expert clinicians, without imposing additional workload on them. These labelled images are then incorporated into an ever-expanding dataset that will be used to constantly improve the performance of the system.

CRediT authorship contribution statement

Riccardo Zese: Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition. **Matteo De Maio:** Writing – review & editing, Validation, Resources, Methodology, Investigation, Data curation, Conceptualization. **Francesca Cremonini:** Writing – review & editing, Validation, Resources, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Research funded by the Italian Ministry of University and Research through PNRR - M4C2 - Investimento 1.3 (Decreto Direttoriale MUR n. 341 del 15/03/2022), Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Union under the NextGeneration EU programme". This work has been partially supported by the Spoke 1 "FutureHPC & BigData" of the Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC) funded by MUR Missione 4 - Next Generation EU (NGEU).

Appendix. Details on the validation of THERE Web application

The 16 questions are:

1. Overall, I am satisfied with how easy it is to use this system.
2. It was simple to use this system.
3. I was able to complete the tasks and scenarios quickly using this system.
4. I felt comfortable using this system.

Table A.6
Distribution of the scores 1–7 assigned by 10 specialists with more than 3 years of experience for each question of the questionnaire, the lower the better. For each score and question also the corresponding percentage is reported (rounded to the nearest integer – so values in rows may not sum exactly to 100%). The last column reports the average score for each question and its standard deviation.

Question #	Answer distribution – Count (%)							Avg ± std. dev.
	1	2	3	4	5	6	7	
1. Overall, I am satisfied with how easy it is to use this system.	8 (80%)	2 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,2 ± 0,42
2. It was simple to use this system.	7 (70%)	2 (20%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,4 ± 0,7
3. I was able to complete the tasks quickly using this system.	8 (80%)	1 (10%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,3 ± 0,67
4. I felt comfortable using this system.	7 (70%)	3 (30%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,3 ± 0,48
5. It was easy to learn to use this system.	9 (90%)	0 (0%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,2 ± 0,63
6. I believe I could become productive quickly using this system.	8 (80%)	2 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,2 ± 0,42
7. The system gave error messages that clearly told me how to fix problems.	5 (50%)	0 (0%)	4 (40%)	0 (0%)	0 (0%)	0 (0%)	1 (10%)	2,4 ± 1,9
8. Whenever I made a mistake using the system, I could recover easily and quickly.	8 (80%)	1 (10%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,3 ± 0,67
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	7 (70%)	3 (30%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,3 ± 0,48
10. It was easy to find the information I needed.	6 (60%)	3 (30%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,71
11. The information was effective in helping me complete the tasks.	7 (70%)	2 (20%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,4 ± 0,7
12. The organization of information on the system screens was clear.	7 (70%)	3 (30%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,3 ± 0,48
13. The interface of this system was pleasant.	6 (60%)	3 (30%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,71
14. I liked using the interface of this system.	6 (60%)	3 (30%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,71
15. This system has all the functions and capabilities I expect it to have.	6 (60%)	2 (20%)	2 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,6 ± 0,84
16. Overall, I am satisfied with this system.	9 (90%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,1 ± 0,32
17. I was able to use the interface without the help of the tutorial	7 (70%)	2 (20%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,4 ± 0,7
18. The tutorial is complete and sufficient to use the interface	8 (80%)	0 (0%)	2 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,4 ± 0,84

5 It was easy to learn to use this system.
6 I believe I could become productive quickly using this system.
7 The system gave error messages that clearly told me how to fix problems.
8 Whenever I made a mistake using the system, I could recover easily and quickly.
9 The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
10 It was easy to find the information I needed.
11 The information was effective in helping me complete the tasks and scenarios.
12 The organization of information on the system screens was clear.
13 The interface of this system was pleasant.
14 I liked using the interface of this system.
15 This system has all the functions and capabilities I expect it to have.
16 Overall, I am satisfied with this system.

We added 2 extra questions explicitly written for our application and following the same 7-point Likert Scale:

17. I was able to use the interface without the help of the tutorial.
18. The tutorial is complete and sufficient to use the interface.

The 3 sub-scales of the questionnaire are computed by :

- System Usefulness averaging items 1–6
- Information Quality averaging items 7–12
- Interface Quality averaging items 13–15

Note that the 16th item is used only in the global score. We added 2 new scores: the first, called System Tutorial score, considering the effectiveness and usefulness of the tutorial, computed as the average of questions 17 and 18; the second is the THERE Overall score, computed averaging the scores of all the question of the questionnaire. All the scores range between 1 and 7, the lower the better.

Tables A.5, A.6, A.7, and A.8 report the detailed distribution of the scores for each question collected respectively by all the people in the sample, specialists with more than 3 years of experience, residents from the School of Specialization in Orthodontics (Specialization Class in Dentistry) of the Department of Orthodontics, and undergraduate students.

Table A.7
Distribution of the scores 1–7 assigned by 57 residents for each question of the questionnaire, the lower the better. For each score and question also the corresponding percentage is reported (rounded to the nearest integer – so values in rows may not sum exactly to 100%). The last column reports the average score for each question and its standard deviation.

Question #	Answer distribution – Count (%)							Avg ± std. dev.
	1	2	3	4	5	6	7	
1. Overall. I am satisfied with how easy it is to use this system.	38 (66.7%)	15 (26.3%)	2 (3.5%)	1 (1.8%)	1 (1.8%)	0 (0%)	0 (0%)	1.46 ± 0.8
2. It was simple to use this system.	39 (68.4%)	14 (24.6%)	2 (3.5%)	1 (1.8%)	0 (0%)	0 (0%)	1 (1.8%)	1.47 ± 0.98
3. I was able to complete the tasks quickly using this system.	37 (64.9%)	14 (24.6%)	3 (5.3%)	1 (1.8%)	0 (0%)	2 (3.5%)	0 (0%)	1.58 ± 1.08
4. I felt comfortable using this system.	35 (61.4%)	17 (29.8%)	3 (5.3%)	0 (0%)	0 (0%)	2 (3.5%)	0 (0%)	1.58 ± 1.03
5. It was easy to learn to use this system.	43 (75.4%)	8 (14%)	4 (7%)	0 (0%)	0 (0%)	1 (1.8%)	1 (1.8%)	1.47 ± 1.14
6. I believe I could become productive quickly using this system.	45 (78.9%)	5 (8.8%)	5 (8.8%)	0 (0%)	1 (1.8%)	0 (0%)	1 (1.8%)	1.44 ± 1.09
7. The system gave error messages that clearly told me how to fix problems.	17 (29.8%)	16 (28.1%)	18 (31.6%)	3 (5.3%)	0 (0%)	1 (1.8%)	2 (3.5%)	2.37 ± 1.37
8. Whenever I made a mistake using the system. I could recover easily and quickly.	26 (45.6%)	22 (38.6%)	8 (14%)	0 (0%)	0 (0%)	1 (1.8%)	0 (0%)	1.75 ± 0.91
9. The information (such as online help. on-screen messages. and other documentation) provided with this system was clear.	35 (61.4%)	14 (24.6%)	5 (8.8%)	1 (1.8%)	1 (1.8%)	1 (1.8%)	0 (0%)	1.63 ± 1.05
10. It was easy to find the information I needed.	35 (61.4%)	15 (26.3%)	6 (10.5%)	0 (0%)	0 (0%)	1 (1.8%)	0 (0%)	1.56 ± 0.91
11. The information was effective in helping me complete the tasks.	37 (64.9%)	15 (26.3%)	4 (7%)	0 (0%)	1 (1.8%)	0 (0%)	0 (0%)	1.47 ± 0.78
12. The organization of information on the system screens was clear.	39 (68.4%)	10 (17.5%)	5 (8.8%)	2 (3.5%)	1 (1.8%)	0 (0%)	0 (0%)	1.53 ± 0.93
13. The interface of this system was pleasant.	35 (61.4%)	10 (17.5%)	10 (17.5%)	1 (1.8%)	0 (0%)	1 (1.8%)	0 (0%)	1.67 ± 1.02
14. I liked using the interface of this system.	38 (66.7%)	10 (17.5%)	7 (12.3%)	1 (1.8%)	0 (0%)	1 (1.8%)	0 (0%)	1.56 ± 0.98
15. This system has all the functions and capabilities I expect it to have.	34 (59.6%)	13 (22.8%)	5 (8.8%)	3 (5.3%)	1 (1.8%)	1 (1.8%)	0 (0%)	1.72 ± 1.13
16. Overall. I am satisfied with this system.	38 (66.7%)	11 (19.3%)	3 (5.3%)	3 (5.3%)	0 (0%)	2 (3.5%)	0 (0%)	1.63 ± 1.17
17. I was able to use the interface without the help of the tutorial	34 (59.6%)	16 (28.1%)	4 (7%)	1 (1.8%)	0 (0%)	2 (3.5%)	0 (0%)	1.65 ± 1.09
16. The tutorial is complete and sufficient to use the interface	40 (70.2%)	11 (19.3%)	4 (7%)	0 (0%)	1 (1.8%)	1 (1.8%)	0 (0%)	1.49 ± 0.98

Table A.8
Distribution of the scores 1–7 assigned by 12 students of the Department of Orthodontics of the University of Ferrara for each question of the questionnaire, the lower the better. For each score and question also the corresponding percentage is reported (rounded to the nearest integer — so values in rows may not sum exactly to 100%). The last column reports the average score for each question and its standard deviation.

Question #	Answer distribution – Count (%)							Avg ± std. dev.
	1	2	3	4	5	6	7	
1. Overall, I am satisfied with how easy it is to use this system.	4 (33,3%)	6 (50%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,83 ± 0,72
2. It was simple to use this system.	5 (41,7%)	5 (41,7%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,75 ± 0,75
3. I was able to complete the tasks quickly using this system.	5 (41,7%)	5 (41,7%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,75 ± 0,75
4. I felt comfortable using this system.	5 (41,7%)	4 (33,3%)	3 (25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,83 ± 0,83
5. It was easy to learn to use this system.	6 (50%)	4 (33,3%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,67 ± 0,78
6. I believe I could become productive quickly using this system.	6 (50%)	4 (33,3%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,67 ± 0,78
7. The system gave error messages that clearly told me how to fix problems.	5 (41,7%)	3 (25%)	4 (33,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,92 ± 0,9
8. Whenever I made a mistake using the system, I could recover easily and quickly.	5 (41,7%)	7 (58,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,58 ± 0,51
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	6 (50%)	5 (41,7%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,58 ± 0,67
10. It was easy to find the information I needed.	7 (58,3%)	4 (33,3%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,67
11. The information was effective in helping me complete the tasks.	4 (33,3%)	6 (50%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,83 ± 0,72
12. The organization of information on the system screens was clear.	6 (50%)	5 (41,7%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,58 ± 0,67
13. The interface of this system was pleasant.	7 (58,3%)	4 (33,3%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,67
14. I liked using the interface of this system.	5 (41,7%)	6 (50%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,67 ± 0,65
15. This system has all the functions and capabilities I expect it to have.	7 (58,3%)	4 (33,3%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,5 ± 0,67
16. Overall, I am satisfied with this system.	6 (50%)	5 (41,7%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,58 ± 0,67
17. I was able to use the interface without the help of the tutorial	4 (33,3%)	3 (25%)	5 (41,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2,08 ± 0,9
18. The tutorial is complete and sufficient to use the interface	5 (41,7%)	5 (41,7%)	2 (16,7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,75 ± 0,75

References

[1] H. Ye, Z. Cheng, N. Ungvijanpunya, W. Chen, L. Cao, Y. Gou, Is automatic cephalometric software using artificial intelligence better than orthodontist experts in landmark identification? BMC Oral. Health 23 (1) (2023) 467.

[2] H. Bao, K. Zhang, C. Yu, H. Li, D. Cao, H. Shu, L. Liu, B. Yan, Evaluating the accuracy of automated cephalometric analysis based on artificial intelligence, BMC Oral. Health 23 (1) (2023) 191.

[3] D. Forsyth, W. Shaw, S. Richmond, C. Roberts, Digital imaging of cephalometric radiographs, part 2: image quality, Angle Orthod. 66 (1) (1996) 43–50.

[4] G. Dot, F. Rafflenbeul, M. Arbotto, L. Gajny, P. Rouch, T. Schouman, Accuracy and reliability of automatic three-dimensional cephalometric landmarking, Int. J. Oral Maxillofac. Surg. 49 (10) (2020) 1367–1378.

[5] A. Polizzi, R. Leonardi, Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews, J. Dent. (2024) 105056.

[6] G. de Queiroz Tavares Borges Mesquita, W.A. Vieira, M.T.C. Vidigal, B.A.N. Travençolo, T.L. Beaini, R. Spin-Neto, L.R. Paranhos, R.B. de Brito Júnior, Artificial intelligence for detecting cephalometric landmarks: a systematic review and meta-analysis, J. Digit. Imaging 36 (3) (2023) 1158–1179.

[7] M.O. Lagravère, C. Low, C. Flores-Mir, R. Chung, J.P. Carey, G. Heo, P.W. Major, Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images, Am. J. Orthod. Dentofacial. Orthop. 137 (5) (2010) 598–604.

[8] J.-H. Kim, S. An, D.-M. Hwang, Reliability of cephalometric landmark identification on three-dimensional computed tomographic images, Br. J. Oral Maxillofac. Surg. 60 (3) (2022) 320–325.

[9] G. Bulatova, B. Kusnoto, V. Grace, T.P. Tsay, D.M. Avenetti, F.J.C. Sanchez, Assessment of automatic cephalometric landmark identification using artificial intelligence, Orthod. Craniofac. Res. 24 (S2) (2021) 37–42, http://dx.doi.org/10.1111/ocr.12542.

[10] H. Kim, E. Shim, J. Park, Y.-J. Kim, U. Lee, Y. Kim, Web-based fully automated cephalometric analysis by deep learning, Comput. Methods Programs Biomed. 194 (2020) 105513, http://dx.doi.org/10.1016/j.cmpb.2020.105513.

[11] S. Yang, Z. Quan, M. Nie, W. Yang, TransPose: Keypoint localization via transformer, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 11782–11792, http://dx.doi.org/10.1109/ICCV48922.2021.01159.

[12] J. Hendrickx, R.S. Gracea, M. Vanheers, N. Winderickx, F. Preda, S. Shujaat, R. Jacobs, Can artificial intelligence-driven cephalometric analysis replace manual tracing? A systematic review and meta-analysis, Eur. J. Orthod. 46 (4) (2024) cjae029.

[13] H.-W. Hwang, J.-H. Moon, M.-G. Kim, R.E. Donatelli, S.-J. Lee, Evaluation of automated cephalometric analysis based on the latest deep learning method, Angle Orthod. 91 (3) (2021) 329–335.

[14] S. Yang, E.S. Song, E.S. Lee, S.-R. Kang, W.-J. Yi, S.-P. Lee, Ceph-net: automatic detection of cephalometric landmarks on scanned lateral cephalograms from children and adolescents using an attention-based stacked regression network, BMC Oral. Health 23 (803) (2023) 1–17.

[15] R. Zese, L. Lombardo, M. De Maio, M. Tamascelli, F. Cremonini, A novel cephalometric tool enhanced by AI assistance (SHORT PAPER), in: F. Calimeri, M. Dragoni, F. Stella (Eds.), Proceedings of the 2nd AIxIA Workshop on Artificial Intelligence for Healthcare (HC@AIxIA 2023) Co-Located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, 08 November 2023, in: CEUR Workshop Proceedings,

vol. 3578, CEUR-WS.org, 2023, pp. 122–129, URL https://ceur-ws.org/Vol-3578/paper10.pdf.

[16] J.R. Lewis, IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use, Int. J. Hum.-Comput. Interact. 7 (1) (1995) 57–78, http://dx.doi.org/10.1080/10447319509526110.

[17] J.J.R. Lewis, J. Sauro, Revisiting the factor structure of the system usability scale, J. Usability Stud. 12 (4) (2017) 183–192.

[18] J.R. Lewis, Psychometric evaluation of the PSSUQ using data from five years of usability studies, Int. J. Hum.-Comput. Interact. 14 (3–4) (2002) 463–488, http://dx.doi.org/10.1080/10447318.2002.9669130.

[19] T.S. Tullis, J.N. Stetson, A comparison of questionnaires for assessing website usability, in: Usability Professional Association (UPA) Conference. Minneapolis, USA,7-11 June 2004, 2004, pp. 1–12.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, CoRR abs/1512.03385. arXiv:1512.03385. URL http://arxiv.org/abs/1512.03385.

[22] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019.

[23] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, TPAMI (2019).

[24] M. Tan, Q.V. Le, EfficientNetV2: Smaller models and faster training, 2021, CoRR abs/2104.00298. arXiv:2104.00298. URL https://arxiv.org/abs/2104.00298.

[25] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society, 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848.

[26] F. Milletari, N. Navab, S. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016, IEEE Computer Society, 2016, pp. 565–571, http://dx.doi.org/10.1109/3DV.2016.79.

[27] P.J. Huber, Robust estimation of a location parameter, in: Breakthroughs in Statistics: Methodology and Distribution, Springer, 1992, pp. 492–518.

[28] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2878–2890, http://dx.doi.org/10.1109/TPAMI.2012.261.

[29] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, et al., A benchmark for comparison of dental radiography analysis algorithms, Med. Image Anal. 31 (2016) 63–76.