applied sciences

MDPI

Article

Cephalometric Landmark Detection in Lateral Skull X-ray Images by Using Improved SpatialConfiguration-Net

Martin Šavc, Gašper Sedej and Božidar Potočnik *

Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška cesta 46, 2000 Maribor, Slovenia; martin.savc@um.si (M.Š.); gasper.sedej@um.si (G.S.)
* Correspondence: bozidar.potocnik@um.si; Tel.: +386-2-220-7484

Abstract: Accurate automated localization of cephalometric landmarks in skull X-ray images is the basis for planning orthodontic treatments, predicting skull growth, or diagnosing face discrepancies. Such diagnoses require as many landmarks as possible to be detected on cephalograms. Today's best methods are adapted to detect just 19 landmarks accurately in images varying not too much. This paper describes the development of the SCN-EXT convolutional neural network (CNN), which is designed to localize 72 landmarks in strongly varying images. The proposed method is based on the SpatialConfiguration-Net network, which is upgraded by adding replications of the simpler local appearance and spatial configuration components. The CNN capacity can be increased without increasing the number of free parameters simultaneously by such modification of an architecture. The successfulness of our approach was confirmed experimentally on two datasets. The SCN-EXT method was, with respect to its effectiveness, around 4% behind the state-of-the-art on the small ISBI database with 250 testing images and 19 cephalometric landmarks. On the other hand, our method surpassed the state-of-the-art on the demanding AUDAX database with 4695 highly variable testing images and 72 landmarks statistically significantly by around 3%. Increasing the CNN capacity as proposed is especially important for a small learning set and limited computer resources. Our algorithm is already utilized in orthodontic clinical practice.

Keywords: detection of cephalometric landmarks; skull X-ray images; convolutional neural networks; deep learning; SpatialConfiguration-Net architecture; AUDAX database

Citation: Šavc, M.; Sedej, G.; Potočnik, B. Cephalometric Landmark Detection in Lateral Skull X-ray Images by Using Improved SpatialConfiguration-Net. *Appl. Sci.* **2022**, *12*, 4644. https://doi.org/10.3390/app12094644

Academic Editors: Ioannis A. Kakadiaris, Michalis Vrigkas and Christophoros Nikou

Received: 14 April 2022
Accepted: 29 April 2022
Published: 5 May 2022

1. Introduction

Cephalometry has been used for many years for the diagnosis of malformations, surgical planning and evaluation, and growth studies. This discipline relies on the identification of craniofacial landmarks [1,2]. Cephalometric analysis, or cephalometrics, is the clinical application of cephalometry to the field of orthodontics. Cephalometrics has been used in orthodontic diagnosis to evaluate the pretreatment dental and facial relationship of a patient, to evaluate changes during treatment, and to assess tooth movement and facial growth at the end of treatment [3]. The first important step in cephalometric analysis is accurate detection of cephalometric landmarks on the cephalogram, i.e., an X-ray image of the craniofacial area (shortly, a skull image). In the cephalometric assessment, certain carefully defined points should be located on the radiographs, and linear and angular measurements are then made from these points [3]. Only accurate measurements and calculations represent diagnostic aids for orthodontists.

There exist lateral and frontal cephalograms. Lateral cephalograms provide a lateral view of the skull, while the frontal cephalograms present an antero-posterior view of the skull. The lateral cephalograms will be utilized in this study. Figure 1 depicts sample lateral cephalograms, captured in a natural head position, which enables the repeatability of image capture and comparison of different cephalometric analyses.

Early attempts for computerized detection of cephalometric landmarks were found around the year 2000. Several (prototype) methods for automatic landmark identification

from skull X-ray images (cephalograms) have emerged, based on heuristic features and rigid rules. These methods were highly dependent on the quality of the input images, and were adapted for a small number of landmarks [1] (the number of landmarks is meant here as the number of different types of landmarks we are looking for in each image). More mature methods, as well as learning-based approaches, emerged after 2010 [4,5]. Lindner et al. [5,6] proposed an efficient detection method based on Haar-like features and random forests (RFs). An RF was trained for each landmark in order to predict the more probable position of that landmark. Each tree in the RF voted for the likely new position. The RF regression-voting mechanism was integrated into the constrained local model framework that optimized a statistical shape model and total votes over all landmark positions. This detection system was adapted for the detection of 19 cephalometric landmarks. A similar method with RF and Haar-like appearance features was proposed by Ibragimov et al. in [4,5,7]. The difference was that a matching of the appearance shape model in a target image was sought by using a game-theoretic optimization framework. The fitted model determined the optimal landmark positions.

Recently, successful methods have emerged based on convolutional neural networks (CNN) and deep learning. We expose the four best, which are comparable in effectiveness. Chen et al., in a conference article [8], proposed the CNN-based architecture that consists of the pretrained VGG-19 net as a feature extraction module, an attentive feature pyramid fusion (AFPF) module, and a prediction module. They fused features from different levels in order to obtain high-resolution and semantically enhanced features in the AFPF module. A self-attention mechanism was utilized to learn corresponding weights for the fusion for different landmarks. Finally, a combination of heat maps and offset maps was employed in the prediction module to perform a pixel-wise regression-voting. The next conference paper is from Li et al. [9], who modeled landmarks as a graph and employed two global-to-local cascaded graph convolutional networks (GCNs) to reposition the landmarks towards the target locations. The graph signals of the landmarks were built by combining local image features and graph shape features. The authors state that their method is able to exploit the structural knowledge effectively and allow rich information exchange between landmarks for accurate coordinate estimation. The first GCN estimated a global transformation of the landmarks, while the second GCN determined local offsets to adjust the landmark coordinates further. Payer et al., in a journal article [10], introduced a CNN architecture that learns to split the localization task into two simpler sub-problems, thus reducing the overall need for large training datasets. Their fully convolutional SpatialConfiguration-Net (SCN) utilized one component to obtain locally accurate but ambiguous candidate predictions, while the other component improved robustness to ambiguities by incorporating the spatial configuration of landmarks. Since our research is based on this method, we will provide details about the SCN in the next sections. Lastly, we expose the method by Song et al. [11]. The authors proposed the usage of an individual model for each landmark, where each model was trained by the ResNet50 architecture. These constructed models were applied to smaller patches extracted from the cephalometric image. The method assumed that each patch that was passed into the model must contain the landmark that was being detected by this model. To ensure this, each testing image was aligned to every training image by using a translational registration. Landmarks from the training image with the best fit after registration were considered as centers for the extracted patches. The results obtained on the database of public cephalograms with 19 landmarks were comparable to other state-of-the-art methods. However, this method does not scale well to a larger number of cephalometric landmarks and training images.

In order for cephalometric analysis to be meaningful and useful as a diagnostic tool, it is necessary to detect as many cephalometric landmarks on the cephalogram as accurately as possible. Usage of lateral cephalograms predominates today in the field of orthodontics; therefore, we also focused on this type of cephalograms in our research (similar to the related works summarized above). The identified shortcomings of early related works indicated that these methods were adapted for a small number of cephalometric landmarks

and for a small number of high-quality input images. State-of-the-art methods [8–10] are practically invariant to brightness/contrast variations, or to situations during cephalograms' capture, respectively. Additionally, an addition of new landmarks that we would like to detect with these methods is relatively simple, as we only need to supplement the learning set and retrain the CNNs (and possibly add some channels). Although state-of-the-art methods have proven to be very effective in locating cephalometric landmarks, it should be noted that these methods have been validated on only 19 landmarks and on just some hundred testing images. Thus, a research question arises as to whether the CNN architectures of these methods have sufficient capacity to localize a larger number of landmarks effectively on a larger set of testing images captured with different X-ray devices. We are tackling a real-world problem from the field of orthodontics in this research; namely, we are developing a detection method as an enhancement of the state-of-the-art, which will be able to detect a large number of cephalometric landmarks (in our study 72) on highly variable testing images. It is understood by variability that testing images are of different sizes (and different spatial resolutions), and that they were captured by using different X-ray devices in different orthodontic clinics (most likely with different device settings). On the other hand, this research also solves one of the concrete problems of the industry (e.g., the AUDAX company). Virtually every orthodontic software includes a module for detecting cephalometric landmarks. A greater number of very precisely localized landmarks of course means better usability of such software. For accurate cephalometric analyses, we need to localize as many landmarks as possible, as only in this way can we diagnose discrepancies or patients' face disharmony, predict skull growth, or plan treatments.

In this study, we will adapt the architecture of the state-of-the-art SCN network in order to detect 72 cephalometric landmarks on highly variable X-ray images. The aim is, on the one hand, to increase the capacity of the CNN (i.e., the ability to learn several different transformation functions), while maintaining approximately the same number of free parameters (degrees of freedom—DoF) as the basic SCN network has. The latter is achieved by expanding the local appearance and spatial configuration components of the SCN network, and not by a raw increase of filters' sizes and numbers of channels. Maintaining DoF while increasing network capacity is important, especially for a small learning set and limited computer resources, which is often the case in healthcare. This, in turn, means a better ability to train such an NN and prevent overfitting. The effectiveness of our proposed SCN-EXT method was confirmed experimentally by detecting 72 cephalometric landmarks on a challenging private database of 4695 cephalograms.

The contribution of this research work is summarized in

1. The development of a sophisticated landmark detection algorithm, where this algorithm is built on the state-of-the-art SpatialConfiguration-Net neural network.
2. Introduction of the most effective algorithm for the detection of 72 cephalometric landmarks on the lateral skull X-ray images.
3. The first study that assesses the effectiveness of the state-of-the-art cephalometric landmark detection algorithms on a large number of landmarks and on a large number of testing images.

This article is structured as follows. A short overview of cephalometric landmarks' classification and employed evaluation databases is given in Section 2. A novel cephalometric landmark detection algorithm based on the SpatialConfiguration-Net architecture is described in detail in Section 3. Some considerations about the proposed method implementation and CNN training are clarified in Section 4. This section also introduces the evaluation metrics used in our experiments. Section 5 presents some of the results obtained on the public and private databases, followed by Section 6, which emphasizes certain aspects of our detection method. Section 7 concludes this paper briefly with some hints about future work.

2. Experimental Methods

2.1. Cephalometric Landmarks

There are two well-known classifications of cephalometric landmarks [3], namely, (1) based on the origin, we distinguish between (i) anatomic and (ii) derived or constructed cephalometric landmarks, and (2) based on the structures involved, we differentiate between (a) hard tissue and (b) soft tissue cephalometric landmarks. Anatomic landmarks represent the actual anatomic structures of the skull (e.g., nasion, point A, point B, ANS, PNS, etc.), while derived or constructed landmarks are obtained secondarily from anatomic structures in a lateral cephalogram (e.g., gnathion, anterior point of occlusion, etc.). On the other hand, the hard tissue cephalometric landmarks represent the actual hard tissue structures of the skull, such as the nasal bone, frontal bone, maxillary bone, etc., while soft tissue landmarks, as their name suggests, are located on the soft tissues (e.g., on the forehead, nose, lips, etc.) [3]. Examples of hard tissue cephalometric landmarks are nasion, temporale, sella, menton, and gonion, while examples of soft tissues landmarks are subnasale, subspinale, stomion, soft tissue pogonion, and soft tissue gnathion [3].

2.2. Evaluation Databases

Two different databases were used to evaluate the effectiveness of the detection methods in this study, namely, the ISBI public image database with 19 annotated cephalometric landmarks on each image, and the AUDAX private image database with 72 landmarks per image.

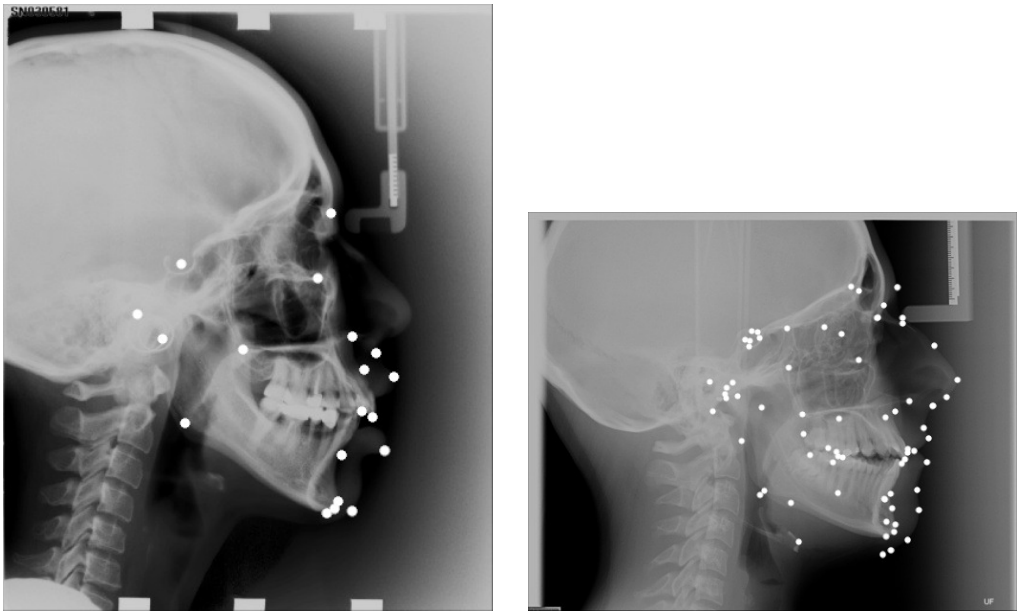
2.2.1. ISBI Public Database

Wang et al. [5] released a public database of 400 cehpalometric images, where 19 of the more common landmarks were annotated on each image. A list of all the annotated landmarks is presented in Table 1. Radiographs were collected from 400 patients ranging from 6 to 60 years old. All cephalograms were captured by the same X-ray device. Every image was annotated manually by two experienced medical doctors. A ground truth was determined as an average of the annotations of both doctors. The images have the same dimension of 1935 × 2400 pixels with 10 pixel/mm spatial resolution.

Table 1. A list of 19 cephalometric landmarks annotated in the ISBI public database. A description of the landmarks and their significance can be found in [3].

−1i—Lower incisal incisior	+1i—Upper incisal incisior	ANS—Anterior Nasal Spine	Ar—Articulare
Gn—Gnathion	Go—Gonion	Li′—Lower lip	Ls′—Upper lip
Me—Menton	N—Nasion	Or—Orbitale	Pg—Pogonion
Pg′—Point Soft Pogonion	PNS—Posterior Nasal Spine	Po—Porion	S—Sella Turcica
Sn′—Subnasale	SS—Subspinale (Point A)	SM—Supramentale (Point B)	

This database is divided into three sets. The first 300 out of 400 images are from the 2015 Automatic Cephalometric X-Ray Landmark Detection Challenge [4]. These 300 images were split into a training set (150 images) and testing set 1 (the remaining 150 images). The 2016 Automatic Cephalometric X-Ray Landmark Detection Challenge brought another 100 images to this public database. These 100 images are denoted as testing set 2. Figure 1a depicts a sample annotated image from this public database. Landmarks are actually pixels, but they are depicted as white circles in this image.



(a) (b)

Figure 1. Sample annotated cephalograms from (**a**) the ISBI public database [5], with 19 landmarks, and (**b**) the AUDAX private database, with 72 landmarks (white circles).

2.2.2. AUDAX Private Database

A private database was constructed during an industrial project between our research group and the Slovenian company AUDAX (https://audaxceph.com (accessed on 13 April 2022)), which is specialized for the development of orthodontic software. This database consists of 4695 unique skull X-ray images. We assumed that each radiograph belongs to a different subject. Information about the image spatial resolution and about the subject in the image (e.g., gender, age, health status) was not provided by AUDAX. The size of images ranged from 355 × 480 pixels (min size) to 4417 × 5963 pixels (max size). There are 287 unique image sizes in this database. On this basis, we concluded that the images were captured with just as many different X-ray devices. The five most common sizes of radiographs were as follows: 2808 × 2148 pixels (1598 images), 1000 × 900 pixels (419), 2685 × 2232 pixels (310), 1804 × 2148 pixels (309), and 1000 × 765 pixels (222). An average image size was 1740 × 2012 pixels.

Seventy-two cephalometric landmarks were annotated on each image by a single experienced orthodontist. A list of all annotated landmarks is gathered in Table 2. Most landmarks are anatomic landmarks, while the rest were constructed relative to anatomic landmarks, or were defined as intersections of particular lines and/or planes, where lines/planes were defined by specific anatomic landmarks or skull structures. An example of a constructed landmark is RT-abo, which is lying on a silhouette, halfway between the landmarks articulare (Ar) and gonion (Go). Based on their expertise, AUDAX classified landmarks into five classes with respect to their importance in cephalometric analyses. The 38 most important landmarks (class 5) are highlighted in Table 2. On the other hand, AUDAX also classified the landmarks into five classes with respect to the difficulty of their determination. The six most difficult to determine landmarks (class 5) are underlined in Table 2. All 72 denoted landmarks were used as the ground truth in our research. Figure 1b depicts a sample image from this private database, with 72 annotated cephalometric landmarks.

The K-fold validation technique was employed by utilizing data from this database to verify the detection methods. The K parameter was set to 3, thus dividing the private database randomly into 3 folds of the same size (i.e., each fold consists of 1565 unique images).

Table 2. A list of 72 cephalometric landmarks annotated in the AUDAX private database. All 19 landmarks from the ISBI public database are also annotated in this database (denoted encircled). The 6 most difficult to determine landmarks are underlined, while the 38 most important landmarks for the cephalometric analyses are bolded. A description of the landmarks and their significance can be found in [3].

−1a–Apex of lower incisor	**−1i—Lower incisal incisor**	**−6a—Apex of lower 1st molar**	**−6c—Cusp of lower 1st molar**
−6d—Distal side of lower 1st molar	**+1a—Apex of upper incisor**	**+1i—Upper incisal incisor**	**+6a—Apex of upper 1st molar**
+6c—Cusp of upper 1st molar	+6d—Distal side of upper 1st molar	**+St′—Upper Stomion**	**A—Point A**
A′—Point Soft A	**ANS—Anterior Nasal Spine**	**APocc—Anterior point of occlusion**	**Ar—Articulare**
B—Point B	**B′—Point Soft B**	**Ba—Basion**	Ci–Clinoidale
Co—Condylion	Col′—Columella	Cp—Condylion posterior	Cs—Condylion superior
D—Point D	DC—Point DC	ER—End Ramus	FMN—frontomaxillary nasal suture
Gl′—Glabella	**Gn—Gnathion**	Gn′—Point Soft Gnathion	**Go—Gonion**
Hy—Hyoid	Ir—Point Ir	L1—L1	**Li′—Lower lip**
LLi—Lower Lip inside	**Ls′—Upper lip**	**Me—Menton**	**Me′—Point Soft Menton**
N—Nasion	**N′—Soft Nasion**	NC—Nasal crown	**Or—Orbitale**
Pg—Pogonion	**Pg′—Point Soft Pogonion**	PM—Suprapogonion	**Pn′—Pronasale**
PNS—Posterior Nasal Spine	**Po—Porion**	**PPocc—Posterior point of occlusion**	**Pt—Pterygoid point**
R1–R1	R3–R3	Rh—Rhinion	RO—Orbital roof of orbital cavity
RT-abo—aboRamalTangent	**S—Sella Turcica**	Se—Entry of Sella	SE—Sphenoethmoidal point
Si—Floor of Sella	**Sn′—Subnasale**	SOr—Supraorbitale	Sp—Dorsum of Sella
−St′—Lower Stomion	Te—Temporale	**tGo—Constructed Gonion (tangent)**	Th′—Throat
U1—U1	ULi—Upper lip inside	W—Walker point	ZyO—Zy Orbit Ridge

3. Computational Methods

3.1. SpatialConfiguration-Net: A Summary

Our proposed landmark detection approach is based on the SpatialConfiguration-Net (SCN) neural network introduced in [10]. The SCN network is a fully convolutional NN and consists of two components, namely (i) local appearance and (ii) spatial configuration components. Both components generate a multidimensional heat map h:

$$h(\mathbf{x}) \in \mathbb{R}^{H \times W \times N}, \tag{1}$$

where $\mathbf{x}$ is a location vector within the heat map, H and W are the height and width of the heat map (also the size of the input image), and N denotes the number of heat map channels (also the number of targeted landmarks). A location of the n-th landmark is predicted as the location of the global maxima in the n-th heat map channel.

The local appearance component is a multi-scale pyramid style network that employs a series of convolutions and downsamplings to extract feature maps. These feature maps are then upsampled and integrated across different scales. An output of this component is the multidimensional or multichannel heat map h of dimension of $H \times W \times N$. Every channel of h can, therefore, be treated as a separate 2D heat map ($H \times W$) that estimates the location of a selected landmark (i.e., N channels for N landmarks).

The spatial configuration component downsamples, by a large factor, the heat map estimated by the local appearance component. It processes this heat map with another series of convolutions with larger kernels, and produces the new multichannel heat map, which is upsampled appropriately at the end. Afterwards, the heat maps from the spatial configuration component, h^{SC}, and from the local appearance component, h^{LA}, are merged

into a new multidimensional heat map h by using the Hadamard product (i.e., element-wise product) as:

$$h(\mathbf{x}) = h^{LA}(\mathbf{x}) \circ h^{SC}(\mathbf{x}). \quad (2)$$

Figure 2 visualizes the above described procedure.

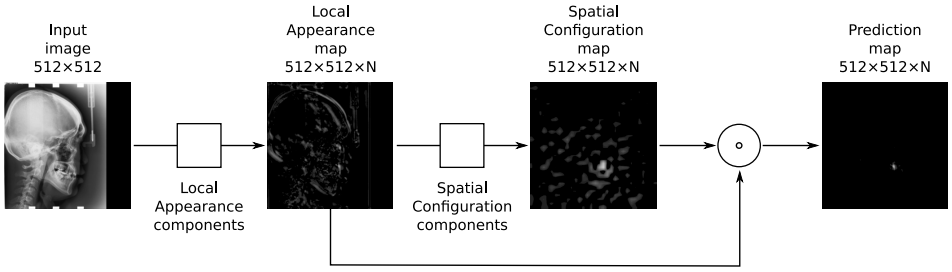


Figure 2. A rough block diagram of the SpatialConfiguration-Net. Depicted are ground plan views of maps (i.e., a single 2D map/channel is shown for a selected landmark).

The local appearance component was designed to learn an accurate landmark position based on local information. On the other hand, the spatial configuration component is aimed to discriminate between possible landmark locations using a larger or global context. An element-wise multiplication of both heat maps is an essential part of the SCN architecture. The latter enables the local appearance component to make multiple estimates for a landmark location across the image, while the spatial configuration component is allowed to selected between these estimates. The local appearance component can, thus, be focused on accurate position estimation without a global discrimination knowledge, while the spatial configuration component does not need to have an accurate landmark's position information, but it is focused on the global discrimination of the landmark's position.

3.2. Proposed SCN-EXT Method

The aim of this research is to develop an effective deep-learning-based method for detecting a large number of cephalometric landmarks from skull X-ray images. State-of-the-art cephalometric landmark detection methods such as [8–11] have proven very effective on a small number of landmarks. Our goal, however, is to upgrade the state-of-the-art appropriately, also for more challenging kinds of detection.

A substantial increase in the number of targeted landmarks requires, typically, an increase in a (convolutional) neural network's capacity. A trivial solution of increasing the number of filters for each convolution layer proved to have two drawbacks. First, doubling the number of filters squares the number of free parameters for most layers. Consequently, the memory requirements grow quadratically. Second, increasing the number of parameters typically makes the learning of an NN with the same training set and similar hyperparameters either unstable or prone to overfitting [12].

The considerable inflation of free parameters is particularly acute for the SCN network, as we have found through experimentation that this network learning has become very unstable. It should also be noted that an exhaustive fine-tuning of the initialization constants for particular SCN layers were carried out. It is expected that an additional fine-tuning of the SCN network would be required by larger expansion of the free parameters. However, the SCN network performed with high accuracy when detecting 19 cephalometric landmarks on testing images from the ISBI public database (see Section 2.2).

We wanted to take advantage of the high detection effectiveness of the SCN network, but, at the same time, we wanted to avoid re-evaluating (i.e., fine-tuning) the initialization constants if the SCN network capacity was increased significantly. Therefore, we propose the following SCN network extension, denoted as SCN-EXT, which increases the capacity of the NN by adding a series of new, but with the same hyperparameters, basic building blocks of the SCN network.

We constructed the SCN-EXT network by introducing J repetitions of the local appearance component into the SCN network, where each of these components was connected with an input image. Figure 3 depicts the basic elements and outputs of the SCN-EXT network. An output of the local appearance component is a multichannel heat map (dimensions of $H \times W \times N$), which is passed on to the input of the new spatial configuration component. We must, therefore, integrate J spatial configuration components into the SCN-EXT network, i.e., one for each local appearance component. The spatial configuration component also returns as an output of the matrix of dimension $H \times W \times N$ (i.e., spatial configuration map). Subsequently, combining the outputs of all J repetitions of a particular component follows. The J outputs of the local appearance components are summed simply into the final local appearance heat map. Similarly, the spatial configuration components' outputs are combined (see Figure 3). Finally, identical to the original SCN network, both the final local appearance and the final spatial configuration heat maps are merged, by using the Hadamard product, into a prediction map, which is then utilized for predicting landmarks' locations. The described procedure for constructing the prediction map h is written formally as:

$$h(\mathbf{x}) = \left(\sum_{j=0}^{J} h_j^{LA}(\mathbf{x}) \right) \circ \left(\sum_{j=0}^{J} h_j^{SC}(\mathbf{x}) \right), \tag{3}$$

where h_j^{LA} and h_j^{SC} denote heat maps of the j-th local appearance and the j-th spatial configuration component, respectively. It should be emphasized once again that, in the SCN-EXT network, we employed the basic components with the same hyperparameters from the SCN network (i.e., components were initialized with the recommended settings from [10]).

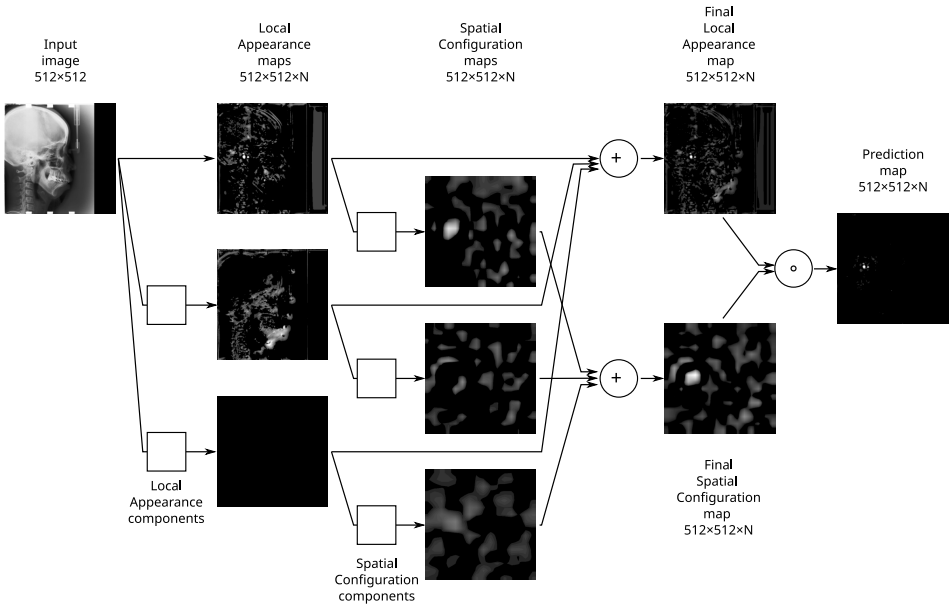


Figure 3. A rough block diagram of the proposed extended SpatialConfiguration-Net (SCN-EXT).

By adding $J - 1$ new local appearance and spatial configuration components, the proposed SCN-EXT network is able to learn $J^2 - 1$ functions more than the original SCN network. Each such component (i.e., neural network) has independent training parameters, and can, thus, learn a subset of the targeted landmarks. On the other hand, compared to the base SCN network, the number of free parameters in SCN-EXT grows linearly with the number of components used.

Landmarks in a training set are not separated into groups (e.g., with respect to an individual anatomical feature or with respect to the neighboring position), so a benefit of utilizing several components in the SCN-EXT is that they can optimize for self-determined

and overlapping groups of landmarks. Each component (neural network) needs to estimate only a fraction of all the targeted landmarks, and multiple networks can cooperate on the same landmark.

An idea in our solution is similar to the so-called grouped convolutions, where the channels in a single convolution layer are grouped together. Each group of channels is processed by a separate set of convolution kernels without overlap between the groups. This has a similar advantage as our proposed approach: a moderate increase of free parameters despite a greater increase of convolutional filters. Increasing the capacity of the CNN network considerably, i.e., the ability to learn several new functions, by a small increase of the number of degrees of freedom (DoF), thus provides more stable learning by using the same training set.

4. Implementation Details and Evaluation Metrics

4.1. Implementation Details and CNN Training

First of all, we will describe image preprocessing and the preparation of training data, followed by an explanation about the training procedure.

Initially, each image was zero-padded along its shorter axis to make it square shaped. Afterwards, it was resampled to a size of 512×512 pixels. A variability in the training set was increased by an augmentation. The training images were augmented "on the fly" by random rotations ($\pm 5^{\circ}$), uniform scaling with a scaling factor selected randomly between 0.6 and 1.2, and intensity changes with a random factor from an interval $[0.75, 1.3]$.

The ground truth heat maps were generated as instructed in [10]. Gaussian kernels were placed at known landmark positions. The Gaussian kernel values were multiplied by a constant $\gamma = 100$ to reduce training instabilities. The standard deviations of the kernels were the training parameters, where they were regularized by using L2-regularization with a weight of 20.

All our own implemented neural networks were trained by using the Adam optimization algorithm [13], with an initial learning rate of 1×10^{-4}. The learning rate was reduced by a factor of 0.5 every 50 epochs without loss improvement on the validation set. The training was limited to a maximum of 150 epochs.

Our software was implemented by using the Python programming language. The constructed and implemented deep neural networks were trained by using the TensorFlow software library. Originally, version 1.15 was employed, but later the code was ported to 2.x libraries (at the end, the models were trained in the 2.4 version library).

All experiments were conducted on a computer system with an AMD Ryzen Threadripper 2920X 12-Core processor, an NVidia Quadro GV100 graphical card with 32 GB of VRAM and 64 GB of physical RAM, and Samsung EVO 970 NVMe 1TB storage.

4.2. Evaluation Metrics

Evaluation metrics and the protocol prescribed for the ISBI public database [4,5] were employed to validate the cephalometric landmark detection methods in this study. The validation was based on the radial error (RE), calculated as the Euclidean distance $d()$ between the estimated, **EST**, and ground-truth landmark location, **GT** (i.e., the 2D point). A basic metric mean radial error (MRE) is derived from this error, where the MRE is calculated as the average of radial errors over L observed landmarks, which is written formally as

$$MRE = \frac{1}{L} \sum_{i=1}^{L} d(\mathbf{EST}_i, \mathbf{GT}_i), \tag{4}$$

where $\mathbf{EST}_i$ in $\mathbf{GT}_i$ denote the estimated and ground-truth locations for the i-th landmark. It should be stressed that L denotes the number of landmarks, and does not necessarily represent the number of different types of landmarks observed in each X-ray image (this is denoted as N in this article).

Two additional statistics of radial error were calculated besides the mean (and the standard deviation) in this research, namely, the median and the 90th percentile of radial error. All the mentioned measures can be estimated per landmark type, per image, or even per all landmark types and all images (i.e., over all landmarks in all images in the database). These metrics are presented either in pixels or in mm if the spatial resolutions of the images are known.

The next metric that has been introduced for the ISBI database is the successful detection rate (SDR), which evaluates the precision (i.e., the positive predictive value) of landmark detection with respect to the radial error. The metric SDR is assessed typically in respect to the radial error up to 2 mm (Class 1), 2.5 mm (Class 2), 3 mm (Class 3), and 4 mm (Class 4) from the ground-truth landmark position. It should be noted that we were unable to determine this metric for the AUDAX private database, because the spatial resolution information was not known for this database.

5. Results

First, we will describe the experiment by which we fine-tuned the SCN-EXT network architecture, and afterwards, we will present the results obtained by the detection of cephalometric landmarks on the ISBI and AUDAX databases.

5.1. SCN and SCN-EXT Architecture Determination

Our research is based on the SCN neural network. The implementation of this network is, to the best of our knowledge, not publicly available; therefore, based on the available information, we recreated the SCN network ourselves. We tested our own implemented SCN on the public ISBI database by using the (hyper)parameters reported in [10]. The SCN network had the following architecture. The local appearance component had 4 layers and 128 filters with 3×3 kernels. The spatial configuration component used a downsampling factor of 16 and included 128 filters of 11×11. In total, this network had around 7.90 million (M) trainable parameters. The SCN network with the described architecture was referred to in the sequel as "our implementation of the method".

Our proposed SCN-EXT solution is a generalization of the SCN architecture, with J-times repetition of local appearance and spatial configuration components (see Section 3.2). We determined the most acceptable SCN-EXT architecture by using the following simple experiment. This experiment was conducted on the AUDAX private database, whereas folds 2 and 3 formed the training set, while fold 1 was utilized as the testing set. According to the presented theory in Section 3, we integrated J repetitions of both components of the SCN network into the SCN-EXT network. If we had employed our fine-tuned SCN for this purpose, then the memory requirements would have become so high (even at small values of J) that this problem could not be solved with today's available hardware. Therefore, we utilized the following simplified SCN architecture for this experiment: (i) local appearance component: 4 layers and 32 filters with 5×5 kernels; and (ii) spatial configuration component: a downsampling factor equal to 16 and 32 filters with 11×11 kernels. Afterwards, the SCN-EXT networks were constructed by changing the number of repetitions of SCN network components, whereas parameter J was varied between 1 and 10 with step 1. It should be stressed that for $J = 1$, we are dealing with the original SCN network.

The results obtained by using different SCN-EXT architectures are summarized in Table 3. The number of repetitions (J) of the SCN architecture components is written next to the method name. For comparison, we also added in this table the results of the fine-tuned SCN architecture (see the first line). Three metrics are shown based on the radial error. All metrics were evaluated across all 72 cephalometric landmarks and across all 1565 testing images. The values are given in pixels, where a lower value indicates the more effective method. We added a number of trainable parameters in the last column. Marked in bold is the SCN-EXT architecture, i.e., SCN-EXT ($J = 6$), which was used in all subsequent experiments. We chose this network because it is a good compromise between effectiveness and training time. At the same time, this network is similar to the fine-tuned SCN with

respect to the DoF (see the "trainable" column). As both networks have similar DoFs, in fact the SCN-EXT network ($J = 6$) has even 1 M lower DoF, all differences in the results can be attributed to changes in the CNN architecture, and not to a raw increase of the DoF (as in the case if we would utilize SCN-EXT with $J = 9$).

Table 3. Effectiveness of different SCN-EXT architectures on cephalometric landmark detection on fold 1 of the AUDAX private database. The column MRE denotes the mean and standard deviation of the radial error, while columns $PCTL_{50}$ and $PCTL_{90}$ denote the 50th (i.e., median value) and 90th percentile of the radial error, respectively. All values are in pixels ("px"). The column "trainable" presents the number of trainable parameters in millions.

Method	MRE (px)	$PCTL_{50}$ (px)	$PCTL_{90}$ (px)	Trainable
SCN †	11.56	6.70	24.91	7.90 M
SCN-EXT, $J = 1$	12.35	7.21	26.44	1.15 M
..., $J = 2$	11.80	6.90	25.25	2.29 M
..., $J = 3$	11.57	6.75	24.72	3.44 M
..., $J = 4$	11.56	6.73	24.68	4.58 M
..., $J = 5$	11.54	6.69	24.57	5.73 M
..., $\boldsymbol{J = 6}$	**11.36**	**6.66**	**24.31**	**6.88 M**
..., $J = 7$	11.42	6.67	24.30	8.02 M
..., $J = 8$	11.35	6.54	24.20	9.17 M
..., $J = 9$	11.26	6.57	24.05	10.31 M
..., $J = 10$	11.48	6.60	24.48	11.46 M

†—Our implementation of the method.

5.2. ISBI Public Database

Initially, the effectiveness of our proposed SCN-EXT method, designed primarily for cephalometric landmark detection, was assessed on the ISBI public database. We used the prescribed methodology and established metrics [5]. The mean and standard deviation of the radial error were calculated over all 19 cephalometric landmarks and over all testing images. In addition, the successful detection rate (SDR) metric was evaluated for the four prescribed classes. The results for testing set 1 are gathered in Table 4, while Table 5 summarizes the obtained results for testing set 2.

Table 4. Effectiveness of cephalometric landmark detection methods on the public ISBI database: testing set 1. The column MRE denotes the mean and standard deviation of the radial error, while the SDR columns denote the successful detection rate (in %) for the four specified classes.

Method	MRE (mm)	SDR (%) 2 mm	SDR (%) 2.5 mm	SDR (%) 3 mm	SDR (%) 4 mm
Li et al. [9]	1.04 ± N/A	88.49	93.12	95.72	98.42
SCN [10] †	1.08 ± 1.08	87.30	91.40	94.25	97.33
Song et al. [11]	1.08 ± N/A	86.40	91.70	94.80	97.80
SCN-EXT	**1.13 ± 1.11**	**85.61**	**90.60**	**93.96**	**97.44**
Chen et al. [8]	1.17 ± N/A	86.67	92.67	95.54	98.53
Chen et al. [8] †	1.30 ± 2.07	83.65	90.70	94.81	97.86
Lindner et al. [5]	1.67 ± 1.48	73.68	80.21	85.19	91.47
SCN [10] ‡	N/A	73.33	78.76	83.24	89.75
Ibragimov et al. [5]	N/A	71.72	77.4	81.93	88.04

N/A—Data not available. †—Our implementation of the method. ‡—Results reported just for the merged testing set 1 and 2.

The effectiveness of the state-of-the-art methods were added to the tables as well. Implementations of these methods were not publicly available; therefore, we just summarized the results published by the authors of the methods. We reimplemented only two state-of-the-art methods successfully. The remaining methods were either basically too ineffective (e.g., methods [6,7]), or it was very difficult to scale them to the problem of 72 cephalometric

landmarks' detection (e.g., method [11]), or method descriptions were not comprehensive enough to be able to reproduce them accurately (e.g., method [9]). Additionally, the results of our methods' implementations are presented in the tables, where they are marked by † next to the method name.

The methods in both tables are arranged according to the decreasing value of the MRE metric. Let us emphasize that a lower MRE value indicates a higher detection effectiveness of the method, which means that the better methods are at the top of the tables.

Table 5. Effectiveness of the cephalometric landmark detection methods on the public ISBI database: testing set 2. See Table 4 for denotations.

Method	MRE (mm)	SDR (%) 2 mm	SDR (%) 2.5 mm	SDR (%) 3 mm	SDR (%) 4 mm
SCN [10] †	1.41 ± 1.40	74.84	81.42	86.89	94.47
Li et al. [9]	1.43 ± N/A	76.57	83.68	88.21	94.31
SCN-EXT	**1.47 ± 1.44**	**74.53**	**82.21**	**87.21**	**93.68**
Chen et al. [8]	1.48 ± N/A	75.05	82.84	88.53	95.05
Chen et al. [8] †	1.65 ± 2.22	71.79	80.32	86.21	93.84
Song et al. [11]	1.54 ± N/A	74.00	81.30	87.50	94.30
Lindner et al. [5]	1.92 ± 1.24	66.11	72.00	77.63	87.42
SCN [10] ‡	N/A	73.33	78.76	83.24	89.75
Ibragimov et al. [5]	N/A	62.74	70.47	76.53	85.11

N/A—Data not available. †—Our implementation of the method. ‡—Results reported just for the merged testing set 1 and 2.

5.3. AUDAX Private Database

The effectiveness of our proposed SCN-EXT method was also assessed on the AUDAX private database. On this database, we applied the threefold validation technique, where there were 1565 images in each fold and 72 cephalometric landmarks in each image. The results obtained on the individual folds were merged, and, afterwards, summarized with various statistics calculated over all images and over all cephalometric landmarks. We calculated the mean radial error and the 50th and 90th percentiles of the radial error. The spatial resolution for the AUDAX database is not known; therefore, all results are given in pixels. The calculated metrics are gathered in Table 6. In addition to our proposed SCN-EXT detection method, this table also presents the results of our implementations of two state-of-the-art methods. The methods in the table are arranged according to the decreasing value of the MRE metric. Based on publicly available information, we also reimplemented the method by Li et al. [9], but, with the calculated MRE of about 34 pixels and the median radial error around 25 pixels, we found that our attempt was completely unsuccessful.

The effectiveness of the methods in Table 6 was also assessed with the nonparametric Friedman's statistical test [14] at a 0.05 significance level. The calculated p-value was equal to 0, which indicates that not all the methods' medians are equal. The proposed SCN-EXT method had the lowest mean rank of 1.88, followed by the SCN method with the mean rank of 1.94, and the method of Chen et al. [8] had the highest mean rank of 2.18. Let us evoke that the lower mean rank correlates with the lower radial error, and, consequently, with the higher effectiveness of the method. Subsequently, we conducted a multiple comparison test of mean ranks, i.e., a pairwise comparison of methods. This analysis pointed out that all three compared methods have significantly different mean ranks. On this basis, we argue that our proposed approach has proven overall to be the most effective detection method on the challenging AUDAX database.

Table 6. Effectiveness of the better cephalometric landmark detection methods on the private AUDAX database. The column MRE denotes the mean and standard deviation of the radial error, while columns $PCTL_{50}$ and $PCTL_{90}$ denote the 50th (i.e., median value) and 90th percentiles of the radial error, respectively. All values are in pixels.

Method	MRE (px)	$PCTL_{50}$ (px)	$PCTL_{90}$ (px)
SCN-EXT	**11.26 ± 17.51**	**6.52**	**24.13**
SCN [10] †	11.57 ± 18.71	6.70	25.10
Chen et al. [8] †	12.19 ± 15.02	8.36	25.00

†—Our implementation of the method.

In the sequel, we extracted the metrics from the obtained results only for those 19 landmarks that are also annotated in the public ISBI database. The mean and standard deviation of the radial error was calculated for each landmark and each compared method separately over all images (i.e., 4695 images). These metrics are accumulated in Table 7. The effectiveness of the methods was then assessed by Friedman's statistical test (0.05 significance level), and by a multiple comparison test of mean ranks. In the table next to the MRE value, we wrote in parentheses the order of methods with respect to the mean rank (value 1 indicates the most effective and value 3 the least effective method), where we denoted by an asterisk whether the differences in results are statistically significant. Our proposed method proved to be the most accurate by 15 landmarks and the second best by 4 landmarks, which is notably better than the compared methods. Improvements were statistically significant for six landmarks. Finally, we calculated the MRE over all 19 landmarks (see the row "all landmarks" in the table). The effectiveness of our proposed detection method was statistically significantly higher by at least 3% than for the compared methods. The SCN method was shown to be the second most effective, followed by the method by Chen et al. [8].

Table 7. Effectiveness of the compared methods on the AUDAX database. Considered are only landmarks from the ISBI database. The mean and standard deviation of the radial error are presented in pixels. A number and * in () denote the method's rank and statistically significant difference. **Better results are marked in bold.**

Landmark	SCN-EXT (px)	SCN [10] † (px)	Chen et al. [8] † (px)
−1i–Lower incisal incisior	$\mathbf{5.06 \pm 7.34}^{(1)}$	$5.09 \pm 7.48^{(2)}$	$7.45 \pm 7.72^{(3)}$
+1i–Upper incisal incisior	$\mathbf{4.55 \pm 6.72}^{(1)}$	$4.55 \pm 6.83^{(2)}$	$8.33 \pm 7.77^{(3)}$
ANS–Anterior Nasal Spine	$\mathbf{8.96 \pm 10.88}^{(1,*)}$	$9.35 \pm 11.55^{(2)}$	$12.16 \pm 19.39^{(3)}$
Ar–Articulare	$\mathbf{8.07 \pm 9.38}^{(1,*)}$	$8.44 \pm 9.69^{(2)}$	$9.56 \pm 8.14^{(3)}$
Gn–Gnathion	$\mathbf{7.11 \pm 6.08}^{(1,*)}$	$7.29 \pm 6.22^{(2)}$	$8.04 \pm 6.17^{(3)}$
Go–Gonion	$9.40 \pm 8.50^{(2)}$	$11.05 \pm 10.15^{(3)}$	$\mathbf{8.81 \pm 7.11}^{(1)}$
Li'–Lower lip	$\mathbf{4.51 \pm 6.77}^{(1)}$	$4.66 \pm 12.77^{(2)}$	$7.01 \pm 6.80^{(3)}$
Ls'–Upper lip	$\mathbf{4.49 \pm 6.30}^{(1)}$	$4.63 \pm 8.68^{(2)}$	$7.16 \pm 6.50^{(3)}$
Me–Menton	$\mathbf{6.77 \pm 6.52}^{(1,*)}$	$6.94 \pm 6.63^{(2)}$	$7.86 \pm 6.17^{(3)}$
N–Nasion	$7.31 \pm 10.21^{(2)}$	$\mathbf{7.29 \pm 10.24}^{(1)}$	$9.12 \pm 10.09^{(3)}$
Or–Orbitale	$12.22 \pm 14.29^{(2)}$	$\mathbf{12.15 \pm 14.20}^{(1)}$	$12.62 \pm 11.92^{(3)}$
Pg–Pogonion	$\mathbf{7.17 \pm 9.34}^{(1)}$	$7.24 \pm 9.49^{(2)}$	$9.62 \pm 8.98^{(3)}$
Pg'–Point Soft Pogonion	$\mathbf{9.02 \pm 15.80}^{(1)}$	$9.30 \pm 31.05^{(2)}$	$9.88 \pm 10.73^{(3)}$
PNS–Posterior Nasal Spine	$\mathbf{9.14 \pm 7.64}^{(1)}$	$9.31 \pm 7.87^{(2)}$	$11.65 \pm 8.61^{(3)}$
Po–Porion	$\mathbf{13.44 \pm 15.34}^{(1)}$	$13.89 \pm 16.21^{(2)}$	$14.89 \pm 12.43^{(3)}$
S–Sella Turcica	$\mathbf{4.85 \pm 3.58}^{(1,*)}$	$4.99 \pm 3.76^{(2)}$	$5.20 \pm 3.60^{(3)}$
Sn'–Subnasale	$\mathbf{5.68 \pm 5.35}^{(1,*)}$	$5.84 \pm 6.77^{(2)}$	$8.02 \pm 6.12^{(3)}$
SS–Subspinale (Point A)	$\mathbf{11.02 \pm 12.70}^{(1)}$	$11.28 \pm 13.41^{(2)}$	$14.16 \pm 12.37^{(3)}$
SM–Supramentale (Point B)	$12.88 \pm 17.02^{(2)}$	$\mathbf{12.92 \pm 16.90}^{(1)}$	$14.08 \pm 15.03^{(3)}$
All landmarks	$\mathbf{7.98 \pm 10.57}^{(1,*)}$	$8.22 \pm 12.82^{(2)}$	$9.77 \pm 10.29^{(3)}$

†—Our implementation of the method.

We gathered in Tables 8 and 9 the ten most accurately and the ten least accurately detected cephalometric landmarks by using our proposed SCN-EXT method. Among the ten best detected landmarks, there are as many as six landmarks (in Table 8 they are circled), which are also annotated in the ISBI database. The Th' point from the soft tissue of the throat was detected with the largest MRE error, which differs significantly from others (see Table 9). The reason is that the throat is not fully visible on all cephalograms and, therefore, an expert annotated the Th' point very inconsistently (i.e., Th' was annotated only approximately). If the Th' point was excluded from the metric calculation, the MRE for the SCN-EXT method decreased by about 0.7 pixels to 10.57 ± 13.93 pixels (see also Table 6). Similarly, the median radial error decreased to 6.44 pixels (previously 6.52) and the 90th percentile of the radial error to 23.21 pixels (previously 24.13).

Table 8. Ten cephalometric landmarks from the AUDAX database detected most accurately by the SCN-EXT method. For denotations, see Tables 2 and 6.

Landmark	MRE (px)	$PCTL_{50}$ (px)	$PCTL_{90}$ (px)
Ls'–Upper lip	4.49 ± 6.30	3.14	8.49
Li'–Lower lip	4.51 ± 6.77	3.30	8.53
+1i–Upper incisal incisor	4.55 ± 6.72	3.00	8.30
Pn'–Pronasale	4.61 ± 7.13	3.46	9.20
S–Sella Turcica	4.85 ± 3.58	3.97	9.67
APocc–Anterior point of occlusion	4.95 ± 5.27	3.66	9.60
Si–Floor of Sella	4.97 ± 4.61	3.90	9.97
−1i–Lower incisal incisor	5.06 ± 7.34	3.55	9.90
B'–Point Soft B	5.22 ± 6.69	3.64	10.06
Sn'–Subnasale	5.68 ± 5.35	4.41	11.58

Table 9. Ten cephalometric landmarks from the AUDAX database detected least accurately by the SCN-EXT method. For denotations, see Table 6.

Landmark	MRE (px)	$PCTL_{50}$ (px)	$PCTL_{90}$ (px)
SOr–Supraorbitale	15.81 ± 20.29	7.99	43.83
ZyO–Zy Orbit Ridge	16.90 ± 15.79	11.84	38.37
Te–Temporale	17.03 ± 18.05	12.39	35.41
Ir–Point Ir	17.09 ± 17.22	12.15	37.25
R1–R1	17.25 ± 14.82	13.25	35.61
Gn'–Point Soft Gnathion	18.59 ± 22.10	10.95	46.41
Gl'–Glabella	19.21 ± 20.64	11.91	46.66
R3–R3	19.58 ± 16.82	14.69	41.66
Rh–Rhinion	24.08 ± 33.69	10.49	81.31
Th'–Throat	60.15 ± 76.76	28.97	177.48

6. Discussion

In this study, we upgraded the state-of-the-art SCN neural network to the SCN-EXT network by adding the J repetitions of both the local appearance (LA) component and the spatial configuration (SC) component into the original SCN architecture. All J replicates of each component were summed up simply, and both sums were, finally, combined by using the Hadamard product. By modifying the architecture in this way, we increased the capacity, as the new SCN-EXT network is able to learn $J^2 - 1$ more transformation functions than the basic SCN network. It is completely trivial that if we add J copies of LA and SC components, then the capacity of such a modified network will, of course, increase compared to the capacity of the original SCN network (if the same LA and SC components are utilized). However, the contribution of our approach is that by J-times repeating and merging the simpler LA and SC components, we can maintain approximately

the same DoF of the new SCN-EXT network as has the original SCN network with the more complex LA and SC components, while we simultaneously increase the capacity and learning ability of the SCN-EXT, respectively. The latter is especially acute if processing and memory resources are limited; namely, training the large models (i.e., with large DoF) requires powerful computing units, a large learning set, and a large primary memory.

This research was focused on the problem of detecting many cephalometric landmarks on diverse lateral skull X-ray images. The SCN-EXT network was designed primarily for this purpose. We have shown experimentally (see the Section 5) that the SCN-EXT network components learn well to predict landmark locations. In our current solution, we do not supervise a training by forcing individual components to learn how to localize a specific subset of landmarks. The latter would be achieved, for example, by adding the $L1$ regularization term for sparsity into the training, which could be one of the future research guidelines.

The final architecture of the SCN-EXT network was determined according to the capacity and DoF of the original SCN network. The SCN network was fine-tuned to detect 19 cephalometric landmarks in the ISBI public database. The LA and SC components utilized there were used as the basis in our work. The goal on the private AUDAX database was to localize 72 cephalometric landmarks; therefore, we modified the architecture of the SCN network only slightly, namely, such that the LA and SC components were able to process inputs with 72 channels. The SCN network that aimed for a detection of 19 cephalometric landmarks (ISBI database) had 6.20 M trainable parameters, while the DoF increased to 7.90 M in the case of detecting 72 landmarks (AUDAX database). The SCN-EXT architecture was determined by a simple experiment on the AUDAX database (see Section 5.1). We varied the number of replicates, J, of the LA and SC components, and monitored the MRE by cephalometric landmarks' detection. Much simpler LA and SC components were applied than in the original SCN. Finally, we chose the SCN-EXT architecture with $J = 6$ repetitions of both components with respect to the hypotheses set out in this study. The SCN-EXT network had 6.88 M trainable parameters when detecting 72 landmarks (AUDAX database), while the DoF decreased to 4.16 M if this architecture was adapted for the ISBI database (i.e., reducing the number of channels). It can be noticed easily that the SCN-EXT network had, on both databases, much fewer trainable parameters than the original SCN.

In order to compare the results of our proposed SCN-EXT method with the results of related works, we reimplemented the SCN method and the method by Chen et al. [8] successfully. We also implemented the method by Li et al. [9], but the results, obtained with our implementation of this method, differed greatly from those reported (see the previous section). We deduced that a reason for the failure to reproduce the method is as follows: the method by Li et al. [9] models each landmark as a graph node. Each node is associated with the landmarks' positions and a feature vector that is extracted from a processed image at that position. The feature vector processing is conducted by using the HRNet18 backbone convolutional network. This method consists of two stages. The first stage estimates a global perspective transformation to align the mean positions of landmarks, constructed from the training data with the specific image. Afterwards, the second stage refines local landmark locations. The estimated global perspective transformation did not improve the landmarks' locations regularly, but, rather, it distorted them. A network that predicted nine free parameters of the perspective transformation matrix was described in [9] explicitly. However, DeTone et al. argued in [15] that such approach is unreliable and difficult to train perspective transformations. Therefore, they suggested applying the four-point estimation approach instead. It is unclear, though, how this four-point estimation would be applied for the landmark detection. The reason for the ineffectiveness of this method was, consequently, sought in the poorly estimated perspective transformations. As mentioned in the Introduction, the method by Song et al. [11] does not scale well to a larger number of cephalometric landmarks and training images. The authors validated their approach on the ISBI public database (i.e., on 19 landmarks and 150 testing images). They reported that a

registration of a single testing image to training images was completed in approximately 20 min. In the AUDAX database, there were 3130 training images per one fold. We estimated that registration in this case would require about 20 times more processing time, i.e., about 400 min per one testing image. In total, this would mean 3 folds × 1565 images × 400 min per image = 1,878,000 min, or around 1304 days, to carry out the registration. The latter, of course, is not acceptable, so we have not implemented this method. The remaining methods from Table 4 were around 40% behind the SCN method in terms of effectiveness, and were, therefore, not included in the comparison on the private AUDAX database.

First, let us analyze the results on the ISBI public database. The effectiveness of the proposed SCN-EXT method is comparable to the effectiveness of state-of-the-art cephalometric landmark detection methods. The SCN-EXT is, on testing set 1, less effective by about 8.65% than the best method by the authors Li et al. [9], and on testing set 2 by about 4.26% than the best SCN method (see Tables 4 and 5). We were unable to reproduce the results of [9], because important implementation details are missing in this method's presentation. Undoubtedly, one of the reasons for the lower effectiveness of our SCN-EXT method is that the architecture was established by using the AUDAX database (and not the ISBI data on which the method was actually applied). It should be noted that the DoF of the SCN-EXT method was almost one-third smaller than the DoF of the SCN method. It can also be seen on testing set 2 that the SCN and SCN-EXT methods have very similar SDR metrics. A great similarity between the methods was also perceived on testing set 1. A reason for the higher MRE of the SCN-EXT method is, therefore, attributed to those landmarks for which the SDR was >4 mm (i.e., incorrectly detected landmarks were detected more erroneously than in the SCN method). Finally, let us emphasize that the ISBI database is a small database with a small learning set (150 images), and with only 250 testing images divided into two sets.

Let us continue with an analysis of the results on the AUDAX private database. This database is very challenging, as it contains 4695 (testing) images, divided into 3 folds, in 287 very different sizes. A goal was to localize 72 cephalometric landmarks in each image. Spatial image resolution data were not available. To the best of our knowledge, this is the first such public or private database with a large number of X-ray images and a larger number of landmarks on which the cephalometric landmark detection methods have been verified. Taking into account all 72 cephalometric landmarks, our proposed SCN-EXT method proved to be superior compared to other state-of-the-art methods. It was more effective than the second-ranked SCN method by about 2.68% (see Table 6). The differences and rankings were confirmed statistically significantly by the nonparametric Friedman's test, and by the multiple comparison test of mean ranks. If we took into consideration from the set of all cephalometric landmarks only those 19 landmarks that were also annotated in the ISBI public database, then the SCN-EXT method this time again proved to be statistically significantly the best method. It surpassed the second-best SCN method by about 2.92% (see Table 7). A similar conclusion was drawn if we compared methods at the level of an individual cephalometric landmark. In this case, the SCN-EXT method was demonstrated to be the more effective method on 15 out of the 19 landmarks, and the second best on 4 landmarks. Afterwards, we arranged the detection effectiveness for the mentioned 19 landmarks with respect to the detection effectiveness for all 72 landmarks on the AUDAX database, where only our SCN-EXT method was observed. It was discovered that as many as 6 landmarks ranked among the top ten (even in the top three, see Table 8), 10 landmarks among the top twenty, and 15 landmarks among the top thirty-five most accurately detected cephalometric landmarks. The less accurately localized were the landmarks point A, orbitale, point B, and porion, as the least accurately detected landmark in 52nd place. On this basis, we argue that the ISBI database consists of 19 relatively easier to detect cephalometric landmarks. On the other hand, the AUDAX database can be said to contain at least 33 cephalometric landmarks, which are more difficult to localize than landmarks in the ISBI database. The latter makes the AUDAX database much more demanding than the ISBI database.

Figure 4 depicts the qualitative result of cephalometric landmarks' detection by using our proposed SCN-EXT method on the AUDAX private database. Seventy-two estimated (denoted by a red x) and ground-truth (blue circle) cephalometric landmarks are superimposed on the skull X-ray image. The predicted and correct location of the landmarks are connected by the green line, where the following applies: the shorter the line, the lower the radial error. It can be noticed that, with the exception of the point on the throat, all the remaining cephalometric landmarks were localized extremely accurately.

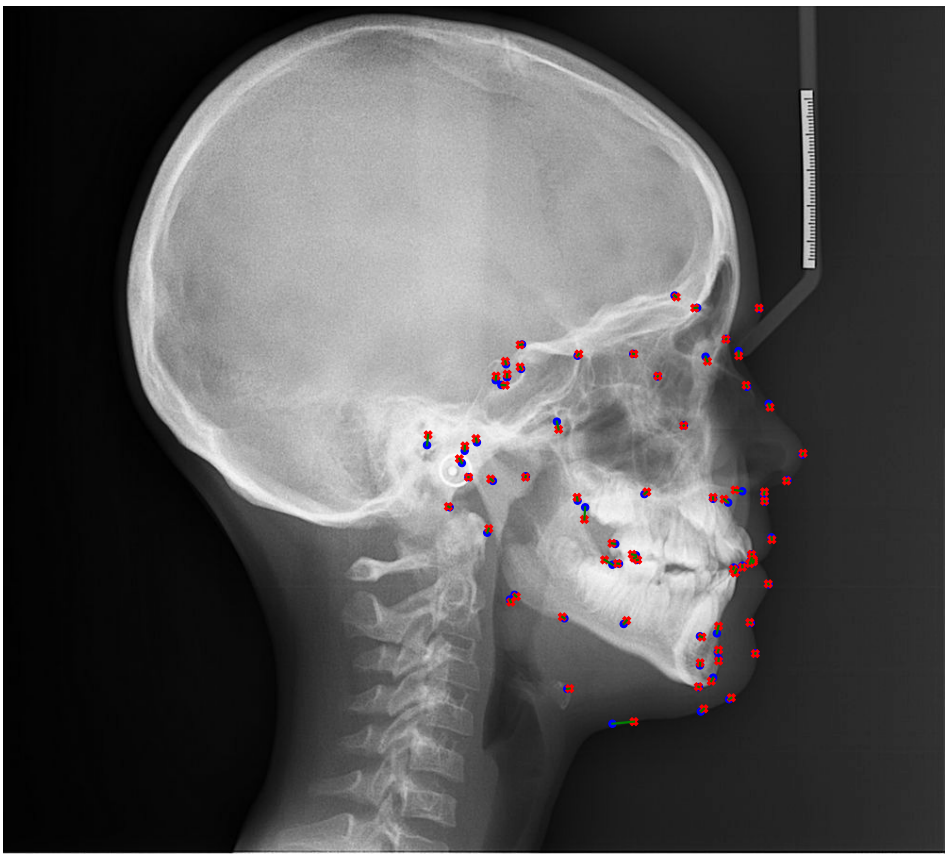


Figure 4. Sample detection result, superimposed on the X-ray image from the AUDAX private database. Cephalometric landmarks were determined by the proposed SCN-EXT method. Estimated landmarks are denoted by a red x, while ground-truth locations are superimposed as blue circles.

The rater's annotations were also analyzed on the AUDAX database. We wanted to find out the positions of which landmarks varied the most on the skull, and, whether the results obtained with our SCN-EXT method were consistent with these findings; accordingly, if the position of the landmark varied slightly on the skull and whether this made our method more accurate, and vice versa. Just a few findings are presented in the sequel, as this analysis is not the main goal of our research. We thus conducted a statistical analysis of skull shapes on the AUDAX database. Seventy-two annotated cephalometric landmarks from all 4695 images were utilized as an input. The aim of this analysis was to determine how the locations of cephalometric landmarks differ (vary, deviate) in the population (i.e., among patients), and how this influenced landmark detection effectiveness. We carried out a so-called generalized Procrustes analysis [16,17]. In each image, the locations of cephalometric landmarks were compensated by translation, scaling, and rotation (i.e., by a rigid transformation), resulting in a mean skull shape (and corresponding mean landmarks' locations) in the Procrustes space. Subsequently, we fitted the Procrustes mean model to the annotated cephalometric landmarks in each image by using an approach from [18], followed by the calculation of the radial error between the fitted model landmarks and the ground-truth landmarks. This error was summarized for each cephalometric landmark

over all images with various statistics (i.e., mean, standard deviation, median, and the 75th percentile). It was discovered that the following 10 cephalometric landmarks have the lowest variability, namely, the landmarks PNS, APocc, W, S, Se, Ci, LLi, +St', −St', and PPocc (see Table 2 for denotations). The ten landmarks with the higher deviation from the Procrustes mean model are the landmarks Go, B, N', Gn', tGo, Ba, Gl', Rh, Hy, and Th', which is the overall highest variability landmark. Both lists remained the same regardless of any statistics (e.g., mean, median, etc.) used in the comparison.

Finally, we evaluated the influence of variability on the cephalometric landmark detection. We calculated the correlation between the landmark variability and detection effectiveness by using the SCN-EXT method. For both quantities, we used data regarding the points order, once in respect to the variability, the second in respect to the detection effectiveness. There was a positive correlation between the two quantities (the correlation coefficient equaled 0.505 with a p value 5.99×10^{-6}). To sum up, the less the landmark varied, the more accurately it was detected, and vice versa. These findings are also consistent with the importance of landmarks for cephalometric analyses as defined by the AUDAX company (see Table 2). With the exception of the Gl' landmark, all the remaining nine poorly localized landmarks (see Table 9) are less important for the cephalometric analyses. Similarly, all 10 accurately localized landmarks (see Table 8) are more important for the cephalometric analyses.

The landmark on the throat soft tissue, Th', with the MRE error of more than 60 pixels, was detected the least accurately. This MRE is almost 2.5 times higher than for the second-least accurately detected landmark, Rh. For the cephalometric analyses conducted by the AUDAX company, the landmark Th' defines just a point where a face profile ends at the bottom. The landmark Th' has no other meaning in these analyses, and, consequently, it was annotated very carelessly. Figure 5 depicts three examples of Th' landmark annotation and localization by the SCN-EXT method. It can be noticed that Th' was annotated on three completely different parts of the throat (see blue circles). Accordingly, this means a poorer ability to learn this landmark and a higher radial error (see the green lines). To illustrate, if we omitted the Th' landmark from the statistics, then the MRE for the SCN-EXT method decreases from 11.26 pixels (see Table 6) to 10.57 pixels, or decreases by 6.13%.

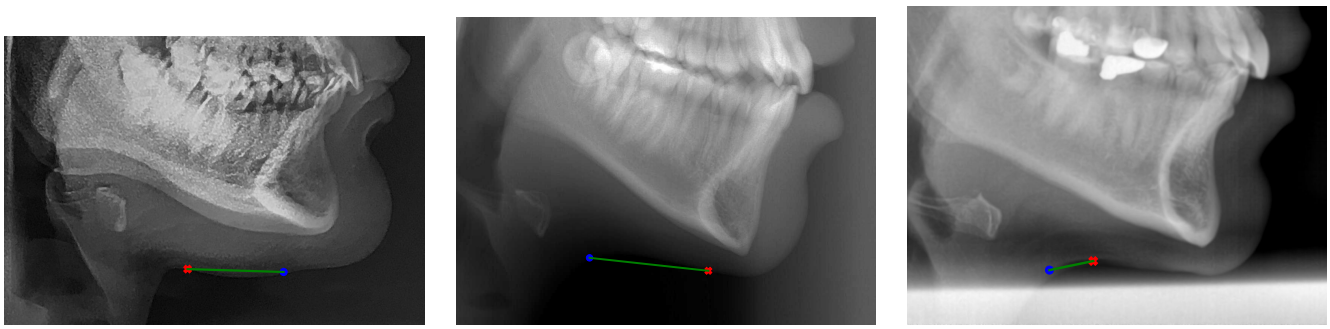


Figure 5. The worst-detected landmark Th' by using the SCN-EXT method: three examples from the AUDAX database. Estimated landmarks are denoted by a red x, while ground-truth locations are superimposed as blue circles.

The CNN training was computationally demanding. The hardware utilized in this study was presented in Section 4.1. On the ISBI database, the training to detect 19 cephalometric landmarks took about 72 min for 150 epochs, or about 29 s per epoch (on GPU). The trained network conducted an inference in around 0.76 s per image on the CPU or in around 0.08 s per image on the GPU. On the AUDAX database, however, the training on GPU took about 2480 min for 150 epochs, or about 992 s per epoch. The trained network localized 72 cephalometric landmarks in around 1.02 s per image on the CPU or in around 0.14 s per image on the GPU.

7. Conclusions

By developing a new method for localizing cephalometric landmarks, we solved a concrete problem from industry in this research. The existing methods have been adapted and tested to detect only 19 landmarks; however, in our work we have addressed the problem of detecting 72 cephalometric landmarks based on industry needs. A large number of accurately detected landmarks on skull X-ray images is a prerequisite for any quality cephalometric analysis. In this study, we upgraded the SpatialConfiguration-Net neural network (SCN), which is one of the state-of-the-art methods for localizing cephalometric landmarks in X-ray images. The SCN architecture was modified by the integration of several repetitions of simpler local appearance and spatial configuration components, with which we increased the capacity of such a modified network (i.e., the SCN-EXT network) with virtually unchanged degrees of freedom (DoF) compared to the original SCN network with the more complex components. Primarily, the SCN-EXT network was designed for localizing a large number of cephalometric landmarks in diverse skull X-ray images.

On the small ISBI public database with 250 testing images, captured by the same X-ray device and with 19 cephalometric landmarks, our, albeit non-tuned SCN-EXT method, was, in terms of effectiveness, just slightly behind the state-of-the-art methods. On the other hand, our fine-tuned SCN-EXT method was statistically significantly the most accurate method on the much more demanding AUDAX database with 4695 highly variable testing images (various X-ray devices!) and with 72 cephalometric landmarks. The improvement of the proposed method was statistically significant, even if we considered out of all 72 cephalometric landmarks only those 19 landmarks that are also in the ISBI database. We also confirmed that the detection accuracy was correlated positively with the importance of landmarks for cephalometric analyses.

An aim of this research was indeed to develop a state-of-the-art cephalometric landmark detection method, but not at the expense of a raw increase of neural network capacity by increasing DoF (e.g., by the addition of more filters, etc.). The presented results in this study were, namely, obtained by using the SCN-EXT network, which had 13% (on the AUDAX database) or 33% (on the ISBI database) fewer free parameters than the original SCN network. Maintaining DoF while increasing network capacity is important, especially for a small learning set and limited computer resources.

Possible improvements to our approach are seen in the use of a more sophisticated augmentation of learning set and in the use of transfer learning. We expect, reasonably, also an improvement in the case if we integrate $J = 9$ or more repetitions of the local appearance and spatial configuration components to the SCN-EXT network, which would indeed increase DoF greatly. For the sake of a fair comparison with the state-of-the-art methods, we have not conducted any of the abovementioned in this study, so these may provide guidelines for future research.

In addition to lateral skull X-ray images, we also have an option of capturing frontal skull X-ray images. This is complementary information that allows complementary cephalometric analyses. One of the future research directions will, therefore, be focused on adapting our method for also localizing cephalometric landmarks on the frontal skull X-ray images.

Finally, let us mention that our detection algorithm is already employed in a clinical practice as a part of a bigger software product. Accurately determined landmarks on the skull X-ray images represent the input for every cephalometric analysis. Automatic localization of 72 cephalometric landmarks undoubtedly disburdens the orthodontist greatly, as manual detection of landmarks means routine and time-consuming work. Nevertheless, he should be aware that, similar to other software tools in clinical practice, our algorithm also does not work 100% accurately. Our trained model is well suited to support and aid manual cephalometric landmarks' annotation, but is not suited for fully automated systems. Manual validation is recommended, and manual correction may be required, based on final application requirements. For this reason, the orthodontist should be able to inspect, and possibly correct, the locations of automatically detected landmarks. Such functionality is, of course, built into the abovementioned software product. The user experiences of

orthodontists with our algorithm are very positive. We conclude with one of the orthodontist's responses: "I conducted the first analysis. I have not used automated tracing for 3 years, but I saw that it is very improved. Landmarks are set at 99% ideally. Very good".

Author Contributions: Conceptualization, B.P. and M.Š.; methodology, B.P. and M.Š.; software, M.Š., G.S. and B.P.; validation, B.P., M.Š. and G.S.; formal analysis, B.P.; investigation, B.P.; resources, B.P.; data curation, B.P.; writing—original draft preparation, B.P., M.Š. and G.S.; writing—review and editing, B.P., M.Š. and G.S.; visualization, G.S. and M.Š.; supervision, B.P.; project administration, B.P.; funding acquisition, B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Slovenian Research Agency (Contract P2-0041).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code used in this paper is publicly available from GitHub https://github.com/MartinSavc/SpatialConfNetExt (accessed on 13 April 2022). The ISBI database [4,5] is available from http://www-o.ntust.edu.tw/~cweiwang/ISBI2015/challenge1/ (accessed on 13 April 2022). The AUDAX database is not publicly available due to data privacy concerns regarding medical data. Data might be available on request from the AUDAX company (https://www.audaxceph.com (accessed on 13 April 2022)).

Acknowledgments: We gratefully acknowledge the valuable scientific and professional contribution of Peter Kobal, from the AUDAX company, Slovenia, who provided us with skull X-ray images and cephalometric landmark annotations (i.e., with the AUDAX database).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Douglas, T. Image processing for craniofacial landmark identification and measurement: A review of photogrammetry and cephalometry. *Comp. Med. Imag. Graph.* **2004**, *28*, 401–409. [CrossRef] [PubMed]
2. Durao, A.; Pittayapat, P.; Rockenbach, M.; Olszewski, R.; Ng, S.; Ferreira, A.; Jacobs, R. Validity of 2D lateral cephalometry in orthodontics: A systematic review. *Prog. Orthod.* **2013**, *14*, 31. [CrossRef] [PubMed]
3. Phulari, B.S. *An Atlas on Cephalometric Landmarks*; Jaypee Brothers Medical Publishers: New Delhi, India, 2013. [CrossRef]
4. Wang, C.; Huang, C.; Hsieh, M.; Li, C.; Chang, S.; Li, W.; Vandaele, R.; Marée, R.; Jodogne, S.; Chen, P.G.C.; et al. Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge. *IEEE Trans. Med. Imaging* **2015**, *34*, 1890–1900. [CrossRef] [PubMed]
5. Wang, C.; Huang, C.; Lee, J.; Li, C.; Chang, S.; Siao, M.; Lai, T.; Ibragimov, B.; Vrtovec, T.; Ronneberger, O.; et al. A benchmark for comparison of dental radiography analysis algorithms. *Med. Imag. Analy.* **2016**, *31*, 63–76. [CrossRef] [PubMed]
6. Lindner, C.; Wang, C.W.; Huang, C.T.; Li, C.H.; Chang, S.W.; Cootes, T. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci. Rep.* **2016**, *6*, 33581. [CrossRef] [PubMed]
7. Ibragimov, B.; Likar, B.; Pernuš, F.; Vrtovec, T. Shape representation for efficient landmark-based segmentation in 3-D. *IEEE Trans. Pattern. Analy. Mach. Intel.* **2014**, *33*, 861–874. [CrossRef] [PubMed]
8. Chen, R.; Ma, Y.; Chen, N.; Lee, D.; Wang, W. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 873–881.
9. Li, W.; Lu, Y.; Zheng, K.; Liao, H.; Lin, C.; Luo, J.; Cheng, C.; Xiao, J.; Lu, L.; Kuo, C.; et al. Structured landmark detection via topology-adapting deep graph learning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX, LNCS 12354, Glasgow, UKm 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 266–283. [CrossRef]
10. Payer, C.; Štern, D.; Bischof, H.; Urschler, M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Med. Imag. Analy.* **2019**, *54*, 207–219. [CrossRef] [PubMed]
11. Song, Y.; Qiao, X.; Iwamoto, Y.; Chen, Y.W. Automatic Cephalometric Landmark Detection on X-ray Images Using a Deep-Learning Method. *Appl. Sci.* **2020**, *10*, 2547. [CrossRef]
12. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, USA, 2017.
13. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; Volume abs/1412.6980.

14. Conover, W. *Practical Nonparametric Statistics*, 3rd ed.; Wiley Series in Probability and Statistics; John Wiley & Sons: New York, NY, USA, 1999.
15. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Deep Image Homography Estimation. *CoRR* **2016**. Available online: http://xxx.lanl.gov/abs/1606.03798 (accessed on 13 April 2022).
16. Goodall, C. Procrustes methods in the statistical analysis of shape. *J. Royal Stat. Soc. B* **1991**, *53*, 285–339. [CrossRef]
17. Rohlf, F.; Slice, D. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. *Syst. Biol.* **1990**, *39*, 40–59. [CrossRef]
18. Cootes, T. *An Introduction to Active Shape Models*; Oxford University Press: Oxford, UK, 2000.