electronics

MDPI

Article

Learning Spatial Configuration Feature for Landmark Localization in Hand X-rays

Gyu-Sung Ham [1] and Kanghan Oh [2,*]

[1] Department of Computer Engineering, Wonkwang University, Iksan 54538, Republic of Korea; ham1231@wku.ac.kr
[2] Department of Computer and Software Engineering, Wonkwang University, Iksan 54538, Republic of Korea
* Correspondence: khoh888@wku.ac.kr; Tel.: +82-63-850-6897

Abstract: Medical landmark localization is crucial for treatment planning. Although FCN-based heatmap regression methods have made significant progress, there is a lack of FCN-based research focused on features that can learn spatial configuration between medical landmarks, notwithstanding the well-structured patterns of these landmarks. In this paper, we propose a novel spatial-configuration-feature-based network that effectively learns the anatomical correlation between the landmarks. Specifically, we focus on a regularization method and a spatial configuration loss that capture the spatial relationship between the landmarks. Each heatmap, generated using U-Net, is transformed into an embedded spatial feature vector using the soft-argmax method and spatial feature maps, here, Cartesian and Polar coordinates. A correlation map between landmarks based on the spatial feature vector is generated and used to calculate the loss, along with the heatmap output. This approach adopts an end-to-end learning approach, requiring only a single feedforward execution during the test phase to localize all landmarks. The proposed regularization method is computationally efficient, differentiable, and highly parallelizable. The experimental results show that our method can learn global contextual features between landmarks and achieve state-of-the-art performance. Our method is expected to significantly improve localization accuracy when applied to healthcare systems that require accurate medical landmark localization.

Keywords: medical landmark localization; spatial configuration feature; soft-argmax

Citation: Ham, G.-S.; Oh, K. Learning Spatial Configuration Feature for Landmark Localization in Hand X-rays. *Electronics* **2023**, *12*, 4038. https://doi.org/10.3390/electronics12194038

Academic Editors: Sathishkumar Easwaramoorthy and Malliga Subramanian

Received: 1 September 2023
Revised: 18 September 2023
Accepted: 22 September 2023
Published: 26 September 2023

1. Introduction

Medical landmark localization involves precisely determining anatomical points or distinctive regions within medical imagery, encompassing modalities such as X-rays, MRIs, or CT scans. This process is of paramount importance due to its pivotal role in guiding medical interventions and strategic treatment planning [1–4]. The accurate localization of landmarks holds indispensable significance for various medical procedures, including surgery, radiation therapy, and image-guided interventions.

However, the manual marking of medical landmarks by a doctor constitutes a laborious task that demands a significant amount of time and is susceptible to errors. This process not only contributes to the overall duration and cost of medical procedures, but also introduces the possibility of inconsistencies and variability among different practitioners. Moreover, the accuracy of manual landmarking heavily relies on the expertise and experience of the healthcare professional, potentially leading to discrepancies, particularly in complex or intricate cases.

To address these difficulties, machine learning-based approaches are predominantly used for automatic medical landmark localization in medical images. In the early days, a traditional strategy for medical landmark localization employed handcrafted local feature responses using prior graphical models, which extract the global spatial configuration of landmarks [5–7]. In these methods, the utilization of prior anatomical spatial features

is popular for comprehending the global context of medical landmarks, and significant progress has been made. However, manually extracting local features is challenging due to the existence of locally similar patterns among medical landmarks. Such ambiguities make it difficult to achieve high performance.

To address this ambiguity problem, the classical machine learning-based approaches [5,8–14] have been used for medical landmark detection and shown to be effective for medical landmark detection. These classical machine learning-based approaches involve regressing the location of landmarks through a regression voting scheme and then using a graphical shape model to optimize the detection results obtained from region proposals. In recent years, deep learning-based approaches [15–21] have outperformed the classical machine learning approaches [5,8–14,22] in terms of landmark detection. The existing deep learning-based approaches [16,17,23–25] have commonly utilized convolutional neural network (CNN)-based frameworks for detecting regional proposals with regard to anatomical landmarks. These regional proposals, including landmark features, were classified into landmark categories. The coordinates-regression-based approaches [26–29] directly estimate the spatial indices of landmarks by employing multiple CNN models.

Additionally, significant advancements have been made through the utilization of heatmap-regression-based approaches [17–19,30–32] using a fully convolutional network (FCN) [33]. In these approaches, the heatmap, consisting of Gaussian distribution for each landmark, was generated using FCN without a dense layer. In the heatmap, the high response of the network's output was concentrated around the center of the target landmarks, while the response of non-landmark regions was suppressed [15]. Therefore, the coordinates of each landmark can be obtained using the non-maximum suppression method. The heatmap-regression-based approaches have demonstrated superior performance compared with coordinate regression, and our methodology is also based on heatmap regression using FCN networks. Even through the heatmap regression approaches have demonstrated notable progress, there is a lack of research on employing the anatomical context information during training.

In medical landmark localization, anatomical information plays a crucial role, because unlike natural images, medical landmarks exhibit a geometrically well-structured pattern. Therefore, the most existing approaches have attempted to incorporate prior knowledge of the anatomical context to improve performance. Hence, most previous approaches have attempted to construct deep networks that comprehend the anatomical contextual information. Payer et al. [18] introduced a pair of cascading neural networks. The initial network is tasked with identifying landmark candidate neighborhood proposals, and the subsequent network pinpoints the exact location of these landmarks. This study exhibited outstanding performance in terms of applying a joint learning method that employs anatomical information. However, the use of multi-stream networks leads to computational complexity and difficulty in controlling hyper-parameters. Oh et al. [15] proposed an anatomical context network that considers the spatial relationship between the landmarks during training. The method learns a much deeper semantic representation of anatomical context, resulting in improved performance. However, there has been a lack of extensive experiments regarding the use of various spatial features.

In this paper, our primary objectives are to develop a spatial-configuration-feature-based network that can learn the anatomical correlation between landmarks. To this end, we focus on the cost function, and introduce a regularization method that captures the spatial relationship between the landmarks. In the proposed regularization term, we also introduce a soft-argmax-based transformation methodology for effectively converting landmark heatmaps into spatial features. The proposed cost function and regularization method can train the network to understand the anatomical correlation between landmarks, which can lead to higher performance in landmark localization.

During the experiment, we evaluated the proposed method using the hand radiograph benchmark [18], which comprises 895 X-ray images containing 37 landmarks. Figure 1 shows the hand X-ray images with ground truth landmarks. The results demonstrate

that our proposed method achieves state-of-the-art performance. In addition, we provide extensive experiments using various spatial features, including Cartesian and Polar coordinates.

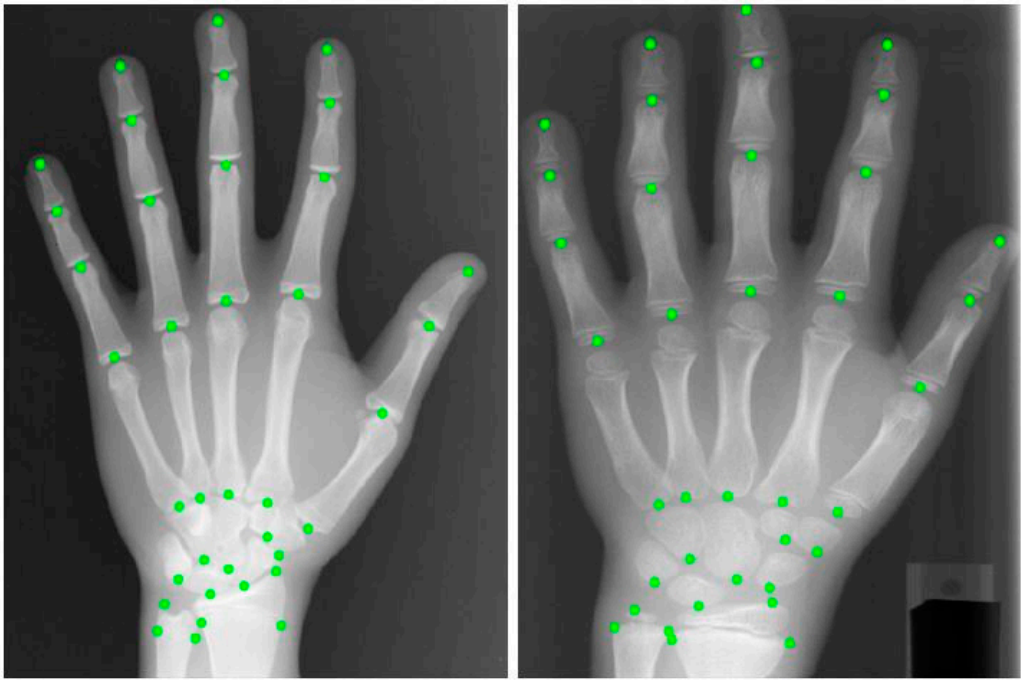

Figure 1. Visualization of 37 landmarks on the digital hand atlas benchmark. Green dots are ground truth landmarks.

This paper is structured as follows: Section 2 provides an overview of related works. Section 3 outlines the proposed methods, providing a detailed explanation. Section 4 presents our experimental results, highlighting the findings. Finally, Section 5 offers a comprehensive discussion and our conclusion.

2. Related Works

In this section, we present an overview of the existing methods for medical landmark detection. In particular, we concentrate on methodologies that utilize X-ray images, such as cephalometric and hand atlas scans.

For classical approaches to medical landmark detection, random forest-based methods [13,14,22] have been proposed for detecting medical landmarks. In these methods, a regression-voting technique was employed to capture contextual features surrounding the target landmark, resulting in remarkable performance. Also, the study [13] employed an iterative framework that refines both appearance information and geometric landmark configuration. The cascade framework was employed in an existing approach [10]. In this method, the region proposals are generated, followed by the incorporation of local features to detect the final landmark locations through the utilization of manually designed rules. Mirzaalian el al. [11] initially extract local features from landmarks and then optimize the geometric attributes between the medical landmarks through the utilization of a spatial graphical model. Although the classical approaches for medical landmark detection have demonstrated commendable performance, they heavily depend on manually engineered features, which consequently yield suboptimal outcomes. In recent years, deep learning-based approaches have solved these drawbacks.

Deep learning approaches can be roughly classified into two mechanisms, namely, spatial coordinate regression [27] and heatmap regression [17–19,30–32]. For the spatial coordinate regression approach, the method directly regresses Cartesian coordinates (x, y) using a CNN-based architecture, whereas the heatmap regression approach generates

a heatmap where the highest response corresponds to the location of a landmark. After acquiring heatmaps, the landmark coordinates can be determined using the non-maximum suppression method. Qian et al. [27] introduced a Faster R-CNN-based approach that directly predicts the spatial coordinates of landmarks. In their method, an undirected graph technique was employed to refine the spatial relationships between landmarks subsequent to the detection of landmark candidate locations. Additionally, during the training phase, multi-scale images were incorporated for enhanced performance. Despite the favorable results reported for this approach, dense layers containing numerous network parameters are employed. This situation could potentially give rise to an overfitting issue.

Unlike spatial coordinate regression, several heatmap-regression-based methods have been studied using FPN-based architectures that exclude dense layers. Park [17] proposed an automated landmark detection FCN model with internally residual connections for cephalometric landmarks. This model was trained to output an archery-target-shaped heatmap when an image patch near the landmark was input. Chen et al. [19] proposed the Spatial Configuration-Net (SCN), a method for decomposing the localization task for anatomical landmark localization into two simple subproblems. In SCN, one component is dedicated to predicting locally accurate, but ambiguous candidates, while the other component incorporates the spatial configuration of the landmark to improve robustness to ambiguity. The heatmap predictions of the two components were multiplied and the network was trained end-to-end to benefit from small datasets. Ao et al. [30] proposed a feature aggregation and refinement network (FARNet). A backbone network pretrained on natural images is used to alleviate the problem of insufficient training data in the medical domain. A multi-scale feature aggregation module and a feature refinement module can cover different resolutions of medical images and achieve improved performance by increasing the resolution of the predicted heatmap. For accurate heatmap regression, they also proposed an exponentially weighted centroid loss to focus on the loss of pixels near landmarks. Kim et al. [31] aimed to develop a fully automated cephalometric analysis method using deep learning and a web-based application that can be used without the need for high-end hardware. They trained a two-step automated algorithm with a stacked hourglass deep learning model and specialized in landmark detection in images.

Although FCN-based heatmap regression methods have made significant progress, there is a lack of FCN-based research focused on features that can learn the spatial configuration between medical landmarks, notwithstanding the well-structured patterns of these landmarks. Existing FCN-based heatmap regression methods have difficulty effectively capturing the global context due to the limitations of the receptive field. These FCN-based networks focus on local features and misinterpret them as heatmaps, even at coordinates where these heatmaps cannot be structurally located within the medical X-ray image. In this paper, we propose a novel spatial-configuration-feature-based network that effectively learns the anatomical correlation between the landmarks. Specifically, we focus on a regularization method and a spatial configuration loss that capture the spatial relationship between the landmarks.

3. Methods

Figure 2 provides an overview of the proposed method for medical landmark localization. First, the hand X-ray image is fed into the U-Net architecture [34,35] to generate landmark heatmaps. Second, each landmark heatmap is transformed into the embedded spatial feature vector by using the soft-argmax method and the spatial feature maps. Third, we generate the correlation map between landmarks based on the spatial feature vector. In the training phase, the loss is minimized by evaluating both heatmaps and the correlation map. Since the proposed approach adopts an end-to-end learning approach, only a single feedforward execution is necessary during the test phase to localize all landmarks.

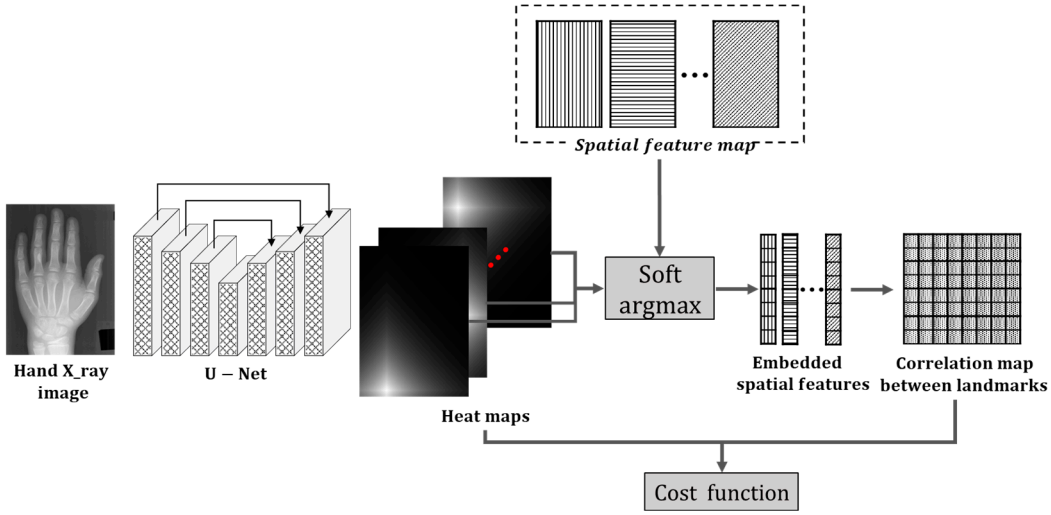

Figure 2. Framework of the proposed method.

3.1. Heatmap Regression Using U-Net

The U-Net architecture, as illustrated in Figure 3, is comprises repeating modules that consist of two 3 × 3 convolutions (with padding), each followed by instance normalization, RELU activation, and 2 × 2 max-pooling (for encoding layers) or up-sampling (for decoding layers). In the encoding layers, the count of feature maps was gradually increased from 64 to 1024, while in the decoding layers, it decreased from 1024 to 64. A skip connection was employed to concatenate identically scaled output maps between the encoder and decoder layers. Within the U-Net outputs, each landmark is represented by an individual Laplace heatmap channel, which is a normalized grayscale image ranging from 0 to 1. The highest intensity within the heatmap indicates the coordinates of the respective landmark. Output heatmaps are two-dimensional matrices with 37 channels, corresponding to the number of landmarks.

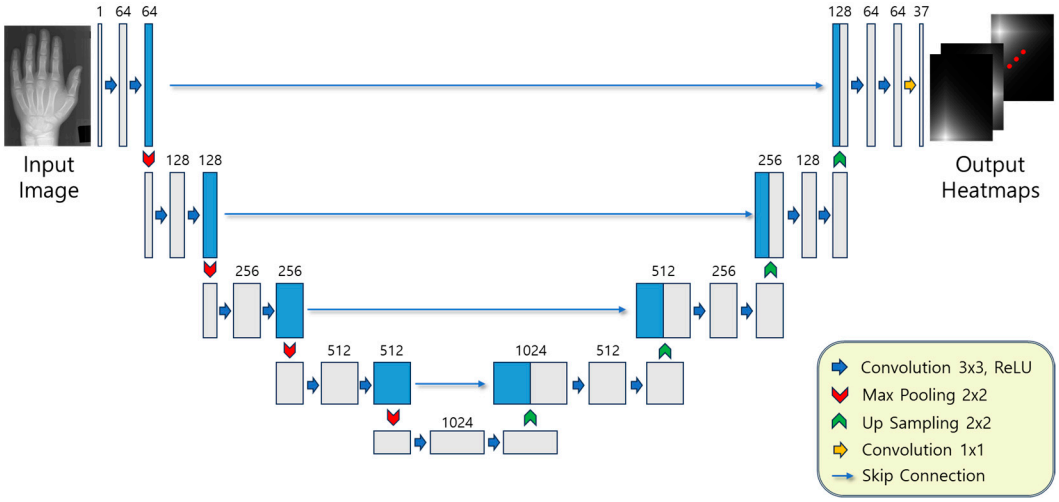

Figure 3. U-Net architecture.

3.2. Spatial Feature Embedding

In this phase, we introduce a transformation method that converts each heatmap into a spatial feature. To achieve this, rather than non-maximum suppression, we utilize the soft-argmax method [36], which guarantees end-to-end differentiability of the networks, in order to transform the heatmap into coordinate values. Given the c-landmark heatmaps $\mathbf{x} \in \mathbb{R}^{h \times w}$ and d-spatial feature maps $\mathbf{f} \in \mathbb{R}^{h \times w \times d}$, we extract the spatial coordinates (x,

y) that approximate the maximum value on the heatmap. By default, both the heatmap and each spatial feature map are represented as a 2D map. The SoftMax function used is defined as follows:

$$\mathbf{s} = \frac{\exp(\mathbf{x} \times \beta)}{\sum_i \exp(\mathbf{x} \times \beta)} \tag{1}$$

where **s** denotes the SoftMax function, and the resulting probability map signifies that the maximum response corresponds to the highest probability. The temperature parameter β is used to regulate the probability distribution. By adjusting β, the SoftMax function can effectively suppress undesirable values that are lower than the maximum value. After normalizing the heatmap using Equation (1), we reshape $\mathbf{x}_{h\times w}$ into a vector $\mathbf{x'}_{1\times \mathrm{h}w}$. Similarly, the d-spatial feature maps $\mathbf{f}_{d\times h\times w}$ are reshaped into $\mathbf{f'}_{hw\times d}$. Then, embedded spatial features can be obtained by calculating the inner product between the normalized heatmap and spatial feature maps:

$$\mathrm{e}_{1\times d} = \mathbf{x'}_{1\times hw} \cdot \mathbf{f'}_{hw\times d} \tag{2}$$

In this study, we employed the two spatial features of Cartesian and Polar coordinates. Figure 4 illustrates the two spatial feature maps employed in our study. From left to right, we can observe both Cartesian coordinates and Polar coordinates. The Cartesian system includes two features that reflect the x- and y-axis maps, while the Polar system includes maps for radius and angle. Before applying soft-argmax pooling, each feature map is normalized using the 2D-instance normalization function provided by the Pytorch API without learnable parameters. In the proposed approach, the feature map normalization process is essential due to its excessively large scale, leading to suboptimal performance.

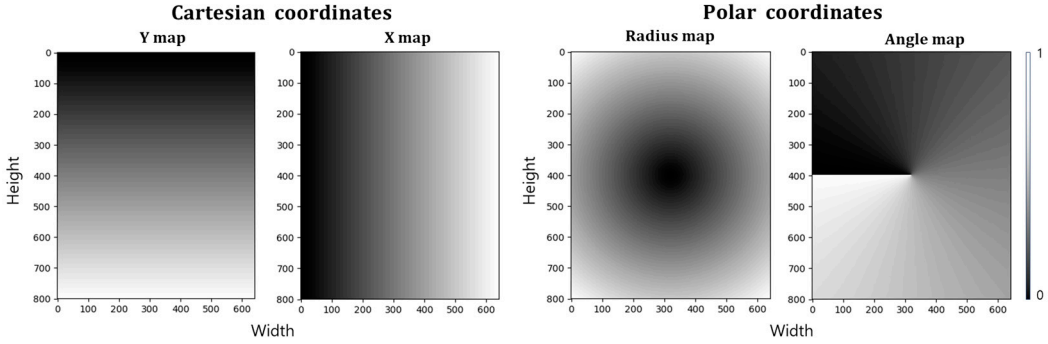

Figure 4. Visualization of spatial feature maps. The size of all coordinate maps is the same as the image size. The X map has values increasing along the x-axis, while the Y map has values increasing along the y-axis. The radius map has values increasing from the center of the map outward, while the angle map has values increasing according to the angle from the center of the map.

3.3. Spatial Configuration Loss

We present a cost function known as the spatial configuration (SP) loss, which aims to learn the spatial relationship between landmarks. As mentioned previously, the proposed method is incorporated into the loss function as a regularization term. The proposed loss function can be defined as:

$$loss(\mathbf{x}, \mathbf{o}) = \|\mathbf{x} - \mathbf{o}\|_2^2 + \gamma \|\mathrm{c}_\mathbf{x} - \mathrm{c}_\mathrm{o}\|_2^2 \tag{3}$$

where **x** is the heatmaps on landmarks, and **o** is the corresponding ground truth heatmaps. We utilized the L2 loss to measure the overall pixel-wise similarity between the network's output heatmaps **x** and the ground truth heatmaps **o**. The second term represents the regularization term used to learn the relationship between landmarks, where c is the correlation map, $\mathrm{c_x}$ and $\mathrm{c_o}$ are calculated from the network output **x** and ground truth

o, respectively, and γ denotes the weight value for the regularization term. Given the n-embedded vectors $e_{n\times d}$, the correlation map is calculated by:

$$c = e_{n\times d} \cdot e_{d\times n}^{T} \tag{4}$$

where n represents the number of landmarks. Figure 5 illustrates the process of constructing a correlation map. As mentioned previously, our work utilizes two spatial features reflecting Cartesian and Polar coordinates. Since both spatial features are simultaneously used, each embedded vector becomes an $e_{1\times 4}$.

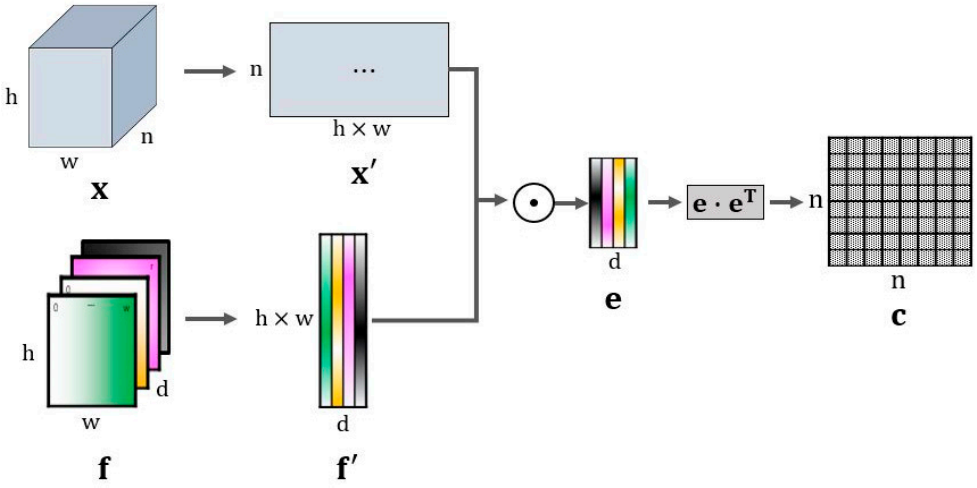

Figure 5. Construction of the correlation map.

4. Results

4.1. Dataset

Our model was evaluated using the publicly available hand radiograph dataset known as the Digital Hand Atlas (DHA) [18]. The DHA dataset comprises 895 images with an average size of 1563 × 2169 pixels, and includes annotations for 37 landmarks, acquired with different X-ray scanners. To evaluate the network, a three-fold cross-validation was performed. Each fold consisted of approximately 600 training images and 300 test images. Table 1 describes the property of the DHA dataset with the evaluation scenario using three-fold cross-validation. Due to variations in the physical pixel resolutions of images, a wrist-width-based normalization factor, as provided by [18], was employed for the purpose of performance evaluation. The wrist-width-based normalization factor $n^{(j)}$ is calculated as follows:

$$n^{(j)} = 50 / \left\| \mathbf{o}_{l\_wrist}^{(j)} - \mathbf{o}_{r\_wrist}^{(j)} \right\|_2 \tag{5}$$

where j is the image for evaluation. We assume a wrist width of 50 mm, which is determined by two landmarks annotated on the wrist as in [18]. The normalization factor $n^{(j)}$ is based on the Euclidean distances of specifically selected landmarks. The normalized point-to-point error is defined as follows:

$$PE_i^{(j)} = n^{(j)} \left\| \mathbf{x}_i^{(i)} - \mathbf{o}_i^{(i)} \right\|_2 \tag{6}$$

where i is a landmark in image j. This allows us to account for variations in physical pixel resolution when calculating the mean and standard deviation of the point-to-point error.

Table 1. Digital hand atlas dataset.

Dataset	Digital Hand Atlas	
Number of landmarks	37	
Number of images	895	
Three-fold cross-validation	Fold 1	Train 597/Test 298
	Fold 2	Train 597/Test 298
	Fold 3	Train 597/Test 298
Resolution	1563 × 2169 (Average)	

4.2. Implementation Details

In this section, we perform both quantitative and qualitative evaluations of the landmark localization performance. The experiments were conducted on a system comprising an Intel Core i7-7800X (Intel, Santa Clara, CA, USA) with a 3.50 CPU, 32 GB of memory, and a Geforce RTX 3090 GPU (Nvidia, Santa Clara, CA, USA). The network was trained and tested using Pytorch 1.8.0 (https://pytorch.org/docs/1.8.0/ accessed on 4 October 2021), a popular deep learning framework. The input X-ray image and output heatmap were downscaled to a size of 800 × 640 pixels, allowing the computing time to be reduced without significant performance loss.

The data augmentation procedure involves two main types of transformations: geometry and intensity transformations. First, the input images and ground truth heatmaps are randomly rotated within the range [−15, 15] degrees, and rescaled within the range [0.8, 1.2]. Second, intensity changes are applied by randomly adjusting the brightness, contrast, saturation, and hue of the images. Image data augmentation is a very popular process in the machine learning field when aiming to improve performance. Note that we adopt on-line data augmentation.

For the U-Net, we minimized the cost function using an Adam optimizer with a mini-batch size of 1, $\beta_1 = 0.9$, $\beta_2 = 0.9$, and initial learning rate of 1×10^{-3}. We trained the U-Net for 800 epochs, reducing the learning rate by 1/10 at the 500th and 700th epochs. For the spatial configuration loss, we utilized four spatial features, which encompassed both Cartesian and Polar coordinates. Also, we set the weight hyper-parameter γ to 1×10^{-3}.

During the soft-argmax phase, the spatial maps were normalized, and we empirically assigned a temperature parameter β of 10 to enhance sensitivity to the maximum value.

4.3. Performance Comparison

The performance of the proposed method was evaluated using two metrics: point-to-point error (PE), and error detection rate (EDR). The PE is the Euclidean distance between the predicted and ground truth landmark points. The EDR measures the rate of misdetection, where a misdetection is counted if the absolute difference between the detected landmark and the reference landmark exceeds z mm. We considered three reference ranges: 2.0, 4.0, and 10.0 mm. Among these, the 2.0 mm range is the clinically accepted reference [15]. Since the ground truth consists of pixel coordinates, in the test phase, we utilized a threshold value of T > 0.95 to generate binary blobs from the predicted heatmap. Subsequently, we determined the centroid of the largest blob as the landmark localization point.

We assessed our approach using the DHA benchmark dataset and compared its performance with that of other state-of-the-art methods. For a quantitative comparison, we chose competitive models based on their citations and the recency of their results. Figure 6 presents the visualization outcomes of the proposed method using the DHA dataset. We can see that the landmarks predicted by our model align with the ground truth landmarks. Table 2 demonstrates that the proposed method outperforms the existing methods, as it achieved a 2 mm range EDR of 3.72% and PE of 0.61. The significance of the proposed method is supported by a statistically significant p-value of <0.05 obtained through a t-test, when compared with the existing methods.

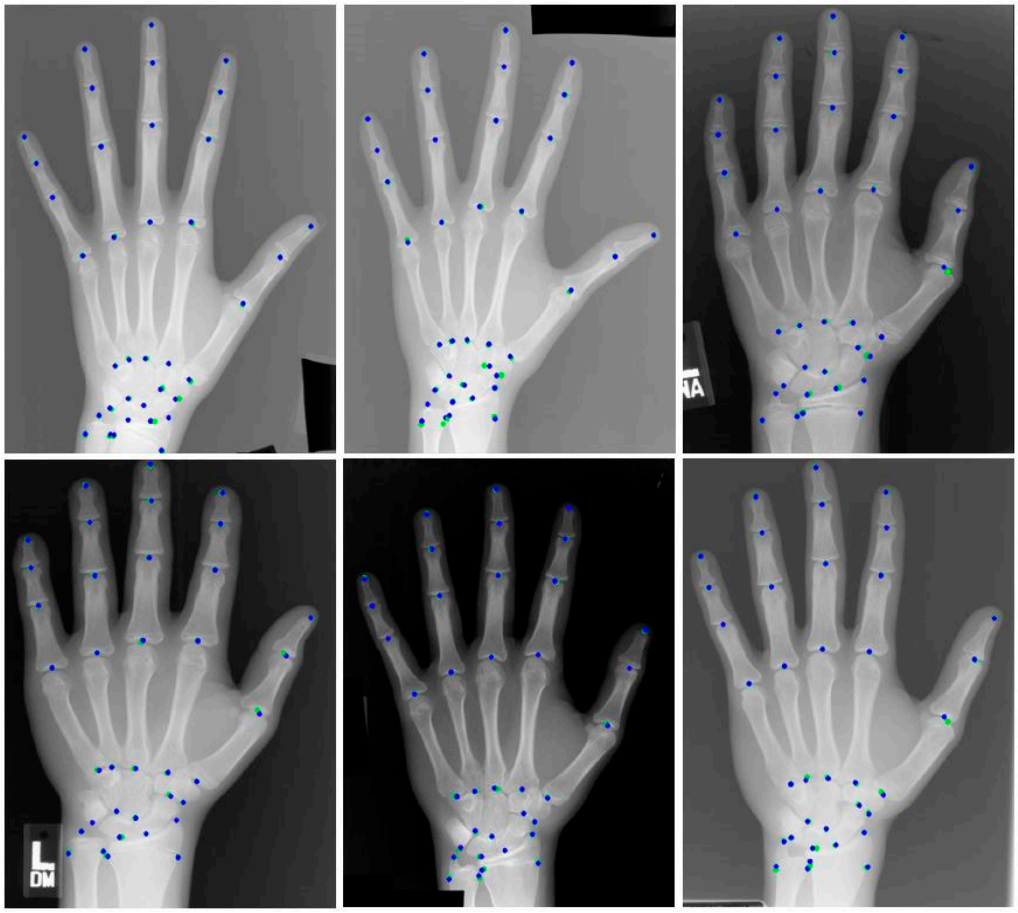

Figure 6. Visualization results of the proposed model in the Digital Hand Atlas dataset. Green dots are the ground-truth, while blue dots are the prediction results.

Table 2. Quantitative comparison with the existing methods on the DHA dataset (p-value < 0.05).

Method	PE (mm)		EDR (%)		
	Mean	**SD**	**>2 mm**	**>4 mm**	**>10 mm**
Štern et al. [6]	0.80	±0.91	7.80%	1.55%	0.05%
Urschler et al. [13]	0.80	±0.93	7.81%	1.54%	0.05%
Payer et al. [18]	0.66	±0.74	5.01%	0.73%	0.01%
Oh et al. [15]	0.63	±0.71	3.93%	0.33%	0.01%
Kang et al. [37]	0.64	±0.64	3.96%	0.34%	0.02%
The proposed method	**0.61**	**±0.61**	**3.72%**	**0.31%**	0.01%

4.4. Ablation Study

In this section, we conducted an ablation study on the validation dataset to investigate the effectiveness of each key component and hyperparameter in the proposed methodology.

Table 3 shows the performances based on the combination of spatial features in the proposed SC loss function. The best performance was observed when both Cartesian and Polar features were used in the proposed loss function. When employing the Polar feature, the performance is slightly higher than when using the Cartesian feature. Furthermore, it is evident that performance improvement has been consistently observed across various types of spatial features when utilizing the SC loss function. These outcomes clearly demonstrate that the proposed SC loss term enhances the network's generalization capability during the training process. Considering the standard deviation, the proposed methods have achieved lower results, compared with the original U-Net (p-value < 0.05).

Table 3. Performance comparison in relation to the combination of spatial features in the proposed SC loss (p-value < 0.05).

Method	PE (mm)		EDR (%)		
	Mean	SD	>2 mm	>4 mm	>10 mm
U-Net	0.66	±0.85	4.24%	0.46%	0.03%
SC Loss (Cartesian)	0.63	±0.63	3.81%	0.35%	0.01%
SC Loss (Polar)	0.62	±0.61	3.79%	0.33%	0.01%
SC Loss (Cartesian + Polar)	**0.61**	±**0.61**	**3.72%**	**0.31%**	0.01%

For the spatial feature embedding phase, Table 4 presents the various performances in relation to the temperature parameter β. Parameter β can adjust sensitivity on the maximum value of the heatmap. For example, a larger β increases the sensitivity to the maximum intensity of the heatmap. In the results, we obtained the best performance with $\beta = 10$, which is moderately sensitive to the values close to the maximum value. Also, a poor performance was observed with $\beta = 20$. This indicates that focusing solely on the largest value while ignoring the surrounding large values leads to disadvantages.

Table 4. Performance comparison in relation to the temperature parameter β.

Temperature Parameter β	PE (mm)		EDR (%)		
	Mean	SD	>2 mm	>4 mm	>10 mm
1	0.63	±0.60	3.74%	0.36%	0.01%
5	0.62	±0.64	3.76%	0.32%	0.01%
10	**0.61**	±**0.61**	**3.72%**	**0.31%**	0.01%
20	0.65	±0.64	3.98%	0.36%	0.02%
30	0.64	±0.66	3.88%	0.35%	0.01%

Table 5 shows the results in relation to the weight parameter γ in the proposed SP loss. As mentioned previously, the proposed SP loss is the regularization term, enabling the consideration of the spatial relationship between landmarks. In the results, a favorable performance was observed for $\gamma = 1 \times 10^{-4}$. Since $\gamma = 0$ completely eliminates the effect of the proposed regularization term, its result is an outcome identical to that of the original U-Net model. We can see that the large weights $\gamma = 1 \times 10^{0}$ and 1×10^{-1} lead to poor performances. The larger weight γ appears to function as a strong regularization term and disrupts the learning process of the pixel-wise heatmap regression task. Therefore, in our study, the parameter γ forces the network to reconstruct the spatial locations of landmarks based on prior anatomical information.

Table 5. Performance comparison in relation to the weight parameter γ.

Weight Parameter γ	PE (mm)		EDR (%)		
	Mean	SD	>2 mm	>4 mm	>10 mm
0	0.66	±0.85	4.24%	0.46%	0.03%
1×10^{0}	0.74	±0.88	8.59%	4.85%	0.99%
1×10^{-1}	0.70	±0.77	6.88%	3.44%	0.15%
1×10^{-2}	0.66	±0.71	4.15%	0.48%	0.02%
1×10^{-3}	0.63	±0.66	3.75%	0.44%	0.01%
1×10^{-4}	**0.61**	±**0.61**	**3.72%**	**0.31%**	0.01%
1×10^{-5}	**0.61**	±0.62	3.73%	0.32%	0.01%
1×10^{-10}	0.67	±0.88	4.31%	0.47%	0.03%

5. Discussion and Conclusions

In this study, we have introduced a novel approach designed for the precise detection of medical landmarks within the hand atlas dataset. Specifically, we focus on a regulariza-

tion method and a spatial configuration loss that capture the spatial relationship between the landmarks. In the proposed regularization term, we introduced a soft-argmax-based transformation methodology for effectively converting landmark heatmaps into spatial features. Each heatmap generated by U-Net is transformed into an embedded spatial feature vector using the soft-argmax method and spatial feature maps, here, Cartesian and Polar coordinates. A correlation map between the landmarks based on the spatial feature vector is generated and used to calculate the loss, along with the heatmap output. This approach adopts an end-to-end learning approach, requiring only a single feedforward execution during the test phase to localize all landmarks. The proposed spatial feature embedding method for extracting spatial features from the heatmaps is computationally efficient, differentiable, and highly parallelizable. Also, this approach is advantageous in that other spatial features can easily be introduced.

In the experimental results, quantitative comparisons show that the proposed method clearly outperforms the existing methods. This demonstrates that our method can learn the global contextual features between landmarks, leading to performance increases. Qian et al. [27] employed the Faster R-CNN-based method and the graph technique for detecting superfluous or undetected landmarks using a repair strategy with Laplacian transformation. However, this method makes it difficult to directly train the network on the spatial configuration between the landmarks. Our method allows the network to learn spatial configurations between the landmarks using correlation maps and the proposed loss. Chen et al. [19] employed a pretrained backbone model to extract multi-scale features and introduced an attentive feature pyramid fusion module. However, due to the linear relationship between the size of the attention enhancement feature map and the number of landmarks, this model increases the number of parameters, memory storage, and computational cost. In contrast, since the proposed method is focused on the cost function, it does not require any additional parameters. Payer et al. [18] introduced a pair of cascading neural networks for landmark localization. Nevertheless, the utilization of multi-stream networks results in computational complexity and challenges in controlling hyper-parameters. As previously mentioned, our approach is specifically designed to be end-to-end. This implies that only a single feedforward run is required during the testing phase to localize all landmarks, thereby rendering it computationally efficient and highly parallelizable.

In the context of medical imaging and diagnostics, the precise identification of anatomical landmarks is of paramount importance. These landmarks serve as critical reference points for healthcare professionals, aiding in the accurate diagnosis, treatment, and monitoring of patients. Our method is expected to significantly improve localization accuracy when applied to healthcare systems that require accurate medical landmark localization. Additionally, since our method achieved an improved performance without increasing the number of parameters, it can be easily applied to existing environments that use U-net for landmark localization.

In future work, we plan to conduct experiments aimed at expanding the application of the proposed method to a broader range of medical landmark detection tasks, encompassing various anatomical structures and datasets. This wider application will enable us to assess the versatility and robustness of our approach across diverse medical contexts. We will also explore the integration of different backbone networks to evaluate their compatibility with our method. Such comparative analysis will help us identify the best backbone architecture for specific medical landmark detection applications, and potentially improve the overall effectiveness of our approach.

Another critical facet of our future research involves the development of an adaptive model that can autonomously learn the temperature parameter β. This innovative feature aims to dynamically optimize the temperature parameter, tailoring it to the specific characteristics of each dataset or task. By doing so, we anticipate a further enhancement in the overall performance and adaptability of our proposed model.

Author Contributions: Conceptualization, G.-S.H. and K.O.; methodology, K.O.; software, G.-S.H. and K.O.; validation, G.-S.H. and K.O.; formal analysis, G.-S.H. and K.O.; investigation, G.-S.H. and K.O.; writing—original draft preparation, G.-S.H. and K.O.; writing—review and editing, K.O.; visualization, K.O.; supervision, K.O.; project administration, K.O.; funding acquisition, K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Digital Hand Atlas: https://ipilab.usc.edu/computer-aided-bone-age-assessment-of-children-using-a-digital-hand-atlas-2/ (accessed on 4 October 2021); https://github.com/christianpayer/MedicalDataAugmentationTool-HeatmapRegression (accessed on 4 October 2021).

Acknowledgments: This paper was supported by Wonkwang University in 2021.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Wang, C.; Huang, C.; Hsieh, M.; Li, C.; Chang, S.; Li, W.; Vandaele, R.; Marée, R.; Jodogne, S.; Geurts, P. 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]
2. Wang, C.; Huang, C.; Lee, J.; Li, C.; Chang, S.; Siao, M.; Lai, T.; Ibragimov, B.; Vrtovec, T.; Ronneberger, O. A benchmark for comparison of dental radiography analysis algorithms. *Med. Image Anal.* **2016**, *31*, 63–76. [CrossRef] [PubMed]
3. Noothout, J.M.; De Vos, B.D.; Wolterink, J.M.; Postma, E.M.; Smeets, P.A.; Takx, R.A.; Leiner, T.; Viergever, M.A.; Išgum, I. Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE Trans. Med. Imaging* **2020**, *39*, 4011–4022. [CrossRef] [PubMed]
4. Al, W.A.; Yun, I.D. Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images. *IEEE Trans. Med. Imaging* **2019**, *39*, 1245–1255.
5. Lindner, C.; Wang, C.; Huang, C.; Li, C.; Chang, S.; Cootes, T.F. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci. Rep.* **2016**, *6*, 33581. [CrossRef] [PubMed]
6. Štern, D.; Payer, C.; Lepetit, V.; Urschler, M. *Automated Age Estimation from Hand MRI Volumes Using Deep Learning*; Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016; Springer: Cham, Switzerland, 2016; Volume 9901, pp. 194–202.
7. Luvizon, D.C.; Tabia, H.; Picard, D. Human Pose Regression by Combining Indirect Part Detection and Contextual Information. *Comput. Graph.* **2019**, *85*, 15–22. [CrossRef]
8. Chu, C.; Chen, C.; Nolte, L.P.; Zheng, G. Fully Automatic Cephalometric X-ray Landmark Detection Using Random Forest Regression and Sparse Shape Composition. Submitted to Automatic Cephalometric X-ray Landmark Detection Challenge. 2014. Available online: https://api.semanticscholar.org/CorpusID:160017622 (accessed on 17 September 2023).
9. Chen, C.; Zheng, G. Fully-automatic landmark detection in cephalometric x-ray images by data-driven image displacement estimation. In Proceedings of the ISBI International Symposium on Biomedical Imaging, Beijing, China, 29 April–2 May 2014.
10. Chen, C.; Wang, C.; Huang, C.; Li, C.; Zheng, G. Fully-Automatic Landmark Detection in Skull X-ray Images. Submitted to Automatic Cephalometric X-ray Landmark Detection Challenge. 2014. Available online: https://api.semanticscholar.org/CorpusID:6412774 (accessed on 17 September 2023).
11. Mirzaalian, H.; Hamarneh, G. *Automatic Globally-Optimal Pictorial Structures with Random Decision Forest Based Likelihoods for Cephalometric X-ray Landmark Detection*; Simon Fraser University: Burnaby, BC, Canada, 2014.
12. Vandaele, R.; Maré, R.; Jodogne, S.; Geurts, P. *Automatic Cephalometric X-Ray Landmark Detection Challenge 2014: A Tree-Based Algorithm*; University of Liege: Liege, Belgium, 2014.
13. Urschler, M.; Ebner, T.; Štern, D. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. *Med. Image Anal.* **2018**, *43*, 23–36. [CrossRef] [PubMed]
14. Ibragimov, B.; Likar, B.; Pernus, F.; Vrtovec, T. *Automatic Cephalometric X-ray Landmark Detection by Applying Game Theory and Random Forests*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1–8.
15. Oh, K.; Oh, I.; Lee, D. Deep anatomical context feature learning for cephalometric landmark detection. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 806–817. [CrossRef]
16. Arık, S.Ö.; Ibragimov, B.; Xing, L. Fully automated quantitative cephalometry using convolutional neural networks. *J. Med. Imaging* **2017**, *4*, 014501. [CrossRef] [PubMed]
17. Park, S.B. Cephalometric Landmarks Detection using Fully Convolutional Networks. Ph.D. Thesis, Seoul National University Graduate School, Seoul, Republic of Korea, 2017.
18. Payer, C.; Štern, D.; Bischof, H.; Urschler, M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Med. Image Anal.* **2019**, *54*, 207–219. [CrossRef]

19. Chen, R.; Ma, Y.; Chen, N.; Lee, D.; Wang, W. *Cephalometric Landmark Detection by Attentive Feature Pyramid Fusion and Regression-Voting*; Medical Image Computing and Computer Assisted Intervention—MICCAI 2019; Springer: Cham, Switzerland, 2019; pp. 873–881.
20. Ourselin, S.; Joskowicz, L.; Sabuncu, M.R.; Unal, G.; Wells, W. *Regressing Heatmaps for Multiple Landmark Localization Using CNNs*; Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016; Springer: Cham, Switzerland, 2016; Volume 9901, pp. 230–238.
21. Liu, X.; Gao, K.; Liu, B.; Pan, C.; Liang, K.; Yan, L.; Ma, J.; He, F.; Zhang, S.; Pan, S. Advances in deep learning-based medical image analysis. *Health Data Sci.* **2021**, *2021*, 786793. [CrossRef]
22. Lindner, C.; Bromiley, P.A.; Ionita, M.C.; Cootes, T.F. Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1862–1874. [CrossRef]
23. Chen, C.; Yang, X.; Huang, R.; Shi, W.; Liu, S.; Lin, M.; Huang, Y.; Yang, Y.; Zhang, Y.; Luo, H. Region proposal network with graph prior and IoU-balance loss for landmark detection in 3D ultrasound. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1–5.
24. Yang, D.; Zhang, S.; Yan, Z.; Tan, C.; Li, K.; Metaxas, D. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, NY, USA, 16–19 April 2015; pp. 17–21.
25. Bayramoglu, N.; Nieminen, M.T.; Saarakkala, S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. *Int. J. Med. Inf.* **2022**, *157*, 104627. [CrossRef] [PubMed]
26. Lee, H.; Park, M.; Kim, J. *Cephalometric Landmark Detection in Dental X-ray Images Using Convolutional Neural Networks*; SPIE: Orlando, FL, USA, 2017; Volume 10134. [CrossRef]
27. Qian, J.; Cheng, M.; Tao, Y.; Lin, J.; Lin, H. CephaNet: An Improved Faster R-CNN for Cephalometric Landmark Detection. In Proceedings of the 2019 IEEE 16th International symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 868–871. [CrossRef]
28. Zhang, J.; Liu, M.; Shen, D. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Trans. Image Process.* **2017**, *26*, 4753–4764. [CrossRef] [PubMed]
29. Zeng, M.; Yan, Z.; Liu, S.; Zhou, Y.; Qiu, L. Cascaded convolutional networks for automatic cephalometric landmark detection. *Med. Image Anal.* **2021**, *68*, 101904. [CrossRef] [PubMed]
30. Ao, Y.; Wu, H. Feature Aggregation and Refinement Network for 2D Anatomical Landmark Detection. *J. Digit. Imaging* **2023**, *36*, 547–561. [CrossRef] [PubMed]
31. Kim, H.; Shim, E.; Park, J.; Kim, Y.; Lee, U.; Kim, Y. Web-based fully automated cephalometric analysis by deep learning. *Comput. Methods Programs Biomed.* **2020**, *194*, 105513. [CrossRef] [PubMed]
32. Lian, C.; Liu, M.; Zhang, J.; Shen, D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 880–893. [CrossRef] [PubMed]
33. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
34. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015 Conference Proceedings, Munich, Germany, 5–9 October 2015; pp. 234–241.
35. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**. [CrossRef]
36. Liu, D.; Zhou, K.S.; Bernhardt, D.; Comaniciu, D. Search strategies for multiple landmark detection by submodular maximization. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2831–2838. [CrossRef]
37. Kang, J.; Oh, K.; Oh, I. Accurate landmark localization for medical images using perturbations. *Appl. Sci.* **2021**, *11*, 10277. [CrossRef]