Accurate Landmark Localization for Medical Images Using Perturbations

Junhyeok Kang\textsuperscript{1,}, Kanghan Oh\textsuperscript{2} and Il-Seok Oh\textsuperscript{1,\ast}

\textsuperscript{1} Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, Korea; k1101jh@jbnu.ac.kr  
\textsuperscript{2} Department of Computer and Software Engineering, Wonkwang University, Iksan 54538, Korea; khoh888@wkku.ac.kr  
\ast Correspondence: isoh@jbnu.ac.kr; Tel.: +82-63-270-3401

Abstract: Recently, various studies have been proposed to learn the rich representations of images during deep learning. In particular, the perturbation method is a simple way to learn rich representations that has shown significant success. In this study, we present effective perturbation approaches for medical landmark localization. To this end, we report an extensive experiment that uses the perturbation methods of erasing, smoothing, binarization, and edge detection. The hand X-ray dataset and the ISBI 2015 Cephalometric dataset are used to evaluate the perturbation effect. The experimental results show that the perturbation method forces the network to extract richer representations of an image, leading to performance increases. Moreover, in comparison with the existing methods that lack any complex algorithmic change of network, our methods with specific perturbation methods achieve superior performance.

Keywords: artificial intelligence; landmark localization; context feature learning; image perturbation

1. Introduction

Landmark localization has been widely used in diverse applications. An early application can be found in face recognition [1]. Thanks to the success in facial analysis, its application has expanded to medical applications such as the analysis of fractures of the humerus [2], precise navigation in magnetic resonance imaging (MRI) scans during neurosurgery [3], and aortic valve localization for pre-operative surgical planning [4]. Other important medical fields include landmark localization for hand X-ray images and cephalometric images. In hand X-ray images, doctors estimate bone age to diagnose growth abnormalities and endocrine disorders [5]. For bone age estimation, they usually use a method such as TW3 [6] that requires accurate landmark localization. Cephalometric landmark localization is an essential task for orthodontic analysis and treatment planning. Based on the geometric features of distances and angles of specific landmarks, dentists create treatment plans. Landmark localization by humans has several drawbacks. First, it is a laborious process that is time demanding. Second, there are high intra-person and inter-person variations [7]. Recently, to solve this problem, research to automate the landmark localization process by deep learning has been actively conducted.

In the early days, landmark localization study was conducted by using probabilistic models, such as the active appearance model (AAM). However, the deep learning method achieved much higher landmark positions accuracy than the classical probabilistic models [8,9]. We introduce two recent papers that achieved state-of-the-art accuracy using deep learning. Payer et al. used two cascaded neural networks [10]. The first network extracts the candidate region of landmarks, whereas the second network determines the exact location of landmarks. The results of the two networks are combined to decide the final landmark locations. In contrast, Oh et al. used a single neural network working entirely in an end-to-end manner [11]. The authors also improved the accuracy by applying the...
anatomical context (AC) loss by considering the statistical distributions of distance and angle between landmarks.

The main reason for the high performance of deep learning lies in learning rich representations [12]. The most straightforward way of representation learning is to use numerous hidden layers, deepening the network. Another effective way of learning richer representations that is popularly used in deep learning is regularization, such as data augmentation and dropout. Recently, a new approach was proposed that perturbs an original input image with the aim of hindering the learner from focusing only on local features, instead enabling the learner to capture semantically global features. There are a variety of techniques for perturbing images, such as adding noise, erasing, and smoothing of patches of an input image. There are abundant reports of successfully applying image perturbation to improve the classification performance [13,14].

Despite the broad applications of image perturbation, only a few such applications can be found in the landmark localization task. To the best of our knowledge, ref. [11] is the first and only attempt to apply smoothing perturbation to localize the landmarks for cephalometric images. This fact motivates us to explore a broader class of perturbation schemes, and to identify the optimal scheme. Our experiment uses region elimination, smoothing, binarization, and edge detection as perturbators. In the experiments on hand X-ray images, the hybrid scheme showed the best accuracy. Randomizing the location and size of the patches to be perturbed enhanced the performance.

Section 2 presents an overview of the related works, whereas Section 3 describes the proposed methods. Section 4 presents our experimental results, and Section 5 discusses and concludes the paper.

2. Related Works

In the first part, we review the efforts that have been developed in the deep learning area to obtain richer representations. The second part gives an overview of landmark localization techniques for various images, such as hand X-ray, cephalometric, and face.

2.1. Rich Representation Learning

It is well known that the generalization ability of deep learning originates from the rich representations extracted through numerous hierarchical layers in the network [12]. Many studies have attempted to improve the generalization ability of the network by developing a better network structure. Network in network, inception, and residual networks are typical of the approach [15–17].

Recent studies have proposed a new approach that forces the learner to solve an additional task. A denoising convolutional auto-encoder (CAE) has been widely employed [18]. It forces the network to learn the representations through the task that eliminates the noise artificially added on purpose. In recent years, there has been an attempt to make models understand the context of the entire image, to extract the rich representations for the semantic segmentation or object localization problems [19–21]. The methods force networks to seek global representation, rather than focusing on the most discriminative local part. To this end, these methods randomly hide important image patches, and then apply the deep learning model using the perturbed images to minimize classification error. Some studies [22,23] have employed an inpainting-based self-supervised learning strategy to learn the rich representation. This strategy randomly eliminates image patches and minimizes pixel-level reconstruction error. Refs. [24,25] used image patch-wise reassembly and matching tasks. The studies divided an image into several patches and solved the problem of aligning the shuffled patches. The above-mentioned papers proved that richer representations improved the performance of the image recognition task.

We note that similar approaches can be applied to the landmark localization to improve the accuracy. However, application to landmark localization can only be found in one paper. Oh et al. applied the smoothing perturbation to a cephalometric image during the
learning process of landmark localization [11]. The paper showed that the perturbation improved the accuracy of landmark localization using the ISBI dataset.

2.2. Landmark Localization

Landmark localization was first developed to analyze face images, and was then expanded to medical images, such as hand X-ray and cephalometric images. Ref. [1] provides an excellent review of face landmark localization. Our review is focused on medical images.

As for classical approaches, for the medical landmark detection the random forests-wise methods [15,26] have been proposed. They employed a regression-voting strategy and showed outstanding performance. The region proposal approaches [12,15,27–29] have applied a cascade process that consists of region proposal detection and landmark point localization. Although these approaches have shown success for the medical landmark localization problem, the computational complexity is increased, and an architecture modification is sometimes required. Moreover, its hyper-parameters of complex architecture make the problem heuristic. In recent years, deep learning-based end-to-end learning strategy [11], which employs the whole image as input, has been outperforming region proposal approaches. This approach incorporates the region proposal and landmark point detection without feature engineering by human knowledge, and this has led to outstanding performance. When considering output designs, the approaches can be divided into coordinates regression [29–31] and heatmap regression [10,11,32–35]. For coordinate regression, the method returns x and y coordinates, which means one pixel of each landmark using CNN with fully connected layers. In contrast, heatmap regression approaches estimate the heatmap of each landmark, which is located on the centroid of the landmark. In such a way, Gaussian distribution is typically used to draw a heatmap shape, and truth landmark points are considered as the centroid of the Gaussian. Recently, heatmap regression-wise approaches have outperformed the coordinate regression methods, and our method is also based on heatmap regression.

3. Materials and Methods

Figure 1 illustrates the overall framework of the proposed model. The input image is perturbed by randomly sampling a patch, applying an operation on the patch, and feeding the resulting image into the network. The landmark points are converted into heatmaps. The U-net is trained using the perturbed image as input, and the heatmaps as output.

![Figure 1. Overall framework of the proposed model. Left side: original image is perturbed before entering the U-Net, Right side: label images where each image represents a landmark.](image-url)
3.1. Datasets

Our model was tested with the public hand radiograph dataset in Table 1, Digital Hand Atlas. To compare with state-of-the-art [10], we followed the same process as [10,26,33]. We used 895 images of the dataset with an average size of $1563 \times 2169$ pixels with 37 landmarks. To evaluate the network, we assumed a wrist width of 50 mm determined by two of the annotated landmarks at the wrist, and performed three-fold cross validation setup by splitting 895 images into approximately 600 training images and 300 test images per fold.

Table 1. Statistics of datasets.

| Property      | Digital Hand Atlas | ISBI 2015 |
|---------------|--------------------|-----------|
| Number of landmarks | 37                | 19        |
| Train set     |                    |           |
| Fold 1        | 597                | 150       |
| Fold 2        | 597                |           |
| Fold 3        | 596                |           |
| Test set      |                    | Test1 (validation): 150 |
| Fold 1        | 298                | Test2 (on-site competition): 100 |
| Fold 2        | 298                |           |
| Fold 3        | 299                |           |
| Resolution    | Variable (on average $1563 \times 2169$) | $1935 \times 2400$ |

In addition, to show that our model works well in landmark localization on X-ray image, we used Cephalometric X-ray data in Table 1 used in the IEEE International Symposium on Biomedical Imaging 2015 challenges (ISBI 2015) [36]. This dataset consists of 400 cephalometric X-ray images of 400 patients with 19 landmarks. Two experienced medical doctors produced the landmarks, and the average point was used as GT. The resolution of the image was $1935 \times 2400$, with pixel size of 0.1 mm $\times$ 0.1 mm. The dataset was divided into 150 training, 150 test1 (validation), and 100 test2 (on-site competition). We trained the network for 2000 epochs using training data.

3.2. Overall Algorithms

Algorithm 1 illustrates a meta-algorithm for the proposed landmark localization method. The label $L_i$ contains the information about k landmark positions, in the form $((x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k))$. A hand X-ray image has 37 landmarks, so $k = 37$. A cephalometric image has 19 landmarks, so $k = 19$. Excluding line 4, Algorithm 1 is the same as the standard BP (back-propagation) learning algorithm. The output of Algorithm 1 is the learned neural network weight.

Algorithm 1. Network training with local feature perturbation

**Input**: Train set $I_i$ (image) and $L_i$ (label set) for $i = 1, 2, \ldots, n$, perturbation operation $q$.

**Output**: Learned network weights.

1. Initialize network weights.
2. While network is not converged
3. For $i = 1$ to $n$.
4. $I_i' = \text{random perturbation of } I_i$ using Algorithm 2 with the operation $q$.
5. $I_i'' = \text{random augmentation of } I_i'$.
6. Perform forward computation with $I_i''$ and compute error $e_i$.
7. Back-propagate $e_i$ and update weights.
Algorithm 2. Image perturbation

Input: Input image I with size H × W, perturbation operation q, patch size range [r_{min}, r_{max}].
Output: Perturbed image I′.

1: Decide the size (h, w) of patch p by uniform random sampling with the range [r_{min}, r_{max}].
2: Decide the center (c_x, c_y) of path p by uniform random sampling with the range [h/2, H − h/2] and [w/2, W − w/2].
3: Apply the operation q within patch p of image I, and save the resulting image into I′.
4: Return I′.

Line 4 deserves attention. The perturbation can be an operation that reduces the information of the input image. In our implementation, we used an operation of smoothing, blackout, whiteout, binarization, or edge detection for the perturbation. The perturbation operation is applied to a patch whose location and size are decided randomly, as explained in Algorithm 2. Algorithm 2 is described in detail in Section 3.3. After altering the patch in line 5, the augmentation is applied to the perturbed image. Rotation, translation, rescale, and color-jitter are applied as augmentation operations. The augmented image is used as input to the network in line 6. Figure 2 illustrates the altered image with each of the 5 perturbation operations.

Figure 2. Visualization of perturbed images, (a) is the original image, (b–f) are the perturbed images, (b) is perturbed with blackout, (c) is perturbed with whiteout, (d) is perturbed with smoothing, (e) is perturbed with binarization, and (f) is perturbed with edge detection.
Note that Algorithm 1 works in the online mode in generating a train sample. The online mode generates a train sample just before inputting the sample into the network. The online mode has the advantage that the size of train set can be considered unlimited.

3.3. Perturbation Operations

In lines 1 and 2 of Algorithm 2, a patch is generated by random sampling of location and size. In line 3, one of the perturbation operations \( q \) is applied to the patch. We summarize the values of parameters used in our experiment in Table 2. The values were determined empirically. We performed the cross validation using ISBI 2015 train set in determining the values. The same values were applied to both ISBI 2015 and Digital Hand Atlas datasets.

Table 2. Values of parameters used in our experiments.

| Perturbation Method | Term | Value |
|---------------------|------|-------|
| Smoothing           | \( \mu \) | 0.2   |
|                     | \( \sigma \) | 0.15  |
| Binarization        | \( \mu \) | 0.2   |
|                     | \( \sigma \) | 0.15  |
| Edge detection      | \( \mu \) | 3.5   |
|                     | \( \Sigma \) | 1.5   |

3.3.1. Blackout and Whiteout

The blackout changes all the pixels in patch \( p \) into black, being value 0. The whiteout changes all the pixels in \( p \) into white, being value 1. These operations are the extreme cases, in the sense that they remove all the information within the patch. The network should look at neighboring regions outside of patch \( p \), so as to learn the landmarks within \( p \). Figure 2b,c show examples of the blackout and whiteout, respectively.

3.3.2. Smoothing

Smoothing can be implemented in several ways. We chose the method that decreases the resolution of patch \( p \), and then increases the resolution back to the original size. The scaling factor \( \alpha \) is determined by sampling from a normal distribution \( N(\mu, \sigma^2) \), where \( \mu \) and \( \sigma^2 \) are the mean and variance. In our experiment, we used \( \mu = 0.2 \) and \( \sigma = 0.15 \). The operation is as follows:

\[
p' = \text{scaleup} \left( \text{scaledown}(p, \alpha), \frac{1}{\alpha} \right)
\]

This algorithm reduces the information in \( p \) to some extent, according to the randomly sampled \( \alpha \). Therefore, the network should learn the representation using both the local information in patch \( p \), and the global information outside \( p \). Figure 2d illustrates an example image perturbed by smoothing.

3.3.3. Binarization

Binarization converts pixels in \( p \) into black (0) or white (1). The threshold value is obtained from the Otsu algorithm [37]. The threshold value in Otsu algorithm is determined such that the value maximizes inter-class variance by separating pixels into two classes, background and foreground. In our perturbation, a value generated from the normal distribution \( N(\mu, \sigma^2) \), where \( \mu = 0.2, \sigma = 0.15 \) is added to the threshold. The resulting threshold is applied to the patch \( p \) to binarize. Figure 2e illustrates an example image perturbed by binarization. As was the case for smoothing, the network should learn the representation using both the reduced local information in patch \( p \), and the global information outside \( p \).
3.3.4. Edge Detection

Edge detection is performed using Canny edge detection algorithm [38], resulting in pixels with 0 (non-edge) or 1 (edge). The sigma value for the Canny edge detector is determined by random sampling from $N(\mu, \sigma^2)$, where $\mu = 3.5$, $\sigma = 1.5$. A larger value of sigma results in coarser scale edges and greater noise suppression effect.

Unlike other operations that only degrade the information in the image, the edges have two opposite facets, degrading or upgrading the information. The image is upgraded in the sense that the positions that change the most in the image are highlighted by being converted to value 1, indicating the edge. As a result, it is expected that the landmarks lying close to the edges will be more easily detected. In contrast, the shading information disappears by edge detection. Therefore, the landmarks far from the edges should rely on global information. We expect that the two opposite effects will make the network more robust to image variations.

3.3.5. Hybrid

The hybrid method uses all the previously described perturbation methods probabilistically. Each perturbation is chosen with equal probability. Each perturbation has different effects, so the hybrid scheme provides more active perturbation.

3.4. Heatmap Regression

Algorithm 1 illustrates that the input of the neural network is an image $I$, and the output is a label set $L$. The label set provides the information about the landmark positions denoted by $L = ((x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k))$, where $k$ is the number of landmarks.

Two approaches are available for representing $L$. The first approach uses the coordinate values $x_i$ and $y_i$ as they are. In this case, the network output layer has 2k output nodes. The second approach converts the coordinate values of $l_i = (x_i, y_i)$ into a heatmap, where a Gaussian distribution is formed centered at $l_i$. The heatmap size is the same as the input image. The heatmap is normalized in the range 0–1. Since each of the landmarks has its own heatmap, $k$ heatmaps are made. The network output layer should output $m \times n \times k$ tensor, where $m \times n$ is the size of the input image, and $k$ is the number of landmarks. The second approach is termed heatmap regression; an approach that has been proven to be superior [10,32,33]. This paper adopts the heatmap regression method.

The literature adopting the heatmap regression usually used the Gaussian distribution to mark a landmark in the heatmap. Figure 3a shows a hand X-ray image, and Figure 3b illustrates the heatmap using Gaussian distribution for the landmark at Figure 3a. Another option is to use the Laplace distribution that is sharper at the center than the Gaussian. Since the goal of the neural network is to determine a landmark position accurately, it seems that the Laplace distribution is better. This paper adopts the Laplace distribution in forming heatmaps. Figure 3c shows a heatmap formed by our method.

---

**Figure 3.** (a) is a visualization of landmarks of a resized sample image in the Digital Hand Atlas dataset. (b–d) are the heatmaps for a landmark of (a).
For the test stage, we should devise a method to convert a predicted heatmap into coordinate values. Figure 3d shows the predicted heatmap for the landmark. To calculate landmark coordinates, we select pixels whose values are larger than 0.85 of the maximum heatmap pixel value. The averages of these pixel positions are regarded as landmark coordinates on the heatmap.

3.5. Network Architecture

Our architecture is based on U-net [39], and U-net with attention module [40]. The contracting path and expanding path consist of the repeated application of two $3 \times 3$ convolutions, each followed by LeakyReLU and a $2 \times 2$ max pooling (for contracting path), or upsampling (for expansive path). The number of feature channels was increased from 64, 128, 256, to 512 in the contracting path, and decreased to 512, 256, 128, to 64 in the expansive path. Attention gates filter the features propagated through the skip connections [40]. We used AC loss as a cost function and minimized cost function by using the Adam optimizer. AC loss is a function proposed by [11]. This function calculates the loss by considering the distance and angle between landmarks. The learning rate is initialized as 0.0001 and annealed by the Cosine annealing schedule.

4. Results

The proposed method was evaluated on hand radiograph data and cephalometric X-ray data to compare with papers [10,11] that achieved the state-of-the-art. The network was trained and tested in PyTorch with Intel Xeon Gold 6126 2.6 GHz CPU, 64 GB of memory and RTX 2080 Ti GPU with 11 GB RAM. For all datasets, we resized input images to 640x800 using the bilinear method. Since our network is based on U-net, it produces heatmap outputs of 640x800 sizes.

The performance of landmark localization methods in Table 3 was evaluated with point-to-point error (PE) and error detection rate (EDR). PE is defined as the Euclidean distance between the ground truth landmark point and the predicted landmark point. The metric “$>n$ mm” in EDR means the ratio of landmarks whose PE is greater than n millimeter (mm).

Table 3. Localization results according to perturbator on the Digital Hand Atlas dataset.

| Method                  | $PE_{all}$ (mm) | EDR (%) |
|-------------------------|-----------------|---------|
|                         | Median | Mean ± SD | >2 mm | >4 mm | >10 mm |
| U-net without perturbator | 0.46   | 0.66 ± 0.84 | 4.27  | 0.47  | 0.05  |
| Blackout                | 0.46   | 0.66 ± 0.68 | 4.13  | 0.39  | 0.03  |
| Whiteout                | 0.47   | 0.69 ± 0.90 | 4.33  | 0.52  | 0.10  |
| Smoothing               | 0.48   | 0.68 ± 0.83 | 4.31  | 0.49  | 0.07  |
| Binarization            | 0.45   | 0.66 ± 0.87 | 4.16  | 0.51  | 0.10  |
| Edge detection          | 0.45   | 0.65 ± 0.77 | 4.19  | 0.46  | 0.06  |
| Hybrid                  | 0.45   | 0.64 ± 0.64 | 3.96  | 0.34  | 0.02  |

4.1. Perturbation Performance Comparison

Table 3 compares the results of perturbation methods on the Digital Hand Atlas dataset. All methods were tested based on U-net. In Table 3, all perturbators except whiteout and smoothing increased performance. Among them, the hybrid perturbator shows significant performance improvement. U-net achieved 2 mm range EDR of 4.27%, and mean PE of 0.66. When hybrid perturbator was applied, 2 mm EDR decreased to 3.96%, and mean PE decreased to 0.64.

Table 4 compares the results of perturbation methods on the ISBI 2015 test1 dataset. U-net achieved 2 mm range EDR of 15.47%, and median PE of 12.03. Significant performance improvement has been observed when using whiteout, binarization, and hybrid
perturbator. When using the edge detection perturbator, the network achieved 2 mm range EDR of 14.42%, and median PE of 11.72.

Table 4. Localization results according to perturbator on the ISBI 2015 dataset (test1).

| Method                      | PE<sub>all</sub> (pxl) | EDR (%) |
|-----------------------------|------------------------|---------|
|                             | Median | Mean ± SD | >2 mm  | >2.5 mm | >3 mm  | >4 mm  |
| U-net without perturbator   | 9.12   | 12.03 ± 11.39 | 15.47  | 8.95    | 5.58   | 2.35   |
| Blackout                    | 9.03   | 11.91 ± 12.00 | 14.88  | 8.88    | 4.98   | 2.42   |
| Whiteout                    | 9.03   | 11.70 ± 10.23 | 14.60  | 8.84    | 5.75   | 2.00   |
| Smoothing                   | 9.32   | 12.10 ± 11.47 | 15.02  | 9.33    | 6.04   | 2.42   |
| Binarization                | 9.28   | 11.83 ± 10.19 | 15.37  | 9.23    | 5.33   | 1.96   |
| Edge detection              | 8.90   | 11.78 ± 12.37 | 14.42  | 8.77    | 5.65   | 2.28   |
| Hybrid                      | 8.94   | 11.72 ± 12.01 | 14.70  | 9.33    | 5.12   | 2.00   |

4.2. Quantitative Performance Comparison with Existing Studies

In this section, we perform a quantitative performance comparison with existing studies by using the best perturbation method for each dataset. We trained two networks, U-net and Attention U-net. In Table 5 using Digital Hand Atlas dataset, the existing study of Payer et al. [33] achieved 2 mm range EDR of 5.01% and mean PE of 0.66. Our hybrid perturbator based on U-net achieved 2 mm range EDR of 3.96% and mean PE of 0.64. The hybrid perturbator based on Attention U-net showed a poorer performance than U-net. It achieved 2 mm range EDR of 4.76% and mean PE of 0.71.

Table 5. Quantitative comparison with existing methods on the Digital Hand Atlas dataset.

| Method                      | PE<sub>all</sub> (mm) | EDR (%) |
|-----------------------------|------------------------|---------|
|                             | Median | Mean ± SD | >2 mm  | >4 mm  | >10 mm |
| Payer [33]                  | 0.91   | 1.13 ± 0.98 | 12.4%  | 1.34%  | 0.04%  |
| Štern [26]                  | 0.51   | 0.80 ± 0.91 | 7.80%  | 1.55%  | 0.05%  |
| Urschler [34]               | 0.51   | 0.80 ± 0.93 | 7.81%  | 1.54%  | 0.05%  |
| Payer [10]                  | 0.43   | 0.66 ± 0.74 | 5.01%  | 0.73%  | 0.01%  |
| Hybrid perturbator (U-net)  | 0.45   | 0.64 ± 0.64 | 3.96%  | 0.34%  | 0.02%  |
| Hybrid perturbator (Attention U-net) | 0.45 | 0.71 ± 1.26 | 4.76%  | 0.96%  | 0.27%  |

In Table 6 using ISBI 2015 dataset (test1), Chen et al. [35] achieved a 2 mm range EDR of 13.33% and median PE of 11.7. Our edge detection perturbator based on Attention U-net achieved 2 mm range EDR of 13.26% and median PE of 11.48. The same perturbator based on U-net achieved 2 mm range EDR of 14.42% and median PE of 11.78.

Table 7 shows the results of the ISBI 2015 on-site competition data. The best existing study is Oh et al. [11], which achieved 2 mm range EDR of 24.11% and median PE of 14.45. Our edge detection perturbator based on U-net achieved 2 mm range EDR of 25.58% and median PE of 14.82. The same perturbator based on Attention U-net achieved 2 mm range EDR of 25.21% and median PE of 15.30. The ISBI 2015 test1 data is different from the test2 data. Previous studies [11,42] have reported that landmarks 3 (Orbitale), 6 (Supramentale), 13 (Upper lip), and 16 (Soft tissue pogonion) have a large performance gap between test1 and test2 data. This factor does not guarantee that experiments in the same environment will produce similar results in the test2 dataset.
Table 6. Quantitative comparison with existing methods on the ISBI 2015 dataset (test1).

| Method               | PE\(_{all}\) (pxl) | EDR (%) | Median | Mean ± SD | >2 mm | >2.5 mm | >3 mm | >4 mm |
|----------------------|--------------------|---------|--------|-----------|-------|---------|-------|-------|
| Lindner [15]         | n/a                | 16.7 ± n/a | 26.32% | 19.79%    | 14.81%| 8.53%   |
| Park [32]            | n/a                | 16.91 ± n/a | 23.82% | 18.60%    | 14.21%| 8.88%   |
| Arik [41]            | n/a                | n/a      | 24.63% | 19.08%    | 15.68%| 11.75%  |
| Qian [29]            | n/a                | n/a      | 17.50% | 13.60%    | 10.70%| 9.40%   |
| Chen [35]            | n/a                | 11.7 ± n/a | 13.33% | 7.33%     | 4.46% | 1.65%   |
| Oh [11]              | n/a                | 11.81 ± n/a | 13.80% | 8.80%     | 5.60% | 2.30%   |
| Edge detection perturbator (U-net) | 8.90 | 11.78 ± 12.37 | 14.42% | 8.77%     | 5.65% | 2.28%   |
| Edge detection perturbator (Attention U-net) | 9.96 | 15.30 ± 12.65 | 25.21% | 18.05%    | 12.95%| 6.79%   |

Table 7. Quantitative comparison with existing methods on the ISBI 2015 dataset (test2).

| Method               | PE\(_{all}\) (pxl) | EDR (%) | Median | Mean ± SD | >2 mm | >2.5 mm | >3 mm | >4 mm |
|----------------------|--------------------|---------|--------|-----------|-------|---------|-------|-------|
| Lindner [15]         | n/a                | 19.20 ± n/a | 33.89% | 28.00%    | 22.37%| 12.57%  |
| Park [32]            | n/a                | 20.58 ± n/a | 32.41% | 25.84%    | 19.84%| 12.32%  |
| Arik [41]            | n/a                | n/a      | 32.32% | 25.84%    | 20.89%| 15.37%  |
| Qian [29]            | n/a                | n/a      | 27.60% | 23.85%    | 20.35%| 14.10%  |
| Chen [35]            | n/a                | 14.80 ± n/a | 24.95% | 17.16%    | 11.47%| 4.95%   |
| Oh [11]              | n/a                | 14.45 ± n/a | 24.11% | 16.64%    | 10.74%| 4.27%   |
| Edge detection perturbator (U-net) | 10.71 | 14.82 ± 13.45 | 25.58% | 17.74%    | 11.53%| 5.42%   |
| Edge detection perturbator (Attention U-net) | 9.6 | 15.30 ± 12.65 | 25.21% | 18.05%    | 12.95%| 6.79%   |

Figure 4a shows mean radial error (MRE) curves for the Digital Hand Atlas dataset according to different perturbations. As Table 3 illustrates, the hybrid perturbation depicted with black dashed line shows the best performance. Figure 4b shows MRE curves for the ISBI 2015 test1 dataset. As Table 4 illustrates, the edge perturbation depicted with red line leads other perturbations. MRE curves on Figure 5 are shown according to different epochs. Between the 20th and 50th epochs, the network performance has improved dramatically. After the 500th epoch, the network performance has improved slightly.
Figure 4. MRE curves according to different perturbations. (a) Digital Hand Atlas dataset (U-Net). (b) ISBI 2015 test1 dataset (U-Net).

Figure 5. MRE curves according to different epochs. (a) Digital Hand Atlas dataset (Hybrid with U-Net). (b) ISBI 2015 test1 dataset (Edge Detection with U-Net).

Figure 6 show the landmarks on the original image by overlapping the ground truths (red dots) and the predictions (blue dots) for the Digital Hand Atlas and ISBI datasets, respectively. The figures illustrate that our model predicted the landmarks very closely to the ground truth.

Figure 6. Visualization of landmarks of a sample test image in the Digital Hand Atlas and ISBI 2015 datasets. Red dots are the ground-truth of image, and green dots are the prediction result of U-Net with the hybrid perturbation model. (a) Digital Hand Atlas. (b) ISBI 2015.

Figure 6. Visualization of landmarks of a sample test image in the Digital Hand Atlas and ISBI 2015 datasets. Red dots are the ground-truth of image, and green dots are the prediction result of U-Net with the hybrid perturbation model. (a) Digital Hand Atlas. (b) ISBI 2015.
5. Discussions and Conclusions

The performance improvement by the perturbations can be explained by the fact that a perturbation functions as a regularizer with data augmentation. Over the entire training phase, the sizes and positions of the patches are chosen randomly, and each training sample is different from one another. The various perturbation parameters described in Section 3.3 enforce further augmentation effect. We suggested several perturbation algorithms and tested them. The perturbation operations cause a loss of the local information of the image. This forces the models to search for global features. The perturbations are different in the amount of information loss. Figure 7 shows the diverse attributes of the perturbators. The blackout and whiteout perturbations erase all local information. Thus, the network has no choice but to localize the landmarks by learning the global features outside the patch. Smoothing reduces the information in the patch, so the network can rely both on the inside and outside of the patch. Binarization and edge detection have two opposite facets, weakening or strengthening the local information of the patch. The weakening facet happens because the details of the patch disappear. The strengthening facet happens because the operations sharpen the region boundaries. So, binarization and edge detection can be regarded as guiding the feature learning.

The guided feature learning, it is worth noting, is an effective way of enriching the feature representation. Many researchers showed that the guided feature learning improved the accuracy of object detection. A notable result can be found in [43] where the dilated convolution layers played the role of the feature guiding. In that work, the feature guidance is learned in the end-to-end manner. On the contrary, our perturbators accomplish the feature learning by using hand-crafted features as a preprocessor. One of our important future works is to embed the perturbation module in the network architecture and learn the perturbator automatically. This extension is expected to improve the accuracy of landmark localization. Another benefit comes from removing the necessity of manually setting the parameters in Table 2.

In Tables 3 and 4, the edge detection that plays the role of the guided feature learning generally leads other operations. The hybrid that randomly chooses edge detection also resulted in significant improvement. These improvements have a positive effect on planning treatment for dentists. Geometric features such as distances and angles are used to classify the patients into several groups. Accurate classification is very important for the treatment. We argue that the accuracy improvement by the edge detection and hybrid perturbators has a significant impact on building a dental treatment system.

The most important future work is to apply the perturbators to other medical tasks described in Introduction such as analyzing fractures of the humerus and aortic valve localization for pre-operative surgical planning.

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

Funding: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT), funding number (No. 2019R1F1A10635221).
Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: ISBI: http://www-ontust.easmat.tw/~cweiwang/ISBI2015/challenge1/, Digital Hand Atlas: https://iplab.usc.edu/computer-aided-bone-age-assessment-of-children-using-a-digital-hand-atlas-2/, https://github.com/christianpayer/MedicalDataAugmentationTool-HeatmapRegression.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, Y.; Ji, Q. Facial landmark detection: A literature survey. Int. J. Comput. Vis. 2019, 127, 115–142. [CrossRef]
2. Negrillo-Cárdenas, J.; Jiménez-Pérez, J.-R.; Cañana-Oya, H.; Feito, F.R.; Delgado-Martínez, A.D. Automatic detection of landmarks for the analysis of a reduction of supracondylyar fractures of the humerus. Med. Image Anal. 2020, 64, 10172. [CrossRef] [PubMed]
3. Edwards, C.A.; Goyal, A.; Russeen, A.E.; Kouzani, A.Z.; Lee, K.H. DeepNavNet: Automated Landmark Localization for Neuronavigation. Front. Neurosci. 2021, 15, 730. [CrossRef] [PubMed]
4. Al, W.A.; Jung, H.Y.; Yun, I.D.; Jang, Y.; Park, H.-B.; Chang, H.-J. Automatic aortic valve landmark localization in coronary CT angiography using colonial walk. PLoS ONE 2018, 13, e0200317. [CrossRef] [PubMed]
5. Pietka, E.; Gertych, A.; Pospišil, S.; Cao, F.; Huang, H.; Ilganz, V. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction. IEEE Trans. Med. Imaging 2001, 20, 715–729. [CrossRef] [PubMed]
6. Tanner, J.M. Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method); Saunders: London, UK, 2001.
7. Kamoen, A.; Luc, D.; Ronald, V. The clinical significance of error measurement in the interpretation of treatment results. Eur. J. Orthod. 2001, 23, 569–578. [CrossRef] [PubMed]
8. Stern, D.; Payer, C.; Lepetit, V.; Urschler, M. Automated Age Estimation from Hand MRI Volumes Using Deep Learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2016; pp. 194–202.
9. Park, J.-H.; Hwang, H.-W.; Moon, J.-H.; Yu, Y.; Kim, H.; Her, S.-B.; Srinivasan, G.; Aljanabi, M.N.A.; Donatelli, R.E.; Lee, S.-J. Automated identification of cephalometric landmarks: Part 1—Comparisons between the latest deep-learning methods YOLOV3 and SSD. Angle Orthod. 2019, 89, 903–909. [CrossRef] [PubMed]
10. Payer, C.; Stern, D.; Bischof, H.; Urschler, M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. 2019, 54, 207–219. [CrossRef] [PubMed]
11. Oh, K.; Oh, I.-S.; Le, T.V.n.; Lee, D.-W. Deep Anatomical Context Feature Learning for Cephalometric Landmark Detection. IEEE J. Biomed. Health Inform. 2020, 25, 806–817. [CrossRef] [PubMed]
12. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798–1828. [CrossRef] [PubMed]
13. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. arXiv 2017, arXiv:1708.04552.
14. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. Proc. Conf. AAAI Artif. Intell. 2020, 34, 13001–13008. [CrossRef]
15. 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] [PubMed]
16. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
18. Masci, J.; Meier, U.; Ciresan, D.; Schmidhuber, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In International Conference on Artificial Neural Networks; Springer: Berlin/Heidelberg, Germany, 2011.
19. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Singh, K.K.; Lee, Y.J. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
21. Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; Huang, T.S. Adversarial complementary learning for weakly supervised object localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
22. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
23. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting With Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
24. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.

25. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2016.

26. Štern, D.; Ebner, T.; Urschler, M. From Local to Global Random Regression Forests: Exploring Anatomical Landmark Localization. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2016.

27. Chu, C.; Chen, C.; Wang, C.-W.; Huang, C.-T.; Li, C.-H.; Nolte, L.-P.; Zheng, G. Fully automatic cephalometric x-ray landmark detection using random forest regression and sparse shape composition. In Proceedings of the ISBI International Symposium on Biomedical Imaging, Beijing, China, 29 April–2 May 2014; pp. 9–16.

28. Mirzaalian, H.; Hamarneh, G. Automatic Globally-Optimal Pictorial Structures with Random Decision Forest Based Likelihoods for Cephalometric X-ray Landmark Detection. In Proceedings of the IEEE International Symposium on Biomedical Imaging Automatic Cephalometric X-ray Landmark Detection Challenge, Beijing, China, 29 April–2 May 2014; pp. 25–36.

29. 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.

30. 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]

31. Zhang, H.; Li, Q.; Sun, Z. Joint voxel and coordinate regression for accurate 3d facial landmark localization. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.

32. Park, S. Cephalometric Landmarks Detection Using Fully Convolutional Networks; College of Natural Sci, Seoul National Univ.: Seoul, Korea, 2017.

33. Payer, C.; Štern, D.; Bischof, H.; Urschler, M. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2016.

34. Urschler, M.; Thomas, E.; Darko, Š. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. 2018, 43, 23–36. [CrossRef] [PubMed]

35. Chen, R.; Ma, Y.; Chen, N.; Lee, D.; Wang, W. Cephalometric Landmark Detection by Attentive Feature Pyramid Fusion and Regression-Voting. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2019.

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

37. Otsu, N. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 1979, 9, 62–66. [CrossRef]

38. Ronneberger, O.; Philipp, F.; Thomas, B. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2015.

39. 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, arXiv:1804.03999.

40. Arik, S.O.; Bulat, I.; Lei, X. Fully automated quantitative cephalometry using convolutional neural networks. J. Med. Imaging 2017, 4, 014501. [CrossRef] [PubMed]

41. Lee, H.; Park, M.; Kim. J. Cephalometric landmark detection in dental x-ray images using convolutional neural networks. In Medical Imaging 2017: Computer-Aided Diagnosis; International Society for Optics and Photonics: Orlando, FL, USA, 2017; Volume 10134.

42. Nie, J.; Anwer, R.M.; Cholakkal, H.; Khan, F.S.; Pang, Y.; Shao, L. Enriched Feature Guided Refinement Network for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9537–9546.