electronics

MDPI

Article

Anatomical Landmark Detection Using a Feature-Sharing Knowledge Distillation-Based Neural Network

Di Huang [1], Yuzhao Wang [2], Yu Wang [2], Guishan Gu [3] and Tian Bai [2,*]

[1] College of Software, Jilin University, Changchun 130012, China; huangdi21@mails.jlu.edu.cn
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China; yuzhaow20@mails.jlu.edu.cn (Y.W.); wangyu2118@mails.jlu.edu.cn (Y.W.)
[3] Bone and Joint Surgery, First hospital of Jilin University, Changchun 130021, China; guguishan@sina.com
* Correspondence: baitian@jlu.edu.cn

Abstract: Existing anatomical landmark detection methods consider the performance gains under heavyweight network architectures, which lead to models tending to have poor scalability and cost-effectiveness. To solve this problem, state-of-the-art knowledge distillation (KD) methods are proposed. However, they only require the teacher model to guide the output of the final layer of the student model. In this way, the semantic information learned by the student model is very limited. Different from previous works, we propose a novel KD-based model-training strategy, named feature-sharing fast landmark detection (FSF-LD), which focuses on intermediate features and effectively transfers richer spatial information from the teacher model to the student model. Moreover, to generate richer and more reliable knowledge, we propose a multi-task learning structure to pretrain the teacher model before FSF-LD. Finally, a tiny and effective anatomical landmark detection model is obtained. We evaluate our proposed FSF-LD on a public 2D hand radiograph dataset, a public 2D cephalometric radiograph dataset and a private 2D hip radiograph dataset. On the 2D hand dataset, our FSF-LD has 11.7%, 12.1%, 12.0,% and 11.4% improvement on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm, r = 4 mm) compared with other KD methods. The results suggest the superiority of FSF-LD in terms of model performance and cost-effectiveness. However, it is a challenge to further improve the detection accuracy of anatomical landmarks and realize the clinical application of the research results, which is also our next plan.

Keywords: knowledge distillation; multi-task learning; landmark detection; teacher–student learning

check for updates

Citation: Huang, D.; Wang, Y.; Wang, Y.; Gu, G.; Bai, T. Anatomical Landmark Detection Using a Feature-Sharing Knowledge Distillation-Based Neural Network. *Electronics* **2022**, *11*, 2337. https://doi.org/10.3390/electronics11152337

Academic Editor: Dah-Jye Lee

Received: 3 July 2022
Accepted: 24 July 2022
Published: 27 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accurate anatomical landmark detection is a primary and vital task in medical image analysis, establishing treatment programs and prognosis, due to its important role in diagnosing various diseases [1–4]. However, manually locating landmarks is time-consuming, and the individual variation between different doctors results in quality deviations. Therefore, the demand for reliable automatic detection of anatomical landmarks has been increasing [5]. Remarkable advances in anatomical landmark detection have been witnessed with the rapid development of deep convolutional neural networks (CNN). Ref. [6] proposed a multi-task learning method, which trains models to predict the landmarks and edges simultaneously. Capturing the resolution between landmarks greatly improved the performance of the models. Ref. [7] proposed a novel CNN architecture and split the landmark detection into two easier substeps: first, locally accurate but ambiguous candidate predictions; and second, refined landmark detection. Ref. [8] applied an end-to-end network named CephaNN that includes two novel parts: the multi-head part and the attention part. Ref. [9] designed a cascaded three-stage network to localize cephalometric landmarks. However, these models are often too large to be deployed on resource-limited devices, which is an obstacle to the wide application of deep learning in clinical medicine. Therefore, the purpose

of this paper is to decrease the scale of the model without model performance degradation, improve the detection accuracy of anatomical landmark detection, and achieve high-quality automatic detection of anatomical landmarks.

As a model compression and acceleration technology, knowledge distillation (KD) has broad applications in computer vision (CV), speech recognition, natural language processing (NLP), etc. KD is often characterized by the so-called 'Student–Teacher' (S-T) learning framework, and its training objective is to transfer the knowledge from a pretrained teacher model to a tiny target model. Based on the principle of KD, we propose a cost-effective model-training strategy for anatomical landmark detection, which decreases the scale of the model without model performance degradation. Unlike natural images, radiograph medical images often have low contrast. The anatomical landmarks from different patients appear diverse in shape, which makes model training difficult. Ref. [10] considered the differences between the features of the teacher and student in different areas and proposed focal and global distillation (FGD) to reduce background interference. Ref. [11] proposed an online KD framework named OKDHP which is designed as a one-stage knowledge distillation model of human body structure.

As Figure 1 shows, different from previous methods [12], we design feature-sharing knowledge distillation (FSF-LD), which enables learning richer information from the teacher and provides more flexibility for performance improvement. Moreover, it is known that a poor teacher is prone to mislead a student model with noise, resulting in poor network performance. Hence, to ensure the feasibility of FSF-LD and improve the performance of the teacher model, we pretrain the teacher model with a multi-task structure. It contains two task branches: a landmark detection task and a segmentation of landmark's local neighborhood task. Considering their similarity, the teacher model will learn more robust and universal feature representations. Moreover, we impose the Non-Local Block (NLB) [13] to process the output of the encoder, which adaptively integrates local features with their global dependencies to capture contexts. Thus, the teacher model obtains more topology and global structure information.

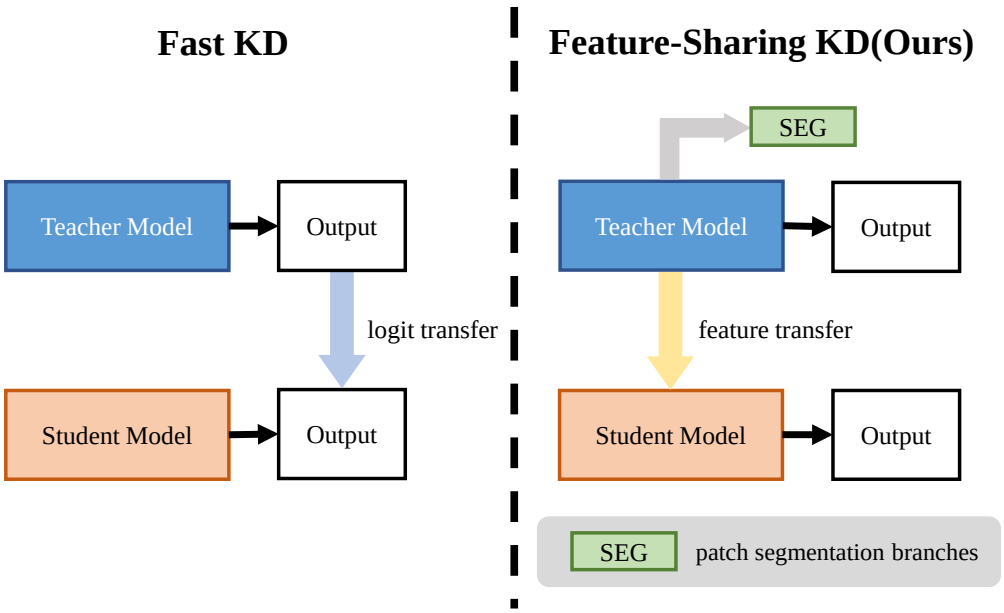

Figure 1. Different from Fast-KD [12], our proposed FSF-LD focuses on intermediate features and transfers them in an effective way. Moreover, to improve the teacher model's performance, we design a multi-task structure to pretrain the teacher model. The details are provided in Section 3.

In summary, our contributions are as follows:

- Focusing on the issue of anatomical landmark detection model deployment, we propose a model-training method named feature-sharing fast landmark detection (FSF-LD), which enables a lightweight model to approximately achieve high performance as good as that of a heavy but strong model. Our proposed FSF-LD outperforms state-of-the-art KD methods on landmark detection.
- Moreover, we propose a multi-task learning (MTL) method to pretrain the teacher network and improve its ability to exploit features and represent knowledge. We carry out some extensive experiments to validate the efficiency and superiority of our MTL methods.

The layout of this paper is as follows: Section 2 describes related work; Section 3 describes the implementation of the algorithm and the details of the models; Section 4 describes the datasets, evaluation methods, and the analysis of experimental results; Section 5 discusses the conclusions and future work.

2. Related Work

2.1. Anatomical Landmark Detection

Anatomical landmark detection plays an important role in medical image analysis. Unfortunately, manual annotation is typically tedious, time-consuming, and subjective. To address these difficulties, many CNN-based methods have been used to automatically localize landmarks in medical images.

Recently, Ref. [7] proposed a novel CNN-based method named Spatial Configuration-Net (SCN), which splits the localization detection task into two simple subproblems. One component makes locally accurate but ambiguous candidate predictions, while the other component improves robustness to ambiguities by incorporating the spatial configuration of landmarks. Inspired by [7], we extract the landmark coordinates from the heatmap images in two branches. However, this method also suffers from inter- and intra-user variability. Ref. [9] proposed cascaded three-stage convolutional neural networks to predict cephalometric landmarks automatically. This model obtains inefficiencies during training and testing because it includes 21 individual CNN models which result in a high cost. Ref. [6] imposed the relative position constraints on each landmark by defining edges among landmarks according to the clinical significance. With multi-task learning, the model can predict the landmarks and edges simultaneously. In this paper,we use a multi-task learning method due to its excellent performance in anatomical landmark detection. Refs. [14–16] proposed an advanced adversarial training method to defend against adversarial examples which are samples created by adding a little noise to the original sample data. The proposed method can correctly classify adversarial examples which will be wrongly classified by a neural network.

However, most existing state-of-the-art methods tend to have very deep and wide cumbersome models, which require large computation and amounts of labeled datasets. These limitations have hindered their clinical application. In this paper, we design a model-training strategy named feature-sharing fast landmark detection (FSF-LD) structure to obtain fast landmark detection models.

2.2. Knowledge Distillation

Knowledge distillation (KD) was originally proposed and generalized in classification tasks [17], and it refers to effective techniques that facilitate the training process of tiny models under the supervision of large models. The knowledge is transferred by minimizing the differences between the knowledge representations they produce. The large model providing knowledge is called the teacher model, and the tiny model learning knowledge is called the student model. The knowledge representations here can refer to logits information, intermediate features, and so on. Ref. [17] used teacher model outputs as soft targets. Ref. [18] captured spatial attention maps and defined them as knowledge representations to transfer. Ref. [19] defined the distilled knowledge to be transferred as the flow between

two layers. To obtain a better student model, Ref. [20] designed an information–theoretic framework for knowledge transfer which formulates knowledge transfer as maximizing the mutual information between the teacher and the student networks.

It is crucial for KD to design the knowledge representation and the method of information transferring [21]. Different from classification tasks which refer to category-level discriminative knowledge, landmark detection requires richer structured information and complex knowledge representation. Ref. [12] proposed a new fast pose distillation training strategy in human pose estimation. It adopted knowledge distillation and provided extra supervision guidance via the mimicry loss function. Ref. [22] presented MoVNet, a 3D real-time human pose estimation model where a heatmap and location map are transferred as knowledge. Therefore, we can conclude that an effective knowledge representation is supposed to express learned information in a more general way [21].

Most existing knowledge distillation methods focus on deep intermediate features; logit distillation methods ignore intermediate features resulting in poor performance. Inspired by [12,21], this paper makes an effort to explore and compare various methods of knowledge representation and transfer in landmark detection. Furthermore, we try to explain their working rationales.

2.3. Multi-Task Learning

As an excellent learning paradigm in machine learning, multi-task learning (MTL) was applied to exploit useful information from related tasks [23]. Benefiting from the extra information, MTL improves its generalization ability and makes latent and effective features easy to capture. Originally, an important motivation of MTL was data sparsity alleviation by aggregating existing knowledge in all the tasks to obtain a more accurate learner for each task. Ref. [23] classifies the MTL structure into five categories. The most widely used MTL structure is a feature-learning approach, which can be implemented by a hard parameter-sharing structure. Considering the similarity of related works, it is reasonable to assume that different tasks share a common feature representation. In [24], they transfer landmark detection tasks into landmark segmentation of the landmark's local neighborhood tasks. Inspired by [24], we make a bold and reasonable assumption that there is a strong similarity between the segmentation of landmark local neighborhoods and landmark detection tasks. They both exploit the local area information around the landmark as a strong identification of the landmark. The common semantic information for the two tasks is universal and effective.

The performance of the student model is influenced by the teacher model.Therefore, in this paper, we optimize the teacher model on both landmark detection tasks and segmentation of landmark local neighborhood tasks to improve the teacher model's performance. In this way, it enables the pretraining of a stronger teacher model. Thus, the student model is easy to learn from the knowledge and improve its performance. We will state our work in detail in Section 3.3.

3. Feature-Sharing Fast Landmark Detection Strategy

3.1. Anatomical Landmark Detection Task

Anatomical landmark detection aims to predict the coordinates of anatomical landmarks on a given medical image. To train a model in a supervised manner, we should have access to a training dataset $\{I^i, G^i\}_{i=1}^N$, which contains N medical images. I^i and G^i are the i-th medical image and corresponding landmarks' coordinates. If medical image I^i with K landmarks, G^i in the image space is defined as

$$G^i = \{g_1^i, ..., g_K^i\} \in \mathbb{R}^{K\times 2} \tag{1}$$

where g_k^i is a landmark of the i-th medical image in a set of k landmarks. The medical image $I^i \in \mathbb{R}^{W\times H\times 3}$, and W, H is the width and height of I^i.

Generally, for landmark detection, each landmark is converted into a confidence map. The landmark detection model takes processed pictures as input and is responsible for predicting and regressing the confidence map.

3.2. Student Model and Original Teacher Model

Considering the outstanding performance of U-net [25] on anatomical landmark detection, we use UNet4 as the student model and UNet5 as the original teacher model. The channels in UNet4 are [32, 64, 128, and 256], and the channels in UNet5 are [64, 128, 256, 512, and 1024].

3.3. Our Proposed FSF-LD Training Procedure

The whole FSF-LD training procedure is shown in the following:

Step 1. Pretrain teacher model: As Figure 2 shows, the teacher model is pretrained with the multi-task structure to make knowledge rich, general, and reliable.

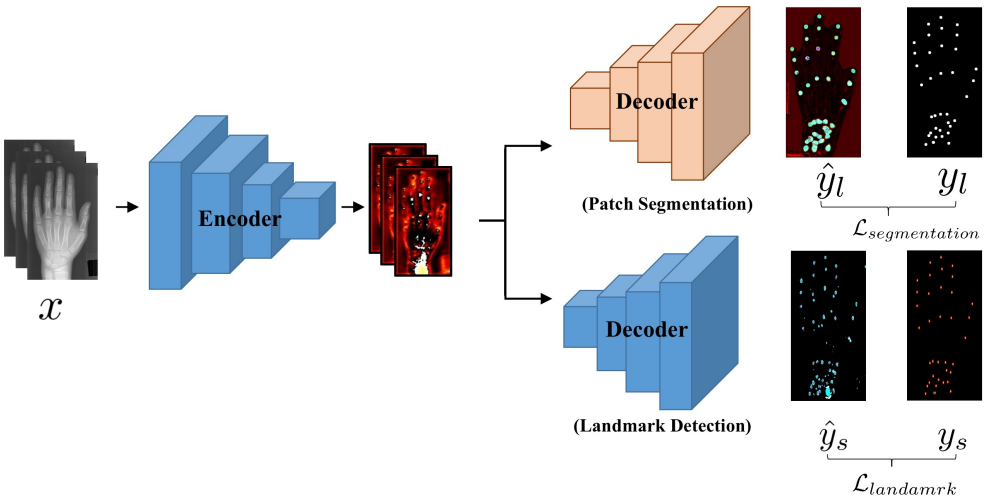

Figure 2. The multi-task learning framework for teacher model pretraining: There are two branches: the segment branch (top) and the landmark detection branch (bottom). The segment branch processes a medical image to predict a segmented mask and the landmark branch predicts a heatmap as shown in Figure 3.

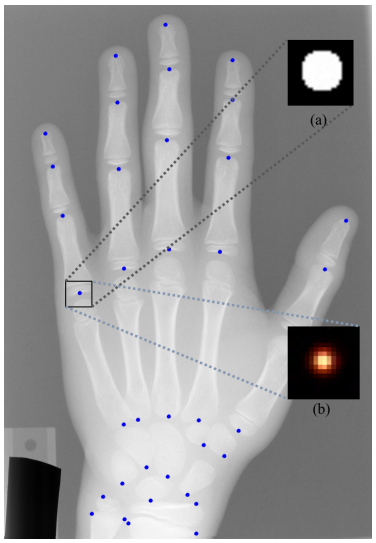

Figure 3. The segmentation mask and ground truth heatmap: (**a**) the segmentation mask, which is a local neighborhood patch P_k^i centered at landmark g_k^i with radius r = 1 pixels; (**b**) the groud truth heatmap, which is a Gaussian distribution centered at landmark g_k^i with the $\sigma = 1.5$.

Step 2. Knowledge distillation: As Figure 4 shows, we extract the intermediate spatial heatmaps from a teacher model as extra supervision for the student. Then we train a target student model UNet4 to locate landmarks and mimic the spatial features with the proposed loss function $\mathcal{L}_{atten}$ (10).

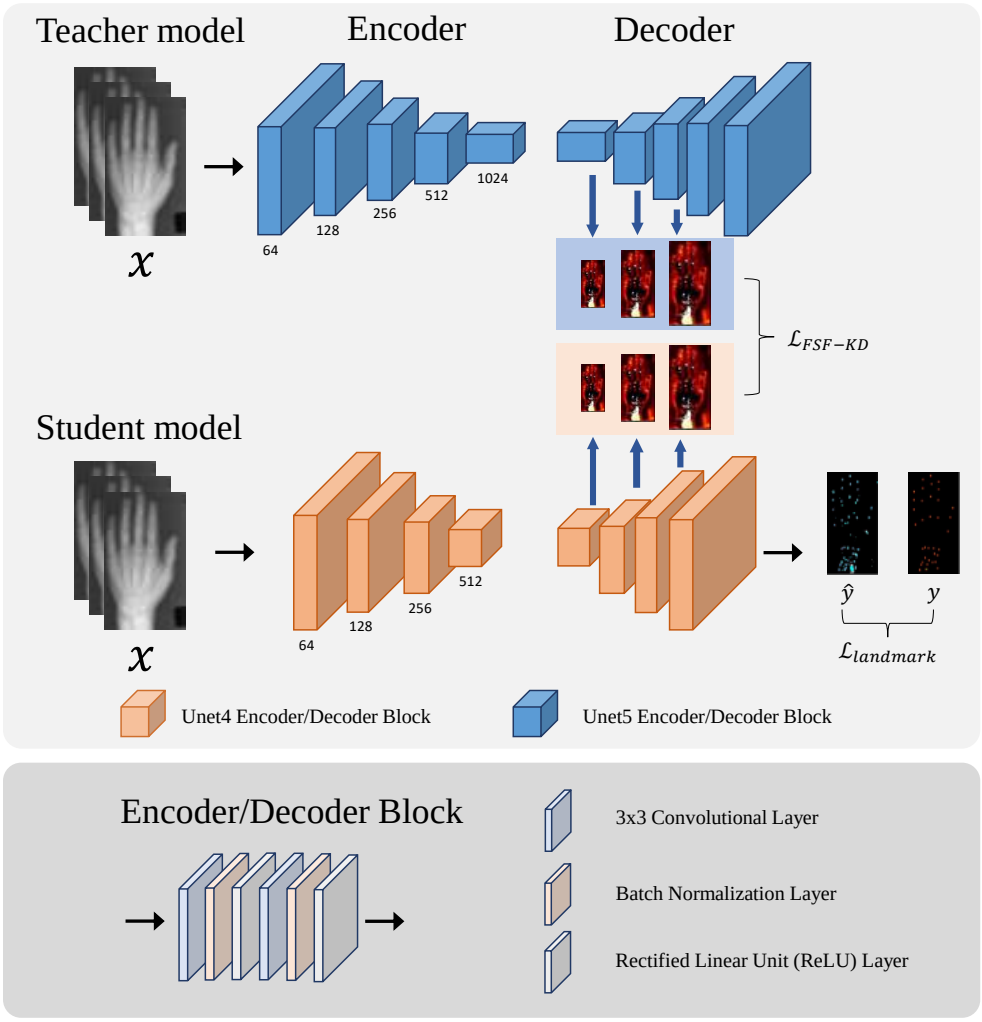

Figure 4. An overview of the feature-sharing fast landmark detection (FSF-LD) model-training strategy: Apart from the ground truth, the pretrained teacher model provides a student model with extra supervision guidance via $\mathcal{L}_{FSF-KD}$. The loss $\mathcal{L}_{FSF-KD}$ imposes the student model to imitate the teachers' representations from intermediate layers.

3.4. Pretrain Teacher Model

In KD, a reliable teacher model is the prerequisite to ensuring the performance of the student model. Consequently, building a well-performing teacher model is essential to provide richer and more general knowledge; otherwise, the student model will be confused and misguided to incorrect learning directions.

In this paper, we propose a novel and effective multi-task learning structure to promote teacher models in exploiting and representing knowledge. In this way, a better teacher model can be obtained, named SEG-UNet5. Figure 2 shows the framework of SEG-UNet5.

As Figure 2 shows, there are two branches: (1) segmentation of the landmark's neighborhood patch branch (Patch Segmentation); and (2) the landmark detection branch (Landmark Detection). Take a 2D hand radiograph dataset as an example to introduce our model. A 2D hand radiograph image has 37 anatomical landmarks and is first processed as 512×256 size. The encoder takes the processed hand image as input to extract high-level features for the following two branches. At the end of encoding, we employ a non-local

module to capture global structure features, which contain some vital topological structure information.

For the landmark detection branch, the landmarks G^i in the i-th image I^i are converted into a confidence map set, named GT^i, as shown in Figure 3.

$$GT^i = \{gt_1^i, ..., gt_K^i\} \tag{2}$$

And gt_k^i is a 2D Gaussian distribution centered at coordinates x_k, y_k from $G_k^i \in \mathbb{R}^{h \times w \times 1}$, defined as

$$gt_k^i = \frac{1}{2\sigma^2} exp(\frac{-[(\frac{x-x_k}{w})^2 + (\frac{y-y_k}{h})^2]}{2\sigma^2}), \quad k = 1, \ldots, K \tag{3}$$

where h and w are the width and height of the input, respectively. Here $h = 512$ and $w = 256$ in the 2D hand radiograph dataset. The hyperparameter σ determines the shape of the distribution. Here we empirically set $\sigma = 1.5$, and K is the total number of landmarks.

Then several convolutional layers are utilized to regress GT^i from the features learned by the encoder and output 37 channel feature maps $lm^i \in \mathbb{R}^{h \times w \times 37}$, where each channel represents a heatmap of a corresponding landmark. For landmark detection, the loss denoted as $\mathcal{L}_{lm}$ is formulated by the mean radial error (MRE), which is consistent with previous literature [26,27].

$$\mathcal{L}_{lm} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \| lm_k^i - gt_k^i \|_2^2 \tag{4}$$

For the segment branch, we mask a circular image patch P_k^i as white, which is centered at landmark g_k^i with a radius of r as the local neighborhood as Figure 3 shows. Then with the landmarks G^i in the i-th image I^i as the center, we can obtain a set of segmentation masks ST^i, which are formulated as

$$ST^i = \{st_1^i, ..., st_K^i\} \tag{5}$$

where $st_k^i \in \mathbb{R}^{h \times w \times 1}$ is defined as

$$st_k^i(a) = \begin{cases} 1 & a \in P_k^i \\ 0 & \text{else} \end{cases} \tag{6}$$

where a is a pixel on st_k^i.

Different from the landmarks branch, these convolution layers serve to regress the segmentation mask ST^i and output 37 channel feature maps sm^i with a size of 512×256, where each channel represents the local neighborhood patch P of the corresponding landmark.

For the segmentation task, we employ dice coefficient loss [28] to optimize the segmentation of mask ST^i. It is named $\mathcal{L}_{seg}$ and defined as

$$\mathcal{L}_{seg} = \frac{1}{N} \sum_{i=1}^{N} [1 - \frac{2 \sum_{i \in \Omega} sm_i \cdot st_i}{\sum_{i \in \Omega} sm_i^2 + \sum_{i \in \Omega} st_i^2}] \tag{7}$$

where Ω is the total pixels in the image, sm^i is 37 channel feature maps, and st^i is one of the set of segmentation masks ST^i.

In summary, in SEG-UNet5, the final objective function L is combined with $\mathcal{L}_{lm}$ and $\mathcal{L}_{seg}$ as follows:

$$\mathcal{L} = \mathcal{L}_{lm} + \lambda * \mathcal{L}_{seg} \tag{8}$$

where λ is a balance factor. Here,we set λ = 0.01. Based on MTL, our proposed method can implicitly improve the ability of teacher model feature extraction during the training process. It contributes to the later knowledge distillation.

3.5. Feature-Sharing Knowledge Distillation

It is vital for KD to represent and transfer knowledge effectively [21]. In fact, the state-of-the-art KD strategy (Fast-KD) [12] imposed student model aligns the teacher model on the output randomly. The knowledge is expressed in the form of the teachers' predictive output heatmap and transferred to the student model by minimizing the proposed mimicry loss function. However, the gap between their learning capabilities is ignored.

To better express and transfer knowledge, we propose two novel and effective knowledge distillation strategies for the landmark detection task. Inspired from [18], we propose a feature-sharing fast landmark detection (FSF-LD) structure, shown in Figure 4. It provides the student model with some spatial feature maps AM instead of the output from the teacher model. Thus, the student model enables our model to learn where the teacher mainly focuses. Then it is trained to capture more important feature information by imitating the spatial feature maps learned by the teacher model. Here, AM consists of some output from the middle layer, defined as

$$AM = \{hp_1, hp_2, hp_3\} \tag{9}$$

In UNet5 and SEG-UNet5, hp_1, hp_2, and hp_3 come from the output of up2, up3, and up4 blocks. In UNet4, hp_1, hp_2, and hp_3 come from the output of the up1, up2, and up3 blocks.

With the FSF-LD method, the objective function of the student model, $\mathcal{L}_{atten}$, is defined as:

$$\mathcal{L}_{atten} = \mathcal{L}_{lm} + \alpha * \mathcal{L}_{ak} \tag{10}$$

where $\mathcal{L}_{lm}$ is the same as the loss function of the SEG-UNet, and α is the knowledge transfer ratio. We set $\alpha = 0.5$, and $\mathcal{L}_{ak}$ is defined as

$$\mathcal{L}_{ak} = \frac{1}{3}\sum_{i=1}^{3}[\|F(thp_i) - F(shp_i)\|_2] \tag{11}$$

where thp_i is the hp_i from AM of UNet5 or SEG-UNet5, and shp_i is the hp_i from AM of UNet4. The function $A(\cdot)$ is defined as

$$F(A) = \frac{1}{C}\sum_{i=1}^{C}|A_i|^2 \tag{12}$$

where $A \in \mathbb{R}^{W\times H\times C}$ and $F(A)$, $A_i \in \mathbb{R}^{W\times H}$, C is the number of the channels. By minimizing the proposed object function $\mathbb{L}_{atten}$, the student can learn about the teacher-training process in detail and focus on features that are more important for landmark detection. Thus, the student can easily master learning skills and improve their performance.

4. Experiments

4.1. Dataset

To illustrate the effectiveness and generalization of our training strategy, we conduct comparative experiments on two publicly available datasets and a private hip dataset.

4.1.1. 2D Hand Radiograph Dataset

We use a public 2D hand radiograph dataset [29] to investigate the number of hyperparameters of the teacher model and the effectiveness of our KD training method. The dataset consists of 895 2D hand radiograph images with an average size of 1563 × 2169 pixels, acquired with different X-ray scanners. Because the images lack information about physical pixel resolution, we assume a wrist width of 50 mm determined by two of the annotated landmarks at the wrist, which is used in [7]. We perform a manual annotation of 37 landmarks on fingertips and bone joints. According to the ratio of 6:2:2, we split the data into

the training, test1, and test2 sets, which contain 537, 179, and 179, respectively. During preprocessing, all images are resized to 512 × 256 pixels.

4.1.2. 2D Cephalometric Radiograph Dataset

We also evaluate our proposed method on a public 2D cephalometric X-ray dataset [30]. The dataset consists of 400 2D cephalometric X-ray images with an average size of 1935 × 1935 pixels. Each X-ray image has 19 landmarks, which were the average of two experienced experts' annotations. According to the ratio of 6:2:2, we split the data into the training, test1, and test2 sets, which contain 240, 80, and 80, respectively. During preprocessing, all images are resized to 512 × 512 pixels.

4.1.3. 2D Hip Radiograph Dataset

To verify the generalization of our training strategy, we apply several supplementary experiments on a private hip dataset. The dataset consists of 210 radiograph images in total. The resolution of an image is 1935 × 2400 pixels. We perform a manual annotation of 10 landmarks. Considering the symmetrical structure of the hip joint, we divide a hip radiograph image into two parts. Thus, the dataset is expanded to 420. Then according to the ratio of 6:2:2, we split the data into training, test1, and test2 sets, which contain 252, 84, and 84, respectively. During preprocessing, all images are resized to 512 × 256 pixels.

4.2. Evaluation Metrics

4.2.1. MRE

The performance of landmark detection methods is evaluated with mean radial error (MRE) and successful detection rate (SDR) metrics [6]. For landmark detection, the loss denoted as $\mathcal{L}_{lm}$ is formulated by the mean square error (MSE), which is consistent with previous literature [26,27]. MRE and MSE are functionally equivalent, and here we express them uniformly in terms of MRE. The MRE is defined as

$$MRE = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{13}$$

where N denotes the number of detected landmarks, and R_i is the Euclidean distance between the predicted landmarks coordinates and the ground truth.

4.2.2. SDR

The SDR (success detection rate) [9] shows the percentage of landmarks successfully localized. For a landmark, if the radical error between it and the ground truth is no greater than r mm (r = 2.0 mm, 2.5 mm, 3.0 mm, 4.0 mm), it is considered a successful detection. The success detection rate for r mm is defined as below:

$$SDR_r = \frac{\mathcal{H}(\{\hat{y}_i : ||\hat{y}_i - y_i||_2 \leq r)\}}{\mathcal{H}(\Omega)} \tag{14}$$

where $\mathcal{H}$ is the cardinal function, and Ω is the set of predictions over all images.

4.2.3. GFLOPs

Giga Floating-point Operations Per Second (GFLOPs) [31] refers to floating-point operands, which can be used to measure the complexity of an algorithm/model. The smaller the GFLOPs, the faster the calculation.

4.3. The Effect of the Multi-Task Pretraining Structure

To validate the validity of our improving-teacher method, we compare the performance of UNet5 and SEG-UNet5 on the 2D hand radiograph of the test1 and test2 datasets. The result is shown in Table 1. In Table 1, MRE and SDR are adopted as the evaluation metric. It is obvious that the SEG-UNet5 outperforms UNet5 in all indicators in Table 1.

Table 1. 2D hand radiograph dataset: The comparison results for our proposed improving-teacher method on the 2D hand radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5. We have bolded the data with the best results.

Model	Test1					Test2					FLOPs(G)	Total Parameters
	SDR (%)				MRE (mm)	SDR (%)				MRE (mm)		
	r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm			
HRNet	44.0	56.2	65.3	76.6	3.0683	41.4	53.4	61.8	73.1	4.1372	**7.9211886**	9,318,595
UNet4	80.0	81.7	82.6	83.3	4.6741	79.5	81.8	82.7	83.5	4.9473	19.9375	**1,948,069**
UNet5	95.4	97.3	98.3	99.3	0.9982	94.4	97.2	**98.5**	**99.4**	0.9683	104.0	31,381,285
SEG-UNet5	**95.8**	**97.7**	**98.6**	**99.4**	**0.8628**	**95.1**	**97.5**	**98.5**	**99.4**	**0.9034**	177.35156	46,022,154

In the 2D hand test1 dataset, the SEG-UNet5 has 0.41%, 0.31%, 0.31%, and 0.1% improvements on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm, r = 4 mm). In the 2D hand test2 dataset, the SEG-UNet5 has 0.74% and 0.31% improvements on SDR (r = 2 mm, r = 2.5 mm), and the other metrics of SEG-UNet5 are close to UNet5. This proves that our proposed segment branch is conducive to obtaining a teacher model with better landmark detection performance.

For the teacher model in KD, we focus on the landmark's detection performance, but also on the feature extraction capacity. Therefore, we investigate several spatial feature maps learned by UNet5 and SEG-UNet5 on the 2D hand radiograph test1 dataset, as shown in Figure 5. We adopt different colors to represent the numerical value, and the darker the color represents a lower value, which means less semantic information in this area. As Figure 5 shows, the color of the feature maps from SEG-UNet5 is lighter, which means that the value is larger. In other words, the spatial feature map SEG-UNet5 learned contains more information that contributes to landmark detection compared with UNet5. These learned spatial maps will be directly or indirectly passed to the student model as knowledge representations.

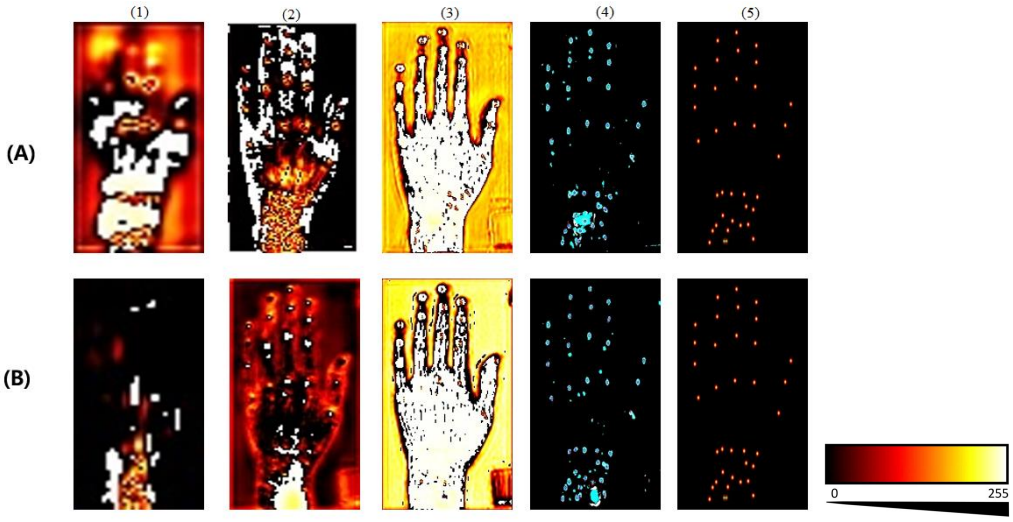

Figure 5. Feature map and output examples come from different models in the 2D hand radiograph dataset: row (**A**), SEG-UNet5; row (**B**), UNet5; columns (**1**–**5**) show the output of the up3 block, the up4 block, the final landmark detection heatmap, and ground truth, respectively. The RGB color value represents the amount of information contained in feature heatmaps.

Moreover, we further apply our method to the 2D cephalometric radiograph dataset and the 2D hip radiograph dataset. The results are shown in Tables 2 and 3. Similarly, we select the MRE and SDR as the evaluation metrics. In the hip test1 dataset, the SEG-UNet5 has 1.49%, 2.62%, and 2.04% improvements on SDR (r = 2.5 mm, r = 3 mm, r = 5 mm). The

other metrics of SEG-UNet5 are close to UNet5. In the hip test2 dataset, the SEG-UNet5 has 2.07%, 1.54%, and 1.38% improvements on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm). In the 2D cephalometric radiograph test1 and test2 dataset, UNet5 and SEG-UNet5 are comparable in performance.

In general, our proposed improving-teacher method is in favor of obtaining a better teacher model. In follow-up experiments, we will prove that with networks applying our improving-teacher method as teachers, the student model will achieve better results under the same knowledge distillation strategy.

Table 2. 2D hip radiograph dataset: The comparison results for our proposed improving-teacher method on the 2D hip radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5.

Model	Test1					Test2					GFLOPs	Total Parameters
	SDR (%)				MRE (mm)	SDR (%)				MRE (mm)		
	r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm			
HRNet	11.8	19.8	29.6	48.2	5.4197	15.4	25.8	35.4	55.2	4.8275	**7.9167938**	9,317,987
UNet4	64.1	74.7	82.4	89.6	2.1110	62.7	70.6	78.6	88.9	2.5568	19.8125	**1,948,069**
UNet5	**70.1**	80.5	86.3	92.8	1.8879	67.5	77.8	84.6	**92.8**	2.2486	103.78125	31,381,285
SEG-UNet5	69.9	**81.7**	**87.7**	**94.7**	**1.8555**	**95.1**	**97.5**	**98.5**	**99.4**	**2.2127**	176.8516	46,022,154

Table 3. 2D cephalometric radiograph dataset: The comparison results for our proposed improving-teacher method on the 2D cephalometric radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5.

Model	Test1					Test2					GFLOPs	Total Parameters
	SDR (%)				MRE (mm)	SDR (%)				MRE (mm)		
	r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm			
HRNet	7.4	13.8	20.6	39.6	70.1222	7.4	12.3	17.8	36.1	72.8118	**7.9187164**	9,318,253
UNet4	52.3	66.7	78.3	89.8	2.1110	61.6	76.1	85.4	93.9	2.8010	19.8125	**1,948,069**
UNet5	63.5	**76.5**	**84.6**	**93.9**	1.9557	72.7	84.8	90.4	**96.6**	1.6668	103.78125	31,381,285
SEG-UNet5	**63.7**	76.0	**84.6**	**93.9**	**1.9274**	**74.0**	**85.5**	**92.1**	**96.6**	**1.6663**	176.8516	46,022,154

4.4. Compared with Other KD Methods

To show the priority of our proposed FSF-LD method, we carry out some comparison experiments on the 2D hand radiograph images from the test1 and test2 datasets. We evaluate FSF-LD by comparing against the art-of-state KD method in landmark detection, named Fast-KD [12]. Moreover, we select Unet5 and SEG-Unet5 as the teacher model, respectively, and Unet4 as the student model.

As Table 4 shows, based on the same teacher model (UNet-5 or SEG-UNet5), our proposed FSF-LD method can achieve much better performance than Fast-KD. We also observe that with the same KD method, using SEG-UNet5 as the teacher model leads to a better student model than UNet5. This indicates that our proposed multi-task structure contributes to improving the effectiveness of knowledge distillation. Moreover, with SEG-UNet5 as the teacher model, the student model applying the FSF-LD method achieves the best performance, with 11.7%, 12.1%, 12.0%, and 11.4% improvements on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm, r = 4 mm), compared with Fast-KD [12].

Table 4. 2D hand radiograph dataset: The comparison results for our proposed KD method on the 2D hand radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5.

Teacher Model	Knowledge Distillation Method	Test1					Test2				
		SDR (%)				MRE (mm)	SDR (%)				MRE (mm)
		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm	
-	HRNet	44.0	56.2	65.3	76.6	3.0683	41.4	53.4	61.8	73.1	4.1372
-	UNet-4	80.1	81.7	82.6	83.3	4.6741	83.0	85.4	86.4	87.3	3.5670
UNet5	Fast-KD [12]	84.4	86.7	87.7	88.7	3.9785	84.7	86.5	87.4	88.4	3.9494
	FSF-LD(ours)	88.0	90.9	92.4	93.8	2.5480	88.3	90.8	92.1	93.2	2.9257
SEG-UNet5	Fast-KD [12]	93.3	96.1	97.3	98.2	1.5257	93.4	95.7	96.9	97.8	1.6445
	FSF-LD(ours)	**94.3**	**97.2**	**98.2**	**98.8**	**1.2912**	**94.1**	**96.2**	**97.3**	**98.1**	**1.3585**

To further determine how our proposed method works, we visualize feature heatmaps learned by some of the models mentioned in Table 4. As Figure 6 shows, without the teacher model's extra supervision, the UNet4 suffers from a limit on parameter capacity and a lack of feature information. It leads to poor performance of UNet4 on landmark detection, and (A), (C), and (D) in Figure 6 prove that the role of KD is to transfer the knowledge learned by the teacher model (e.g., feature information, even noise) to the student model. We also deploy the HRNet model which is the classical algorithm for landmark detection tasks on the same datasets, and the results show that it does not work well on the HRNet network because of the small amount of data, and our method can achieve good results on a small amount of data. Moreover, comparing (A), (B), and (C) in Figure 6, our FSF-LD tends to help the teacher model transfer profuse and more important spatial feature information to students. The differences between the true and predicted values of the five landmark detection methods on the hand radiograph images and hip radiograph images, respectively, are shown in Figure 7 and Figure 8. Correspondingly, some noise is also introduced. Fortunately, it brings little interference for landmark detection.

To verify our inference, we further apply our method to the 2D cephalometric radiograph dataset and the 2D hip radiograph dataset. The results are presented in Tables 5 and 6. UNet applied FSF-LD based on SEG-UNet5 outperforms all other models on both test datasets.

Table 5. 2D hip radiograph dataset: The comparison results for our proposed KD method on the 2D hip radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5.

Teacher Model	Knowledge Distillation Method	Test1					Test2				
		SDR (%)				MRE (mm)	SDR (%)				MRE (mm)
		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm	
-	HRNet	11.8	19.8	29.6	48.2	5.4197	15.4	25.8	35.4	55.2	4.8275
-	UNet-4	64.1	74.7	82.4	89.6	2.1110	62.7	70.6	78.6	88.9	2.5568
UNet5	Fast-KD [12]	63.4	74.0	80.2	89.4	2.4609	62.2	74.9	80.0	91.1	2.3546
	FSF-LD(ours)	61.2	72.0	81.0	89.4	2.3665	63.6	74.2	81.2	91.8	2.1146
SEG-UNet5	Fast-KD [12]	64.3	75.6	81.7	90.6	2.2257	64.7	76.0	82.1	90.6	2.2347
	FSF-LD(ours)	**66.3**	**77.1**	**83.9**	**91.8**	**1.9821**	**66.7**	**76.1**	**83.1**	**91.3**	**1.9710**

Table 6. 2D cephalometric radiograph dataset: The comparison results for our proposed KD method on the 2D cephalometric radiograph dataset. The student model is UNet4, while the teacher model is UNet5 and SEG-UNet5.

Teacher Model	Knowledge Distillation Method	Test1					Test2				
		SDR (%)				MRE (mm)	SDR (%)				MRE (mm)
		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm		r = 2 mm	r = 2.5 mm	r = 3 mm	r = 4 mm	
-	HRNet	7.4	13.8	20.6	39.6	70.1222	7.4	12.3	17.8	36.1	72.8118
-	UNet-4	52.3	66.7	78.3	89.8	3.5213	61.6	76.1	85.4	93.9	2.8010
UNet5	Fast-KD [12]	54.7	68.9	79.5	90.7	2.2899	62.0	77.6	86.3	93.6	2.5698
	FSF-LD(ours)	55.1	69.5	79.5	91.2	2.2267	63.2	78.1	87.4	94.4	2.1396
SEG-UNet5	Fast-KD [12]	55.3	72.7	81.1	91.5	2.2752	64.1	78.2	86.7	94.7	2.0041
	FSF-LD(ours)	**62.0**	**77.6**	**86.4**	**93.6**	**2.1899**	**64.6**	**78.7**	**87.4**	**95.2**	**1.9798**

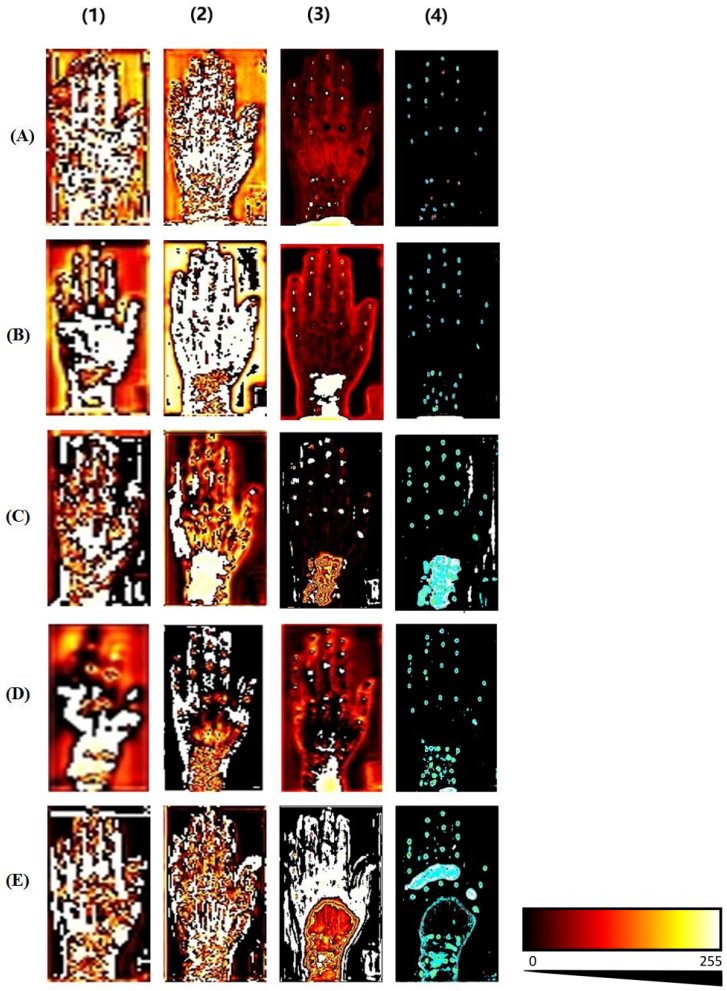

Figure 6. Feature map examples in 2D hand radiograph: rows (**A**–**E**) are respectively from UNet4, UNet4 with Fast-KD on UNet5, UNet4 with Fast-KD on SEG-UNet5, and UNet4 with FSF-LD on UNet5, UNet4 with FSF-LD on SEG-UNet5; columns (**1**–**4**) represent the output of the up3 block, the up4 block, and the final landmark detection heatmap, respectively. The RGB color value represents the amount of information contained in feature heatmaps. It intuitively shows the effect of our proposed improving-teacher method and FSF-LD. Pseudo color values: First, the feature map of the network output is normalized to between 0 and 1, and then mapped to 0–255, with each value representing a color.

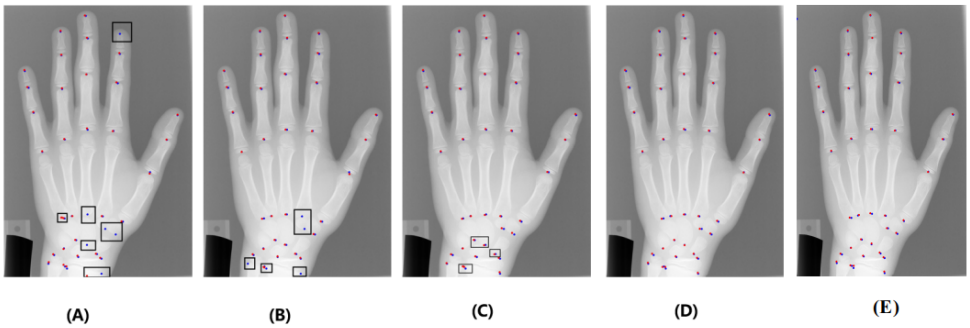

Figure 7. Landmark detection examples on hand radiograph images. The blue points represent the ground truth, and the red points represent prediction from different models: column (**A**), Unet4; column (**B**), Unet4 (Fast-KD on UNet5); column (**C**), Unet4 (FSF-LD on UNet5); column (**D**), Unet4 (Fast-KD on SEG-UNet5); column (**E**), Unet4 (FSF-LD on SEG-UNet5).

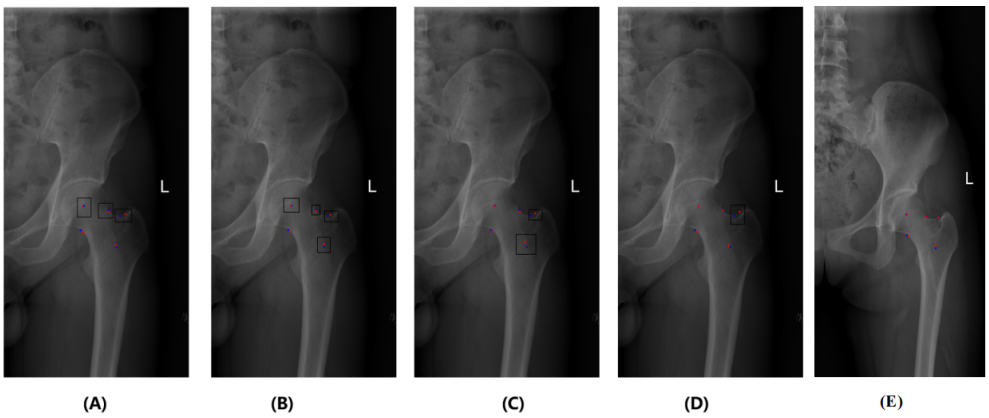

Figure 8. Landmark detection examples on hip radiograph images. The blue points represent the ground truth, and the red points represent predition from different models: column (**A**), Unet4; column (**B**), Unet4 (Fast-KD on UNet5); column (**C**), Unet4 (FSF-LD on UNet5); column (**D**), Unet4 (Fast-KD on SEG-UNet5); column (**E**), Unet4 (FSF-LD on SEG-UNet5).

5. Conclusions

In this paper, we propose a model-training method named Feature-Sharing Fast Landmark Detection (FSF-LD). In contrast to most existing anatomical landmark detection models, the FSF-LD aims to obtain a tiny and effective anatomical landmark detection model, which is easily deployed in clinical practice. First, we build a well-performing large teacher model by the proposed multi-task learning method. Thus, the teacher model enables the provision of richer and more general knowledge for the student model. Moreover, different from Fast-KD, the FSF-LD we proposed focuses on intermediate features and transfers knowledge in a more effective way. To verify our proposed methods, we carried out some experiments on a public 2D hand radiograph dataset and a private 2D hip radiograph dataset. On the 2D hand dataset, our FSF-LD had 11.7%, 12.1%, 12.0%, and 11.4% improvement on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm, r = 4 mm), compared with other KD methods. On the 2D hip dataset, our FSF-LD has which has 4.57%, 4.19%, 4.61%, 2.68% improvement on SDR (r = 2 mm, r = 2.5 mm, r = 3 mm, r = 4 mm) compared with other KD methods. We validated the model on three medical datasets, which gains better results. The results suggest the superiority of FSF-LD in terms of model performance and cost-effectiveness. It validates the model on several medical datasets, which gains better results. In the future, we will focus on how to improve the detection accuracy of anatomical landmarks and the robustness of models, such as modifying the loss function and implementing some data augmentation strategies. We hope that the next steps can

achieve better results in anatomical landmark detection of other human bones and apply our methods to practical clinical applications to save time and space resources.

Author Contributions: Software, D.H.; writing—original draft preparation, Y.W. (Yu Wang); writing—review and editing, Y.W. (Yuzhao Wang) and D.H.; project administration, T.B.; resources, G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (U21A20390), the Development Project of Jilin Province of China (YDZJ202101ZYTS128) and the Fundamental Research Funds for the Central University, JLU.

Data Availability Statement: The 2D hand radiograph dataset https://ipilab.usc.edu/research/baaweb/ (accessed on 27 July 2017). The 2D cephalometric radiograph dataset http://www-o.ntust.edu.tw/~cweiwang/ISBI2015/challenge1/ (accessed on 19 April 2015). The 2D hip radiograph dataset is not publicly available.

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

References

1. Beichel, R.; Bischof, H.; Leberl, F.; Sonka, M. Robust active appearance models and their application to medical image analysis. *IEEE Trans. Med Imaging* **2005**, *24*, 1151–1169. [CrossRef] [PubMed]
2. Heimann, T.; Meinzer, H.P. Statistical shape models for 3D medical image segmentation: A review. *Med. Image Anal.* **2009**, *13*, 543–563. [CrossRef] [PubMed]
3. Johnson, H.J.; Christensen, G.E. Consistent landmark and intensity-based image registration. *IEEE Trans. Med. Imaging* **2002**, *21*, 450–461. [CrossRef] [PubMed]
4. Štern, D.; Likar, B.; Pernuš, F.; Vrtovec, T. Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images. *Phys. Med. Biol.* **2011**, *56*, 7505. [CrossRef] [PubMed]
5. Kwon, H.J.; Koo, H.I.; Park, J.; Cho, N.I. Multistage Probabilistic Approach for the Localization of Cephalometric Landmarks. *IEEE Access* **2021**, *9*, 21306–21314. [CrossRef]
6. Liu, W.; Wang, Y.; Jiang, T.; Chi, Y.; Zhang, L.; Hua, X.S. Landmarks Detection with Anatomical Constraints for Total Hip Arthroplasty Preoperative Measurements. In Proceedings of the 2020–23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lima, Peru, 4–8 October 2020; pp. 670–679.
7. 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] [PubMed]
8. Qian, J.; Luo, W.; Cheng, M.; Tao, Y.; Lin, J.; Lin, H. CephaNN: A Multi-Head Attention Network for Cephalometric Landmark Detection. *IEEE Access* **2020**, *8*, 112633–112641. [CrossRef]
9. 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]
10. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and Global Knowledge Distillation for Detectors. *CoRR* **2021**, *abs/2111.11837*. Available online: http://xxx.lanl.gov/abs/2111.11837 (accessed on 26 November 2021).
11. Li, Z.; Ye, J.; Song, M.; Huang, Y.; Pan, Z. Online Knowledge Distillation for Efficient Pose Estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 11720–11730. [CrossRef]
12. Zhang, F.; Zhu, X.; Ye, M. Fast human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3517–3526.
13. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
14. Kwon, H.; Kim, Y. BlindNet backdoor: Attack on deep neural network using blind watermark. *Multimed. Tools Appl.* **2022**, *81*, 6217–6234. [CrossRef]
15. Kwon, H. Medicalguard: U-net model robust against adversarially perturbed images. *Secur. Commun. Netw.* **2021**, *2021*, 5595026:1–5595026:8. [CrossRef]
16. Kwon, H.; Lee, J. AdvGuard: Fortifying Deep Neural Networks against Optimized Adversarial Example Attack. *IEEE Access* **2020**, *4*, 2016. [CrossRef]
17. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
18. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
19. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.

20. Ahn, S.; Hu, S.X.; Damianou, A.; Lawrence, N.D.; Dai, Z. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9163–9171.
21. Wang, L.; Yoon, K.J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3048–3068. [CrossRef] [PubMed]
22. Hwang, D.H.; Kim, S.; Monet, N.; Koike, H.; Bae, S. Lightweight 3D human pose estimation network training using teacher-student learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2020; pp. 479–488.
23. Zhang, Y.; Yang, Q. A Survey on Multi-Task Learning. *arXiv* **2021**. [CrossRef]
24. Liu, C.; Xie, H.; Zhang, S.; Mao, Z.; Sun, J.; Zhang, Y. Misshapen Pelvis Landmark Detection With Local-Global Feature Learning for Diagnosing Developmental Dysplasia of the Hip. *IEEE Trans. Med. Imaging* **2020**, *39*, 3944–3954. [CrossRef] [PubMed]
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
27. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 483–499.
28. Liu, Y.C.; Tan, D.S.; Chen, J.C.; Cheng, W.H.; Hua, K.L. Segmenting hepatic lesions using residual attention U-Net with an adaptive weighted dice loss. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3322–3326.
29. Viterbi School of Engineering Digital Hand Atlas. 2017. Available online: https://ipilab.usc.edu/research/baaweb/ (accessed on 27 July 2017).
30. Wang, C.W.; Huang, C.T.; Hsieh, M.C.; Li, C.H.; Chang, S.W.; Li, W.C.; Vandaele, R.; Maree, 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]
31. Zhu, L. THOP: PyTorch-OpCounter. Available online: https://github.com/Lyken17/pytorch-OpCounter (accessed on 2 April 2018).