Landmark Localization from Medical Images with Generative Distribution Prior

Zixun Huang, Rui Zhao, Frank H.F. Leung, Sunetra Banerjee, Kin-Man Lam, Yong-Ping Zheng, Sai Ho Ling

***Abstract*— In medical image analysis, anatomical landmarks usually contain strong prior knowledge of their structural information. In this paper, we propose to promote medical landmark localization by modeling the underlying landmark distribution via normalizing flows. Specifically, we introduce the flow-based landmark distribution prior as a learnable objective function into a regression-based landmark localization framework. Moreover, we employ an integral operation to make the mapping from heatmaps to coordinates differentiable to further enhance heatmap-based localization with the learned distribution prior. Our proposed Normalizing Flow-based Distribution Prior (NFDP) employs a straightforward backbone and non-problem-tailored architecture (i.e., ResNet18), which delivers high-fidelity outputs across three X-ray-based landmark localization datasets. Remarkably, the proposed NFDP can do the job with minimal additional computational burden as the normalizing flows module is detached from the framework on inferencing. As compared to existing techniques, our proposed NFDP provides a superior balance between prediction accuracy and inference speed, making it a highly efficient and effective approach. The source code of this paper is available at `https://github.com/jacksonhzx95/NFDP`.**

***Index Terms*— landmark localization, normalizing flows, density estimation, regression, heatmap-based localization**

I. Introduction

LANDMARK localization aims to extract salient points from a given image, which is an essential task in medical image analysis [1]–[3]. In clinical applications, manually localizing the anatomical landmarks is tedious and time-consuming, which may contain large inter-observer and intra-observer variations. An automatic landmark localization algorithm is of great interest in artificial intelligence-assisted healthcare systems.

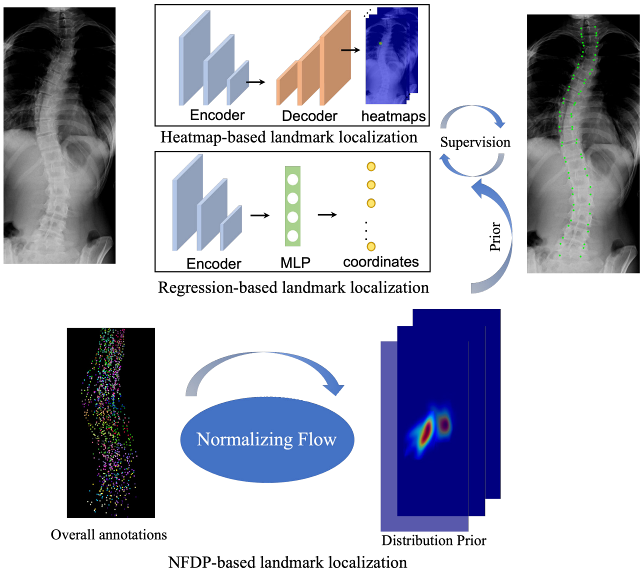


Fig. 1: An illustration of the heatmap-based approach, regression-based approach, and our proposed NFDP.

Owing to the rapid development of deep learning technologies, landmark localization has been extensively studied over the past decade, and has shown great performance in different scenarios. Currently, landmark localization methods can be roughly categorized into two main groups, i.e., heatmap-based approaches [3]–[5] and regression-based approaches [6]–[8]. Heatmap-based approaches dominate the field because of their powerful performance. Specifically, heatmap-based frameworks predict a likelihood heatmap to locate each landmark point. The landmark point is identified as a local or global maxima in the likelihood heatmap. Most state-of-the-art landmark detectors follow this method to infer the landmark coordinates. One of the main disadvantages of these detectors is that the ground-truth heatmaps need to be manually constructed. Although the heatmap can be easily generated according to the landmark coordinates, there are many hyperparameters that need to be predefined, such as the Gaussian kernel size, Gaussian standard deviation as well as heatmap resolution. The Gaussian Kernel has a high impact on the performance of heatmap-based localization algorithms. As a result, it may be necessary to select a new kernel for optimal performance when performing heatmap-based localization in a new dataset. Assuming that the ground-truth landmarks follow a distribution, the Dirac delta distribution [9] and Gaussian distribution [10] are the common choices to construct the

This study was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R5017-18 & No. B-Q86J), and the Research Institute for Smart Ageing, The Hong Kong Polytechnic University.

Zixun Huang, Frank H.F. Leung, and Kin-Man Lam work under the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. (zixun.huang@connect.polyu.hk, {frank-h-f.leung, enkmlam}@polyu.edu.hk)

Rui Zhao works under Hisilicon Semiconductor Huawei Technologies, Wuhan, China. (rick10.zhao@connect.polyu.hk)

Sunetra Banerjee, Sai Ho Ling work under the School of Electrical and Data Engineering, University of Technology Sydney, Australia. (sunetra.banerjee@student.uts.edu.au, steve.ling@uts.edu.au)

Yong-Ping Zheng works under the Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong. (yong-ping.zheng@polyu.edu.hk)

heatmaps. We expect the chosen landmark distribution to be an inaccurate estimation of the real distribution. This inaccurate estimation behind the manually constructed heatmaps will introduce bias when learning from the training samples.

Regression-based methods, on the other hand, directly predict the landmark coordinates from the given signal, which is a more natural and straightforward approach to estimating salient points. Regression-based frameworks output the point coordinates instead of constructing the likelihood heatmaps, thus skipping the step of manually designing the heatmaps in learning. However, regression-based methods require penalizing the distance between the estimated coordinates and the ground-truth coordinates. Currently, the commonly used criteria are the standard L_1 and L_2 losses, which assume that the landmarks follow a Laplacian and a Gaussian distribution, respectively. Although these two penalties generally perform well in existing regression frameworks [7], [8], [11], the assumption of the adopted distribution can still be inaccurate for modeling complex landmark distributions.

As we can see, both heatmap-based and regression-based methods need to implicitly explore the underlying distribution of landmarks. However, current studies only employ simple assumptions to characterize the landmarks. This phenomenon usually is due to the fact that the landmarks behind natural images are too complex to be learned efficiently. Nevertheless, different from natural images, medical images usually focus on a specific region, e.g., brain, chest, eye, etc. Thus, the structural information is relatively uniform among samples, leading to strong prior knowledge of their landmark distribution. An example of spine landmarks is shown in Fig. 2. In this paper, we argue that a landmark localization framework for medical images can greatly benefit from the strong prior knowledge to improve localization accuracy. An overview of our proposed method is shown in Fig. 1. In regression-based methods, we can directly employ the learned landmark distribution as a penalty function to supervise the regression learning. In heatmap-based methods, we further introduce a differentiable integral operation to project the heatmaps into coordinates. Thus, the learned regression penalty can also be used to enhance the heatmap-based frameworks.

To leverage the structural knowledge behind landmarks in medical images for better localization, we have to estimate the density of the underlying real distribution of landmarks. Generative models, such as variational autoencoders [12], generative adversarial networks [13], and normalizing flows [14], have been widely studied in the field of density estimation. They have achieved promising performance in real-world applications. Among these generative models, normalizing flows show great potential in our task owing to their flexible structures and more accurate estimation results. Normalizing flows characterize an unknown distribution by transforming it into a traceable distribution through multiple invertible mappings [14]. By observing the ground-truth landmarks from the training data, normalizing flows can construct a density function to describe the landmark structures, which indicates a desirable penalty function to supervise the regression learning. To integrate the generative model with a standard landmark detector, we make the landmark detector predict the deviation of the distribution, i.e., the scale and the shift, conditioned on the specific input image. Therefore, the whole framework can be trained in an end-to-end manner. Moreover, as the generative model represents a penalty function in learning, it can be detached from the framework when deployed in applications.

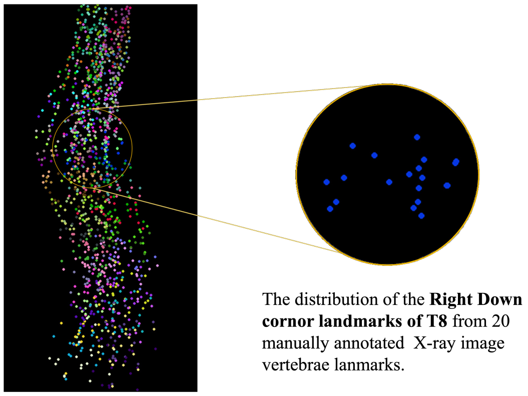

The distribution of the **Right Down cornor landmarks of T8** from 20 manually annotated X-ray image vertebrae lanmarks.

Fig. 2: Visualization of the distribution of the vertebra landmarks from 20 expert annotated X-ray vertebra images. Circles of the same color represent the same corner landmark of various vertebrae across different individuals in the spinal X-ray dataset.

At the early stage of the training, the predictions are chaotic under a randomly initialized penalty function, making the learning unstable and hard to converge. To address this issue, we further introduce the concept of residual learning into our framework and construct the desirable landmark distribution by the difference between the real distribution and a traceable distribution, such as the Gaussian distribution. Consequently, the whole framework can be trained under a standard L_2 criterion at the beginning, which greatly stabilizes the learning process. The main contributions of this paper can be summarized as follows:

- We propose a novel framework to solve the landmark localization problems from a regression perspective without a problem-tailored design. Specifically, we introduce the structural distribution knowledge as a learnable penalty function into the proposed framework to improve landmark localization with minimal additional computational burden to the baseline framework.
- We employ normalizing flows with residual learning to characterize the underlying distribution of the medical landmarks, and integrate this density estimator with the regression network for parallel optimization. Moreover, we adopt a differentiable integral operation for heatmap-based frameworks, so as to introduce the learned distribution knowledge into heatmap-based methods that further enhance their performance.
- We conduct extensive experiments to demonstrate that the proposed frameworks with only some simple backbones can achieve good performance in landmark localization tasks for medical images.

II. Related Works

A. Landmark Localization from Medical Images

Over the past few decades, numerous frameworks for medical landmark localization have been investigated. Few of them considered the prior structural knowledge behind the medical landmarks. Since the landmarks in medical images tend to contain rich structural information, handcrafted graphical models are used in the literature to capture the landmark structure. Lindner et al. [15] employed principal component analysis to coarsely reduce the search space based on the landmarks distribution, and utilized random forests to accurately regress the landmark coordinates. Urschler et al. [16] integrated both the spatial information and the landmark configuration into a random forest to perform robust landmark localization.

Recently, owing to the superiority of deeply learned features, deep learning frameworks have become increasingly popular, achieving good performance in medical landmark localization tasks. Deep-learning-based landmark localization frameworks can be roughly divided into regression-based methods and heatmap-based methods. Regression-based methods directly predict the landmark coordinates. However, as medical images are generally noisy and of low quality, directly predicting the coordinates makes them prone to predict outliers in some challenging scenarios. To address this issue, many algorithms have been proposed to improve the robustness of landmark identification. Sun et al. [17] proposed a multi-task framework to jointly learn the vertebra landmark localization and the Cobb angle assessment by exploiting the dependency between the two tasks. Wu et al. [11] proposed BoostNet that eliminated the deleterious outliers in the feature space to achieve robust vertebra landmark localization. Zeng et al. [18] proposed a three-stage cascaded network to perform coarse-to-fine cephalometric landmark localization.

Instead of directly predicting the coordinates, heatmap-based methods perform landmark localization via pseudo-likelihood heatmaps. Specifically, they utilize a majority voting strategy, i.e., argmax operation, to produce the landmark coordinates from heatmaps. Following this protocol, Yang et al. [19] proposed to predict vertebra landmark locations through a heatmap-based framework and introduced a Markov Random Field model to refine the landmarks with missing responses based on their neighboring landmarks. Payer et al. [20] utilized an extra CNN to model the spatial configuration of the landmarks and incorporated it with the ambiguous local landmarks to improve the robustness. Instead of generating an individual heatmap for every landmark, Yi et al. [21] separated the spine landmark detection task into three sub-tasks, i.e., center localization, center offset regression, and corner offset regression. Then, a CNN with three output heads was utilized to accomplish the three sub-tasks simultaneously. Based on [21], Guo et al. [22] introduced a vertebra segmentation task and incorporated a key point transformer to capture the relationships between local vertebrae and the global spine structure.

The aforementioned methods improve the performance of the localization model at the cost of increased computational complexity during both the training and testing phases owing to the multi-stage framework, extra modules, and/or post-processing. In this paper, we propose to integrate the landmark distribution prior for the landmark localization model into the objective function; our proposed NFDP solely impacts the training phase and the normalizing flow model will be discarded during inferencing, introducing minimal additional computational burden. Therefore, our proposed method requires relatively low computational resources compared to existing methods. The increase in computational demand, when transitioning from a baseline method to our proposed method, is negligible.

B. Integral Heatmap-based Landmark Localization

Different from traditional heatmap-based methods that utilize the argmax function to obtain landmark coordinates from the predicted heatmaps, integral heatmap-based algorithms adopt an integral operation to make the mapping from the heatmaps to coordinates differentiable. As stated in [23], direct integral heatmap regression can handle latent distributions but will lead to lower accuracies. Hence, the previous works mainly focused on augmenting supervision to enhance integral learning. For instance, Sun et al. [24] trained the IntegralNet with both heatmap labels and coordinates, while Nibali et al. [25] introduced a regularization strategy in the heatmap feature layer. Iqbal et al. [26] proposed a Gaussian prior loss to expedite training and enhance localization performance. Gu et al. [23] suggested a two-stage framework for performing coarse-to-fine localization. Most of the previous algorithms introduced external supervision such as Gaussian distribution loss on the heatmap layer or relied on complex network structures to enhance the landmark localization learning, which means that they need to specifically adjust their learning strategy or learning parameters, such as feature scale, channel number, and feature interaction, to fit different network structures. Moreover, in single iterative supervised training, integral heatmap-based approaches focus solely on a specific input image and its corresponding landmarks, which means that the representations learned by integral heatmap-based models are optimized primarily for predicting the target variable rather than capturing the overall structure or distribution of the data. In contrast, our proposed method utilizes normalizing flows to learn the underlying distribution of the landmarks from the manual annotation across all the training samples. We introduce the structural distribution knowledge into a network to improve landmark localization learning with minimal additional computational burden.

C. Distribution Learning for Regression

Previous studies have demonstrated that learning the output distribution can greatly promote the regression performance of regression-based landmark localization. Specifically, Gao et al. [27] utilized label ambiguity for label distribution learning, which achieved a significant improvement in classification and prediction tasks. From the distributional perspective, Bellemare et al. [28] designed a new algorithm for reinforcement learning, and showed the potential of distribution learning to prevent overfitting and improve representations. Studies focusing on knowledge distillation [29], [30] reveal that a smaller student network can achieve competitive performance by learning from a reliable teacher network. Imani et al. [31] proposed an output distribution learning algorithm to improve model generalization using a histogram loss function. The aforementioned algorithms show the superiority of distribution learning for enhanced regression learning, which motivates us to explore distribution learning in medical image landmark localization.

D. Normalizing Flows for Density Estimation

Dinh et al. [32] proposed real-valued nonvolume preserving (Real-NVP) for density estimation. The density of a landmark (i.e., the likelihood) can be computed by sampling the landmark distribution and applying invertible transformations to transform the landmark back to some original pre-defined distribution. A normalizing flow model consists of a series of bijective transformation matrices that convert a simple pre-defined distribution into a complex target distribution [14], [33], [34]. Recently, normalizing flows were utilized for constructing the kinematic prior of 3D human pose estimation [35]–[38]. Zanfir et al. [35] proposed to utilize a normalizing flow-based human pose prior for weakly supervised 3D human pose and shape estimation. Xu et al. [36] introduced a kinematic prior based on normalizing flows for the 3D human pose and shape reconstruction. Wehrbein et al. [37] utilized normalizing flows to directly model the posterior distribution of 3D poses conditioned on the input image. Specifically, Li et al. [38] proposed a learning paradigm that introduces the normalizing flow-based distribution prior for enhancing regression-based human pose estimation. Normalizing flows are also attracting increasing attention in facial data applications. Aksan et al. [39] proposed LiP-Flow that utilized a prior model with normalizing flows to reduce the training-inference asymmetry for 3D facial reconstruction. Liang et al. [40] proposed TalkFlow to model the difference variance of the data distribution for talking facial landmark generation. In human pose estimation [35]–[38], kinematic prior knowledge, which aids pose estimation, is often sourced from external motion capture repositories, i.e., CMU [1]. This reliance complicates the direct transfer of their methods to tasks that lack datasets with strong prior knowledge, as is the case with many medical datasets. LiP-Flow [39] learned the prior between the training data and the inference data, which is not flexible in clinical applications. TalkFlow [40] introduced additional scale to network layers which will significantly increase the computational complexity. Notably, LiP-Flow [39] is tailored for 3D avatar reconstruction, and TalkFlow [40] is tailored for generating talking faces synced with temporal speech data. The working pipelines of these algorithms are not suitable for landmark localization. In this paper, we propose to learn the landmark distribution prior directly from the manual annotations of the training dataset, eliminating the dependence on external datasets. This self-sufficiency renders our Normalizing Flow-based Distribution Prior (NFDP) a plug-and-play tool that can directly embed into different network structures and different datasets.

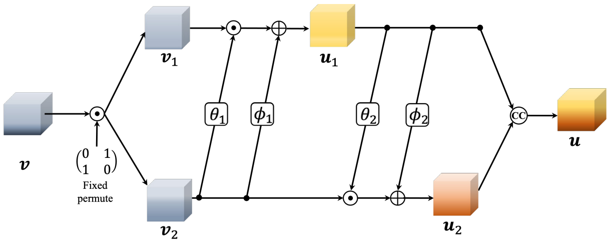


Fig. 3: An illustration of the employed double-side affine coupling (DAC) module.

III. Methodology

In this section, we present the details of the proposed framework for medical landmark localization. We first introduce the detailed design for modeling the underlying landmark distribution with normalizing flows. Then, we explicitly discuss the strategies to utilize the learned distribution prior for facilitation of both the regression-based and the heatmap-based landmark localization.

A. Estimating Landmark Distribution with Normalizing Flows

We propose to characterize the landmark distribution with normalizing flows. Normalizing flows can transform a simple distribution (such as a Gaussian distribution $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$) into a complex distribution through multiple invertible matrices [14]. In this paper, we consider the anatomical landmarks $\boldsymbol{l}$ sampled from an unknown distribution. To characterize this distribution, we estimate its density function by maximizing the log-likelihood $\log p_\delta(\boldsymbol{l})$ over the whole training set $\boldsymbol{V}$, with respect to a flow model f_δ, whose parameter δ can be optimized as follows:

$$\max_{\delta} \sum_{\boldsymbol{l} \in \boldsymbol{V}} \log p_\delta(\boldsymbol{l}). \tag{1}$$

Moreover, we further consider $\boldsymbol{l} = f_\delta(\boldsymbol{z})$, where $f_\delta(\cdot)$ denotes a component-wise invertible mapping, aiming to construct the target distribution from a traceable distribution.

With this approach, we can rewrite the log-likelihood under a change of variables as follows:

$$\begin{aligned} \log p_\delta(\boldsymbol{l}) &= \log p_\delta(f_\delta(\boldsymbol{z})) \\ &= \log p_\gamma(\boldsymbol{z}) + \log \left|\det J(f_\delta^{-1}(\boldsymbol{l}))\right|, \end{aligned} \tag{2}$$

where $f_\delta^{-1}(\cdot)$ denotes the inverse of $f_\delta(\cdot)$ and $\boldsymbol{z} = f_\delta^{-1}(\boldsymbol{l})$. $J(f_\delta^{-1}(\boldsymbol{l}))$ denotes the Jacobian matrix of $f_\delta^{-1}(\boldsymbol{l})$. $p_\gamma(\boldsymbol{z})$ describes a pre-defined traceable distribution. For example, if $\boldsymbol{z}$ follows a standard Gaussian distribution, its density function $p_\gamma(\boldsymbol{z})$ can be expressed as $\frac{1}{2\pi} e^{-\frac{1}{2}\|\boldsymbol{z}\|^2}$. Given an arbitrary $\boldsymbol{l}$, the corresponding distribution can be estimated through Eq. (2) by feeding $\boldsymbol{z}$ into the learned normalizing flows. Theoretically, $p_\delta(\boldsymbol{l})$ is able to fit any distribution, as long as $f_\delta(\cdot)$ is complex enough.

[1] CMU graphics lab motion capture database. 2009. http: //mocap.cs.cmu.edu/

As we discussed in Eqs. (1) and (2), the target density function $p_\delta(\boldsymbol{l})$ can be learned by maximizing the log-likelihood over the whole training set, which is equivalent to minimizing the objective function as follows:

$$\begin{aligned}\mathcal{L}_{ml} &= -\log p_\delta(\boldsymbol{l})|_{\boldsymbol{l}=\boldsymbol{\mu_g}} \\ &= -\log p_\gamma(f_\delta^{-1}(\boldsymbol{\mu}_g)) - \log\left|\det J(f_\delta^{-1}(\boldsymbol{\mu}_g))\right|,\end{aligned} \tag{3}$$

where $\mathcal{L}_{ml}$ denotes the maximum likelihood penalty. $\boldsymbol{\mu_g}$ represents the expert annotated landmark locations. It is worth noting that the optimal δ relies on both $\log\left|\det J(f_\delta^{-1}(\boldsymbol{\mu}_g))\right|$ and$f_\delta^{-1}(\boldsymbol{\mu}_g)$ in Eq. (3). Therefore, the resultant flow model $f_\delta(\cdot)$ covers the distribution of $\boldsymbol{\mu}_g$ over all the training samples.

On the other hand, we also need to construct effective invertible mapping for density estimation with normalizing flows. In practice, a complex mapping function can be decomposed into several simple mappings successively, i.e., $\boldsymbol{l} = f_\delta(\boldsymbol{z}) = f_{\delta L}(f_{\delta(L-1)}(\dots(f_{\delta 1}(\boldsymbol{z})))\dots)$, where L is the number of invertible transformations. In this paper, the employed invertible network consists of three stacked double-side affine coupling modules (DACs). The affine coupling module is the key component for invertible transformation, which contains one pre-defined permutation matrix and an invertible affine transform block, as shown in Fig. 3. In this figure, given an input vector, $\boldsymbol{v} \in \mathbb{R}^{K\times 2}$, where K denotes the number of landmarks, we have two corresponding coordinates in the vertical and horizontal directions. DAC first permutes the input vector $\boldsymbol{v}$ to chain transformations for efficient learning [35]. After the vector is divided into two parts, i.e. $\boldsymbol{v}_1 \in \mathbb{R}^{K\times 1}$, $\boldsymbol{v}_2 \in \mathbb{R}^{K\times 1}$, the learned scale and shift vectors are applied to both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ through affine transformations, as follows:

$$\begin{aligned}\boldsymbol{u}_1 &= \boldsymbol{v}_1 \odot e^{(\theta_1(\boldsymbol{v}_2))} + \phi_1(\boldsymbol{v}_2), \\ \boldsymbol{u}_2 &= \boldsymbol{v}_2 \odot e^{(\theta_2(\boldsymbol{u}_1))} + \phi_2(\boldsymbol{u}_1),\end{aligned} \tag{4}$$

where $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are the transformed vectors, $\odot$ denotes the point-wise product; $\theta(\cdot)$ and $\phi(\cdot)$ are learnable scale and shift functions, respectively, which can be represented by arbitrary neural networks. Specifically, we employ a three-layer fully connected sub-network with N_n neurons to construct both $\theta(\cdot)$ and $\phi(\cdot)$, in which each layer is followed by a Leaky-ReLU activation. Finally, the output vector $\boldsymbol{u}$ is obtained by concatenating $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in the direction dimension. The aforementioned DAC modules introduce sufficient nonlinearity into the invertible mapping, which facilitates our flow model to more effectively fit the complex target distribution.

B. Distribution Prior for Regression-based Landmark Localization

To introduce the learned landmark distribution into the landmark regression framework, we assume that all the underlying distributions share the same density function family, but with different mean and variance, conditioned on the input medical image $\mathcal{B}$. The working pipeline is shown in Fig. 4. First, the flow model $f_\delta(\cdot)$ estimates the density of a zero-mean deformed distribution $p_\delta(\bar{\boldsymbol{l}})$ from a zero-mean initial distribution $\overline{\boldsymbol{z}} \sim \mathcal{N}(0, \boldsymbol{I})$. Then, the regression model ϵ predicts both $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ to control the position and scale of the distribution of the anatomical landmarks, respectively. The final density $p_{\epsilon,\delta}(\boldsymbol{l}|\mathcal{B})$ is obtained by shifting and rescaling $\bar{\boldsymbol{l}}$ to $\boldsymbol{l}$, where $\boldsymbol{l} = \bar{\boldsymbol{l}}\cdot\hat{\boldsymbol{\sigma}} + \hat{\boldsymbol{\mu}}$. Therefore, the learned objective function with the landmark distribution prior can be formulated as:

$$\begin{aligned}\mathcal{L}_{ml} &= -\log p_{\epsilon,\delta}(\boldsymbol{l}|\mathcal{B})|_{l=\boldsymbol{\mu}_g} \\ &= -\log p_\delta(\overline{\boldsymbol{\mu}}_g) - \log\left|\det J(\overline{\boldsymbol{\mu}}_g)\right| \\ &= -\log p_\delta(\overline{\boldsymbol{\mu}}_g) + \log\hat{\boldsymbol{\sigma}},\end{aligned} \tag{5}$$

where $\overline{\boldsymbol{\mu}}_g = (\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})/\hat{\boldsymbol{\sigma}}$, $J(\overline{\boldsymbol{\mu}}_g)$ is the Jacobian matrix of $\overline{\boldsymbol{\mu}}_g$, and $J(\overline{\boldsymbol{\mu}}_g) = 1/\hat{\boldsymbol{\sigma}}$. In summary, Eq. (5) indicates that the flow model focuses on learning the distribution of $\boldsymbol{\mu}_g$, while the regression model focuses on learning the deviation to the expert annotated mean from the input medical image.

However, at the early stage of the learning, the learned landmark distribution is far from the real distribution, which makes the training difficult to converge. Inspired by the hypothesis in ResNet [41] that optimizing residual mappings is easier than the original unreferenced learning, a shortcut is developed in the normalizing flow model to stabilize the training process. Thus, the objective function for the whole framework is reformulated as follows:

$$\begin{aligned}\mathcal{L}_{ml} &= -\log p_\delta(\overline{\boldsymbol{\mu}}_g) + \log\hat{\boldsymbol{\sigma}} \\ &= -\log(\overline{\boldsymbol{z}} \cdot \frac{p_\delta(\overline{\boldsymbol{\mu}}_g)}{\overline{\boldsymbol{z}}}) + \log\hat{\boldsymbol{\sigma}} \\ &= -\log\overline{\boldsymbol{z}} - \log\frac{p_\delta(\overline{\boldsymbol{\mu}}_g)}{\overline{\boldsymbol{z}}} + \log\hat{\boldsymbol{\sigma}},\end{aligned} \tag{6}$$

where $\overline{\boldsymbol{z}} \sim \mathcal{N}(0, \boldsymbol{I})$ represents the original zero-mean distribution, and $\frac{p_\delta(\overline{\boldsymbol{\mu}}_g)}{\overline{\boldsymbol{z}}}$ denotes the residual likelihood. Consequently, the flow model aims to perform the residual maximum likelihood estimation, instead of the entire distribution estimation.

C. Distribution Prior for Heatmap-based Landmark Localization

Different from regression-based methods, heatmap-based methods generally utilize the argmax function to obtain landmark coordinates from the predicted heatmaps. As the argmax function is non-differentiable, our proposed flow-based landmark distribution prior cannot be used in the heatmap-based framework directly. Inspired by IntegralNet [24] and DSNT [25], which use an integral operation to solve the non-differentiable problem, we adopt a soft-argmax function to produce the landmark coordinates from their corresponding heatmap in a differentiable manner as follows:

$$\begin{aligned}\mathcal{C}_i &= \int_{\boldsymbol{b}\in\mathbb{R}^{H\times W}} \boldsymbol{b}\bar{\boldsymbol{M}}_i(\boldsymbol{b}) \\ &= \sum_{y=1}^{H}\sum_{x=1}^{W} b_{xy}\cdot\bar{\boldsymbol{M}}_i(b_{xy}),\end{aligned} \tag{7}$$

where $\bar{\boldsymbol{M}}_i \in \mathbb{R}^{H\times W}$ denotes the i-th normalized heatmap and b_{xy} is the heatmap value at the location (x, y). The

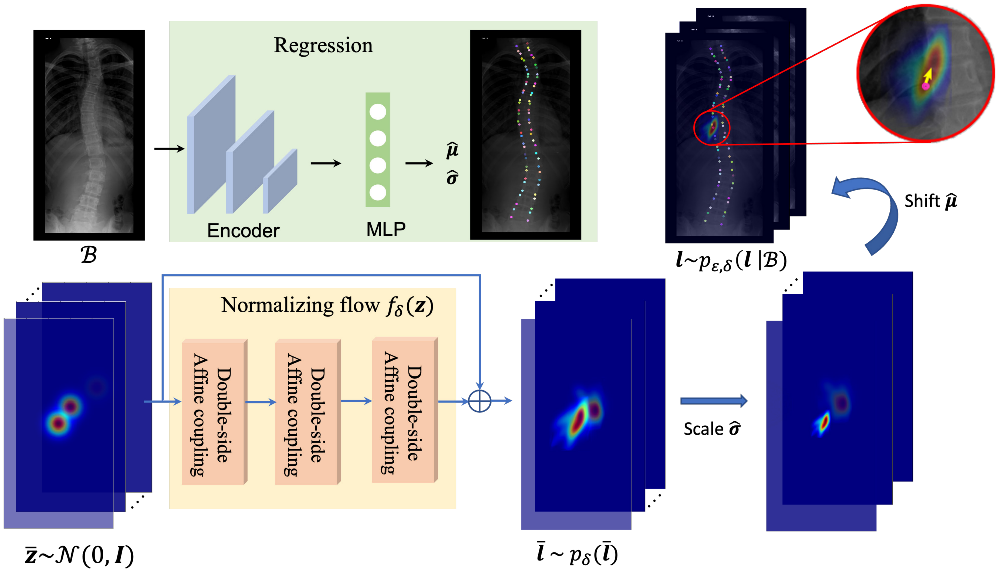

Regression
Encoder
MLP
$\hat{\boldsymbol{\mu}}$
$\hat{\boldsymbol{\sigma}}$
$\mathcal{B}$
$\boldsymbol{l}$~$p_{\varepsilon,\delta}(\boldsymbol{l}\,|\mathcal{B})$
Shift $\hat{\boldsymbol{\mu}}$
Normalizing flow $f_\delta(\boldsymbol{z})$
Double-side Affine coupling
Double-side Affine coupling
Double-side Affine coupling
Scale $\hat{\boldsymbol{\sigma}}$
$\bar{\boldsymbol{z}}$~$\mathcal{N}(0,\boldsymbol{I})$
$\bar{\boldsymbol{l}} \sim p_\delta(\bar{\boldsymbol{l}})$

Fig. 4: An overview of the proposed regression-based landmark localization framework with normalizing flows for landmark distribution learning.

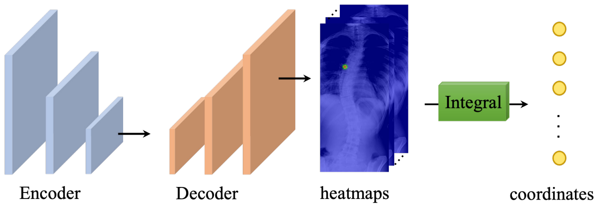


Fig. 5: An illustration of the integral operation for the heatmap-based approach.

predicted coordinate is the integration of all locations $\boldsymbol{b}$. The working pipeline of the integral operation is shown in Fig. 5. In this paper, we adopt the softmax function to normalize the heatmaps to ensure that all the values of the heatmap are non-negative with their sum being equal to 1. With the differentiable soft-argmax function, heatmap-based frameworks can benefit from the learned landmark distribution to improve their performance. Moreover, different from IntegralNet [24] which learns from both heatmap labels and coordinates, and DSNT [25] which adopts a regularization strategy in the heatmap feature layer, we directly utilize the coordinates to train the heatmap-based model. This is because we have already constructed the landmark distribution with normalizing flows.

D. Training Scheme

Our proposed landmark distribution prior works as a learnable penalty function in our framework. Therefore, the whole framework can be trained in an end-to-end manner under the supervision defined in Eq. (6), which optimizes the regression and the flow models simultaneously. It is worth noting that the normalizing flows serve as a penalty function, which only affects the training stage, and will be detached from the framework on inferencing. Thus, it adds minimal additional computational burden to real-world applications. In the inference stage, the regression model directly estimates the coordinates, i.e. $\hat{\boldsymbol{\mu}}$, of the given image. In other words, our proposed algorithm introduces the distribution prior without changing the structure of the original model. This property makes our algorithm easy to embed into various regression and heatmap-based frameworks.

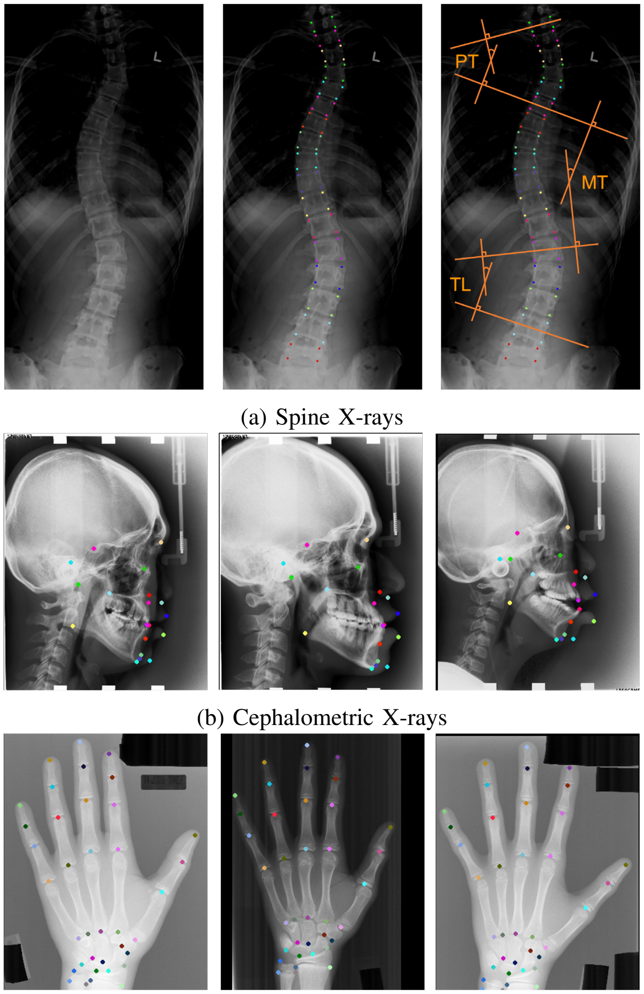
IV. Experiments

In this section, we introduce the implementation details of the proposed framework with the normalizing-flow-based distribution prior (NFDP). We also present the datasets used in our experiments. We further discuss and analyze the experimental results based on the proposed NFDP framework. Specifically, we evaluate our framework on three publicly available datasets, which are collected from the X-ray modality with different focusing regions, namely spine, head, and hand. Fig. 6 shows some representative examples from these three datasets.

A. X-ray Spinal Landmark Localization

1) Dataset: We use the Accurate Automated Spinal Curvature Estimation (AASCE[2]) MICCAI2019 challenge dataset 16 to compare our proposed framework with various regression-based methods, as well as the state-of-the-art heatmap-based methods. The AASCE dataset consists of 609 spinal X-ray images in the anterior-posterior direction. It is further split into two groups, i.e., 481 training samples and 128 testing samples. Each sample image contains 68 landmarks for locating 17 vertebrae from the thoracic and lumbar regions, and each vertebra is located by four corner landmarks. After performing landmark localization, the detected landmarks are

[2] https://aasce19.github.io/



(a) Spine X-rays

(b) Cephalometric X-rays

(c) Hand X-rays

Fig. 6: Visualization of sample images from the three datasets. Circles denote the reference annotations. (a) Anterior-posterior (AP) X-ray images in the spine region, in which the vertebrae are labeled by circles in different colors (4 circles of the same color for each vertebra). The image on the right computes the Cobb angles of the proximal thoracic (PT), main thoracic (MT), and thoracolumbar (TL) regions, by using the measurement in [8]. (b) Lateral cephalograms of three different patients with 19 expert annotated landmarks. (c) Left-hand radiographs with 37 annotated landmarks.

used to compute the Cobb angle based on the algorithm provided by the challenge organizers. The images from this dataset are of different sizes, but basically of a resolution of $2,500 \times 1,000$. To spatially normalize the images and reduce the computational complexity, we resize them into 512×256 for both training and testing.

2) Implementation Details: We implement our framework with PyTorch. As for the regression-based framework, we adopt ResNet18 [41] as the encoder network for feature extraction. Different from the vanilla ResNet18 that only uses the features from the last convolutional block, we follow the feature pyramid strategy [42] and fuse the intermediate features from all convolutional blocks to obtain a comprehensive representation of the image. Finally, we employ two regression heads to predict the shift $\hat{\boldsymbol{\mu}}$ and the scale $\hat{\boldsymbol{\sigma}}$. Both heads consist of a dense layer with $2 \times K$ nodes, where $K = 68$ is the number of spinal landmarks. In terms of normalizing flows, the neuron number N_n is empirically set to 64 for both $\theta(\cdot)$ and $\phi(\cdot)$. More details can be found in Sec. III-A. As for the heatmap-based framework, we also adopt ResNet18 as the backbone, followed by a Feature Pyramid Network (FPN) to fuse multi-scale features for heatmap regression.

During training, we employ Adam to jointly optimize the localization networks and the normalizing flows under the supervision defined in Eq. (6). We train the network for 300 epochs with the learning rate linearly decreasing from 8×10^{-4} to 10^{-5}. The batch size is set to 8. Random shifting with a range of [-0.2, 0.2], random scaling with a range of [-0.15, 0.15], and random rotating with a range of [-0.3, 0.3] are performed for data augmentation. The training and the inference efficiency test (measured in Frames per Second) were performed on a PC with Intel Core i9-9900K, 64GB of RAM, and one Nvidia GeForce RTX 3090 GPU.

3) Metrics: To evaluate the performance of different spinal landmark localization algorithms, we follow the settings in the AASCE challenge, and use symmetric mean absolute percent error (SMAPE) to calculate the accuracy of the Cobb angles based on the detected landmarks. It is formulated as follows:

$$\text{SMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{3}(|\hat{c}_{ij} - c_{ij}|)}{\sum_{j=1}^{3}(\hat{c}_{ij} + c_{ij})}, \tag{8}$$

where N is the number of testing images; c_{ij} and $\hat{c}_{ij}$ are the ground-truth Cobb angle and the estimated Cobb angle of the i-th image, respectively; $j = 1$, 2 and 3 because the Cobb angles concern three regions, i.e., proximal thoracic (PT), main thoracic (MT), and thoracolumbar (TL). Furthermore, we adopt mean radial error (MRE) to evaluate the localization accuracy of the landmarks, which is defined as follows:

$$\text{MRE} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \parallel \boldsymbol{x}_{ij} - \hat{\boldsymbol{x}}_{ij} \parallel_2, \tag{9}$$

where N and M are the number of images and the number of landmark categories, respectively; $\boldsymbol{x}_{ij} \in \mathbb{R}^2$ and $\hat{\boldsymbol{x}}_{ij} \in \mathbb{R}^2$ represent the j-th ground-truth landmark coordinates and the estimated landmark coordinates of the i-th image, respectively; $\parallel \cdot \parallel_2$ denotes the Euclidean distance.

4) Comparisons: To show the superiority of our proposed NFDP, we compare our framework with other state-of-the-art (SOTA) methods. They include heatmap-based methods such as landmark detection network (LDN) [21], HybridNet [22] and HRNet [43], integral heatmap-based methods such as IntegralNet [24] and DSNT [25], and regression-based methods such as Multi-View Extrapolation Net (MVENet) [8] and BoostNet [11]. Since the dataset used in LDN [21], HRNet [43], MVENet [8], IntegralNet [24], DSNT [25], and BoostNet [11] are different from ours, we compare our results with the results of LDN [21] reported in [22] and further re-implement the other methods. As the source code of MVENet [8] is not publicly available, we re-implement it based on the description

TABLE I: Quantitative results on the X-ray Spinal dataset in terms of symmetric mean absolute percent error (SMAPE %) and mean radial error (MRE). The best results are in **bold** and the second-best results are $\underline{\text{underlined}}$.

Structure	Methods	resolution	Params ↓	GFLOPs ↓	FPS ↑	$SMAPE \downarrow$	$SMAPE_{PT} \downarrow$	$SMAPE_{MT} \downarrow$	$SMAPE_{TL} \downarrow$	$MRE \downarrow$
Regression	MVENet [8]	512 × 256	18.8M	9.7	24.4	18.34	11.3	25.59	32.74	73.31
	BoostNet [11]	256 × 128	31.2M	**2.9**	**27.4**	$25.35^{\ddagger}$	$23.13^{\ddagger}$	$25.47^{\ddagger}$	$33.53^{\ddagger}$	$93.54^{\ddagger}$
	NFDP (ours)	512 × 256	**11.5M**	$\underline{4.8}$	$\underline{25.5}$	$\underline{8.92}^{\dagger}$	$6.31^{\ddagger}$	$\mathbf{14.90}^{\dagger}$	$\underline{20.52}^{\dagger}$	$\underline{55.58}^{\ddagger}$
Heatmap	LDN [21]	1024 × 512	24.2M	85.8	10.3	$9.71^{\ddagger}$	$\underline{6.22}^{\ddagger}$	$\underline{15.39}^{\ddagger}$	$22.39^{\ddagger}$	$61.90^{\ddagger}$
	HybridNet [22]	1024 × 512	20.1M	51.5	8.5	**8.40**	**5.51**	15.67	**18.00**	$\underline{55.56}$
	HRNet [43]	512 × 256	28.5M	20.6	14.1	17.55	11.39	25.57	31.35	59.12
	IntegralNet [24]	512 × 256	38.7M	42.2	11.8	$16.68^{\dagger}$	$10.23^{\ddagger}$	$22.3^{\dagger}$	$30.56^{\ddagger}$	$55.82^{\ddagger}$
	DSNT [25]	512 × 256	29.8M	24.4	13.2	14.74	9.80	22.70	29.55	56.90
	NFDP (ours)	512 × 256	$\underline{12.1M}$	6.8	20.8	14.58	8.87	20.94	29.44	**48.92**

†, ‡ denote p-values smaller than 0.1 and 0.01 respectively.

in its original paper. HRNet[3] [43], IntegralNet[4] [24], DSNT[5] [25] and BoostNet[6] [11] are established with their official implementations. All the aforementioned methods are trained under the same settings for fair comparison.

The results are tabulated in Table I. It can be seen that heatmap-based approaches generally achieve better landmark localization accuracy than regression-based approaches. This is because heatmap-based methods employ high-resolution features to predict the likelihood of heatmap for each landmark independently, making them more robust than the regression-based methods. Our proposed NFDP with regression learning can produce comparable or even better results than the SOTA heatmap-based methods. More importantly, our heatmap-based NFDP achieves the lowest mean radial error, i.e., 48.92, among all the competitors, which demonstrates the effectiveness of the proposed distribution prior in landmark localization learning.

On the other hand, the regression-based NFDP surpasses heatmap-based NFDP by a large margin in terms of Cobb angle measurement using SMAPE, which indicates that the regression-based NFDP can better capture the curvature of the whole spine. The reason is that the landmarks are predicted by the last fully connected layer of the regression-based NFDP, in which the predicted landmark coordinates are controlled by a limited number of neurons. Therefore, the dependency between different landmarks can be easily captured. To more clearly show the advantages of the regression-based NFDP, we visualize three cases from the testing set. As shown in Fig. 7, although the heatmap-based NFDP can predict more accurate landmark positions, the similar representations of different landmark heatmaps may confuse the detector in some landmark categories, which results in lower accuracy in terms of the Cobb angle measurement. In contrast, the regression-based NFDP predicts all the landmarks with their structure dependency, making it more favorable for the spine curvature measurement.

As shown in Table I, HybridNet [22] generally produces the best results in terms of SMAPE among all the competitors. However, it suffers from high computational complexity, as it requires about 51.5 GFLOPS to process one testing sample. Similarly, heatmap-based approaches like IntegralNet [24], DSNT [25], LDN [21] and HRNet [43] also require large computational resources, i.e., model size and GFLOPS, on deployment. Nevertheless, our proposed regression-based NFDP achieves comparable results on both MRE and SMAPE to those of the SOTA methods with much lower computational complexity, i.e., 11.5M learnable parameters and 4.8 GFLOPS. Considering the regression-based track, we can see that the regression-based NFDP outperforms the two previous regression-based methods, i.e., MVENet [8] and BoostNet [11], by a large margin in terms of both SMAPE and MRE. As compared with the existing methods, our proposed NFDP provides a superior balance between prediction accuracy and inference speed, as shown in Fig. 8.

To statistically evaluate the significance of the obtained comparison results, we use the paired sample t-test [44] to analyze the experiment results. Specifically, we compare the adopted evaluation metrics, i.e., SMAPE and MRE of our proposed heatmap-based NFDP with those of the proposed regression-based NFDP and other compared methods, including BoostNet [11], LDN [21] and IntegralNet [24]. The obtained p-value results with a threshold of 0.01 for validating the statistical significance are tabulated in Table I. The results indicate the differences in performance are statistically significant, which means that our proposed method improves the landmark localization performance of those benchmark methods. Although some results only have a threshold p-value of 0.1, those results still provide some evidence against the null hypothesis.

B. X-ray Cephalogram Landmark Localization

1) Dataset: To further analyze our proposed framework, we evaluate it with the dataset of ISBI 2015 Cephalometric X-ray Image Analysis Challenge [51]. The dataset consists of 400 cephalograms from 400 subjects, each of which is associated with 19 annotated landmarks. The images are of a uniform size of $1,935 \times 2,400$ pixels with spacing of $0.1mm \times 0.1mm$. We resize them into 512×512 for both training and testing. We follow the training-testing split strategy in [51], which selects 150 images for training, 150 images for first testing (testing data 1), and 100 images for second testing (testing data 2). Similarly, we train our framework for 300 epochs with a learning rate linearly decreasing from 8×10^{-4} to 10^{-5}. The implementation details and training strategies follow the settings in Sec. IV-A.2.

2) Metrics: To measure the performance of different localization algorithms, we use the same metrics from previous

[3] https://github.com/HRNet/

[4] https://github.com/JimmySuen/integral-human-pose

[5] https://github.com/anibali/dsntnn

[6] https://github.com/lulufa390/Spine-Landmark-Detection

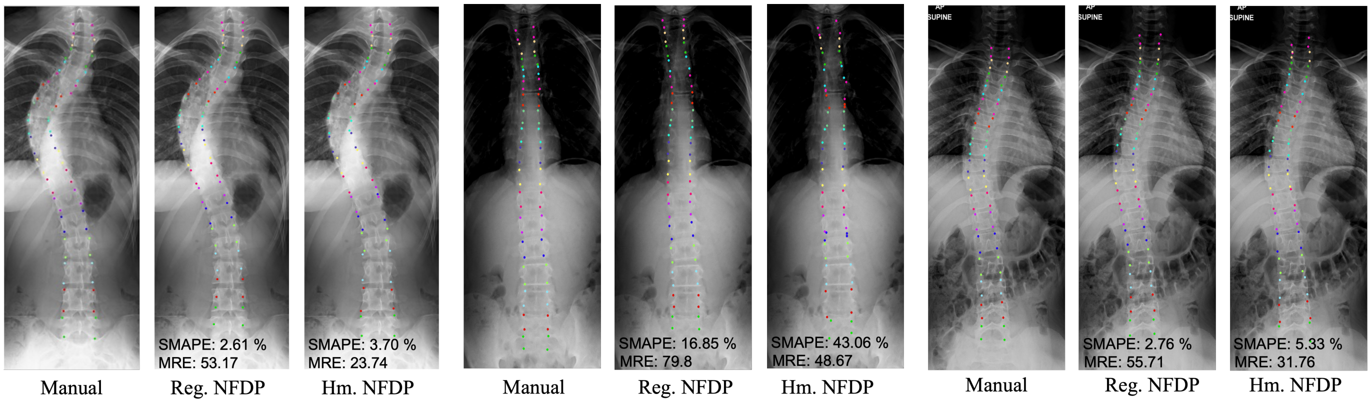


Fig. 7: Visualization of the spine landmark localization results based on the regression-based NFDP and heatmap-based NFDP.

TABLE II: Quantitative landmark localization results over the 19 landmarks of X-ray Cephalograms dataset in terms of mean radial error (MRE (in mm)) and successful detection rate (SDR (%)). The best results are in **bold** and the second-best results are <u>underlined</u>.

Structure	Methods		Testing data 1					Testing data 2				
			MRE ↓	SDR ↑				MRE ↓	SDR ↑			
				2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
Single-stage Regression	Zhou et al. [45]		2.13 ± 1.41	55.43	70.32	80.57	90.49	2.49 ± 1.95	47.68	62.53	73.72	84.42
	Nie et al. [6]		2.43 ± 1.79	47.63	64.70	74.84	84.40	2.94 ± 2.21	39.73	55.26	67.21	78.89
	NFDP (ours)		1.98 ± 1.29	60.01	73.72	83.54	92.98	2.36 ± 1.74	50.58	65.05	75.32	86.58
Multi-stage Regression	Zeng et al. [18]		1.34 ± <u>0.92</u>	81.37	89.09	93.79	97.86	1.64 ± <u>0.91</u>	70.58	79.53	86.05	93.32
	Lee et al. [46]	initial	2.63 ± 1.47	41.19	55.23	67.02	83.12	-	-	-	-	-
		final	1.19 ± **0.80**	86.42	92.00	<u>95.68</u>	<u>98.46</u>	-	74.58	81.79	87.53	94.26
Heatmap	Oh et al. [47]		1.18 ± 1.01	86.20	91.20	94.40	97.70	**1.46 ± 0.82**	**75.90**	<u>83.40</u>	**89.30**	94.70
	Chen et al. [48]		<u>1.17</u>	<u>86.67</u>	**92.67**	95.54	**98.53**	1.48	75.05	82.84	88.53	<u>95.05</u>
	McCouat et al. [49]		1.20	83.47	89.16	92.60	96.49	**1.46**	74.63	**83.58**	87.21	93.79
	Sun et al. [24]		1.23 ± 1.09	84.73	91.22	94.55	97.57	1.55 ± 1.44	73.25	80.72	87.48	94.26
	Nibali et al. [25]		1.35 ± 1.06	81.48	90.53	94.32	97.82	1.68 ± 1.38	69.52	79.53	86.00	93.74
	Chen et al. [50]		1.35 ± 1.10	79.57	88.63	93.72	96.81	1.61 ± 1.35	70.47	80.21	87.53	93.68
	NFDP (ours)		**1.14** ± 0.93	**87.02**	<u>92.38</u>	**95.76**	98.35	**1.46** ± 1.31	<u>75.11</u>	82.48	<u>89.16</u>	**95.16**

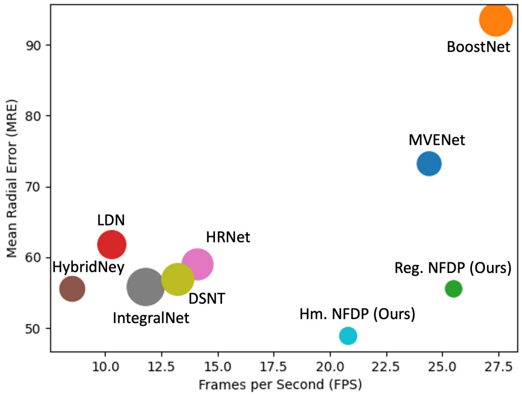


Fig. 8: Comparison of different methods based on Frames per Second (FPS), Mean Radial Error (MRE), and model size. The model size is reflected by the diameter of the circle.

studies [18], [46]–[49], i.e., mean radial error (MRE) and successful detection rate (SDR), where MRE is defined in Eq. (9). For SDR, consider a predicted landmark $\hat{\boldsymbol{l}}_i$. It is regarded as a successful detection if the Euclidean distance between it and its ground-truth landmark $\boldsymbol{l}_i$ is smaller than a pre-defined distance β. Therefore, SDR is formulated as follows:

$$\mathrm{SDR}_\beta = \frac{\#(\{\hat{\boldsymbol{l}}_i : \| \hat{\boldsymbol{l}}_i - \boldsymbol{l}_i \|_2 < \beta\})}{\#(L)}, \tag{10}$$

where $\#(\cdot)$ denotes the cardinal function, and L is a set of estimated landmarks over the whole testing set.

3) Comparisons: We compare our proposed NFDP with other SOTA methods, including single-stage regression-based algorithms [6], [45], multi-stage regression-based algorithms [18], [46], heatmap-based algorithms [47]–[50], and integral heatmap-based algorithms [24], [25]. The quantitative results are tabulated in Table II. From this table, our proposed heatmap-based NFDP achieves the best performance in terms of MRE on both the first and second testing sets, which shows that the learned landmark distribution can significantly promote landmark localization learning. In terms of successful detection rate, the heatmap-based NFDP, benefiting from the learned landmark structure, can generally achieve the best performance among the other SOTA algorithms.

As for the regression-based algorithms, owing to the lack of high-resolution features, our proposed regression-based NFDP fails to accurately locate the cephalogram landmarks. To solve this problem, [46] and [18] proposed to first predict the initial positions of all the landmarks and then utilize a series of encoder networks to refine the locations of the corresponding landmarks. Although this coarse-to-fine regression-based framework achieves promising performance, both methods in [46] and [18] need to train 19 refined networks to accurately estimate all the landmarks, which is tedious and computationally expensive. When comparing with the single-stage regression-based framework, our proposed regression-based NFDP significantly outperforms those in [45], [6] and the initial regression network proposed in [46], which illustrates the superiority of the learned landmark distribution prior.

To evaluate the robustness of our proposed NFDP, we conducted disturbance tests by i) introducing Gaussian noise with σ of 9, ii) applying average blur filtering of size 25×25, iii) random shifting with a factor of ± 0.1 in both horizontal and vertical directions, and iv) random rotation with ± 15 degrees to the test images from the X-ray Cephalograms dataset. In our evaluations, we merged testing data 1 and testing data 2 into a single testing dataset for simple comparison to simplify the comparison and analysis. We included benchmark comparisons with state-of-the-art algorithms for cephalogram landmark localization, such as context feature learning network (ContextNet) [47], structural-award network (SA-Net) [50], contour-hugging network (CH-Net) [49], as well as multi-stage regression (Multi-Reg) [46]. We established CH-Net [7] [49] using the source code released by the authors and re-implemented ContextNet [47], SA-Net [50], and Multi-Reg [46] based on their respective papers. Our assessments utilized the Mean Radial Error (MRE, in mm) and Frames per Second (FPS) to evaluate the accuracy and inference speed of our NFDP versus the other candidates. Quantitative results are presented in Table III. Notably, when Gaussian noise was introduced, all algorithms witnessed a performance drop, with SA-Net [50] experiencing a particularly pronounced 1mm decrease in MRE. While our NFDP was not immune to the noise, it still achieved the best MRE, reflecting its robustness. A similar trend was observed with the results after applying a smoothing filter. Since the data augmentation techniques used during the training phase included image shifting and rotation, the trained models of all the tested algorithms show strong robustness to the disturbance of random shifting and random rotation. Specifically, our proposed NFDP shows the least performance drop, indicating superior robustness as compared to other methods. Regarding inference efficiency, our proposed NFDP, due to its straightforward architecture, reached an impressive FPS of 12.7. Meanwhile, the incorporation of attention modules [47], [49], [50] and multi-stage network structures [46] in the compared algorithms caused them to lag behind our NFDP in inference efficiency.

To highlight the applicability of our proposed NFDP, we trained it using different numbers of training samples, ranging from 15 to 150. We then benchmarked its performance against leading algorithms in cephalogram landmark localization [46], [47], [49], [50]. The findings are detailed in Table IV. As indicated in Table IV, our proposed NFDP consistently outperforms others across different training sample sizes. Notably, when trained with a limited dataset of just 15 samples, NFDP significantly outperforms all other candidates. Furthermore, to illustrate the flexibility of our proposed NFDP, we integrated it with different localization backbones, i.e., EfficientNet [52], HRNet [43] and ResNet [41]. The quantitative results are also listed in Table IV. It can be seen that our proposed NFDP is beneficial to the landmark localization performance, as the MRE improves by about 0.15, 0.1, and 0.24 as compared to the ResNet [41], HRNet [43], and EfficientNet [52] baseline, respectively.

TABLE III: Quantitative landmark localization disturbance test results over the 19 landmarks of X-ray Cephalograms testing data 1 and testing data 2 (total 250 samples) in terms of mean radial error (MRE (in mm)) and Frames per Second (FPS). The best results are in **bold**.

Method	MRE (mm) ↓					FPS ↑
	original	Noise	Smooth	Rotation	Shift	
ContextNet [47]	1.45	1.64	1.70	1.52	1.55	6.9
SA-Net [50]	1.35	2.32	1.55	1.42	1.47	6.2
CH-Net [49]	1.46	1.68	1.67	1.56	1.58	8.2
Multi-Reg [46]	1.40	1.78	1.59	1.46	1.53	3.4
Hm. NFDP (ours)	**1.27**	**1.57**	**1.44**	**1.31**	**1.33**	**12.7**

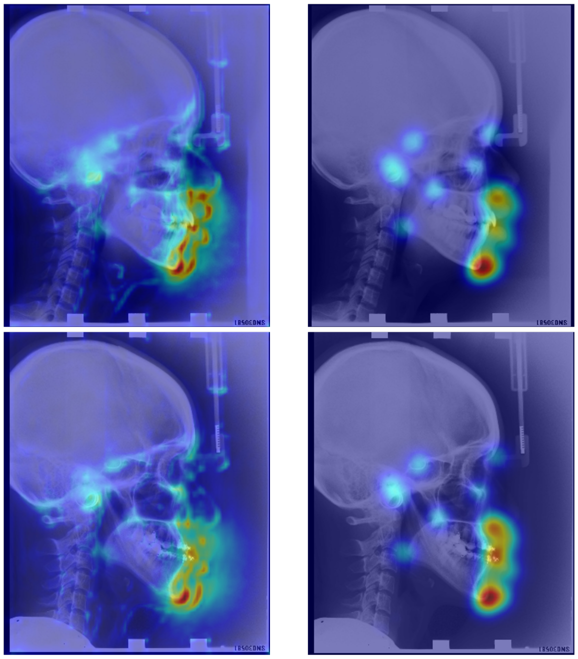
TABLE IV: Quantitative landmark localization results over the 19 landmarks of X-ray Cephalograms testing data 1 and testing data 2 (total 250 samples) based on different training samples in terms of mean radial error (MRE (in mm)). The best results are in **bold**.

Method	No. of Training Samples					
	15	30	60	90	120	150
ContextNet [47]	2.53	2.27	1.66	1.58	1.51	1.45
SA-Net [50]	2.83	2.17	1.55	1.45	1.39	1.35
CH-Net [49]	2.25	1.95	1.59	1.51	1.48	1.46
Multi-Reg [46]	2.42	2.04	1.63	1.57	1.48	1.40
EfficientNet-b0 [52]	4.6	3.43	2.07	1.79	1.59	1.63
EfficientNet-b0-NFDP	3.2	2.89	1.78	1.52	1.43	1.39
HRNet-w18 [43]	3.92	2.83	1.90	1.64	1.57	1.48
HRNet-w18-NFDP	2.54	2.4	1.75	1.56	1.47	1.38
ResNet18 [41]	3.84	2.34	2.01	1.68	1.49	1.42
ResNet18-NFDP	**2.02**	**1.71**	**1.53**	**1.38**	**1.31**	**1.27**

To more clearly illustrate the effectiveness of the proposed NFDP, we visualize the produced heatmaps based on the proposed heatmap-based NFDP and the traditional heatmap-based methods, on two testing samples by Grad-CAM [53]. The results are shown in Fig. 9. It can be seen that, different from traditional heatmap-based algorithms that indiscriminately synthesize a Gaussian heatmap for every landmark, our proposed heatmap-based NFDP aims to learn the distribution for each landmark. More importantly, we can see that the learned distribution is far from the pre-defined Gaussian or Laplacian distributions. This indicates that the traditional heatmap-based approaches inherently introduce bias when using manually designed heatmaps in learning. The examples above further demonstrate the superiority of our proposed NFDP in landmark localization tasks.

Also, we utilize Shapley Additive Explanations (SHAP) [54] to show the contribution of each input image feature to the predicted landmarks. The visualizations of the SHAP values according to the corresponding landmarks are shown in Fig. 10. As demonstrated by the SHAP values in Fig. 10, the significance of the input feature aligns well with the corresponding landmark. This reflects the effectiveness of

[7] https://github.com/jfm15/ContourHuggingHeatmaps/



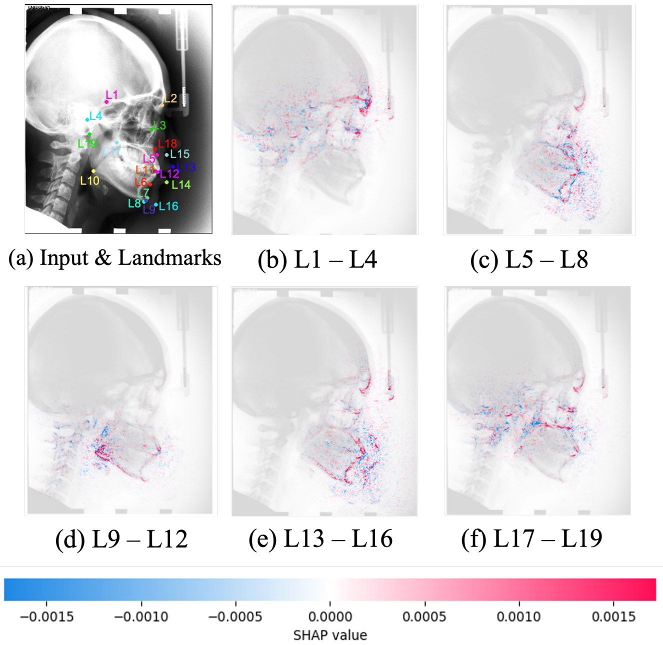
(a) Learning with NFDP (b) Learning with heatmap

Fig. 9: Visualizations of the feature maps extracted from the output layer of the heatmap-based network. (a) The heatmap produced by our proposed heatmap-based NFDP; (b) The heatmap produced by the model trained by heatmap label.

the decision-making process of our proposed heatmap-based NFDP.

C. X-ray Hand Landmark Localization

1) Dataset: To further validate our proposed NFDP, we apply it to a commonly used dataset for landmark detection from hand radiographs [8]. This dataset consists of 895 left-hand images, each of which contains 37 annotated landmarks on bone joints and fingertips. Since the radiographs are acquired from different scanners, the images are of different spatial sizes, but basically have a resolution of 1536×2169. Following the same normalization strategy in [16], [20], [55], [56], the wrist width of each image is assumed to be 50mm. For fair comparison, we employ the 3-fold cross-validation with the same split provided by [20], and utilize mean radial error (MRE) and success detection rate (SDR) with $\beta = 2, 4, 10mm$ to assess the performance of different algorithms. Each sample image is resized to 512×512 for both training and testing. We follow the same training strategies in Sec. IV-A.2 to train the network for 200 epochs with the learning rate linearly decreasing from 8×10^{-4} to 10^{-5}.



(a) Input & Landmarks (b) L1 – L4 (c) L5 – L8

(d) L9 – L12 (e) L13 – L16 (f) L17 – L19

Fig. 10: Visualizations of the SHAP value of the heatmap-based NFDP. (a) The input image and the predicted landmarks; (b) - (f) The SHAP value according to the corresponding landmark prediction.

TABLE V: Quantitative results of X-ray Hand dataset in terms of mean radial error (MRE (in mm)) and successful detection rate (SDR (%)).

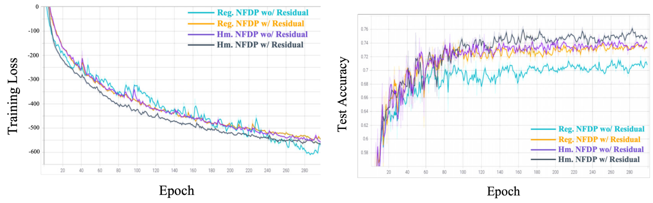
Structure	Methods	MRE ↓	SDR ↑		
			2mm	4mm	10mm
Regression	Zhou et al. [45]	1.26 ± 0.92	82.41	96.46	98.52
	Nie et al. [6]	1.31 ± 0.98	81.31	95.49	98.23
	NFDP (ours)	1.04 ± 0.87	88.27	98.89	99.97
Random Forest	Šternetal et al. [55]	0.80 ± 0.91	92.20	98.45	99.83
	Lindner et al. [15]	0.85 ± 1.01	93.68	98.95	99.94
	Urschler et al. [16]	0.80 ± 0.93	92.19	98.46	99.95
Heatmap	Zhu et al. [56]	0.84	**95.40**	99.35	99.75
	Nibali et al. [25]	0.85 ± 1.07	93.20	99.10	99.67
	Payer et al. [20]	**0.66** ± **0.74**	94.99	99.27	**99.99**
	Sun et al. [24]	0.79 ± 0.94	93.82	99.32	99.96
	NFDP (ours)	0.72 ± **0.74**	94.38	**99.49**	**99.99**

2) Comparisons: We compare our proposed NFDP with regression-based algorithms [6], [45], random forest-based algorithms [15], [16], [55], and other state-of-the-art algorithms, [20], [24], [25], [56]. The results are tabulated in Table V. Our proposed heatmap-based NFDP produces the best SDR with $\beta = 4mm$ and $10mm$. It also achieves comparable performance in terms of SDR with $\beta = 2mm$ and MRE to the other algorithms. Although our regression-based NFDP achieves promising accuracy in terms of SDR with $\beta = 4mm$ and $10mm$, a significant drop of about 6.1% in SDR with $\beta = 2mm$ can be seen when compared with the heatmap-based NFDP. We consider the reason is that the regression-based NFDP lacks high-resolution features in its landmark prediction layer, while the heatmap-based NFDP takes full advantage of high-resolution features in its framework.

D. Ablation Study

To show the impact of different distribution assumptions in the landmark regression frameworks, we compare the results by using different density loss functions on the spinal X-ray dataset. The result is shown in Table VI. The L1 and L2 loss functions imply that the landmarks follow the Laplace and Gaussian distribution, respectively. Table VI reports the mean radial error (MRE) for landmark localization and symmetric mean absolute percent error (SMAPE) for Cobb angle

[8] www.ipilab.org/BAAweb



(a) Training process (b) Validation comparison

Fig. 11: Training processes and validation comparison of different network structures on the X-ray spinal dataset.

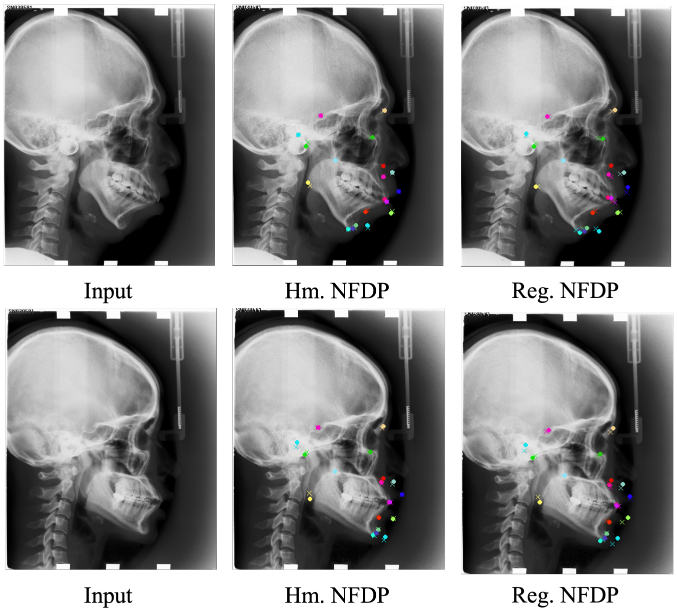
measurement. It can be seen that a significant improvement is achieved when using our proposed NFDP to learn the landmark distribution. We also present the number of model parameters and floating-point operations per second (FLOPs) to assess the computational complexity of different methods. As tabulated in Table VI, these values are not changed by the NFDP. We can conclude that the NFDP introduces no computational cost to the framework, making it more valuable in clinical applications.

We also analyze the effectiveness of the learned landmark distribution for heatmap-based methods. We conducted ablation experiments with 1) direct heatmap, 2) integral heatmap without the NFDP, 3) the NFDP without residual learning and with integral, and 4) the NFDP with residual learning and with integral. The quantitative results are tabulated in Table VI. When directly regressing the heatmap without applying the normalizing flow-based distribution prior, we observed a decrease in accuracy. This result reflects the significance of utilizing the proposed distribution prior in our approach. Though the integral heatmap showed reasonable performance, it was still not as effective as our full model. For our full model that combines the NFDP, residual learning, and integral heatmap, the results demonstrate superior performance as compared to the other settings.

To further show the importance of the residual learning in the proposed density estimation, we eliminate the shortcut of residual learning in the framework of both regression-based and heatmap-based structures, denoted as "NFDP without residual", and compare the results with the whole framework. As listed in Table VI, the residual learning strategy is beneficial to landmark localization, as both SMAPE and MRE are improved by a large margin for the regression-based and heatmap-based network structures. To better demonstrate the stabilizing effect of our proposed residual learning on the training process, we visualize the training progress both with and without residual learning in Fig. 11. As seen from this figure, the incorporation of residual learning leads to a more stable decrease in training loss and improved validation accuracy.

TABLE VI: Quantitative landmark localization and Cobb angle measurement results of the X-ray Spinal dataset in terms of symmetric mean absolute percent error (SMAPE (%)) and mean radial error (MRE (in pixel)) based on different distribution assumptions and different network structures.

Structure	Methods	Params	GFLOPs	$SMAPE\downarrow$	$MRE\downarrow$
Regression	Laplace (L1)	11.5M	4.8	13.59	62.52
	Gaussian (L2)	11.5M	4.8	14.95	72.11
	NFDP w/o residual	11.5M	4.8	12.06	56.06
	NFDP	11.5M	4.8	**8.93**	55.58
Heatmap	Direct heatmap	12.1M	6.8	15.08	58.33
	Integral heatmap w/o NFDP	12.1M	6.8	14.72	55.26
	NFDP w/o residual w/ Integral	12.1M	6.8	14.81	53.42
	NFDP w/ residual w/ Integral	12.1M	6.8	14.58	**48.92**



Input Hm. NFDP Reg. NFDP

Input Hm. NFDP Reg. NFDP

Fig. 12: Visualization of the X-ray head landmark localization results based on the regression-based NFDP and heatmap-based NFDP. (circle: predicted, cross: manual landmarks)

E. Summary and Discussion

We have validated the proposed NFDP algorithm on three widely used benchmark datasets. Specifically, for spinal landmark localization, although the heatmap-based NFDP achieves the best localization accuracy, the regression-based NFDP shows its superiority in capturing the spinal curvature features. Moreover, our method achieves competitive performance to the state-of-the-art heatmap-based algorithms in terms of both landmark localization and Cobb angle measurement, with lower computational consumption and fewer model parameters.

When evaluating our proposed NFDP in cephalogram landmark localization and hand landmark localization tasks, our proposed regression-based NFDP performs unremarkably, while the heatmap-based NFDP achieves good performance. In our proposed methods, the difference between the regression-based NFDP and heatmap-based NFDP is that the heatmap-based NFDP has an additional decoding network, which enables the model to utilize high-resolution features for landmark location estimation. We argue that the degradation is because the encoder architecture uses only the high semantic features, while accurate localization for every landmark also relies on the high-resolution features. Furthermore, without manually constructing heatmaps in learning, our proposed heatmap-based NFDP outperforms the other heatmap-based methods, which indicates that heatmap labels may not be necessary for

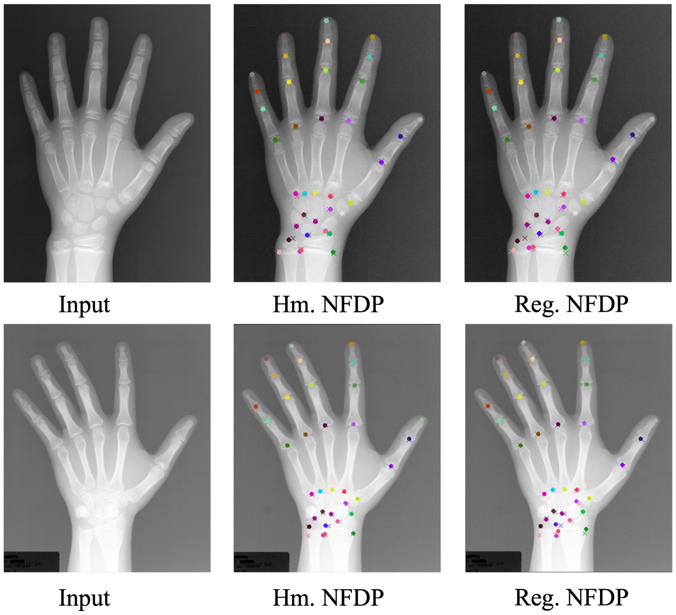


Fig. 13: Visualization of the X-ray hand landmark localization results based on the regression-based NFDP and heatmap-based NFDP. (circle: predicted, cross: manual landmarks)

training heatmap-based frameworks.

The regression-based NFDP utilizes a Multilayer Perceptron (MLP) for prediction. This ensures that all the landmark predictions are based on the same features, resulting in strong correlation among the predicted landmarks. In the spine X-ray dataset, this correlation makes the regression method more robust to the prediction of spinal curvature in scoliosis detection. This is because along the entire spine, the landmarks of different vertebrae from T1 to L5 have strong location characteristics. For example, the landmarks of L2 are always between L1 and L3, and so on.

As shown in Fig. 7, although the heatmap-based NFDP is more accurate in the prediction of a single landmark, the characteristics of different spinal vertebrae are very similar, which makes the heatmap-based method prone to confusion, resulting in a large deviation in predicting the degree of curvature of the whole spine.

For the other two datasets, as shown in Fig. 12 and Fig. 13, since there is no strong correlation and similarity between the landmarks, the prediction accuracy of the regression-based algorithm drops significantly as compared with that of the heatmap-based NFDP owing to the lack of high-resolution features in the prediction of landmarks. Yet, as the network structure of the regression-based algorithm is relatively simple, the regression-based algorithm has a higher inference efficiency and lower computational requirement.

In summary, both algorithms have their own advantages. Under the tasks of strong correlation between landmarks and strong similarity of landmark features, if there is a requirement for the overall shape and trend of the landmarks, the regression-based algorithm should be more robust. In scenarios where there is high independence among the features of various landmarks, and the precision of a single landmark is substantial, the performance of the heatmap-based method is superior.

Although our proposed normalizing flow-based distribution prior has achieved promising results in the tested datasets, in practical applications, because of the limited number of marked landmarks, the learned landmark distribution may not appreciatively match the actual underlying distribution. Considering the fact that our proposed method only needs the annotated landmark coordinates for constructing the underlying landmark distribution, one potential solution is to utilize landmark annotations from different modalities to ensure that the learned distribution matches the underlying actual distribution.

V. Conclusion

In this paper, we propose to introduce the landmark distribution prior into the landmark localization frameworks to promote landmark localization learning. Our proposed NFDP utilizes normalizing flows to characterize the landmark distribution to facilitate regression learning. In the regression-based frameworks, we formulate the learned distribution prior as a penalty function to supervise the regression learning, while in the heatmap-based frameworks, we further incorporate an integral operation to project the heatmaps into coordinates in a differentiable manner. Specifically, our proposed NFDP employs a straightforward backbone and a non-problem-tailored architecture, and achieves high-fidelity results across three X-ray landmark localization datasets with minimal additional computational burden. Moreover, the computational burden is further reduced since the normalizing flow module will be detached from the framework on inferencing. Consequently, the NFDP presents a superior trade-off between prediction accuracy and inference efficiency as compared to existing techniques, establishing it as a highly efficient and effective approach in clinical applications.

References

[1] S. Liu and S. Ostadabbas, "Seeing under the cover: A physics guided learning approach for in-bed pose estimation," in *MICCAI*. Cham: Springer International Publishing, 2019, pp. 236–245.

[2] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. 65, no. 20, p. 20TR01, 2020.

[3] C. Liu, H. Xie, S. Zhang, Z. Mao, J. Sun, and Y. Zhang, "Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3944–3954, 2020.

[4] J. Xu, H. Xie, C. Liu, F. Yang, S. Zhang, X. Chen, and Y. Zhang, "Hip landmark detection with dependency mining in ultrasound image," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3762–3774, 2021.

[5] T.-H. Wu, C. Lian, S. Lee, M. Pastewait, C. Piers, J. Liu, F. Wang, L. Wang, C.-Y. Chiu, W. Wang *et al.*, "Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2022.

[6] X. Nie, J. Feng, J. Zhang, and S. Yan, "Single-stage multi-person pose machines," in *ICCV*, 2019, pp. 6951–6960.

[7] J. M. H. Noothout, B. D. De Vos, J. M. Wolterink, E. M. Postma, P. A. M. Smeets, R. A. P. Takx, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning-based regression and classification for automatic landmark localization in medical images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4011–4022, 2020.

[8] L. Wang, Q. Xu, S. Leung, J. Chung, B. Chen, and S. Li, "Accurate automated cobb angles estimation using multi-view extrapolation net," *Medical Image Analysis*, vol. 58, p. 101542, 2019.

[9] S. K. Zhou and Z. Xu, “Landmark detection and multiorgan segmentation: Representations and supervised approaches,” in *Handbook of MICCAI*. Elsevier, 2020, pp. 205–229.

[10] X. Chen and C. L. et al., “Fast and accurate craniomaxillofacial landmark detection via 3d faster r-cnn,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3867–3878, 2021.

[11] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, “Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet,” in *MICCAI*. Springer, 2017, pp. 127–135.

[12] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi, “Learning a probabilistic model for diffeomorphic registration,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2165–2176, 2019.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” *NIPS*, vol. 27, 2014.

[14] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *ICML*. PMLR, 2015, pp. 1530–1538.

[15] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, “Robust and accurate shape model matching using random forest regression-voting,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1862–1874, 2014.

[16] M. Urschler, T. Ebner, and D. Štern, “Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization,” *Medical image analysis*, vol. 43, pp. 23–36, 2018.

[17] H. Sun, X. Zhen, C. Bailey, P. Rasoulinejad, Y. Yin, and S. Li, “Direct estimation of spinal cobb angles by structured multi-output regression,” in *International conference on information processing in medical imaging*. Springer, 2017, pp. 529–540.

[18] M. Zeng, Z. Yan, S. Liu, Y. Zhou, and L. Qiu, “Cascaded convolutional networks for automatic cephalometric landmark detection,” *Medical Image Analysis*, vol. 68, p. 101904, 2021.

[19] D. Yang, T. Xiong, D. Xu, Q. Huang, D. Liu, S. K. Zhou, Z. Xu, J. Park, M. Chen, T. D. Tran *et al.*, “Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization,” in *International conference on information processing in medical imaging*. Springer, 2017, pp. 633–644.

[20] C. Payer, D. Štern, H. Bischof, and M. Urschler, “Integrating spatial configuration into heatmap regression based cnns for landmark localization,” *Medical image analysis*, vol. 54, pp. 207–219, 2019.

[21] J. Yi, P. Wu, Q. Huang, H. Qu, and D. N. Metaxas, “Vertebra-focused landmark detection for scoliosis assessment,” in *ISBI*. IEEE, 2020, pp. 736–740.

[22] Y. Guo, Y. Li, X. Zhou, and W. He, “A keypoint transformer to discover spine structure for cobb angle estimation,” in *ICME*. IEEE, 2021, pp. 1–6.

[23] K. Gu, L. Yang, M. B. Mi, and A. Yao, “Bias-compensated integral regression for human pose estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[24] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, “Integral human pose regression,” in *ECCV*, 2018, pp. 529–545.

[25] A. Nibali, Z. He, S. Morgan, and L. Prendergast, “Numerical coordinate regression with convolutional neural networks,” *arXiv preprint arXiv:1801.07372*, 2018.

[26] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, “Hand pose estimation via latent 2.5 d heatmap regression,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 118–134.

[27] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, “Deep label distribution learning with label ambiguity,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[28] M. G. Bellemare, W. Dabney, and R. Munos, “A distributional perspective on reinforcement learning,” in *ICML*. PMLR, 2017, pp. 449–458.

[29] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[30] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[31] E. Imani and M. White, “Improving regression performance with distributional losses,” in *ICML*. PMLR, 2018, pp. 2157–2166.

[32] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real NVP,” in *ICLR*, 2017. [Online]. Available: https://openreview.net/forum?id=HkpbnH9lx

[33] I. Kobyzev, S. J. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[34] E. G. Tabak and C. V. Turner, “A family of nonparametric density estimation algorithms,” *Communications on Pure and Applied Mathematics*, vol. 66, no. 2, pp. 145–164, 2013.

[35] A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, “Weakly supervised 3d human pose and shape reconstruction with normalizing flows,” in *ECCV*. Springer, 2020, pp. 465–481.

[36] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, “Ghum & ghuml: Generative 3d human shape and articulated pose models,” in *CVPR*, 2020, pp. 6184–6193.

[37] T. Wehrbein, M. Rudolph, B. Rosenhahn, and B. Wandt, “Probabilistic monocular 3d human pose estimation with normalizing flows,” in *ICCV*, 2021, pp. 11 199–11 208.

[38] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, “Human pose regression with residual log-likelihood estimation,” in *ICCV*, 2021, pp. 11 025–11 034.

[39] E. Aksan, S. Ma, A. Caliskan, S. Pidhorskyi, A. Richard, S.-E. Wei, J. Saragih, and O. Hilliges, “Lip-flow: Learning inference-time priors for codec avatars via normalizing flows in latent space,” in *European Conference on Computer Vision*. Springer, 2022, pp. 92–110.

[40] S. Liang, Z. Zhou, R. Li, J. Zhang, and H. Bao, “Talkingflow: Talking facial landmark generation with multi-scale normalizing flow network,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4628–4632.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[42] T. Y. Lin and P. D. et al., “Feature pyramid networks for object detection,” in *CVPR*, 2017, pp. 2117–2125.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *CVPR*, 2019, pp. 5693–5703.

[44] H. A. David and J. L. Gunnink, “The paired t test under artificial pairing,” *The American Statistician*, vol. 51, no. 1, pp. 9–12, 1997.

[45] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” *arXiv preprint arXiv:1904.07850*, 2019.

[46] M. Lee, M. Chung, and Y.-G. Shin, “Cephalometric landmark detection via global and local encoders and patch-wise attentions,” *Neurocomputing*, vol. 470, pp. 182–189, 2022.

[47] K. Oh and I.-S. e. a. Oh, “Deep anatomical context feature learning for cephalometric landmark detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 806–817, 2021.

[48] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, “Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting,” in *MICCAI*. Springer, 2019, pp. 873–881.

[49] J. McCouat and I. Voiculescu, “Contour-hugging heatmaps for landmark detection,” in *CVPR*, 2022, pp. 20 597–20 605.

[50] R. Chen, Y. Ma, N. Chen, L. Liu, Z. Cui, Y. Lin, and W. Wang, “Structure-aware long short-term memory network for 3d cephalometric landmark detection,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1791–1801, 2022.

[51] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger *et al.*, “A benchmark for comparison of dental radiography analysis algorithms,” *Medical image analysis*, vol. 31, pp. 63–76, 2016.

[52] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[53] R. R. Selvaraju and M. C. et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *ICCV*, 2017, pp. 618–626.

[54] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[55] D. Štern, T. Ebner, and M. Urschler, “From local to global random regression forests: exploring anatomical landmark localization,” in *MICCAI*. Springer, 2016, pp. 221–229.

[56] H. Zhu, Q. Yao, L. Xiao, and S. K. Zhou, “You only learn once: Universal anatomical landmark detection,” in *MICCAI*. Springer, 2021, pp. 85–95.