Sagittal Cervical Spine Landmark Point Detection in X-Ray Using Deep Convolutional Neural Networks

ALI POURRAMEZAN FARD\textsuperscript{1,3}, JOE FERRANTELLI\textsuperscript{2}, ANNE-LISE DUPUIS\textsuperscript{2} and MOHAMMAD H. MAHOOR\textsuperscript{1,3}, (Senior Member, IEEE)

\textsuperscript{1}Ritchie School of Engineering and Computer Science, University of Denver, 2155 E Wesley Ave, Denver, CO 80208, USA (e-mail: {ali.pourramezanfard, mmahoor}@du.edu)
\textsuperscript{2}PostureCo, Inc, Trinity, FL 34655, USA (email: {joe.ferrantelli, annelise.dupuis}@postureco.com)
\textsuperscript{3}DreamFace Technologies LLC, Centennial, CO USA

Corresponding author: Ali Pourramezan Fard (e-mail: ali.pourramezanfard@du.edu)

This work was supported by PostureCo.

ABSTRACT Sagittal cervical spine alignment measured on X-Ray is a key objective measure for clinicians caring for patients with a multitude of presenting symptoms. Despite its applications, there has been no research available in this field yet. This paper presents a framework for automatic detection of the Sagittal cervical spine landmark point. Inspired by UNet, we propose an encoder-decoder Convolutional Neural Network (CNN) called PoseNet. In developing our model, we first review the weaknesses of widely used regression loss functions such as the L1, and L2 losses. To address these issues, we propose a novel loss function specifically designed to improve the accuracy of the localization task under challenging situations (extreme neck pose, low or high brightness and illumination, X-Ray noises, etc.) We validate our model and loss function on a dataset of X-Ray images. The results show that our framework is capable of performing precise sagittal cervical spine landmark point detection even for challenging X-Ray images.

INDEX TERMS Neck Landmark Point Detection, Landmark Point Detection, Intensity Aware Loss, Custom Loss Function, Medical Image Processing, X-Ray Image Processing, Convolutional Neural Network, Computer Vision, Deep Learning

I. INTRODUCTION
The sagittal cervical spine has been both modeled and validated for biomechanical alignment with regard to a normal range [1]–[4]. Using these normal ranges of alignment as guides, clinicians are better equipped to gauge the health of their patient’s spine as related to their presenting symptoms. For example, reversal of the expected normal cervical lordosis (known as cervical kyphosis) is a spinal deformity and is correlated with disability and pain [5], [6]. Furthermore, abnormal cervical sagittal spine alignment is also correlated with symptoms such as headache [7]–[10], migraine [10], [11], as well as related to increased incidence of cervical radiculopathy and myelopathy [12], [13]. Furthermore, a forward head posture which is also measured on X-Ray is significantly related to neck pain as well as distorted nervous system function both sensorimotor and autonomic [14]. Consequently, spinal radiography is vital for the clinical evaluation of sagittal cervical spine alignment to address functional losses, pain, and weakness [15].
Heatmap regression studied by [16]–[23] is one of the most successful and widely used methods among the proposed solutions for landmark point detection tasks. In heatmap-based regression, for each landmark point, we generate a channel (a two-dimensional matrix) using a Gaussian distribution centered at the location of the landmark point. We can model heatmap-based regression as $F(X) = Y$, where $X \in \mathbb{R}^{W \times H \times 3}$ is an input colored-image with the width and height of $W$ and $H$, respectively. Likewise, $Y \in \mathbb{R}^{W_{hm} \times H_{hm} \times k}$ is the predicted heatmap which is a $k$-dimensional tensor where the width, height and number of channels are $W_{hm}$, $H_{hm}$ and $k$ (the number of landmark point), respectively. Finding $F()$, a function that maps the input image to the set of heatmaps is the goal of the landmark localization task. In other words, the regression task is to predict the value of each pixel in each heatmap channel.

L2 loss is widely used in heatmap-based regression [19], [22], [24]–[26]. However, as L2 loss is insensitive to small errors, it is not the most suitable loss function for heatmap-based regression landmark point detection. Thus, we propose a novel loss function, called ILoss, in which for each heatmap channel we define 3 different regions (see Fig. 3): 1- Core-Foreground (CF) region, which contains the pixels with the highest intensity values. This region is considered as the most crucial region since any inaccuracy in the prediction of the intensity values of the pixels in this region, will likely result in inaccurately predicted coordinates. 2- Background (BG) region, which is the region containing the pixels with the smallest intensity values. Regressing the intensity value of the pixels in this region can be taken as a less-important task for the network. In other words, since the coordinates of the facial landmark points heavily rely on the pixels that are located in the CF region, finding the exact intensity value of the pixels in the BG region does not result in a more accurate landmark localization task. 3- Foreground (FG) region, a region which can be considered as a boundary between the BG and CF regions. We penalize the model to predict the intensity value of the pixels in this region more accurately than the BG region but not as accurate as the CF region. Accordingly, the model can learn the Gaussian distribution of heatmap channels.

Although almost all (to the best of our knowledge) of the previous work in landmark point detection considered landmark localization as a regression task, our proposed loss function implicitly considers such task as classification as well. We propose the categorical loss, called CLoss, which utilizes the cross-entropy (CE) loss to guide the model to classify the pixels as $CF$, $BG$, or $FG$. More specifically, for each pixel that belongs to $FG$, and $BG$ regions, CLoss guides the model to learn the region they belong to. Consequently, to guide the model to focus on predicting the intensity value of the pixels located in the $CF$ region more accurately, CLoss stops penalizing the model as the model correctly classifies the pixels in the $BG$ and $FG$ regions. Consequently, CLoss guides the model to learn the proposed regions rather than the intensity value of the pixels. After learning the distribution of the defined regions, the loss function only penalizes the model for accurately predicting the intensity value of the pixels located in the $CF$ region. Such characteristics prevent the network to put a huge effort into predicting the exact intensity value of the pixels located in the $BG$ and $FG$ regions, which can result in a more accurate prediction of the intensity values of the pixels belonging to the $CF$ region.

The contributions of our approach are summarized as follows:

- We propose ILoss, which spans each heatmap channel to the $BG$, $CF$ and $FG$ regions based on the intensity value of the pixels.
- We propose CLoss, which consider the landmark localization task as a classification problem and guides the model to classify the pixels as $CF$, $BG$, or $FG$.
- We propose IC-Loss as a combination of ILoss and CLoss for localizing the sagittal cervical spine landmark point more accurately.
- We propose a new network architecture, called PoseNet, and train it using our proposed IC-Loss.

The remainder of this paper is organized as follows. Sec. II reviews the related work in landmark localization. Sec. III describes our proposed loss function and how it improves the accuracy of heatmap-based regression for the sagittal cervical spine landmark point detection task. Sec. IV-D provides the experimental results and analysis. Then, in Sec.IV-E we provide an extensive ablation study to investigate the effect of each component of our proposed loss function. Finally, Sec. V concludes the paper with some discussions on the proposed method and future research directions.

II. RELATED WORK

To the best of our knowledge, this is the first work on sagittal cervical spine landmark point detection in X-Ray images. However, face alignment as well as body joint tracking tasks, can be taken as similar to sagittal cervical spine landmark point detection. landmark point detection has a wide variety of applications both in medical and non-medical image processing. Accordingly, we first review the application of landmark points on medical images and then go through the non-medical applications including facial alignment and body joint tracking.

A. LANDMARK POINT DETECTION IN MEDICAL CONTEXT

In this section, we review some of the previously proposed landmark localization methods in the medical image processing context.

Lee et al. [27] is among the first who applied CNN to cephalometric landmark detection, by training 38 independent CNN structures to regress 19 landmark points. Likewise, a 2-stage UNet [28] framework proposed by Zhong et al. [29] for automatic detection of anatomical landmarks in cephalometric X-Ray images. A CNN architecture proposed by Qian et al. [30] inspired by Faster R-CNN [31]
to improve the accuracy of cephalometric landmarks localization. Using Bayesian CNN [32], Lee et al. [33] proposed a framework to localize cephalometric landmarks while providing a confidence region of the identified landmarks considering model uncertainty. Dai et al. [34] proposed a new automated cephalometric landmark localization method under the framework of GAN to learn the mapping from features to the distance map of a specific target landmark. Zeng et al. [35] proposed a cascaded three-stage CNN for localization of cephalometric landmark point. A context-guided CNN was proposed by Zhang et al. [36] for joint landmark localization and segmentation of cranio-maxillofacial bone. Zhong et al. [37] proposed a Coarse-to-Fine CNN heatmap-based regression method that is capable of localizing Adolescent idiopathic scoliosis assessment from X-Ray CT images. Ma et al. [38] proposed Loc-Net for fast and memory-efficient 3D landmark detection and applied it to computed tomography angiography scans of the sagittal cervical spine to localize the bifurcation of the left and right carotid arteries. Chen et al. [39] proposed Adaptive Error Correction Net (AEC-Net) to estimate the Cobb angles from spinal X-rays as a high-precision regression. Most of the mentioned work has focused on designing or modifying CNN to make the models better fit the corresponding application or make the prediction accuracy better. We proposed our model architecture, PoseNet, inspired by UNet [28]. Although there might be more accurate CNN for landmark localization task such as Stacked hourglass [40] or HRNet [41] models, we choose to design PoseNet using the idea behind UNet [28] to keep the model efficient in terms of memory and CPU.

On Contrary, the following work mostly proposed a custom methodology or an algorithm to improve the landmark point detection accuracy. To improve the accuracy of the 3D cephalometric landmark point, Huang et al. [42] proposed a sigmoid-based intensity transform that uses the nonlinear optical property of X-Ray films to increase image contrast. HeadLocNet [43] was proposed for localization and classification of inner ear images which can be used to estimate a point-based registration with the atlas image. Urschler et al. [44] proposed a random-forest that uses a combination of image appearance information and geometric landmark configuration for anatomical landmark localization. Likewise, Urschler et al. [44] proposed a heatmap-based regression CNN for anatomical landmark localization, to split the localization task into two simpler sub-problems to reduce the need for large training datasets. Poltaretskyi et al. [45] employed a statistical shape model to predict the premordial anatomy of the proximal humerus. A two-stage algorithm composed of a rigid image-to-image and a deformable surface-to-image registration is proposed by Brehler et al. [46] for detecting a set of landmarks in the knee joint. Negrillo et al. [47] proposed a geometrically-based algorithm to detect the landmarks of the humerus for further analysis of a reduction of supracondylar fractures. Likewise to the mentioned research, we choose to design a new loss function that can cope well with the challenges of landmark localization task. In addition, in Sec. III we discuss the weakness of the widely used L1, and L2 loss and try to design the IC-Loss to perform the sagittal cervical spine localization task more accurately.

B. LANDMARK POINT DETECTION

In general, there are three main methods for landmark point detection. We have classified and reviewed the previously proposed methods in the following.

Classical models (aka template-based methods) are among the first methods that are designed for landmark point detection specifically face alignment. Active Shape Model (ASM) [48] and Active Appearance Model (AAM) [49], [50] which utilize Principal Components Analysis (PCA) to simplify the problem and learn parametric features of faces to model facial landmarks variations in an iterative manner. Further, Martins et al. [50] proposed a 2.5D AAM that combines a 3D metric Point Distribution Model (PDM) with a 2D appearance model to match a 3D deformable face model to 2D images. In addition, Cristinacce and Cootes [51] introduced the Constrained Local Model (CLM) which models the face shapes with Procrustes analysis and principal component analysis. Although CLM models and their various extensions including [52]–[55], are among the most promising methods for face alignment, they are sensitive to occlusion as well as illumination when detecting landmarks in unseen datasets. Robust Cascade Pose Regression (RCPR) [56] was introduced to detect occlusions explicitly and use robust shape-indexed features. Another computationally lightweight method was Local Binary Features (LBF) [57].

Coordinate-based regression models predict the landmark coordinates vector from the input image directly. Mnemonic Descent Method (MDM) [58] has utilized a recurrent convolutional network to detect facial landmarks. Feng et al. [59] introduced Wingloss, a new loss function that is capable of overcoming the widely used L2 loss in conjunction with a strong data augmentation method as well as a pose-based data balancing (PDB). To ease the parts variations and regresses the coordinates of different parts, Two-Stage Re-initialization Deep Regression MODEL (TSR) [60] splits the face into several parts. Zhang et al. [61] proposed Exemplar-based Cascaded Stacked Auto-Encoder Network (ECSAN) for face alignment, which is utilized to handle partial occlusion in the image. To cope with self-occlusions and large face rotations, Valle et al. [62] proposed a face alignment algorithm based on a coarse-to-fine cascade of ensembles of regression trees, which is initialized by robustly fitting a 3D face model to the probability maps produced by a pre-trained convolutional neural network (CNN). Fard et al. [63] proposed a lightweight multi-task network for jointly detecting facial landmark points as well as the estimation of face pose. In another work, a student-teacher network was proposed by Fard et al. [64] for extracting facial landmark points using more accurately using lightweight CNN models. More recently, ACR Loss [65] proposed a loss function to estimate the level of difficulty in predicting each landmark point for the network and hence improve the accuracy of the...
Heatmap-based regression models are another widely used method for landmark point detection tasks. First, the likelihood heatmaps for each landmark point are created, and then the network is trained to generate the heatmaps for each input image. A two-part network proposed by Yang et al. [24], including a supervised transformation to normalize faces and a stacked hourglass network [40], is designed to predict heatmaps. LAB [16] proposed by Wu, expressed that the facial boundary line contains valuable information. Hence, they utilized boundary lines as the geometric structure of a face to help facial landmark detection. Furthermore, for a better initialization to Ensemble of Regression Trees (ERT) regressor, Valle et al. [66] proposed a simple CNN to generate heatmaps of landmark locations. As well as that, JMFA [67] leveraged a stacked hourglass network for multi-view face alignment, and it achieved state-of-the-art accuracy and demonstrated more accuracy than the best three entries of the Menpo Challenge [68]. Moreover, HRNet being introduced by Sun et al. [22] is a high-resolution network that is applicable in many Computer Vision tasks such as facial landmark detection and achieves a reasonable accuracy. Besides that, to deal with shape variations in facial landmark detection, Iranmanesh et al. [20] proposed an approach that provides a robust algorithm that aggregates a set of manipulated images to capture robust landmark representation. Gaussian heatmap vectors proposed by Xiong et al. [19] to be used instead of the widely used heatmap for facial landmark point detection. More recently, [69] proposed a two-paired cascade subnetwork to generate heatmaps and accordingly the coordinates of the facial landmark point to deal with the more challenging faces.

Since the accuracy of the heatmap-based landmark localization is more than the Coordinate-based method, we decided to follow the former approach.

Custom Loss function is a remarkable approach for improving the performance of both regression-based and classification-based [70] models. Similarly, Custom loss functions play an important role in improving the accuracy of landmark localization. However, there is only a few research that has proposed and studied the effect of using a task-related loss function in this context. Fard et al. [63] have proposed ASMNet, a lightweight CNN as well as ASMLoss which is designed to guide the network toward learning a smoother version of the facial landmark point. KD-Loss proposed by Fard et al. [64] is a loss function designed to use the knowledge gained by two teacher networks to improve the accuracy of a student network for face alignment task. GoDP [71] proposed a distance-aware softmax loss to assign a large penalty on incorrectly classified positive samples. Then, they gradually reduce the penalty on the misclassified negative samples as the distance from nearby positive samples decreases. The Wing loss [59] is a modified log loss for coordinate-based face alignment which is designed to amplify the influence of small errors. However, since Wing loss [59] is very sensitive to small errors in the background pixels it is not applicable to heatmap-based regression landmark point localization. To cope with this issue, Wang et al. [21] proposed Adaptive Wing Loss, a loss function that can adapt its curvature to different ground truth pixel values. Consequently, it can be sensitive to small errors on foreground pixels while it is tolerant to small errors on background pixels.

Contrary to the previously proposed loss functions, our proposed IC-Loss is designed to firstly spans each heatmap channel to the BG, CF and FG considering the intensity value of each pixel and penalizing the model accordingly. More clearly, IC-Loss penalizes the model for accurately predicting the intensity value of the pixels located in the CF region, while it can tolerate more errors when it comes to the pixels located in either of the BG or the FG regions.

III. METHODOLOGY
In this section, we first introduce our proposed network architecture, PoseNet. Next, we illustrate how to generate the heatmap attention map. Then, we explain ILoss, CLoss, and the proposed IC-Loss.

A. POSENET ARCHITECTURE
Inspired by UNet [28], we propose an Encoder-Decoder network called PoseNet for the sagittal cervical spine landmark point localization task. As Fig. 1 shows, PoseNet contains 4 main modules including Head, Down-Sampling, Keep, and Up-Sampling.

The Head module contains 3 sets of 2-dimensional convolution (Conv2D) layers followed by a ReLU and a batch-normalization layer (see Fig. 2). Each Conv2D layer has a $3 \times 3$ kernel size, 64 filters and stride as 2. We design the Head module to extract the feature from the input image and downscale it to $\frac{1}{2}$ of the size of the original input image.

As Fig. 2 shows, each Down-Sampling module has 2 sets of $3 \times 3$ Conv2D layers having the same number of filters and stride equals 1, followed by a ReLU and a batch-normalization layer. We define 2 outputs for the Down-Sampling module called Skip and Out. Skip is a skip-connection used to pass the information and features to the corresponding Up-Sampling module. Besides, Down is followed by a MaxPooling layer (with pool_size=2), which is used to downscale the input by the scale of 2. In the Down-Sampling modules, we multiply the number of the filter of the Conv2D layers by a factor of 2 as we downscale the image.

The first Down-Sampling module has 64 filters, the second, third, and fourth have 128, 256, and 512 filters, respectively.

In the Keep module, we only use a $3 \times 3$ Conv2D layer with 1024 filters and stride as 1, followed by a ReLU and a batch-normalization layer. Then, we have a $2 \times 2$ Deconvolution (DeConv2D) layer having 512 filters with stride equals to
2 followed by a ReLU and a batch-normalization layer (see Fig. 2). Since the input is in its smallest scale at this module, we use a Conv2D layer with 1024 filters to extract more information from the image.

The Up-Sampling modules are designed to firstly extract information from the input, and then scale up the inputs to further generate the heatmaps. Thus, as Fig. 2 represents, in each Up-Sampling module we use 2 sets of $3 \times 3$ Conv2D with stride=1 followed by a ReLU and a batch-normalization layer. Then, a $2 \times 2$ DeConv2D with stride=2 is used to upscale the input by the factor of 2. The number of filters we use for both Conv2D layers and the DeConv2D are the same. While the first Up-Sampling module has 512 filters, we decrease the filters by a factor of 2, so the second, third, and fourth have 256, 128, and 64 filters respectively.

**B. HEATMAP ATTENTION MAP**

We define a heatmap attention map based on the intensity value of the ground-truth heatmap channels. Heatmap attention maps define the importance of each pixel in the ground-truth heatmap. For an input image $Img$, we define $H_{W_{hm} \times H_{hm} \times N}$ and $H'_{W_{hm} \times H_{hm} \times N}$ as the corresponding ground-truth and the predicted heatmaps, respectively. The $W_{hm}$, $H_{hm}$, and $N$ are the width, the height and the number of the channels of the heatmaps respectively. For each heatmap channel, we define the $BG$, $FG$ and $CF$ regions based on the intensity value of the corresponding pixels.

As Fig 3 shows, a pixel $p_{ij}$ belongs to $BG$ if its intensity value $I_{ij}$ is in $[0, \theta_{BG}]$. Likewise, $p_{ij}$ belongs to $FG$ if $I_{ij}$ is in $(\theta_{BG}, \theta_{FG}]$, and to $CF$ in case $I_{ij}$ is in $(\theta_{FG}, 1]$. Accordingly, we define $w_{ij}$, the attention for $p_{ij}$ according to Eq. 1:

$$w_{ij} = \begin{cases} 
1 & \forall p_{ij} \in BG \\
\omega_{FG} & \forall p_{ij} \in FG \\
\omega_{CF} & \forall p_{ij} \in CF
\end{cases} \quad (1)$$

where $\omega_{FG}$ and $\omega_{CF}$ are the hyper parameters that define the magnitude of the attention map for the $BG$, $FG$ and $CF$ regions, respectively.

We define each heatmap channel using 2 dimensional Gaussian function using Eq. 2:

$$H = -exp\left(\frac{(x - p_x)^2}{2\sigma^2} + \frac{(y - p_y)^2}{2\sigma^2}\right) \quad (2)$$
where $P = (p_x, p_y)$ is the location of the landmark point and $\sigma$ is the standard deviation of the distribution. Following the typical practices, we define $\sigma = 1.5$. Since predicting the intensity value of the pixels belonging to $CF$ region accurately is very crucial for generating the coordinates of the landmark point, we define this region to be smaller compared to the other regions. Thus, we empirically set $\theta_{FG} = 0.9$, which means the pixels with intensity values between 0.9 and 1 belong to this region. Likewise, we set $\theta_{BG} = 0.4$.

Considering the ratio of the pixels exists in each of the $CF$, $FG$, regions compared to the pixels in the $BG$ region in each heatmap channel $H$, we set the values of $\omega_{CF}$, and $\omega_{FG}$, respectively, using Eq. 3:

$$
\omega_{FG} = \frac{BG^*}{FG^*} \quad \text{and} \quad \omega_{CF} = \frac{BG^*}{CF^*}
$$

where $X^*$ indicates the number of pixels that exist in the $X$ region. To clarify, since the number of pixels that exist in the $BG$ region is much greater than the pixels in $CF$, and $FG$, the model should take a huge effort on predicting the intensity value of the pixels in the former region, while predicting those values accurately does not pay a crucial role in the final accuracy of the landmark point detection task. Thus, we introduce the heatmap attention map which guides the network to put more focus on predicting the accurate intensity value of the pixels which are important for the landmark point detection task.

C. ILOSS

As Fig.3 shows (also discussed in Sec.I), predicting the accurate intensity value of the pixels which belong to the $CF$ region can heavily affect the final accuracy of the landmark localization task. Accordingly, we propose an intensity aware loss function called ILoss, which is designed to penalize the localization task. Accordingly, we propose an intensity aware loss function called ILoss, which is designed to penalize the model with the number of pixels that exist in each region in a heatmap channel $H$.

Thus, we introduce the heatmap attention map which guides the network to put more focus on predicting the accurate intensity value of the pixels which are important for the landmark point detection task.

Moreover, we define the Intensity Map function, $IM()$, using Eq. 5:

$$
IM_{m,i,j,k}(l_b, h_b) = \begin{cases} 
1 & \text{if } I_{m,i,j,k} \in [l_b, h_b] \\
0 & \text{otherwise}
\end{cases}
$$

where $l_b$ and $h_b$ are the minimum and the maximum intensity values. In other words, if the intensity value of the ground truth pixel located at $m, i, j, k$ position is between $l_b$ and $h_b$, the output of the function is 1, and otherwise, its output is 0. We use $IM()$ to define ILoss for each of the proposed regions.

1) ILoss for the BG Region

For the pixels belonging to the $BG$ region, there is no need for a highly accurate prediction of the intensity values. Thus, we define a threshold and consider the prediction as accurate enough if the prediction error is smaller than the threshold. The intensity value of each pixel in a heatmap channel $H$ can vary between 0 and 1, so we consider the prediction as accurate enough if $\Delta \leq \phi_{BG}$. The influence of the loss function should be high if the prediction error is more than the defined threshold, $\phi_{BG}$, while it should drop dramatically as soon as the prediction error becomes smaller than the threshold. Having said that, ILoss can be a quadratic ($y = ax^2$) function in the $BG$ region. Consequently, we define $Loss_{BG}$ as a quadratic function using Equations 6 and 7:

$$
Loss_{BG} = \frac{M}{N} \sum_{m=1}^{M} \sum_{j=1}^{H_{hm}} \sum_{k=1}^{N} IM_{m,i,j,k}(\theta_{BG}) LBG_{m,i,j,k}
$$

where $C_1$ is a constant used to smoothly connect 2 pieces of $Loss_{BG}$ together, and $\alpha$ is a hyperparameter that can set the magnitude and the influence of the loss function. We empirically define $\alpha = 0.5$ (and accordingly $C_1 = \frac{1}{2}$) which make the influence of the loss function very low while the prediction is considered as good enough. Moreover, we define $\phi_{BG} = 0.5$ which means for each pixel in the $BG$ region, the prediction is accurate if the error is less than 0.5.

As Fig. 4-A shows, we define $Loss_{BG}$ such that the influence of the loss decreases as the magnitude of the loss decreases. This characteristic of $Loss_{BG}$ guides the model to put more focus on the prediction of the intensity value of the pixels belonging to the $FG$, and $CF$ regions as soon as the model predict the intensity value of the pixels in $BG$ accurately enough.

2) ILoss for the FG Region

Compared to the $BG$ region, it is much more important that the model to predict the intensity value of the pixels belonging to the $FG$ region since learning the Gaussian distribution of a heatmap channel can further help the model to predict
the intensity value of the pixels in the CF region. The same as the proposed loss function for the BG region, we define the error prediction threshold as $\Delta \leq \phi_{FG}$. The influence of the loss function should be high if the prediction error is greater than the threshold, so we define the loss as a quadratic function for this condition. Then, when the prediction error is smaller than the threshold, we define the loss function as a linear function. Contrary to Loss\(_{BG}\), the gradient of the linear part of Loss\(_{FG}\) is a constant value, and accordingly it penalizes the model independently from the magnitude of the prediction error. This characteristic of Loss\(_{FG}\) guides the model to predict the intensity value of the pixels belonging to CF more accurately. we define Loss\(_{FG}\) as a pieces-wise function using Equations 8 and 9:

$$\text{LFG}_{m,i,j,k} = \begin{cases} \Delta_{m,i,j,k}^2 & \text{if: } \Delta_{m,i,j,k} > \phi_{FG} \\ \Delta_{m,i,j,k} + C_2 & \text{otherwise} \end{cases}$$

$$\text{Loss}_{FG} = \sum_{m=1}^{M} \sum_{i=1}^{W_h} \sum_{j=1}^{H_h} \sum_{k=1}^{N} \omega_{FG} \text{LFG}_{m,i,j,k} / M N W_h H_h$$  \hspace{1cm} (8)

where $C_2$ is a constant used to smoothly connects 2 pieces of Loss\(_{FG}\) together. As Fig. 4-B shows, the influence of the quadratic part of Loss\(_{FG}\) is related to the magnitude of the loss function which guides the model to quickly learn the distribution of the heatmap values and predict the intensity value of the pixels such that they can be considered as accurate enough. Then, the linear part of Loss\(_{FG}\), having a constant influence value, penalizes the model to improve the prediction accuracy independently of the magnitude of the loss function. Following the value of $\phi_{BG}$, we define $\phi_{FG}$ as 0.5.

3) Loss for the CF Region

The accuracy of the landmark point detection task has a significant reliance on how accurately the model predicts the intensity value of the pixels in the CF region. Similarly to Loss\(_{BG}\), Loss\(_{FG}\), a quadratic loss function ($y = x^2$) is a suitable option for huge errors since the influence of the loss depends on the magnitude of the loss, and accordingly a quadratic loss function can penalize the model dramatically for huge errors. For the small prediction errors where $\Delta \leq \phi_{1_{CF}}$, the influence of the quadratic loss function reduces dramatically, and hence the model gets penalized much less. We introduce a logarithmic loss function ($y = \ln(1 + x)$) where the influence of the loss function increases as the magnitude of the loss decreases. In other words, the logarithmic loss can force the model to put a huge effort into accurately predicting the intensity value of the pixels in the CF region.

Although the logarithmic loss function can guide the model towards learning the very accurate intensity values of the pixels in the CF region, putting too much effort into this region has a negative consequence on the accuracy of the other regions. Thus, we define a threshold and call it $\phi_{2_{CF}}$ and consider the prediction as accurate enough if the prediction error is smaller than this threshold. We use a linear loss function ($y = x$) if the prediction error is smaller than $\phi_{2_{CF}}$. Since the gradient of the linear loss is a constant number ($y' = 1$), the influence of the loss is independent to the magnitude of the loss function and thus the loss function keeps penalizing the model to improve the accuracy. We define Loss\(_{CF}\) as a piece-wise function using Equations 10 and 11:

$$\text{LCF}_{m,i,j,k} = \begin{cases} \Delta_{m,i,j,k}^2 & \text{if: } \Delta_{m,i,j,k} > \phi_{1_{CF}} \\ \beta \ln(1 + \Delta_{m,i,j,k}) + C_3 & \text{if: } \phi_{2_{CF}} < \Delta_{m,i,j,k} \leq \phi_{1_{CF}} \\ \Delta_{m,i,j,k} + C_3 & \text{otherwise} \end{cases}$$

$$\text{Loss}_{CF} = \sum_{m=1}^{M} \sum_{i=1}^{W_h} \sum_{j=1}^{H_h} \sum_{k=1}^{N} \omega_{CF} \text{LCF}_{m,i,j,k} / M N W_h H_h$$  \hspace{1cm} (10)

$$\text{Loss}_{CF} = \sum_{m=1}^{M} \sum_{i=1}^{W_h} \sum_{j=1}^{H_h} \sum_{k=1}^{N} \omega_{CF} \text{LCF}_{m,i,j,k} / M N W_h H_h$$  \hspace{1cm} (11)
where \( C_3 = \phi_{2,CF} - \beta \ln(1 + \phi_{2,CF}) \) and \( C_4 = \beta \ln(1 + \phi_{1,CF}) - \beta \ln(1 + \phi_{2,CF}) - \phi_{3,CF} + \phi_{2,CF} \) are constants used to smoothly connect pieces of Loss function together. We introduce the hyper parameter \( \beta \) to adapt the curvature of the logarithmic piece. Fig. 4-C shows, different values of \( \beta \), and \( \phi_{2,CF} \) change the influence and the magnitude of the loss function. The accurate prediction of the intensity value of the pixels belonging to the \( CF \) region affects the final landmark point detection accuracy dramatically. Thus, we conduct experiments on different values of \( \beta \), and \( \phi_{2,CF} \) and accordingly define \( \beta = 4 \) and \( \phi_{2,CF} = 0.05 \) (see Sec. IV-E).

**D. CLOSS**

We classify any arbitrary pixel \( P_{m,i,j,k} \) located at \( i \in H_{hm}, j \in W_{hm} \) and \( k \in N \) to be in one of the \( BG \), \( FG \), or \( CF \) regions. Hence, we can consider the sagittal cervical spine landmark localization task as a classification problem too, where we penalize the model to learn the region each pixels belongs to. Since predicting the very accurate intensity values of the pixels belonging to the \( BG \) region, the \( FG \) region do not affect the final landmark point detection accuracy, we can stop penalizing the model as soon as the model classify each pixels correctly. We define 2 classes called \( C_{BG} = 0 \) and \( C_{FG} = 1 \) and accordingly label any pixel located in \( BG \) and \( FG \) as \( C_{BG} \) and \( C_{FG} \) respectively. We define \( \mathcal{L} \) operator as:

\[
\mathcal{L}(V) = \begin{cases} 
0 & \text{if } V \in [0, \theta_{BG}] \\
1 & \text{if } V \in [\theta_{BG}, \theta_{FG}] 
\end{cases}
\]

where \( V \) is the intensity value of the input pixel. We introduce \( \mathcal{L} \) operator to define the class label of each pixel located in the \( BG \) and \( FG \) regions. Then, we use CE loss function and define CLoss for each pixel using Equations 13, and 14:

\[
CE_{m,i,j,k} = -\mathcal{L}(I_{m,i,j,k}) \ln(I_{m,i,j,k}) + \mathcal{L}(I'_{m,i,j,k}) \ln(I'_{m,i,j,k})
\]

\[
\Omega_{m,i,j,k} = \begin{cases} 
0 & \text{if } CE_{m,i,j,k} = 0 \\
1 & \text{if } \text{otherwise}
\end{cases}
\]

To explain, for any pixel \( P_{m,i,j,k} \) with the ground truth and predicted intensity values \( I_{m,i,j,k} \) and \( I'_{m,i,j,k} \) respectively, we first use \( CE_{m,i,j,k} \) to figure out if the model has classified the pixel correctly or not. Then, we introduce \( \Omega_{m,i,j,k} \) as a binary weight, which is 0 if the classification is correct and is 1 otherwise.

Then, we define categorical intensity aware loss function for pixels belonging to the \( BG \) and \( FG \) regions using Equations 15, and 16:

\[
\text{CLoss}_{BG} = \sum_{m=1}^{M} \sum_{i=1}^{W_m} \sum_{j=1}^{H_m} \sum_{k=1}^{N} I_{m,i,j,k} \ln(O_{m,i,j,k} \beta \ln(1 + \phi_{2,CF}) + L_{BG}(m,i,j,k) \Omega_{m,i,j,k})
\]

\[
\text{CLoss}_{FG} = \sum_{m=1}^{M} \sum_{i=1}^{W_m} \sum_{j=1}^{H_m} \sum_{k=1}^{N} I_{m,i,j,k} \ln(O_{m,i,j,k} \beta \ln(1 + \phi_{2,CF}) + L_{FG}(m,i,j,k) \Omega_{m,i,j,k})
\]

The proposed \( \text{CLoss}_{BG} \) and \( \text{CLoss}_{FG} \) are designed to stop penalizing the model to predict the intensity of a pixel as soon as the model correctly classifies it. Consequently, the model puts more effort into predicting the intensity value of the pixels located at \( CF \) more accurately. We show in Sec.IV-E that using the proposed categorical \( \text{CLoss}_{BG} \) and \( \text{CLoss}_{FG} \) compared to their base models \( \text{Loss}_{BG} \) and \( \text{Loss}_{FG} \) improves the accuracy of the final landmark point detection task.

**E. PROPOSED LOSS FUNCTION**

As Eq.17 shows we define our proposed loss function as the sum of the loss functions proposed for the \( BG \), \( FG \), and \( CF \) regions and call it Intensity-aware Categorical Loss, IC-Loss.

\[
\text{IC-Loss} = \text{CLoss}_{BG} + \text{CLoss}_{FG} + \text{Loss}_{CF}
\]
IV. EXPERIMENTAL RESULTS

In this section, we first provide a detailed description of our dataset. Afterward, we explain the implementation detail of our proposed method. Then, we illustrate the evaluation metrics which are used to assess the performance of our proposed method for the sagittal cervical spine landmark detection task. Finally, we provide our results.

A. DATASET

Our dataset contains 24,419 sagittal cervical spine X-Ray images that are labeled by expert humans. As Fig. 5 depicts, each sagittal cervical spine image in the dataset can be categorized to 3 different types with respect to its pose including normal (LC), extension (LCE), and flexion (LCF).

As Table 1 shows, the number of LC images (19,599) is by far greater than the LCE (2,495) and LCF (2,325) images. We split our dataset into 2 independent sets, a training set (which is used for the training purpose), and a test set (which is used to evaluate our proposed method). We create our test set by randomly selecting 10% of the total number of LC, 5% of LCE, and 5% of LCF images. Then, we use the rest of the images as the training set. Table 1 shows the detail of our generated training and test set. Moreover, in Table 2, Table 3, and Table 4 we provide a detailed information about the year range of X-Rays, the gender and the age of the subjects respectively.

To obtain the dataset of X-Rays, 5 chiropractic clinics were included that are met the basic criteria of being certified in a chiropractic technique called Chiropractic BioPhysics. This was done to make sure the digitization of anatomical landmarks was consistent for the dataset. The offices then signed and agreed to data sharing and business associate agreements with the company PostureCo, Inc. to be compliant with HIPAA guidelines. All X-Ray data was anonymized and selected from the years 2008-2021. From the dataset, the patient population broke down to males making up 3.72% of lateral cervical extension (LCE) X-Rays, 3.56% of lateral cervical flexion X-Rays (LCF), and 31.23% of neutral lateral cervical radiographs (LC). Females contributed 6.49% LCE, 5.96% LCF, and 49.03% LC. Collectively 80.26% of the X-Rays were LC, with 10.22% LCE and 9.52% LCF respectively. Age demographics revealed that age 5-20 comprised 11.16% of the X-Rays included with the 20-40yrs age bracket making up 36.66% and >40 yrs comprising the remaining 52.18%.

1) Preprocessing

Since the sagittal cervical spine X-Ray images in our dataset have been captured with different devices and technologies, there exists 2 different types of X-Rays in the dataset: white-foreground images and white-background (see Fig. 6), while we consider the bones as the foreground and anything else as the background. We introduce the white-foreground X-Rays as the images in which the intensity value of the pixels in the foreground segments are greater than the other segments, while in the white-background X-Rays, the intensity value of the pixels in the bone segments is less than the other segments. As the discrepancy in types of X-Ray images can impact the accuracy of the model negatively, we propose a preprocessing algorithm to deal with this issue and eventually improve the accuracy of the model.

As Fig. 7 shows, we first convert each input X-Ray image to a gray-scale image by calculating the mean of the RGB colors of each pixel. Then, we apply a threshold to the gray-scale image to separate the foreground from the background. Finally, we apply a morphological opening operation to remove any remaining background noise.

| TABLE 1. Total number of X-Ray images in the training and test set. |
|-----------------|---------|---------|---------|---------|
| Dataset         | LC      | LCE     | LCF     | Total   |
| 19,599          | 2,495   | 2,325   | 24,419  |
| Training Set    | 17,639  | 2,371   | 2,209   | 22,219  |
| Test Set        | 1,960   | 124     | 116     | 2,200   |

| TABLE 2. X-Ray date range per view type. All numbers are in % format. |
|----------------|---------|---------|---------|---------|
| Range          | LC      | LCE     | LCF     | Total   |
| 2008-2010      | 4.46    | 0.97    | 0.97    | 6.40    |
| 2010-2014      | 18.65   | 2.45    | 2.41    | 23.51   |
| 2010-2018      | 29.91   | 3.28    | 2.90    | 36.08   |
| 2018-2021      | 27.24   | 3.52    | 3.24    | 34      |

| TABLE 3. X-Ray gender per view type. All numbers are in % format. |
|----------------|---------|---------|---------|---------|
| Gender         | LC      | LCE     | LCF     | Total   |
| Male           | 31.23   | 3.27    | 3.56    | 38.52   |
| Female         | 49.03   | 6.49    | 5.96    | 61.48   |

| TABLE 4. X-Ray age per view type. All numbers are in % format. |
|----------------|---------|---------|---------|---------|
| Age            | LC      | LCE     | LCF     | Total   |
| <20            | 8.69    | 1.28    | 1.20    | 11.16   |
| 20-40          | 28.83   | 4.02    | 3.80    | 36.66   |
| >40            | 42.74   | 4.92    | 4.52    | 52.18   |
Maximum coordination of the landmarks points in both X and Y axes. Then, we resize each X-Ray image to have width and height of 128, and 256 respectively. We augmented the X-Ray images in terms of rotation (from -45 to 45 degrees), and contrast, brightness and color modification to add robustness to the model.

Furthermore, since generating heatmaps having the same width and height as the input cropped images is memory and processor-consuming, we created our heatmaps to be four times smaller than the cropped input images, having width and height equal to 32 and 64 respectively. We used the Adam optimizer [72] for training the networks while choosing the learning rate $10^{-2}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\text{decay} = 10^{-5}$. We then trained the networks for about 150 epochs with a batch size of 60. We implemented our networks using the TensorFlow library and ran them on an NVidia 1080Ti GPU.

C. EVALUATION METRICS

We follow landmark localization tasks and employ normalized mean error (NME), in Eq.18, to measure our model’s accuracy.

$$NME = \frac{1}{m \cdot n \cdot d} \sum_{i=1}^{m} \sum_{j=1}^{n} \sqrt{(P_{x_{ij}} - G_{x_{ij}})^2 + (P_{y_{ij}} - G_{y_{ij}})^2}$$

(18)

where $m$ is the number of samples in the test set, and $n$ is number of the landmark point, $P_{x_{ij}}, P_{y_{ij}}$ are the predicted points ($j^{th}$ landmark point of the $i^{th}$ sample) while $G_{x_{ij}}, G_{y_{ij}}$ are the corresponding Ground-truths. We also define $d$ as the normalizing factor for presenting the results in a unified manner (the accuracy of the model will not be affected by the size of the image, the cropping margin, etc.). Inspired by previously proposed work in facial alignment where the normalizing factor is selected as the distance between the pupils of the human eyes (or the distance between the outer corner of the eyes), we choose $d$ as the distance between the landmark point number 0 and 4 (see Fig 11). In addition, we calculate NME for each vertebra to better analyze the accuracy of the model. Furthermore, we calculate failure rate (FR), defined as the proportion of failed detected sagittal cervical spines, for a maximum error of 0.08 and 0.1. Cumulative Errors Distribution (CED) curve and the area-under-the-curve (AUC) [73] are reported as well.

D. RESULTS ANALYSIS

In this section, we first evaluate the accuracy of our proposed network architecture, PoseNet. Then, we evaluate the accuracy of our proposed loss function. Moreover, we study the performance of our proposed PoseNet.

1) Evaluation of PoseNet

Since this is the first work in the field, we first compare the accuracy of our PoseNet trained with the widely used L2 loss with 2 very popular baseline models, MobileNetV2 [74] called $mnv2$, and ResNet50 [75], called $res50$. Moreover,
since PoseNet is a heatmap-based regression, for better comparison we create 2 encoder-decoder models using MobileNetV2 [74], and ResNet50 [75] as the encoders respectively, and introduce 3 sets of DeConv2D layers, with filters=256, kernel=3, and stride=2, followed by a ReLU and a batch-normalization layer, as the decoder. We call the former model mnn2-hm, and the latter model res50-hm.

As Table 5 shows, the accuracy of the sagittal cervical spine landmark point detection using heatmap-based regression models is much better than the coordinate-based regression models. Moreover, while the NME for the mnn2-hm is 5.60%, 6.42%, and 10.56% for the LC, LCE, and LCF subsets respectively, these values are reduced to 4.79% (about 0.81% reduction), 5.20% (about 1.22% reduction), and 7.68% (about 2.88% reduction) respectively using res50-hm as the model. Using the PoseNet, the NME reduces to 4.75%, 5.21%, and 7.48% indicating about 0.85%, 1.21%, 3.08% decreases compared to resn50-hm.

As we discussed in Sec. III, the influence of L2 loss is very sensitive to large errors while it is very insensitive to small errors. Conversely, the influence of L1 loss is a constant value. Our provided experiments in Table 7 also shows empirically that L1 loss performs better than L2 loss. Since NME and FR are sensitive to outliers, we depict the CED curve of the sagittal cervical spine landmark point detection using PoseNet as the model and train the model using L1, L2, and IC-Loss. As Fig 8 shows, the performance of the landmark localization is much better when we train the model using IC-Loss compared to L1, and L2 loss.

E. ABLATION STUDY

In this section, we first study the effect of the hyperparameters defined in Equations 10 and 11. Then, we conduct experiments to measure how positively the proposed CLoss can improve the sagittal cervical spine landmark point detection task.

1) Studying the Effects of $\beta$ and $\phi$

As mentioned in Sec. III, different values of the hyperparameters $\beta$ and $\phi_{C-F}$ in Equations 10 and 11 can affect the accuracy of sagittal cervical spine landmark point detection task. Accordingly, we conduct 2 independent experiments to study the effect of each hyperparameter. Needless to say, finding the best combination of these 2 hyperparameters is almost impossible.

In the first experiment, we assign 5 different values to $\beta$ and calculate the NME of the landmark point detection task. As Fig 9 shows choosing $\beta = 1$ results NME to be 4.82%, 5.37% and 7.62% for LC, LCE, and LCF subsets respectively. These values are almost the same choosing $\beta = 2$. Then, choosing $\beta = 3$ reduces the NME more and we achieve 4.66%, 5.19%, and 7.19% for LC, LCE, and LCF subsets respectively. We followed the trend and increase $\beta$ to 4, and achieved the least NME. However, choosing $\beta = 5$ increases the NME again, and accordingly, we choose the hyperparameter $\beta = 4$ in our experiments.

In the second experiment where we define $\beta = 4$, we assign 3 different values to $\phi_{C-F}$ and calculate the NME of the landmark point detection task. As Fig 10 shows, the value of $\phi_{C-F}$ affects the accuracy of the model very slightly and the lowest NME is achieved by choosing $\phi_{C-F} = 0.05$.

2) Studying the Effects of the Categorical CLoss

To better understand the effect of our proposed Categorical Loss, we conduct 2 different experiments. In the first experiment, called IC-LOSS$_{No-CAT}$, we train the model without using the Categorical Loss. In the second experiment, called
IC-Loss\textsubscript{All-CAT} we apply the CLoss to all the 3 regions, \textit{BG}, \textit{FG} and \textit{CF}.

As Table 8 shows, removing the proposed CLoss from IC-Loss increases the value of NME about 0.37\% (from 4.38\% to 4.75\%), 0.55\% (from 4.76\% to 5.31\%), and 0.90\% (from 6.50\% to 7.40\%) for \textit{LC}, \textit{LCE}, and \textit{LCF} subsets respectively. In addition, using IC-Loss\textsubscript{All-CAT}, where we use the CLoss for all the regions devastate the accuracy of the model and NME increases dramatically by about 3.8\% (from 4.38\% to 8.18\%), 5.2\% (from 4.76\% to 9.96\%), and 7.46\% (from 6.50\% to 13.96\%) for \textit{LC}, \textit{LCE}, and \textit{LCF} subsets respectively. As discussed in Sec III, using CLoss for the pixels belonging to the \textit{CF} region stops penalizing the model as soon as the model classifies the pixels correctly. However, the accurate values for the pixels in \textit{CF} are vital for the ultimate goal of the model which is the localization of the sagittal cervical spine landmark point.

V. CONCLUSION

In this paper, we proposed PoseNet, a CNN inspired by UNet [28] for detecting sagittal cervical spine landmark points in X-Ray images. We discussed that the widely used L1 and L2 loss functions are not the best options for heatmap-based regression landmark point detection. Consequently, we proposed IC-Loss, an intensity aware categorical-based loss function. IC-Loss set the magnitude of the loss and its influence according to our defined \textit{BG}, \textit{FG} and \textit{CF} regions. In addition, our novel IC-Loss simplifies the regression task for the pixels located in \textit{BG}, \textit{FG} to a classification task, which improves the accuracy of the sagittal cervical spine landmark point detection task. Our proposed experiments show that PoseNet trained using IC-Loss can perform a very accurate sagittal cervical spine landmark point localization task. Moreover, both PoseNet and IC-Los are capable of being used in similar landmark localization tasks too.
FIGURE 11. Examples of the sagittal cervical spine landmark point detection using PoseNet trained with IC-Loss.

REFERENCES

[1] D. D. Harrison, S. J. Troyanovich, D. E. Harrison, T. J. Janik, and D. J. Murphy, “A normal sagittal spinal configuration: a desirable clinical outcome,” Journal of manipulative and physiological therapeutics, vol. 19, no. 6, pp. 398–405, 1996.

[2] D. D. Harrison, T. J. Janik, S. J. Troyanovich, D. E. Harrison, and C. J. Colloca, “Evaluation of the assumptions used to derive an ideal normal cervical spine model,” Journal of manipulative and physiological therapeutics, vol. 20, no. 4, pp. 246–256, 1997.

[3] D. D. Harrison, T. J. Janik, S. J. Troyanovich, and B. Holland, “Comparisons of lordotic cervical spine curvatures to a theoretical ideal model of the static sagittal cervical spine,” Spine, vol. 21, no. 6, pp. 667–675, 1996.

[4] D. D. Harrison, D. E. Harrison, T. J. Janik, R. Cailliet, J. R. Ferrantelli, J. W. Haas, and B. Holland, “Modeling of the sagittal cervical spine as a method to discriminate hypolordosis: results of elliptical and circular modeling in 72 asymptomatic subjects, 52 acute neck pain subjects, and 70 chronic neck pain subjects,” Spine, vol. 29, no. 22, pp. 2485–2492, 2004.

[5] H. Y. Seong, M. K. Lee, S. R. Jeon, S. W. Roh, S. C. Rhim, and J. H. Park, “Prognostic factor analysis for management of chronic neck pain: can we predict the severity of neck pain with lateral cervical curvature?” Journal of Korean Neurosurgical Society, vol. 60, no. 4, p. 456, 2017.

[6] K. Han, C. Lu, J. Li, G.-Z. Xiong, B. Wang, G.-H. Lv, and Y.-W. Deng, “Surgical treatment of cervical kyphosis,” European spine journal, vol. 20, no. 4, pp. 523–536, 2011.

[7] C. Fernandez-de Las-Penas, C. Alonso-Blanco, M. Cuadrado, and J. Pareja, “Forward head posture and neck mobility in chronic tension-type headache: a blinded, controlled study,” Cephalalgia, vol. 26, no. 3, pp. 314–319, 2006.

[8] A. Nagasawa, T. Sakakibara, and A. Takahashi, “Roentgenographic findings of the cervical spine in tension-type headache,” Headache: The Journal of Head and Face Pain, vol. 33, no. 2, pp. 90–95, 1993.

[9] M. M. BRAAF and S. ROSNER, “Trauma of cervical spine as cause of chronic headache,” The Journal of Head and Face Pain, vol. 33, no. 2, pp. 90–95, 1993.

[10] H. Vernon, I. Steiman, and C. Hagino, “Cervicogenic dysfunction in muscle contraction headache and migraine: a descriptive study,” Journal of manipulative and physiological therapeutics, vol. 15, no. 5, pp. 441–446, 1975.

[11] G. N. Ferracini, T. C. Chaves, F. Dach, D. Bevilaqua-Grossi, C. Fernández-
[53] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “3d constrained local model for rigid and non-rigid facial tracking,” in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 2610–2617.

[54] J. M. Saragih, S. Lucey, and J. F. Cohn, “Deformable model fitting by regularized landmark mean-shift,” International journal of computer vision, vol. 91, no. 2, pp. 200–215, 2011.

[55] Y. Wang, S. Lucey, and J. F. Cohn, “Enforcing convexity for improved alignment with constrained local models,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008, pp. 1–8.

[56] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, “Robust face landmark estimation under occlusion,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1513–1520.

[57] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.

[58] G. Trigeorgis, S. Snape, M. A. Nikolou, E. Antonakos, and S. Zafeiriou, “Mnemonic descent method: A recurrent process applied for end-to-end face alignment,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4177–4187.

[59] Z.-H. Feng, J. Kittler, M. Awaiss, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2235–2245.

[60] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, “A deep regression architecture with two-stage re-initialization for high performance facial landmark detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3317–3326.

[61] J. Zhang and H. Hu, “Exemplar-based cascaded stacked auto-encoder networks for robust face alignment,” Computer Vision and Image Understanding, vol. 171, pp. 95–103, 2018.

[62] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, “Face alignment using a 3d deeply-initialized ensemble of regression trees,” Computer Vision and Image Understanding, vol. 189, p. 102846, 2019.

[63] A. P. Fard, H. Abdollahi, and M. Mahoor, “Asmnet: A lightweight deep neural network for face alignment and pose estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1521–1530.

[64] A. P. Fard and M. H. Mahoor, “Facial landmark points detection using knowledge distillation-based neural networks,” Computer Vision and Image Understanding, p. 103316, 2021.

[65] A. P. Fard and M. Mahoor, “Acr loss: Adaptive coordinate-based regression loss for face alignment,” arXiv preprint arXiv:2203.15835, 2022.

[66] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, “A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 585–601.

[67] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, “Joint multi-view face alignment in the wild,” IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3636–3648, 2019.

[68] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, “The menpo facial landmark localisation challenge: A step towards the solution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 170–179.

[69] S. Mahpud, R. Das, E. Maorana, Y. Keller, and P. Campisi, “Facial landmarks localization using cascaded neural networks,” Computer Vision and Image Understanding, p. 103317, 2021.

[70] A. P. Fard and M. H. Mahoor, “Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild,” IEEE Access, vol. 10, pp. 26756–26768, 2022.

[71] Y. Wu, S. K. Shah, and I. A. Kakadiaris, “Gdock: Globally optimized dual pathway deep network architecture for facial landmark localization in-the-wild,” Image and Vision Computing, vol. 73, pp. 1–16, 2018.

[72] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[73] H. Yang, X. Jia, C. C. Loy, and P. Robinson, “An empirical study of recent face alignment methods,” arXiv preprint arXiv:1511.05049, 2015.

[74] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[75] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

MOHAMMAD H. MAHOOR received the MS degree in Biomedical Engineering from Sharif University of Technology, Iran, in 1998, and the Ph.D. degree in Electrical and Computer Engineering from the University of Miami, Florida, in 2007. Currently, he is a professor of Electrical and Computer Engineering at the University of Denver. He does research in the area of computer vision and machine learning including visual object recognition, object tracking, affective computing, and human-robot interaction (HRI) such as humanoid social robots for interaction and intervention of children with autism and older adults with depression and dementia. He has received over 57M in research funding from state and federal agencies including the National Science Foundation and the National Institute of Health. He is a Senior Member of IEEE and has published over 158 conference and journal papers.

JOSEPH R. FERRANTElli graduated with honors in 1995 from Florida State University with a Bachelor of Science in Biology. In 1999 Dr. Ferranteelli graduated with honors from Life University College of Chiropractic with his Doctor of Chiropractic (D.C.) degree. In 2010 he co-founded PostureCo, Inc, a technology company focusing on Posture and Movement assessment software as well as spinal x-ray biomechanical mensuration EMR products for healthcare professionals. Since 2019, he continues to serve on the board for CBP Non-Profit, a non-profit spinal research foundation where he has co-authored manuscripts published in numerous health care journals.

ANNE-LISE DUPUIS received her Bachelor of Science degree in Computing Science from the University of Alberta in 1980. She was awarded a substantial scholarship from the National Research Council of Canada in 1980 towards a master’s degree which she declined after working as a computer consultant for large firms. Having worked in various scientific, process control, business, and computer environments as a Project leader, Database Administrator, Systems Analyst, and Programmer, Anne-Lise has a rich background to draw upon. Currently, she resides as the lead programming architect for application development at PostureCo.