Automatic Assessment of Pectus Excavatum Severity From CT Images Using Deep Learning

Bruno Silva, Inês Pessanha, Jorge Correia-Pinto, Jaime C. Fonseca, and Sandro Queirós

***Abstract*—Pectus excavatum (PE) is the most common abnormality of the thoracic cage, whose severity is evaluated by extracting three indices (Haller, correction and asymmetry) from computed tomography (CT) images. To date, this analysis is performed manually, which is tedious and prone to variability. In this paper, a fully automatic framework for PE severity quantification from CT images is proposed, comprising three steps: (1) identification of the sternum's greatest depression point; (2) detection of 8 anatomical keypoints relevant for severity assessment; and (3) measurements' geometric regularization and extraction. The first two steps rely on heatmap regression networks based on the Unet++ architecture, including a novel variant adapted to predict 1D confidence maps. The framework was evaluated on a database with 269 CTs. For comparative purposes, intra-observer, inter-observer and intra-patient variability of the estimated indices were analyzed in a subset of patients. The developed system showed a good agreement with the manual approach (a mean relative absolute error of 4.41%, 5.22% and 1.86% for the Haller, correction, and asymmetry indices, respectively), with limits of agreement comparable to the inter-observer variability. In the intra-patient analysis, the proposed framework outperformed the expert, showing a higher reproducibility between indices extracted from distinct CTs of the same patient. Overall, these results support the feasibility of the developed framework for the automatic, accurate and reproducible quantification of PE severity in a clinical context.**

***Index Terms*—Pectus excavatum, computed tomography, deep learning, keypoint/landmark detection.**

I. Introduction

PECTUS excavatum (PE) is the most common abnormality of the thoracic cage, characterized by the inward displacement of sternum and adjacent costal cartilages. This abnormality may be present at birth or only start to develop during puberty, causing psychological effects on the patients, potentially impacting patients' physical activity due to reduced lung capacity, and even leading to cardiopulmonary complications when severe [1], [2].

To determine whether repair surgery is needed, a CT scan is usually performed, and measurements of the patient's rib cage (Fig. 1) are extracted to evaluate the PE severity by calculating the Haller, correction, and asymmetry indices [3], [4]. The Haller index (HI) assesses the severity of the sternum's depression, being defined by eq. (1) as the ratio of the transverse chest distance (TCD, *i.e.*, horizontal length of the inside of the rib cage) and the anteroposterior distance (APD, *i.e.*, shortest distance between the sternum's posterior part and the spine's anterior face) [4]. The correction index measures the amount of correction needed [3], [5], and is given by eq. (2) as the relative percentage between the virtual correction distance (VCD, *i.e.*, maximum distance between the spine and the expected sternum's position after correction) and the APD. The asymmetry index is used to characterize the degree of PE asymmetry [3], and is calculated by eq. (3) as the ratio between left and right anteroposterior distances (LAPD and RAPD).

$$Haller\,Index = TCD/APD \tag{1}$$

$$Correction\,Index = [(VCD - APD)/VCD] \times 100 \tag{2}$$

$$Asymmetry\,Index = RAPD/LAPD \times 100 \tag{3}$$

In clinical practice, such analysis is made manually, which is tedious, time-consuming and prone to intra- and inter-observer variability. Indeed, given the CT's 3D nature, the expert must first identify the axial slice in which to perform the measurements. This slice must be located at the point of greatest depression of the chest anterior wall, so one must navigate through the volume and assess, using reconstructed sagittal slices, where this point lies. On the other hand, the positioning of the points for distance measuring is intimately linked to the experience and subject to the interpretation of the clinician. Despite the relevance of these

Manuscript received February 24, 2021; revised May 18, 2021; accepted June 17, 2021. Date of publication June 21, 2021; date of current version January 5, 2022. This work was supported in part by National Funds, through the Foundation for Science and Technology (FCT, Portugal) - projects UIDB/50026/2020 and UIDP/50026/2020. The work of Sandro Queirós was supported in part by Grant CEECIND/03064/2018. *(Corresponding author: Sandro Queirós.)*

Bruno Silva and Sandro Queirós are with the Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho Braga 4710-057, Portugal, and also with ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães Braga 4710-057, Portugal (e-mail: a78786@alunos.uminho.pt; sandroqueiros@med.uminho.pt).

Inês Pessanha is with the Department of Pediatric Surgery, Hospital Pediátrico, Centro Hospitalar e Universitário de Coimbra, Coimbra 3000-602, Portugal (e-mail: ipessanha3@gmail.com).

Jaime C. Fonseca is with the Algoritmi Center, School of Engineering, University of Minho, Guimarães 4800-058, Portugal (e-mail: jaime@dei.uminho.pt).

Jorge Correia-Pinto is with the ICVS, School of Medicine, University of Minho, Braga 4710-057, Portugal, with ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal, and also with the Department of Pediatric Surgery, Hospital de Braga, Braga 4710-243, Portugal (e-mail: jcp@med.uminho.pt).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2021.3090966, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2021.3090966

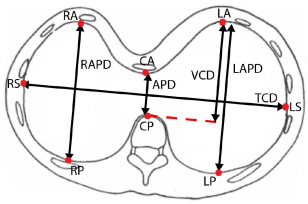

Fig. 1. Illustration of the PE measurements and relevant keypoints. CA: center anterior; CP: center posterior; RS: right side; LS: left side; RA: right anterior; RP: right posterior; LA: left anterior; LP: left posterior.

indices in the context of PE, there is currently no solution to automate this process.

Hereto, we address this problem by proposing a novel fully automatic deep learning (DL)-based system using a two-step keypoint detection approach, aiming to analyze the patient's CT and accurately extract the PE indices. In addition to the integrated framework itself, the present work describes a novel 1D heatmap regression network based on the Unet++ architecture [6] that outperforms its 2D and/or Unet-based counterparts. Moreover, we extensively validate the proposed system and its inner modules in a large database with 269 CTs, ultimately comparing it to clinical experts.

II. Related Work

A. Computed Tomography

Although no past work has focused on the automatic extraction of clinically used CT-based PE indices, some studies have suggested new methodologies to assess PE from these images. On the one hand, some works propose to automatically extract new metrics, different from those used in clinical practice. Kim *et al.* [7], [8] proposed four new indices to quantify the chest wall deformity, having also developed a fully automatic system to calculate these indices through various image processing techniques. The proposed indices showed a good correlation with HI. Later, the same group proposed two new indices that provided a linear output for complex chest wall deformities [9]. On the other hand, some authors argue that indices may not fully represent the chest and therefore propose to bypass their calculation and assess the severity qualitatively from the CT volumes directly. Following this idea, Lai *et al.* [10] proposed a multi-class classification model based on the VGG network to classify each CT slice between normal, mild PE, or severe PE. The final diagnosis is reached by majority voting across the entire CT volume. The 2D network achieved an accuracy of 94.76% in images from 42 patients, with the full volume classification reaching 97.62% accuracy. Although useful for PE diagnosis, the system only provides a qualitative assessment of the severity of the chest wall deformity.

B. Alternative Imaging Modalities

CT is the most frequently employed modality for PE assessment. However, exposure to this ionizing radiation may increase the likelihood of developing a radiation-related pathology, a fact with increased relevance given that most patients are of pediatric age. In an effort to decrease this exposure, alternative assessment methodologies have been studied.

One of these methods is optical imaging, where a scan is made to the patient's torso using lasers or structured (white) light, resulting in a virtual representation of the patient's torso topology from which measurements can be made to quantify the deformity of the chest wall. Glinkowski *et al.* [11] and Hebal *et al.* [12] used this method to calculate an external Haller index (EHI), obtaining a statistically significant correlation with the CT-derived HI. Despite this result, EHI is not interchangeable with HI as it presents a lower average value, thus requiring the definition of new severity thresholds. Moreover, Scalabre *et al.* [13] showed that EHI only presents a significant correlation with HI for symmetric PE cases. In its turn, optical and CT-derived asymmetry indices showed good correlation independently of the PE symmetry [13]. In a similar effort, Uccheddu *et al.* [14] reported good correlations between the external (optical-based) correction index and the (CT-based) correction index. Notwithstanding, one key limitation of optical methods is that they are based on the external topology of the patient's body, causing the quantified indices to vary according to the patient's body fat. To overcome these limitations, Taylor *et al.* [15] developed a predictive model to estimate the HI from EHI and patient's biometric data, reporting a median error of 8.11% against radiographic HI. Also using optical imaging, Gomes-Fonseca *et al.* [16] presented a new methodology to quantitatively follow-up patients after bar removal following the Nuss procedure.

Another alternative to CT is magnetic resonance imaging (MRI), which has been shown to be feasible for quantitative assessment of PE [17]–[19]. Using a fast MRI protocol, Piccolo *et al.* [17] showed a good correlation against CT for both Haller and asymmetry indices. Similar results were achieved by Birkemeier *et al.* [18], with no significant differences between CT- and MRI-derived HI. However, both studies were conducted with a small study size. More recently, Viña *et al.* [19] validated their findings in a larger cohort, demonstrating that both Haller and correction indices taken from a cardiac MRI show an excellent agreement with CT measures. Despite these results, the key limitation of MRI-based methods is the increase in costs, in required imaging time and in sensitivity to patient's movement during acquisition.

III. Methodology

The proposed framework is divided in three blocks (Fig. 2):

1) the sagittal block (Section III-A), that aims to locate the axial slice with the sternum's greatest depression by feeding a sagittal slice reconstructed from the CT (III-A1) to a novel 1D variant of a heatmap regression network based on the Unet++ architecture [6] (III-A2);
2) the axial block (Section III-B), that aims to detect the eight relevant keypoints (see Fig. 1) in each axial slice inside a region of interest (ROI) centered on the axial slice detected by the previous block (III-B1) using a 2D (Unet++) heatmap regression network (III-B2);
3) the post-processing block (Section III-C), that computes the keypoints' positions from the predicted heatmaps,

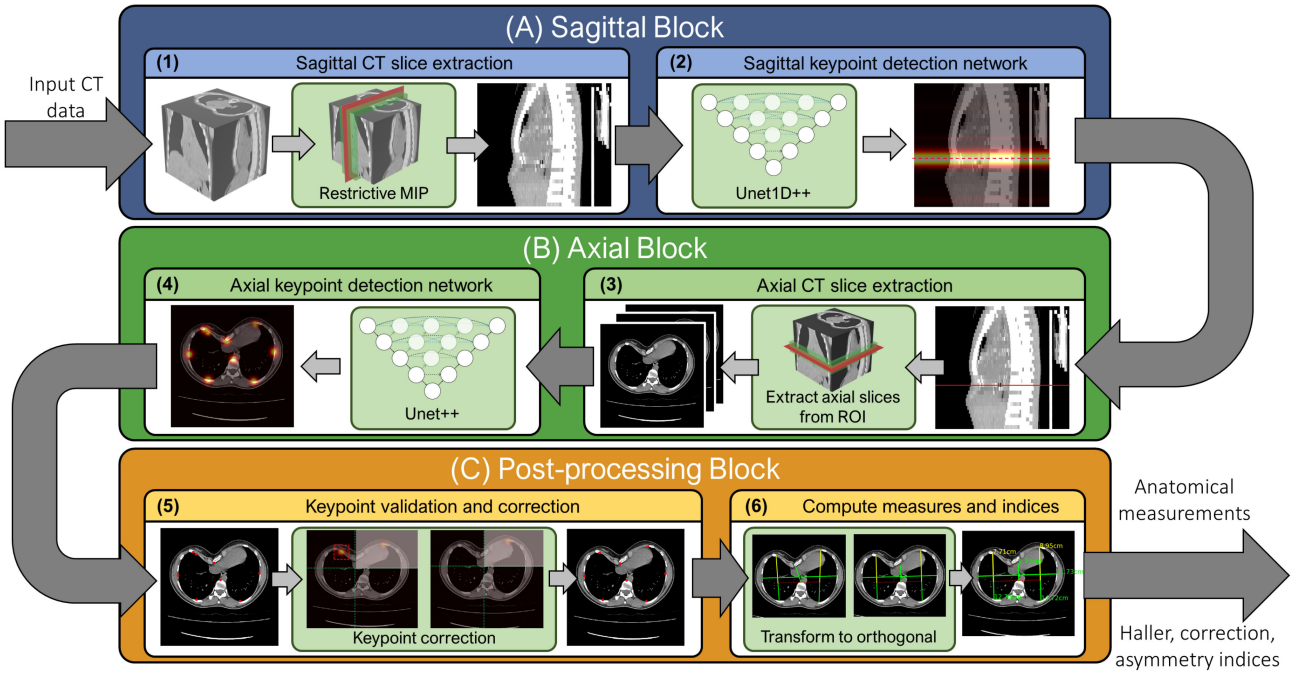

Fig. 2. Overview of the proposed fully automatic framework for PE severity quantification.

validates and corrects them if needed, and then extracts the PE measurements and indices from the axial slice with the lowest APD distance, while also ensuring the orthogonality of the measured distances.

A. Sagittal Block

1) Preprocessing: This module aims to prepare a sagittal image to be fed to the keypoint detection network. A problem of using a 2D image only is the fact that the patient's positioning in the CT bed is unknown, and thus the chosen cut may completely miss the sternum's depression, resulting in a non-representative image of the patient's chest being fed to the network. To mitigate this issue, and inspired by [20], [21], one proposes to convert the 3D CT volume into a 2D image through the use of a maximal intensity projection (MIP). The MIP creates a 2D image by projecting the voxels that contain the greatest intensity in the line of the chosen projection plane (*i.e.* the sagittal plane), thus flattening the 3D volume into a 2D image but retaining relevant high-intensity information (like the bones). Instead of projecting the entire CT, a restrictive MIP is here proposed to obtain an image with a clearer view of the sternum. In other words, the MIP is only applied to a ROI in the middle of the CT, avoiding the appearance of the lateral ribs in the resulting image (Fig. 3). Since the depression is often relatively centered in the CT, the ROI chosen for the MIP is 10 cm wide (Fig. 2A-1, the green highlight is the ROI, and the red plane the CT center plane).

Since a significant variability in slice thickness is found among CT images, the pixel spacing of the 2D MIP image is normalized (using nearest neighbor interpolation) to 0.5×0.5 mm^2, guaranteeing the consistency of the network's input. In addition, to normalize and highlight useful information in the

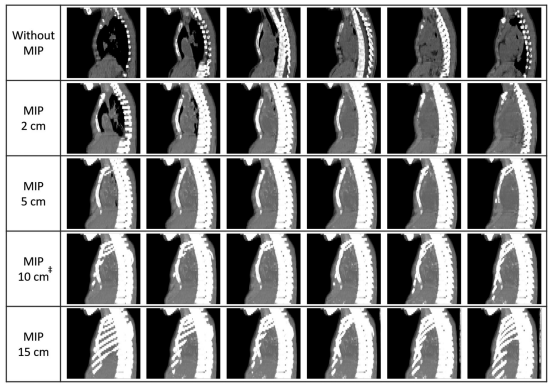

Fig. 3. Influence of ROI width and location during MIP extraction. Taking a single CT as example, each row represents a different ROI width, and each column assumes a different ROI center in the CT volume (to simulate off-centered acquisitions). ± = proposed ROI size.

slice, a windowing function between 50 Hounsfield units (HU) and 400 HU (a typical soft tissue window) was applied, eliminating the details within the lungs but allowing visualization of both bones and cartilages present at the end of the sternum, as well as the surrounding tissues and organs.

2) Architecture: The proposed heatmap regression network is based on the Unet++ architecture [6] (Fig. 4A), adapted to predict 1D heatmaps only (henceforth named Unet1D++). The input is the restricted MIP (scaled to 256×256 pixels, whose values are normalized to zero mean and unit variance) and the output are two vectors with size 1×256: a 1D Gaussian map that represents the network's confidence in the location of the sternum's greatest depression point and a 1D background map (*i.e.*, the inverse of the first vector, as in [21]).

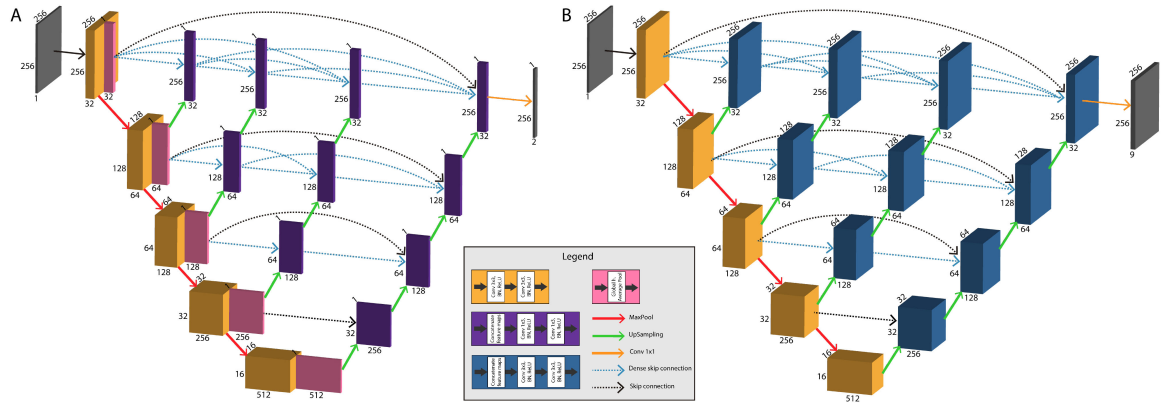

Fig. 4. Keypoint detection network used in the (A) sagittal and (B) axial blocks. The network is based on the Unet++ architecture [6].

In short, the Unet++ is an encoder-decoder type of architecture, consisting of a contraction path and a symmetrical expansion path, with skip pathways containing convolution layers to reduce the semantic gap between paths, as well as dense skip connections to better propagate the high resolution feature maps from the encoding to the decoding path. In opposition to [6], no deep supervision is employed.

The network's contraction and expansion paths are each composed of 5 levels, each containing two sequences of convolution (kernel size of 3×3 and 1×3, respectively), batch normalization (BN) and ReLU. An average pooling layer (pool size of 256×1) is attached to the end of each contraction level to horizontally flatten its output to 1×256 feature maps. This dimensionality reduction block (pink block in Fig. 4A) allows the network to focus on predicting the keypoint's y-coordinate, as its position along the x-axis does not affect the selected axial slice. The number of kernels per convolution layer is doubled after each level, with the initial value set to 32. Between contraction levels, a 2×2 max-pooling layer (red arrows in Fig. 4) is applied to halve the feature maps' resolution. Similarly, after each expansion level and on the skip pathways, 1×2 upsampling is applied to double the resolution of the feature maps (green arrows in Fig. 4). Each block in the skip pathways is also composed by the same convolution+BN+ReLU sequence (1×3 kernel), whose input results from the concatenation of feature maps from the skip connections (black arrows in Fig. 4), dense skip connections (blue arrows in Fig. 4) and the upsampled 1D feature maps from the lower level (green arrows in Fig. 4).

Lastly, one 1×1 convolution layer is used to regress the two-vector heatmap output (orange arrow in Fig. 4).

3) Implementation Details: Mini-batch gradient descent was used (batch size of 8) together with mean squared error (MSE) as loss and Adam [22] as optimizer. The weights were initialized from a normal distribution as proposed in [23], and a l2-regularization (with weight 1×10^{-5}) was added to the loss. The initial learning rate was set to 1×10^{-3}. This value is reduced to one third when no new loss minimum is found for 5 epochs. The training limit was set to 150 epochs, with early stopping set for 10 consecutive epochs without a new loss minimum. The proposed model has ≈6.2 M parameters.

The ground-truth 1D heatmaps were generated following:

$$g_\sigma(y) = e^{-\frac{(y-center_y)^2}{2\sigma^2}} \tag{4}$$

where y is a pixel's y-coordinate, $center_y$ is the labeler's keypoint y-coordinate, and σ is the Gaussian sigma value, set to 12.5 mm (converted to pixels following input resizing).

During training, data augmentation was applied on-the-fly to artificially increase the dataset size. Besides translation (between $\pm 15\%$ of the image size), rotation (from $-5°$ to $5°$) and scaling (with a factor between 0.95 and 1.05) transformations, one also augmented the images by simulating various slice thicknesses (from 0.5 to 10 mm). The goal is to mimic the variability seen in CT acquisitions, and decrease the network's generalization error. This is accomplished by vertically downsample the image and then upsample it back using bicubic and nearest neighbor interpolation, respectively. Each transformation has a 50% probability of being applied, and the generated sample is verified to ensure that the keypoint is still inside the augmented image (or otherwise re-augmented).

B. Axial Block

1) Preprocessing: This module is responsible for the extraction and preparation of all axial slices contained in a 2-cm wide ROI centered on the position detected by the previous block (Fig. 2B-3; the green highlight represents the ROI, and the red plane the axial slice passing through the detected point). Upon applying the spatial and intensity normalization described in Section III-A1, each slice is fed to the axial network.

2) Architecture: This module aims to detect, from each extracted axial slice, the eight keypoints necessary to compute the measurements used in the PE indices calculation.

Once again, the Unet++ architecture is employed in a heatmap regression strategy. Since one intends to predict 2D heatmaps, the dimensionality reduction blocks are now dropped, kernels are set to 3×3, and all pool/upsample sizes are set to 2×2 (Fig. 4B). The input images are resized to 256×256 pixels, and the values normalized to zero mean and unit variance. The network outputs nine heatmaps (of the same size as the input), with eight of them representing, through a 2D Gaussian, the

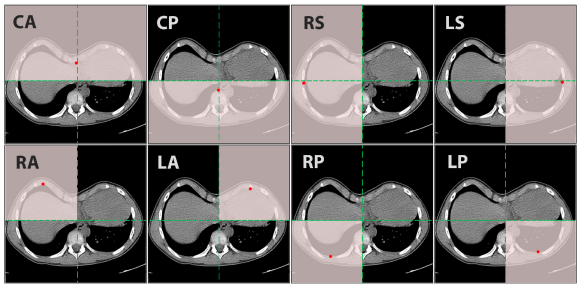

Fig. 5. Illustration of the valid region for each keypoint.

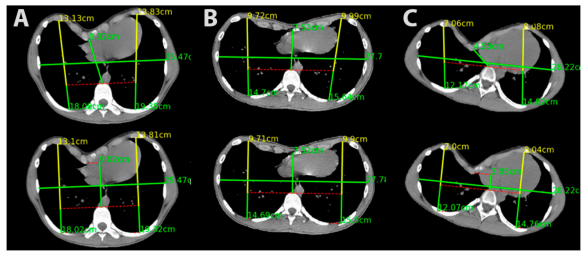

Fig. 6. Keypoints' transformation to ensure orthogonality. Top row: predicted keypoints and measures; bottom row: constrained ones.

network's confidence on the location of a given keypoint, and the ninth being the background map (*i.e.*, the inverse of the sum of the other eight heatmaps) [21].

3) Implementation Details: The weight initialization and regularization, optimizer, loss and training scheme are the same as in the previous block, and the network has ≈9.2 M parameters. Similarly, the ground-truth 2D heatmaps were generated from each labeler's keypoint following eq. (5).

$$g_{\sigma,k}(x,y) = e^{-\frac{(x-center_{x,k})^2+(y-center_{y,k})^2}{2\sigma^2}} \tag{5}$$

where x and y represent a pixel coordinates, $center_{x,k}$ and $center_{y,k}$ the ground-truth coordinates of keypoint k, and σ the Gaussian sigma value, being in this case equal to 7.5 mm.

A data augmentation scheme similar to the previous block was used, accounting for spatial transformations only and ensuring the presence of all eight keypoints inside the image.

C. Post-Processing Block

This block (Fig. 2C) comprehends two algorithms that validate (correcting if needed) the detected keypoints, and select, among the axial slices in the ROI, the one from which to extract the measures and calculate the indices.

Although the axial keypoint detection network should predict heatmaps with a single high-confidence peak, if presented with an input image that differs significantly from those in the training set, the network may predict a scattered heatmap with small confidence values or one with two separate high-confidence peaks (see simulated example in Fig. 2C-5), potentially failing to correctly predict the keypoint.

To solve this issue, one proposes an algorithm that divides the image into four distinct regions and automatically detects the misplacement of the predicted keypoints. These four zones are set based on a coordinate system defined by the RS-LS axis and the center of mass (CoM) of the 8 keypoints (Fig. 5), whose coordinates are calculated using eq. (6), with m_k being the keypoint k's score (highest value in heatmap k), and x_k and y_k the keypoint k's coordinates on the x- and y-axes.

$$CoM_x = \frac{\sum_{k=0}^{8} m_k\, x_k}{\sum_{k=0}^{8} m_k} \qquad CoM_y = \frac{\sum_{k=0}^{8} m_k\, y_k}{\sum_{k=0}^{8} m_k} \tag{6}$$

After establishing each keypoint's valid region with respect to the coordinate system (Fig. 5), if any keypoint is not within its region, its heatmap is zero-masked with a squared window (of size 2 times the Gaussian sigma value, *i.e.* 15 mm) centered on the wrongly detected peak, therefore deleting it (Fig. 2C-5, the applied mask is dashed in red). After this, the CoM is re-calculated and the keypoints re-validated, being this process repeated until all keypoints are in their valid regions.

Finally, following the experts' practice, one must guarantee that the measured distances are orthogonal to each other (with the RS-LS axis serving as reference). This is usually not the case, both due to the patients' rib cage geometrical variation and due to the network's own errors. To guarantee the measurements' orthogonality, a set of geometrical rules are used to calculate new points on every axial slice (Fig. 6).

Based on the corrected keypoints, one computes the APD distance for all axial slices in the ROI, and the one with the lowest APD (*i.e.* the one with the sternum's deepest depression) is selected to extract the measurements and compute the three indices (eqs. (1) to (3)). Note that, to compute the correction index, one must compute the VCD, which may differ depending on the chosen hemithorax. Thus, one computes two new points, termed mid-RP and mid-LP, that result from the interception of the line parallel to the RS-LS axis that passes through point CP and the lines perpendicular to it that pass through points RA and LA (Fig. 6). Upon determining these two points, one computes the distance towards the respective anterior keypoints (RA and LA, respectively) and the maximum value corresponds to the VCD.

IV. Experiments and Results

The framework's training was carried out on a workstation with a NVIDIA RTX 2080 Ti GPU, with 11 GB of VRAM.

A. Dataset

A total of 269 thoracic CTs, from 92 patients who underwent PE correction surgery, were retrospectively gathered and used to implement and evaluate the proposed framework. All CTs were acquired, using distinct CT scanners over the span of several years, as part of the preoperative planning performed prior to the surgery. The data gathering was approved by the responsible ethical board and informed consent was waived given the study's retrospective nature. Each CT was captured with an image resolution and size ranging from 0.3346 to 0.8223 mm and 512×512 to 768×768 pixels, respectively. The slice thickness varied between 0.5 mm and 10 mm, with a total number of slices between 24 and 824. Of the 92 patients (mean age, 15.5 ± 3.8

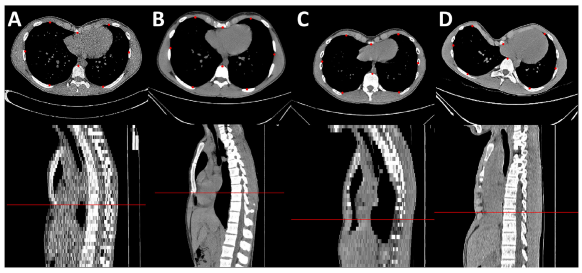

Fig. 7. Example axial and sagittal CT slices from 4 patients, and expert-annotated keypoints. The axial slices represent the slice with greatest depression of the sternum, and the sagittal ones cut through the middle of the CT volume. A clear variability in slice thickness (sagittal), PE severity (axial) and overall image quality (both) can be observed.

yrs, range, 10-36 yrs), 77 were male and 15 female. The dataset includes 17, 23 and 40 patients with a clinical assessment of mild, moderate and severe PE, respectively (no information available for the others).

Fig. 7 exemplifies the dataset's variability in terms of subjects' positioning, PE severity, and image acquisition.

All CTs were processed and analyzed using custom MATLAB (MathWorks Inc., USA) scripts, and were annotated by a pediatric surgeon using the Labelbox platform (Labelbox, CA, USA). In short, each volume was first analyzed to identify the sagittal slice that roughly corresponded to the center of the depression. Here, the expert identified the greatest depression point. Then, the region delimited by the most prominent area of the ribs' cage was identified, and multiple sagittal slices within this region (spaced 1 cm apart) were extracted. The keypoint provided by the expert was considered the ground truth for all sagittal cuts. By extracting multiple sagittal slices from one CT, one artificially enlarges the dataset. In addition, given that the network receives as input the sagittal slice from the middle of the volume (which may not be the center of the PE deformity), this approach increases the network's ability to deal with images from patients with asymmetric PE or those not centrally positioned in the CT bed (Fig. 7C).

A similar strategy was followed for the axial slices. The valid region for PE bar placement (from the base of the heart to the base of the manubrium) was identified, and axial slices spaced 1 cm apart were extracted. For each slice, the same expert identified the eight relevant keypoints (Fig. 7).

The dataset was divided, in a patient-disjoint manner, into two groups: (1) train/validation set used for architecture design and hyper-parameter tuning (in a 5-fold cross validation), with 74 patients (∼80%); and (2) test set, with 18 patients (∼20%).

B. Evaluation Metrics

The keypoint distance error (KDE) was used to assess the performance of the sagittal keypoint detection network. In short, for each predicted slice, one measures the distance between the predicted axial slice and the ground truth one. This distance was measured in mm to objectively account for the slice thickness of each CT.

Similarly, for the axial keypoint detection network, one averages the KDE for all keypoints in a given slice (in mm, as a 2D Euclidean distance). This metric was named as mean keypoint distance error (MKDE) and is formally expressed by eq. (7). The network's performance was reported as the average MKDE over the entire set of predicted images.

$$MKDE = \frac{\sum_k^N \|P_k - GT_k\|}{N} \tag{7}$$

where $P_k = \{x_k, y_k\}$ and $GT_k = \{x_{k,gt}, y_{k,gt}\}$ represent the predicted and ground truth coordinates of keypoint k, N is the number of keypoints (8), and $\|\cdot\|$ is the Euclidean distance.

C. Sagittal Block

The proposed Unet1D++ network obtained a median KDE of 4.5 mm over the 5-fold cross validation.

Fig. 8 presents the influence of key algorithmic choices (network architecture, type of dimensionality reduction block, and ROI width for MIP) on the proposed model's performance.

Regarding the network architecture (Fig. 8A), the proposed Unet1D++ significantly outperformed all tested variants, with the exception of the Unet2D1D++ (*i.e.*, single pooling block after the expansion path) that obtained a similar performance. However, the former is computationally lighter (∼67% of the weights). With respect to the dimensionality reduction block used in the proposed architecture (Fig. 8B), the best average KDE was achieved while using average pooling blocks, with a significant difference to most tested variants (with the exception of the $N \times 1$ convolution layer). Finally, this experiment also confirmed that the ROI width used for MIP calculation is a critical parameter, with all variations presenting a significant difference against the proposed value (10 cm, Fig. 8C).

On the test set, the proposed architecture achieved a median KDE of 5.00 mm. Fig. 9 presents some detection results for test samples, when compared to the expert annotations.

D. Axial Block

Table I summarizes the 5-fold cross-validation performance of the axial keypoint detection network. The results obtained when replacing the proposed architecture (Unet++) by five other common architectures are also presented. The details of each architecture are presented in the supplementary files. All networks were trained under the same conditions (input size, heatmaps' sigma size, weight initialization, data augmentation, optimizer, loss and training scheme). Overall, the proposed model presented the best performance (*i.e.* significantly lower average MKDE), with consistent results across all keypoints.

As in Section IV-C, an analysis of key algorithmic choices (type of windowing function used for image normalization and type of regression) was carried out to assess the method's sensitivity to the defined choices (Fig. 10). In terms of input normalization, three different windowing functions (soft tissues, bones, and lungs) or no windowing normalization (unchange) were assessed (Fig. 10A). Overall, the use of a

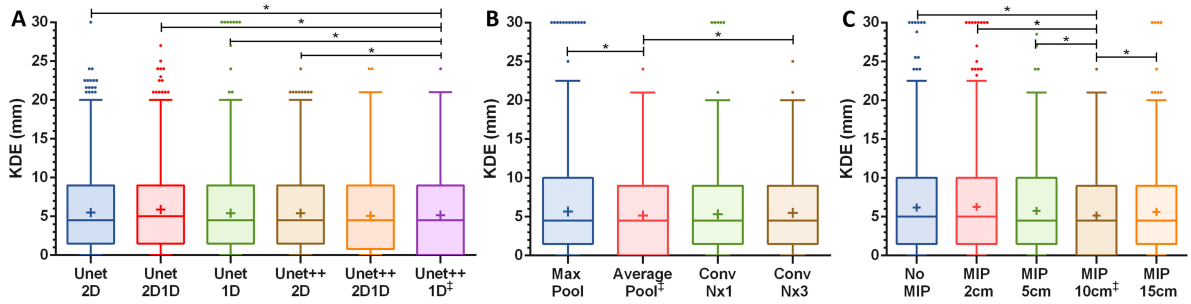

Fig. 8. Influence of (A) network architecture, (B) type of dimensionality reduction block, and (C) MIP's ROI width on the performance of the sagital block (assessed in a 5-fold cross validation). * $p < 0.05$, in a Wilcoxon matched-pairs signed rank test against the proposed parameter's value (marked with ±). 2D: 2D heatmap prediction; 2D1D: 1D heatmap prediction using a single global horizontal average pooling layer before the last 1×1 convolution layer; 1D: one global horizontal average pooling layer per contraction level (see Fig. 4).

TABLE I
PERFORMANCE OF THE PROPOSED AXIAL BLOCK IN A 5-FOLD CROSS VALIDATION, AND COMPARISON AGAINST OTHER ARCHITECTURES

Model	MKDE	KDE_{CP}	KDE_{CA}	KDE_{RS}	KDE_{LS}	KDE_{RA}	KDE_{RP}	KDE_{LA}	KDE_{LP}
UNet++ (proposed)	2.99±1.18	1.29±0.82	2.90±2.56	3.45±5.58	4.15±3.30	3.09±2.46	2.81±2.19	3.15±3.66	3.09±2.46
UNet	3.18±1.64*	1.32±1.42	3.31±2.69	3.90±7.07	4.28±3.26	3.25±2.45	2.85±2.15	3.44±5.11	3.10±2.45
MobileNetV2	3.17±1.21*	1.42±0.94	3.25±2.82	3.67±2.80	4.35±3.37	3.32±4.35	2.88±2.10	3.23±3.67	3.22±2.48
MobileNetV1	3.40±1.47*	1.34±0.89	3.80±4.23	4.45±3.32	4.55±3.59	3.31±2.58	2.98±3.75	3.43±5.42	3.30±2.68
VGG16	3.64±1.56*	2.18±1.15	3.89±2.83	4.23±4.84	4.68±3.40	3.77±5.52	3.20±2.12	3.71±5.02	3.44±2.34
VGG19	3.62±1.51*	2.22±1.16	3.68±2.71	4.30±6.47	4.70±3.33	3.56±2.45	3.42±5.16	3.61±2.65	3.49±2.36

Values are mean ± standard deviation, presented in mm. * $p < 0.05$, in a Wilcoxon matched-pairs signed rank test against the proposed architecture.

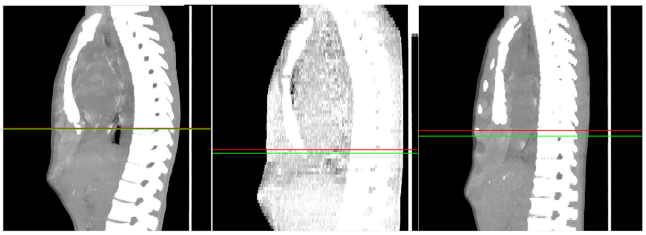

Fig. 9. Sagittal keypoint detection examples for three representative samples from the test set (MIP image is shown). The red line represents the proposed model's output, and the green line the ground truth.

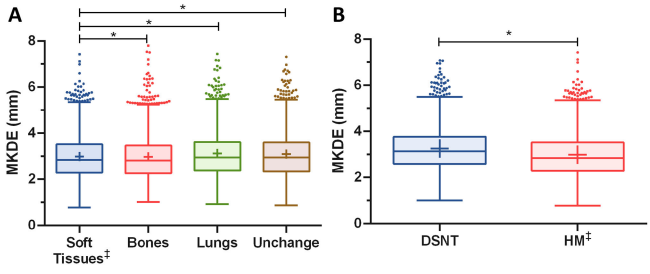

Fig. 10. Influence of (A) windowing normalization function, and (B) type of regression on the performance of the axial block (assessed in a 5-fold cross validation). * $p < 0.05$, in a Wilcoxon matched-pairs signed rank test against the proposed choice (marked with ++). DSNT: differentiable spatial to numerical [24] layer; HM: heatmap matching.

soft tissue function outperformed the other variants. Considering the regression method (Fig. 10B), the heatmap matching approach (proposed) obtained a significantly lower error when compared to the numerical coordinate regression [24] approach (*i.e.*, addition of a DSNT layer at the network's output to directly regress the keypoints' coordinates from the heatmaps).

On the test set, the proposed axial block achieved an average MKDE of 3.43 ± 1.69 mm. Fig. 11A presents the errors' distribution for each keypoint (and their average), and Fig. 11B illustrates some detection results for test samples.

E. Clinical Validation

This section intends to validate the proposed framework as a whole, including the post-processing block and the selection of the correct axial slice for measurements' extraction.

All CTs from the test set (52 volumes from 18 patients) were used to carry out this experiment. Using the Labelbox platform, for each CT, the expert identified the sternum's greatest depression point in a sagittal image, and then annotated the 8 relevant keypoints in the selected axial slice. The identified keypoints were then regularized to guarantee the measurements' orthogonality, before computing the indices. The above procedure was done twice by the same expert to obtain data for an intra-observer analysis. A second observer also repeated the annotation to assess the inter-observer variability.

Table II compares the indices extracted by our framework and the expert, summarizing the linear regression (correlation coefficient, slope and two-tailed paired *t*-test between indices) and Bland-Altman (bias and limits of agreement; LOA) results. Overall, the automatic indices have a remarkable agreement against the expert's values. To understand the results' sensitivity to each module of the proposed framework, an ablation study is included in Appendix B of the supplementary files.

To further understand the framework's performance, the agreement with the expert was also compared against the inter-

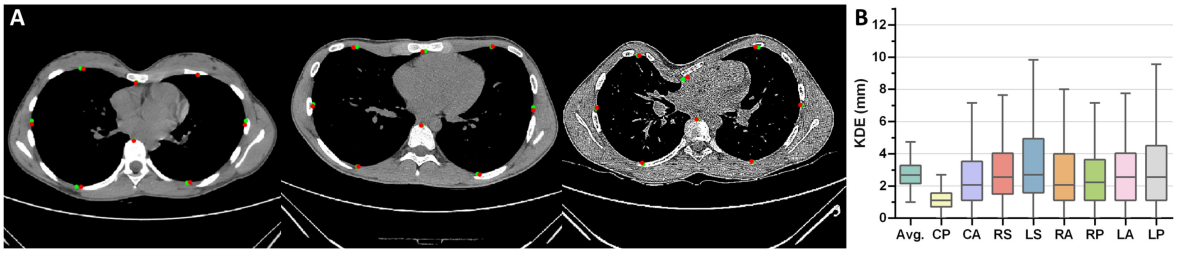

Fig. 11. (A) Axial keypoint detection examples for three representative test samples (red: proposed model; green: ground truth). (B) Performance for each keypoint individually, and averaged, on the test set. Outliers were omitted for better visualization.

TABLE II
LINEAR REGRESSION AND BLAND-ALTMAN ANALYSES BETWEEN AUTOMATIC AND MANUAL INDICES

	Haller index	Correction index	Asymmetry index
Proposed[1]	4.29 ± 1.44 [2.46; 8.53]	39.04 ± 11.20 [21.57; 62.31]	96.84 ± 6.13 [81.07; 107.61]
Expert[1]	4.35 ± 1.73 [2.29; 10.15]	38.65 ± 12.64 [17.81; 67.77]	96.43 ± 6.60 [76.02; 106.51]
Person's r	0.984	0.967	0.969
Slope	0.820	0.856	0.901
***t*-test**	0.263	0.412	0.080
Bias	-0.06	-0.39	0.41
LOA	[-0.86; 0.73]	[-6.27; 7.05]	[-2.81; 3.63]

[1] Values are mean ± standard deviation [range].

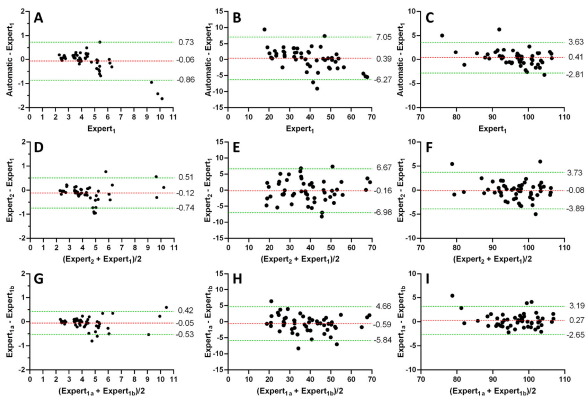

Fig. 12. Bland–Altman analysis of proposed vs. expert (A–C), inter- (A–C) and intra-observer variability (E–H), for the Haller (A, D, and G), correction (B, E, and H) and asymmetry (C, F, and I) indices.

and intra-observer variability (Fig. 12). In short, the proposed framework obtained statistically similar performance when compared to the intra- and inter-observer variability ($p > 0.05$ in a two-tailed F-test, except against the intra-observer agreement for the Haller index).

Additionally, an intra-patient analysis was performed to assess the framework's robustness with respect to the CT used as input, and compare it to the variability obtained by the expert. The intra-patient analysis consisted in calculating, for each patient, the error between the indices computed for each CT and the average value across that patient's CTs. As a goal, these variations are expected to be very small since the information of the patient's rib cage is the same across all CTs. Importantly,

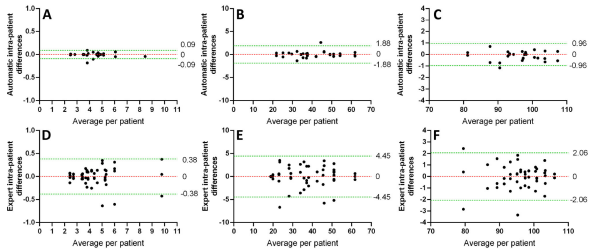

Fig. 13. Bland-Altman analysis for proposed (A–C) and expert's (D–F) intra-patient variability for the Haller (A and D), correction (B and E) and asymmetry (C and F) indices.

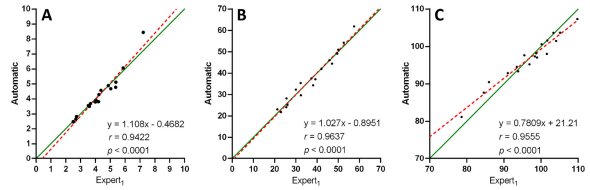

Fig. 14. Linear regression analysis between the automatically extracted PE indices and those manually extracted by the expert in a clinical environment. Analysis for the Haller (A), correction (B), and asymmetry (C) indices. Green line: identity; red dashed line: linear regression result.

the developed framework demonstrated a significantly higher reproducibility (in a two-tailed F-test) for all indices when compared to the expert's results (Fig. 13).

Finally, the automatic values were compared against the measures originally extracted by the expert in practice (extracted from the surgeon's logs). In the latter, and using a DICOM viewer, the expert freely navigates the volume and, after selecting the axial slice with the sternum's greatest depression point, draws several lines to extract the necessary distances to calculate the indices (without guarantees of the measures' orthogonality). For patients with more than one CT available, the average value of each index extracted by the proposed system was considered (*i.e.* sample size of 18). The results are shown in Fig. 14. The automatic strategy presented a great correlation with the manually extracted values for all indices (between 0.94 and 0.96), with an average relative absolute error of 4.41%, 5.22% and 1.86% for the Haller, correction and asymmetry indices, respectively.

V. Discussion

In this study, we sought to evaluate the feasibility, accuracy and reproducibility of a novel fully automatic DL-based framework for PE severity quantification from CT data. By employing a two-step keypoint detection strategy and a geometrical-based post-processing method for measurements' extraction, the system demonstrated its ability to identify the relevant anatomical landmarks and quantify the PE indices with a performance similar to expert clinicians.

Despite the successful application of similar keypoint detection methods in the medical imaging field [20], [21], [25]–[31], to the authors' best knowledge, no other system (similar or not) existed in the context of PE. Considering the literature, our results, as well as the peculiarities and constraints of the application, three major algorithmic design choices seem to support the reported system's feasibility and accuracy.

First, given the limited data available and the notable anatomical variability between PE patients, we chose to employ 2D convolution neural networks (CNNs) in a two-step pipeline instead of a volumetric one. Besides reducing the system's computational requirements, this strategy effectively reduces the amount of data needed given the lower complexity of 2D CNNs and the ability to use multiple slices of a CT as independent samples (as described in Section IV-A).

Second, the usage of region-based techniques, such as the restricted MIP image or the ROI-based analysis at the axial block, help mitigate the issues introduced by a slice-based processing. Indeed, Fig. 8C demonstrated the usefulness of the MIP image, a finding corroborated by the perceived influence of its ROI width on the generated sagittal image (Fig. 3). A narrow (or single-slice) ROI is sensitive to the body positioning on the CT bed (with off-center slices potentially having insufficient information about the sternum and its greatest depression point; see first row), while a wider ROI leads to reconstructed images with overlapping details (particularly the lateral ribs obstructing the visibility of the sternum; see last row). Hence, a median value is preferred to guarantee enough information to increase robustness to body positioning without compromising the visibility of the sternum. In regard to the ROI used for axial analysis, its usage together with the measurement-based axial selection strategy (Section III-C) enables the optimization of the chosen slice on which to extract the indices and increases the framework's accuracy and reproducibility over a single-slice assessment (Appendix B).

Third, the Unet++ architecture used, including our 1D variant, obtained a superior accuracy compared to other models (see Fig. 8A, Table I and Fig. 10). In the sagittal block, the proposed Unet1D++ surpassed its 2D counterpart, a finding simultaneously explained by the task at hand (slice localization only) and by the technique used to create the annotated dataset itself (Section IV-A). The latter may originate samples whose keypoint (when defined by a 2D coordinate) appears positioned within the thorax (for slices deviated from the depression's center), negatively impacting the 2D network's convergence. Nevertheless, if such technique would not be employed, the dataset size (for the same human effort) would decrease significantly and the performance also (data not shown). The proposed variant also surpassed the Unet-based counterpart (as the one used in [21]), which was later corroborated in the axial block for the 2D variants (Table I). As stated in [6], the performance gain may be explained by the modified skip pathways present in Unet++, that reduce the semantic gap between encoding and decoding paths, decreasing the complexity of the optimization problem and leading to a better convergence. When compared to other architectures (Table I), the lower average MKDE may be linked to the larger output size, mitigating issues linked with post-inference interpolation operations. The latter may also explain the superior accuracy when compared to the numerical coordinate regression approach (applied in previous works [25], [26]), as Unet-like models already output a high-resolution heatmap and therefore overcome the need for better spatial generalization or output resolution as enabled by the DSNT layer [24] (Fig. 10B).

When comparing the accuracy for different keypoints (Fig. 11B), larger errors were found for keypoints not associated with specific anatomical landmarks (such as RS and LS), possibly resulting from the expert's lower precision in positioning them. Nevertheless, it is interesting to note that the results obtained by the axial keypoint detection network originated no case of keypoint inversion (*i.e.*, swap of a left/right or anterior/posterior keypoint pair) or other substantial positioning error on the test set, which ultimately avoided the triggering of the keypoint correction algorithm (Section III-C) when validating the full framework (Section IV-E).

In this validation experiment, the system's automatically extracted indices showed high correlation to expert values (Table II), with small biases and narrow LOAs (Fig. 12A–C). Interestingly, these LOAs were comparable to the reported inter-observer variability (Fig. 12D–F) and close to the intra-observer one (Fig. 12G–I, with statistical significance in a F-test only for HI). Note, however, the slightly narrower range of values extracted by our system when compared to the expert (see, for example, the outliers for the cases with a remarkably high HI in Fig. 12A). Interestingly, in the intra-patient analysis (Fig. 13), the system proved to be remarkably reproducible in extracting the PE indices, clearly outperforming the expert.

Also relevant is the fact that these results were also observed when the proposed framework was validated against values extracted by the expert during clinical practice (a real-world scenario with a distinct measuring workflow, Fig. 14).

Finally, given that novel indices for PE assessment and management are regularly proposed [32]–[34] and may eventually be mainstreamed into clinical practice, it is pertinent to highlight that the proposed framework may easily be adapted to extract other clinically useful indices as long as they rely on identifiable anatomical landmarks. These may be added to the annotated dataset, the proposed model(s) modified to predict the additional heatmaps and the post-processing block adapted to regularize and extract the associated indices.

The proposed framework has nevertheless some limitations. First, and despite the implemented region-based techniques discussed above, the proposed CNNs models are inherently 2D. Even though our database is larger than those used in similar

works from the literature (Section II), this limitation is primarily linked to the lack of sufficient data for training 3D models. Second, being a single center study, an independent validation with data from a distinct center would be valuable to understand the generalization capacity of the proposed system. It is though relevant to highlight the large variability in image quality present in our database (Fig. 7), with images acquired over a large period with multiple CT machines. Finally, this work targets CT data, which exposes patients to ionizing radiation. The adaptation of the proposed framework to work with MRI data would mitigate this limitation.

VI. Conclusion

In summary, a novel fully automatic framework for PE severity quantification in CT data was presented and validated. Experiments showed the feasibility of the proposed two-step keypoint detection strategy, and demonstrated the superior performance of the presented Unet1D++ architecture for one-dimensional keypoint detection. Automatically-extracted PE indices showed excellent agreement with those extracted by experts, comparable to the inter-observer variability and with higher reproducibility in an intra-patient analysis. Given its automatic nature, the proposed system has the potential to replace the current tedious and variability-prone manual procedure, which would ultimately save healthcare professionals' time and increase measurements' reproducibility.

Acknowledgment

The authors would like to acknowledge Labelbox, Inc. (CA, USA) with the granting of an academic license to their collaborative training data platform.

References

[1] E. Coln, J. Carrasco, and D. Coln, "Demonstrating relief of cardiac compression with the nuss minimally invasive repair for pectus excavatum," *J. Pediatr. Surg.*, vol. 41, no. 4, pp. 683–686, 2006.

[2] A. C. Koumbourlis and C. J. Stolar, "Lung growth and function in children and adolescents with idiopathic pectus excavatum," *Pediatr. Pulmonol.*, vol. 38, no. 4, pp. 339–343, 2004.

[3] J. A. Sujka and S. D. S. Peter, "Quantification of pectus excavatum: Anatomic indices," *Proc. Seminars Pediatr. Surg.*, Elsevier, vol. 27, pp. 122–126, 2018.

[4] J. A. Haller, S. S. Kramer, and S. A. Lietman, "Use of CT scans in selection of patients for pectus excavatum surgery: A preliminary report," *J. Pediatr. Surg.*, vol. 22, no. 10, pp. 904–906, 1987.

[5] P. M. Poston *et al.*, "The correction index: Setting the standard for recommending operative repair of pectus excavatum," *Ann. Thoracic Surg.*, vol. 97, no. 4, pp. 1176–1180, 2014.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet : Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[7] H. C. Kim *et al.*, "Development of automatized new indices for radiological assessment of chest-wall deformity and its quantitative evaluation," *Med. Biol. Eng. Comput.*, vol. 46, no. 8, pp. 815–823, 2008.

[8] H. C. Kim *et al.*, "Fully automatic initialization method for quantitative assessment of chest-wall deformity in funnel chest patients," *Med. Biol. Eng. Comput.*, vol. 48, no. 6, pp. 589–595, 2010.

[9] H. C. Kim *et al.*, "New computerized indices for quantitative evaluation of depression and asymmetry in patients with chest wall deformities," *Artif. Organs*, vol. 37, no. 8, pp. 712–718, 2013.

[10] L. Lai, S. Cai, L. Huang, H. Zhou, and L. Xie, "Computer-aided diagnosis of pectus excavatum using CT images and deep learning methods," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.

[11] W. Glinkowski, R. Sitnik, M. Witkowski, H. Kocon, P. Bolewicki, and A. Górecki, "Method of pectus excavatum measurement based on structured light technique," *J. Biomed. Opt.*, vol. 14, no. 4, p. 044041, 2009.

[12] F. Hebal, E. Port, C. J. Hunter, B. Malas, J. Green, and M. Reynolds, "A novel technique to measure severity of pediatric pectus excavatum using white light scanning," *J. Pediatr. Surg.*, vol. 54, no. 4, pp. 656–662, 2019.

[13] A. Scalabre *et al.*, "Utility of radiation-free imaging for initial evaluation of pectus excavatum," *Interactive Cardiovasc. Thoracic Surg.*, vol. 29, no. 4, pp. 503–509, 2019.

[14] F. Uccheddu *et al.*, "A novel objective approach to the external measurement of pectus excavatum severity by means of an optical device," *Ann. Thoracic Surg.*, vol. 106, no. 1, pp. 221–227, 2018.

[15] J. S. Taylor *et al.*, "Three-dimensional optical imaging for pectus excavatum assessment," *Ann. Thoracic Surg.*, vol. 108, no. 4, pp. 1065–1071, 2019.

[16] J. Gomes-Fonseca *et al.*, "A new methodology for assessment of pectus excavatum correction after bar removal in nuss procedure: Preliminary study," *J. Pediatr. Surg.*, vol. 52, no. 7, pp. 1089–1097, 2017.

[17] R. L. Piccolo *et al.*, "Chest fast MRI: An imaging alternative on preoperative evaluation of pectus excavatum," *J. Pediatr. Surg.*, vol. 47, no. 3, pp. 485–489, 2012.

[18] K. L. Birkemeier, D. J. Podberesky, S. Salisbury, and S. Serai, "Limited, fast magnetic resonance imaging as an alternative for preoperative evaluation of pectus excavatum: A feasibility study," *J. Thoracic Imag.*, vol. 27, no. 6, pp. 393–397, 2012.

[19] N. A. Viña *et al.*, "Evaluation of pectus excavatum indexes during standard cardiac magnetic resonance: Potential for single preoperative tool," *Clin. Imag.*, vol. 53, pp. 138–142, 2019.

[20] S. Belharbi *et al.*, "Spotting L3 slice in CT scans using deep convolutional network and transfer learning," *Comput. Biol. Med.*, vol. 87, pp. 95–103, 2017.

[21] F. Kanavati, S. Islam, E. O. Aboagye, and A. Rockall, "Automatic L3 slice detection in 3D CT images using fully-convolutional networks," 2018, *arXiv:1811.09244*.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

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

[25] F. Galbusera *et al.*, "Fully automated radiological analysis of spinal disorders and deformities: A deep learning approach," *Eur. Spine J.*, vol. 28, no. 5, pp. 951–960, 2019.

[26] B. S. Andreassen, F. Veronesi, O. Gerard, A. H. S. Solberg, and E. Samset, "Mitral annulus segmentation using deep learning in 3-D transesophageal echocardiography," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 994–1003, Apr. 2020.

[27] X. Wang, X. Yang, H. Dou, S. Li, P.-A. Heng, and D. Ni, "Joint segmentation and landmark localization of fetal femur in ultrasound volumes," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2019, pp. 1–5.

[28] P. Astudillo *et al.*, "Automatic detection of the aortic annular plane and coronary ostia from multidetector computed tomography," *J. Interventional Cardiol.*, vol. 2020, May 2020.

[29] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, May 2019.

[30] J. Zhang, M. Liu, and D. Shen, "Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4753–4764, Oct. 2017.

[31] A. Alansary *et al.*, "Evaluating reinforcement learning agents for anatomical landmark detection," *Med. Image Anal.*, vol. 53, pp. 156–164, Apr. 2019.

[32] W. Zhong *et al.*, "Effects of pectus excavatum on the spine of pectus excavatum patients with scoliosis," *J. Healthcare Eng.*, vol. 2017, Jan. 2017.

[33] C. W. Snyder, S. M. Farach, C. N. Litz, P. D. Danielson, and N. M. Chandler, "The modified percent depth: Another step toward quantifying severity of pectus excavatum without cross-sectional imaging," *J. Pediatr. Surg.*, vol. 52, no. 7, pp. 1098–1101, 2017.

[34] C. Capunay *et al.*, "Sternal torsion in pectus excavatum is related to cardiac compression and chest malformation indexes," *J. Pediatr. Surg.*, vol. 55, no. 4, pp. 619–624, 2020.