RibSeg v2: A Large-scale Benchmark for Rib Labeling and Anatomical Centerline Extraction

Liang Jin, Shixuan Gu, Donglai Wei, Kaiming Kuang, Hanspeter Pfister, Jiancheng Yang, Ming Li, and Bingbing Ni

***Abstract*— Automatic rib labeling and anatomical centerline extraction are common prerequisites for various clinical applications. Prior studies either use in-house datasets that are inaccessible to communities, or focus on rib segmentation that neglects the clinical significance of rib labeling. To address these issues, we extend our prior dataset (*RibSeg*) on the binary rib segmentation task to a comprehensive benchmark, named *RibSeg v2*, with 660 CT scans (15,466 individual ribs in total) and annotations manually inspected by experts for rib labeling and anatomical centerline extraction. Based on the *RibSeg v2*, we develop a pipeline including deep learning-based methods for rib labeling, and a skeletonization-based method for centerline extraction. To improve computational efficiency, we propose a sparse point cloud representation of CT scans and compare it with standard dense voxel grids. Moreover, we design and analyze evaluation metrics to address the key challenges of each task. Our dataset, code, and model are available online to facilitate open research at https://github.com/M3DV/RibSeg.**

***Index Terms*— rib segmentation, rib labeling, rib centerline, point cloud, computed tomography.**

I. Introduction

RIB labeling and anatomical centerline extraction are of significant clinical value for facilitating various clinical applications. For example, it is critical for detecting rib fractures, which can identify chest trauma severity that accounts for $10\% \sim 15\%$ of all traumatic injuries [1]. Besides, the structure and morphology of rib bones are stable references for multiple analysis and quantification tasks such as lung volume estimation [2] and bone abnormality quantification [3]. Based on rib anatomical centerlines, internal coordinate systems can localize organs for surgery planning and postoperative evaluation [4], as well as registering pathologies such as lung nodules [5]. Moreover, automatic rib labeling and centerline extraction is the key to developing visualization tools of unfolded rib cages [6]–[8], significantly reducing the burden of rib interpretation for clinicians.

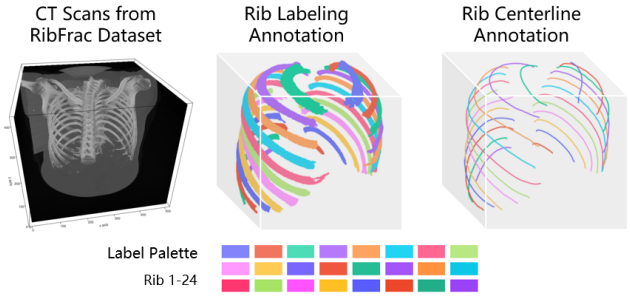

Fig. 1: ***RibSeg v2* Dataset**. *RibSeg v2* extends the annotations of 660 CT scans from the existing *RibFrac* dataset [9], containing labeled rib segmentation and anatomical centerlines. The color palette indicates the label assigned to each rib.

Despite the high contrast of ribs, rib labeling and centerline extraction is challenging. Ribs in human bodies are typically elongated and oblique across numerous CT sections; In other words, a large number of CT slices must be evaluated sequentially by radiologists. Ribs are anatomically close to the scapula and clavicle, and some ribs might be connected by the metal implant, which is hard to label. Besides, rib centerline extraction is subject to image noise, artifacts, and the quality of rib labels.

Previous studies on this topic only focus on rib segmentation [10] and trivialize rib labeling as a counting process [11], [12], which underestimates the challenges of false merging and neglects that anatomical labeling of ribs is clinically more desirable. Besides, tracing-based rib segmentation and centerline extraction is highly sensitive to initially detected seed points and vulnerable to local ambiguities [13], [14]. Although the deep learning-based method is robust as it learns hierarchical visual features from raw voxels [15], it does not consider the sparsity and elongated geometry of ribs. Moreover, there is no public dataset on this topic, making it difficult to benchmark existing methods and develop downstream applications.

Manuscript submitted October 7, 2022.

Liang Jin and Shixuan Gu have contributed equally.

Jiancheng Yang, Ming Li, and Bingbing Ni are the corresponding authors.

Liang Jin is with Radiology Department, Huadong Hospital, affiliated to Fudan University, and also with Shanghai Key Lab of Forensic Medicine, Key Lab of Forensic Science, Ministry of Justice, China (jin_liang@fudan.edu.cn).

Shixuan Gu is with Carnegie Mellon University, PA, USA, and also with Harvard University, MA, USA (shixuang@andrew.cmu.edu).

Donglai Wei is with Boston College, MA, USA (donglai.wei@bc.edu).

Hanspeter Pfister is with Harvard University, MA, USA (pfister@seas.harvard.edu).

Kaiming Kuang is with Dianei Technology, Shanghai, China.

Jiancheng Yang is with Shanghai Jiao Tong University, and Dianei Technology, Shanghai, China, and also with EPFL, Switzerland (jekyll4168@sjtu.edu.cn).

Ming Li is with Radiology Department, Huadong Hospital, affiliated to Fudan University, Shanghai, China (minli77@163.com).

Bingbing Ni is with Shanghai Jiao Tong University, and also with Huawei Hisilicon, Shanghai, China (nibingbing@sjtu.edu.cn).

To tackle these problems, We first extended our previous benchmark for rib segmentation [16] to develop a benchmark for rib labeling and anatomical centerline extraction, named *RibSeg v2*, including annotations of 660 chest-abdomen CT scans (15,466 individual ribs) from *RibFrac* dataset [9]. In addition, we formulated rib labeling as the task of segmenting ribs from CT scans and labeling the binary segmentation into 24 instances, and benchmark the *RibSeg v2* with a pipeline including a deep learning-based method for rib labeling and a TEASAR-based [17] method for rib anatomical centerline extraction. Besides, we compared the data representations of CT scans as dense voxel grids and sparse point clouds, respectively, and proposed various metrics for each task to provide comprehensive evaluations. Finally, by detailed quantitative and qualitative analysis of the challenging cases, we explored the key challenges of each task, which are valuable to guide future studies on this topic.

Contributions. 1) The first large public benchmark for rib labeling and anatomical centerline extraction, which facilitates the development of downstream applications and method comparison. 2) Challenging cases categorized and strong baseline methods for rib labeling and anatomical centerline extraction. 3) Metrics for rib segmentation, labeling, and anatomical centerline extraction, providing a comprehensive method evaluation.

II. Related Works

A. Automatic Rib Analysis

Rib segmentation and labeling. A few studies have addressed rib segmentation and labeling [13], [14] before the era of deep learning, where rib tracing with initial seed point detection is the key method. Supervised deep learning-based segmentation [11] from CT volumes is robust as it adopts 3D-UNet [18] to learn hierarchical visual features from raw voxels. MDU-Net [19] is proposed to segment clavicles and ribs from CT scans, which combines multiscale feature fusion with the dense connection [20].

Rib anatomical centerline extraction. A few non-learning studies work on rib centerline extraction by modeling the ribs as elongated tubular structures and conducting rib voxel detection by structure tensor analysis [10], [21]. And rib tracing-based method is also introduced to centerline extraction [22]. There are also deep learning-based studies focusing on rib centerline extraction instead of full rib segmentation, *e.g.*, rib centerlines are extracted by applying morphological methods such as deformable template matching [15] and rib tracing method [12] to the rib cages detected by deep learning method.

B. Deep Learning Models for 3D CT Volumes

In this study, we represented CT scans as dense voxel grids and sparse point clouds for method comparison.

Voxel grids. 3D-UNet [18] is first introduced to work on sparsely annotated volumetric data. VoxSegNet [23] is further proposed as an effective volumetric method for 3D shape part segmentation, which extracts discriminative features encoding detailed information under limited resolution. PVCNN [24] and PointGrid [25] integrate representations of points and voxels to enhance feature extraction and model efficiency.

Point clouds. Deep learning for point cloud analysis [26] is pioneered by PointNet [27] and DeepSet [28]. Later studies also introduce sophisticated feature aggregation based on spatial graphs [29], [30] or attention [31]. In medical imaging scenarios, point cloud matching has been applied to 3D volumes [8] since 2014. The transformer mechanism [32] is further introduced for medical point cloud analysis [33], and point cloud-based methods are adapted to various medical applications such as nodule detection [34] and vessel reconstruction [35].

C. Semantic Segmentation-guided 2-Stage Methods

Considering the sparsity of ribs in CT volumes, we first perform foreground-background segmentation to roughly segment the ribs and label them by multi-class segmentation with a second model. The semantic segmentation-guided method is common in fine-grain classification tasks such as pedestrian detection [36], [37] where semantic segmentation is first performed to obtain complementary higher-level semantic features. A similar pipeline is also used in the medical scenario such as intracranial aneurysm segmentation [38], where vessel segments with aneurysms are detected from the whole CT scan, and segmented by a second model. This pipeline essentially tackles the sparsity issue and eases the follow-up tasks.

D. Skeletonization Methods

Besides the studies of rib anatomical centerline extraction, other skeletonization methods for elongated objects are potentially applicable to this topic.

Learning-based skeletonization. Most studies of learning-based skeletonization work on 2D images, *e.g.*, DeepFlux predicts a two-dimensional vector field to map scene points to extract the skeleton [39], and the skeleton can also be extracted by integrating image and segmentation to obtain complementary information [40]. For 3D skeleton extraction, the previous study utilizes normalized gradient vector flow on volume data [41], and most studies of 3D skeleton focus on human recognition and re-identification [42], *e.g.*, PointSkelCNN is proposed to extract 3D human skeleton from point clouds [43].

TEASAR method. The *Tree-structure Extraction Algorithm for Accurate and Robust Skeletons* (TEASAR) [17], [44] is originally proposed to skeletonize binary discretized 3D occupancy maps of tree-like structures, such as neurons [45]. The pipeline of the original TEASAR is summarized as follows: 1) first locate a root point on the rib volume, 2) and then serially trace the shortest path via a penalty field [46] to the most distant unvisited point. 3) After each passing, a circumscribing cube is applied to expand around the vertices in the path, marking the visited regions. 4) Repeat the process above until the whole volume is traversed.

TABLE I: **Data Division and Stats of *RibSeg v2* Dataset**. The table includes the number of total cases, individual ribs, cases with the incomplete rib cage, and unqualified cases for each subset. The unqualified cases refer to the cases that 1) miss annotations of labels or centerlines, and 2) have flaws in rib label annotations. The file names and details of abnormal cases are categorized into a dataset description file, which will be made available together with *RibSeg v2* dataset.

Subset	CT Scans	Individual Ribs	Incomplete Rib Cages	Unqualified Cases
Training	420	9,961	28	27
Development	80	1,780	13	9
Test	160	3,725	32	11

III. *RibSeg V2* Dataset

A. Dataset Overview

Most prior studies on rib segmentation or rib centerline extraction use small in-house datasets [19], which makes it inconvenient to conduct comparative studies and develop new methods. To address this issue, we previously developed the rib segmentation benchmark, *RibSeg* [16], by annotating CT scans from the *RibFrac* dataset [9]. *RibFrac* contains 660 cases, while *RibSeg* only contains 480 cases that are relatively easy to annotate. In this study, we extended *RibSeg* to a comprehensive benchmark for rib labeling and anatomical centerline extraction by 1) adding the 170 remaining cases with rib segmentation, 2) thoroughly labeling the cases except for the low-quality one, and 3) providing annotations of rib centerline for all cases. The resultant *RibSeg v2* dataset contains 660 cases, with 15,466 ribs in total. Considering the clinical practicality, we further categorize the cases that are hard to annotate as challenging cases. Fig. 1 gives an overview of the *RibSeg v2* dataset.

Data source. *RibSeg v2* Dataset uses the public computed tomography (CT) scans from the *RibFrac* dataset [9], an open dataset with 660 chest-abdomen CT scans for rib fracture segmentation, detection, and classification. The CT scans are saved in NIFTI (.nii) format with volume sizes of $N \times 512 \times 512$, where 512×512 is the size of CT slices, and N is the number of CT slices (typically $300 \sim 500$). Most cases are confirmed with complete rib cages (24 ribs) and manually annotated with at least one rib fracture by senior radiologists.

Dataset division and statistics. The data split of the *RibSeg v2* dataset is summarized in Tab. I: training set (420 cases), development set (80 cases), and test set (160 cases). The division of *RibSeg v2* training, development, and test sets are from those of the *RibFrac* dataset respectively, facilitating the development of downstream applications such as rib fracture detection. In Tab. I, we also report the number of cases with incomplete rib cages and unqualified cases. Specifically, the cases with incomplete cages only cover the upper chest-abdomen region, while the unqualified cases refer to the cases whose annotations are missed or contain potential flaws, including 4/4/3 (training/development/test) cases that miss annotations of labels or centerline due to the CT scans quality degradation, and 23/5/8 cases with label crossing in annotations. The file names and details of all the abnormal cases are categorized into a dataset description file, which will be made available together with the *RibSeg v2* dataset.

B. Dataset Annotation

Annotating rib labels and anatomical centerlines from CT scans is labor-intensive due to the elongated and oblique shape of ribs. To ease the workload and facilitate the annotation, we develop a morphological pipeline, and for abnormal cases where the method fails, we use a deep learning-based method with heavy post-processing. Based on the high-quality labels, we use a TEASAR-based method to extract anatomical centerlines. Each step contains manual checking and refinement, and final annotations are confirmed by senior radiologists.

Rib labeling. Rib labeling contains segmenting ribs from CT scans and labeling the segmentation. We describe the primary steps as follows. For each volume, we first filter out non-target voxels by thresholding at 200 HU [47] and then separate the ribs from the vertebra through morphological methods (*e.g.*, dilation, and erosion). For cases that remain parts of the clavicle and scapula, we manually locate and remove them according to the coordinates of their connected components [48], [49]. The resultant rib segmentation is labeled from top to bottom and left to right. This pipeline generates high-quality rib labels for most cases, and for the cases in which it fails, we turn to the deep learning-based method. Specifically, we train a customized PointNet++-based model on the annotated cases, utilize it to predict labeled segmentation, and post-process the results with morphological methods. All results are manually checked and refined in a human-in-the-loop procedure to ensure high quality. The cases beyond repair are denoted as unqualified in Tab. I, and categorized into a dataset description file, which will be made available together with the *RibSeg v2* dataset.

Rib anatomical centerline extraction. Based on the labeled rib segmentation, we extract the rib anatomical centerlines by implementing a variant of the TEASAR method. The resultant centerline could be tortuous since the rib components are hollow inside, and the sizes of the point set composing centerlines are different. Hence, we post-process the centerlines by smoothening and upsampling them to 500 points. The proposed pipeline can generate high-quality rib anatomical centerlines in most cases. For the failed cases, senior radiologists manually annotate their centerlines.

Manual proofreading. The abnormal cases, along with the pursuit of high annotation quality, incentivize us to perform laborious checking and refinement after the annotation stages. For instance, in quite a few cases, the floating ribs are too short or sparse that segmentation and centerlines vanish after the morphological procedure. Hence, we manually check and refine the annotation case by case. To recover and annotate missed ribs, we turn back to manually ensure the segmentation completeness by modifying the corresponding connected components voxel by voxel. To ensure high quality, all final segmentations and centerlines are manually checked, refined, and confirmed by senior radiologists based on visual assessment and consensus review.

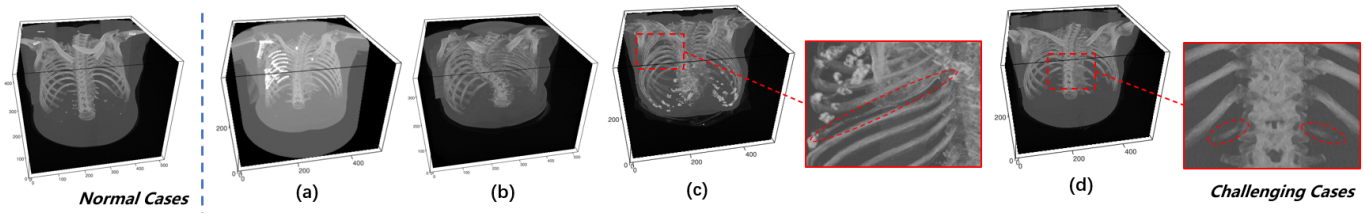

Fig. 2: **Challenging Cases**. 76 cases in *RibSeg v2* are categorized as challenging cases. a) The case suffers HU deviation, and the left 3rd∼9th ribs are connected by metal plates. b) The case contains serious scoliosis. c) The case contains serious fractures in the sternum and the left 5th rib. d) The floating ribs are abnormally short (the 12th pair of ribs).

C. Challenging Cases

We report specific challenges of rib labeling and centerline extraction by analyzing and categorizing the abnormal cases, whose modalities are relatively rare. In clinical, however, the diagnosis of these cases is time-consuming, while for normal cases, even computer-assisted intervention is less needed. Hence, the discussion and categorization of these cases are valuable. Based on our visual assessment, 99 cases in *RibSeg v2* are categorized as challenging cases (47/19/33 in training/development/test), which is contained in the dataset description file.

Challenging case categories. We categorized 4 challenging situations: 1) The adjacent bones are connected by the growing callus or metal implants, as depicted in Fig. 2 (a). Algorithmically, such cases will also cause morphological false merge, *i.e.*, a single connected component contains several ribs. 2) The cases with metal implants like Fig. 2 (a) also tend to suffer severe HU deviation, i.e., the HU value of the bone is higher/lower than their normal HU value. In such cases, the rib cage is wrapped by a 'noise shell', which is hard to filter. 3) The cases that are partly scanned or suffer severe bone lesions such as scoliosis in Fig. 2 (b) and fractures like Fig. 2 (c), which are hard to segment and label. 4) The floating ribs are missing or too vague to segment, as depicted in Fig. 2 (d), and there might also exist a third stubby little floating rib (the 13th pair of ribs).

IV. Methodology

The clinical significance of rib labeling and centerline extraction, along with the limitations of morphological methods, urge for a more robust method. Hence, we benchmark *RibSeg v2* by formulating rib labeling as a multi-class segmentation task and proposing a deep learning-based method for rib labeling, and a skeletonization-based method for centerline extraction.

A. Pipeline Overview

As depicted in Fig. 3, the pipeline is divided into three steps. 1) CT denoising: we first preprocessed the input CT scans to obtain the denoised CT volumes through morphological methods. 2) Point-based rib labeling: we converted the CT volume to point clouds and applied a 2-stage point-based method to first obtain the binary rib segmentation and then segment individual ribs with a second model. Then the resultant label prediction is post-processed for centerline extraction. 3) Rib centerline extraction: based on the labeled segmentation, the centerlines are extracted through a variant of the TEASAR method.

B. CT Denoising

Considering the sparsity of ribs in 3D volumes ($< 0.5\%$ voxels) and the high HU value of bones in CT scans ($>$ 200 HU), we filter the non-target parts of CT volumes in a coarse-to-fine manner. Specifically, we first filter the non-bone voxels roughly by setting a threshold of 200 HU on CT volumes, which is the common CT attenuation value for bones. Although the resultant binarized volumes may contain many noises covering the rib cage, we keep the noises and propagate the volumes to the model training procedure to improve model robustness. While in the inference stage, we remove most noises by sorting out and eliminating the connected components of small volumes. We denote such connected components-based denoise procedure as *Connected-Components-Denoising* (CCD). In Sec. V, we report that CCD is crucial to obtaining high-quality rib labels, especially for the cases suffering HU deviation where the roughly filtered volumes will contain a huge number of noises.

C. Point Cloud Baseline for Rib Labeling

Problem formulation. We formulate it as a 25-class part segmentation problem to segment and label ribs from CT scans (24 classes for 24 ribs and 1 class for other bones and backgrounds). However, in this scenario, the target parts (24 ribs) are extremely sparse in the input volume ($< 0.7\%$ voxels after HU thresholding), which is different from conventional part segmentation tasks such as PASCAL-Part [50] and PartNet [51], where the segmentation parts of target objects have a rather balanced distribution.

Frameworks. To alleviate the sparsity issue, we propose a 2-stage framework for rib labeling: 1) first perform binary segmentation to obtain ribs from CT scans, and 2) perform multi-class segmentation to segment individual ribs from binary segmentation. The label predictions are further post-processed by CCD to remove mis-segmented noise points for centerline extraction. For comparison, we also test the single-stage method which directly predicts rib labels from CT volumes via multi-class segmentation in Sec. V-A.

Data representation for CT scans. Most learning-based methods model CT scans as 3D volumes, and work on dense

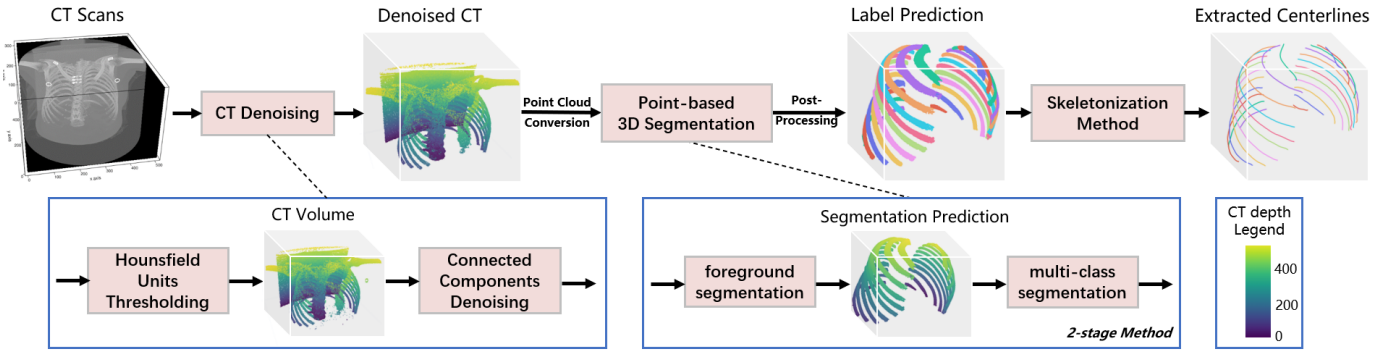

Fig. 3: **The Pipeline of Rib Labeling and Anatomical Centerline Extraction.** The pipeline is divided into 3 steps: 1) **CT Denoising:** CT scans are thresholded by HU value and denoised by *Connected-Components-Denoising* (CCD) to remove most non-bone voxels. 2) **Point-based Rib Labeling:** The denoised CT volume is converted to point clouds for a 2-stage point-based segmentation method, which first predicts binary rib segmentation, and then segments individual ribs with a second model. The label prediction is further denoised by CCD for centerline extraction. 3) **Rib Centerline Extraction:** The rib centerlines are extracted from the label prediction via a TEASAR-based skeletonization method. **Color:** For clear visualization, all binary volumes are colored by axial depth.

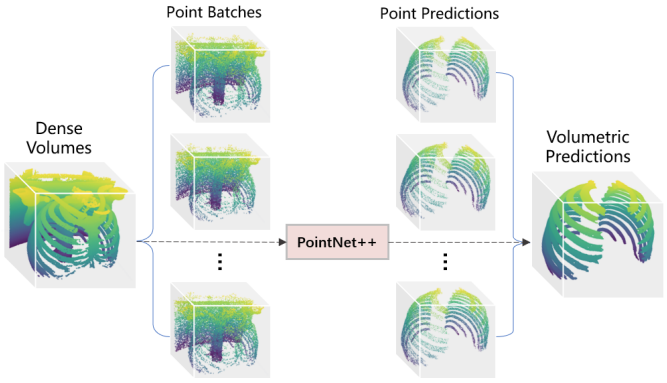

Fig. 4: **Point clouds generation from dense volumes**. The CT volumes are converted to a point cloud and divided into equally-sized point batches as input, and the multi-batch point predictions are concatenated into volumetric predictions.

voxel grids, which is computationally inefficient [12], [15]. To address the memory issue, we convert the dense 3D volumes to sparse point clouds [16] and adopt the fundamental and efficient PointNet++ [29] as the backbone model. For comparison, we also test the voxel-based method with 3D-Unet [18] in Sec. V-A.

Point clouds conversion. In Fig. 4, we take binary rib segmentation as an example to show the procedure of point cloud conversion. Specifically, we first convert the CT volumes into a dense point cloud and divide the point cloud into equally-sized batches of 30K points. For the batch with insufficient points, we 'ceil' it up by randomly sampling points from other batches, and applying majority voting to the prediction of repeated points. Finally, the point predictions of all batches are concatenated to obtain the voxel prediction.

Good practices for point-based representation. To alleviate the information loss during the point clouds conversion from the volume, as well as enriching the local geometric representation, we combine the absolute coordinates and the relative coordinates with respect to their neighboring points in the volume as model input, denoted as *Absolute-and-Relative-Positioning* (ARP), which is inspired by absolute and relative embedding module for point cloud representation enrichment [31]. Specifically, for each point, we first sample and group its neighboring points from the volume by ball query searching, and then we calculate the relative coordinates by subtracting their absolute coordinates. Besides, to enhance the robustness of the model, we also apply online *Points-Augmentation* (PA) which includes scaling, translation, and jittering. As a benchmark work, we don't want to complicate this study by introducing sophisticated architectures or designs with a limited performance boost, so we don't dig deep into the tricks. We report the enhancement of model performance by adapting ARP and PA in Tab. II and leave them as potential improvements which can also be generalized to other point cloud-based methods.

D. TEASAR-based Rib Anatomical Centerline Extraction

During the annotation procedure, we found that the TEASAR-based method can guarantee visually satisfying results as long as the segmentation is correctly labeled. Hence, we simply cascade the TEASAR-based method to the learning-based pipeline as an end-to-end baseline method for centerline extraction. Specifically, 1) we first apply morphological operations to eliminate the mislabeled regions, and obtain connected components of individual ribs according to the volume. 2) Then for each rib, a raster scan is applied to locate an arbitrary foreground voxel, and its furthest point is denoted as the root point (it lies on the end of the connected component). 3) By implementing the Euclidean distance transform, a penalty field is defined [46] to guide the centerline passing through the center of the rib volume. 4) Then Dijkstra's shortest path is implemented to derive the path from the root point to the most geodesically distant point from it (it lies on the other end of the connected component), and the resultant path is the

extracted rib centerline. 5) Finally, we smoothen the result and upsample the centerline to 500 points by linear interpolation for the convenience of evaluation. Although such a pipeline can achieve high-quality centerlines for most cases, the time consumption is not cheap (over 80s for a case), and it's sensitive to the quality of rib label predictions, which urges a more computationally efficient and robust method.

TABLE II: **Rib Labeling Metrics on *RibSeg v2* Test Set**. The metrics include average *Label-Dice* and *Label-Accuracy* of all / first / intermediate / twelfth rib pairs (A/F/I/T). Both 1-stage and 2-stage methods are included with voxel and point-based data representation. The point-based method includes different settings, where ARP, PA, and CCD denote *Absolute-and-Relative-Positioning*, *Points-Augmentation*, and *Connected-Components-Denoising*, respectively.

	Methods			$Dice_{avg}^{(L)}$	*Label Accuracy* (A/F/I/T)
Single-stage	Voxel-based Method			70.9%	71.3% / 78.9% / 73.3% / 58.9%
	Point-based Method			**75.9%**	**76.5%** / **86.2%** / **76.2%** / **68.3%**
	ARP	PA	CCD		
	-	✓	✓	72.8%	72.1% / 80.5% / 72.3% / 64.5%
	✓	✓	✓	**75.9%**	**76.5%** / **86.2%** / **76.2%** / **68.3%**
Two-stage	Voxel-based Method			82.8%	78.1% / 84.5% / 80.4% / 60.8%
	Point-based Method			**87.1%**	**86.4%** / **90.7%** / **87.3%** / **73.0%**
	ARP	PA	CCD		
	-	-	-	84.4%	80.6% / 87.2% / 81.5% / 62.7%
	-	-	✓	86.6%	84.8% / 89.7% / 85.6% / 70.0%
	-	✓	-	85.2%	83.0% / 87.5% / 84.2% / 63.9%
	-	✓	✓	86.8%	85.4% / **90.7%** / 86.0% / **73.0%**
	✓	-	-	84.6%	81.9% / 88.1% / 82.3% / 63.5%
	✓	-	✓	86.7%	83.7% / 90.1% / 84.3% / 68.4%
	✓	✓	-	85.8%	83.8% / 86.5% / 85.2% / 64.3%
	✓	✓	✓	**87.1%**	**86.4%** / 90.1% / **87.3%** / 72.6%

V. Experiments

A. Experiments on Rib labeling

1) Evaluation Metrics: We first define the *Label-Dice* [52], [53] of rib i as:

$$Dice_i^{(L)} = \frac{2 \cdot |y_i \cdot \hat{y}_i|}{|y| + |\hat{y}|}, \tag{1}$$

where y and $\hat{y}$ indicate the label prediction and ground truth, respectively. For quantitative analysis, we evaluate the performance by reporting the average *Label-Dice* of the 24 ribs, denoted as $Dice_{avg}^{(L)}$. While in qualitative analysis, we reflect the performance degradation by reporting the minimal *Label-Dice* amongst the 24 ribs, denoted as $Dice_{min}^{(L)}$. Moreover, we report the *Label-Accuracy* of individual ribs to evaluate the method's clinical applicability. Specifically, an individual rib i is counted as correctly labeled if $Dice_i^{(L)} > 0.7$, and the accuracy can be calculated with ease. Considering that labeling first and twelfth rib pairs tend to be more difficult as they are shorter and curvier than other ribs, we report the *Label-Accuracy* of all / first / intermediate / twelfth rib pairs, respectively.

2) Quantitative Analysis: We first evaluate the methods on *RibSeg v2* test set, comparing the 1-stage and 2-stage methods with different settings. As depicted in Tab. II, all 2-stage methods significantly outperform 1-stage methods. The interpretation is that the rib in the input volume of the 1-stage method is too sparse (16%) to provide sufficient features while the 2-stage method removes most background noises for the label prediction. Besides, the point-based method (PointNet++) outperforms the voxel-based method (3D-UNet), as it takes the shape of the whole volume as inputs, leading to rich feature representation. The method with ARP has a significant improvement on the 1-stage method where the rib is sparse in the input, with 3.1% higher Label Dice and 4.4% higher label accuracy. The methods with PA have a slight boost, while methods with CCD enjoy 1.3%∼2.2% higher label Dice values and at most 4.2% higher label accuracy. For inference speed, the point-based method is more efficient for taking sparse point clouds as inputs, as reported in Tab. III, whereas the point-based method is $12 \sim 15\times$ faster than the voxel-based method.

TABLE III: **Speed Comparison**. The table reports model forward time in seconds, including 1-stage and 2-stage methods with different data representations. Post-processing time is not included as it heavily depends on the implementation.

Methods		Forward Time (s)
Single-stage	Voxel-based 3D-UNet [18]	30.36
	Point-based PointNet++ [29]	2.46
Two-stage	Voxel-based 3D-UNet [18]	62.73
	Point-based PointNet++ [29]	5.12

3) Qualitative Analysis: For robustness analysis, we compare the inference results of normal and challenging cases, respectively, and visualize the predictions on challenging cases.

Analysis on challenging cases. To evaluate the robustness of the method, we tested the best model on all/normal/challenging cases from the test set, respectively, as reported in Tab. IV. The model enjoys a state-of-the-art performance in the normal cases while suffering a significant drop in the challenging cases: 24.8% lower on average *Label-Dice* and 34.4% lower on *Label-Accuracy*. To further investigate the performance degradation, we also analyze the minimal instance-wise *Label-Dice*, which is unsatisfying even in normal cases (79.7%) and suffers a huge drop by 50% in the challenging cases. Note that the minimal *Label-Dice* occurs on rib 12 or 24 (the 12th pair of ribs). It is exactly these challenging cases that are clinically time-consuming to diagnose, and the floating ribs with various lesions are algorithmically difficult to segment, hence, a more robust method of tackling the challenging cases is desired.

Visualization of performance degradation. As the metrics may not necessarily reflect the prediction quality in detail, we further visualize the results in challenging cases in Fig. 5. Specifically, for cases where adjacent ribs are connected by metal implants such as Fig. 5 (a), where left 2 and 3 ribs are connected to a pacemaker, the prediction suffers serious cross-labeling. For cases that suffer HU deviation like Fig. 5 (b), the prediction suffers cross-labeling and contains too many noises. For cases missing the floating ribs as Fig. 5 (c), the model seems to impose 24 classes of rib labels on the rib cage, and the false ribs (the 8th to 12th pairs of ribs) are mislabeled. While in Fig. 5 (d), the rib 5 left is severely

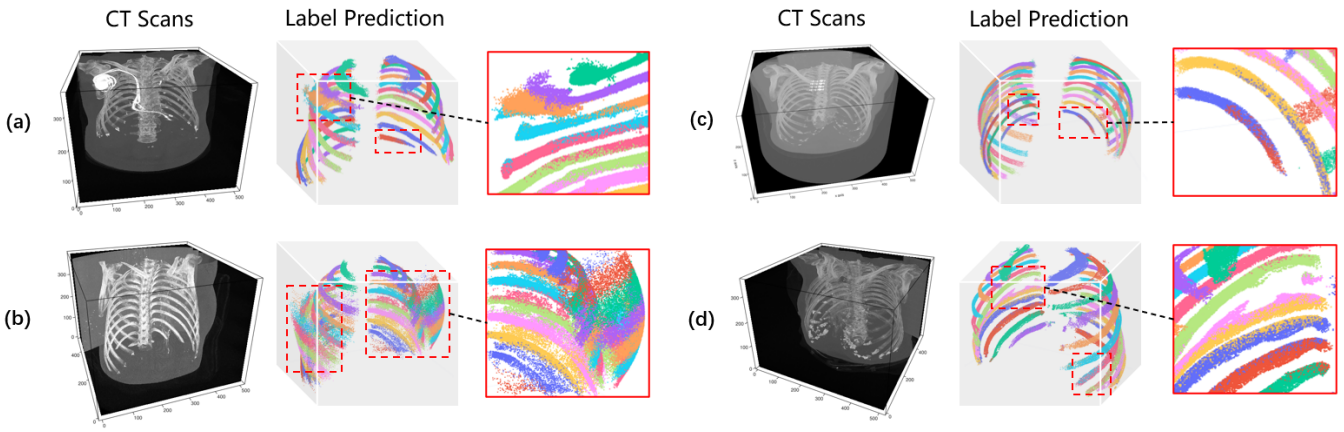

Fig. 5: **Visualization of Label Predictions on Challenging Cases**. (a) The case contains a pacemaker, which partly remains in the label prediction. (b) The case suffers HU deviation, leading to severe cross-labeling. (c) The case misses the 12th pair of floating ribs, causing the cross-labeling. (d) The case has a serious fracture in the left 5th rib, causing the cross-labeling.

TABLE IV: **Performance Comparison on Challenging Cases**. The baseline method is tested on all / normal / challenging cases from *RibSeg v2* test set, respectively. The metrics include average *Label-Dice*, minimal *Label-Dice*, and *Label-Accuracy* over all pairs of ribs.

Cases	$Dice_{avg}^{(L)}$	$Dice_{min}^{(L)}$	$LabelAccuracy$
All	87.1%	65.6%	86.4%
Normal	93.6%	79.7%	92.1%
Challenging	68.8%	14.2%	57.7%

damaged, also causing cross-labeling. In brief, despite the visually satisfying predictions in most cases, the performance degradation in a few challenging cases is huge. Considering the clinical practicality, it urges a more robust method for rib labeling with the ground truth provided by the *RibSeg v2* dataset.

B. Experiments on Rib Centerline Extraction

1) Evaluation Metrics: For centerline evaluation, we used three metrics to measure the quality of extracted centerline. We first report the Chamfer Distance [54] between the extracted centerline and the centerline annotation, denoted as *Line-line-Chamfer-Distance* (LLCD):

$$LLCD(\mathbf{L},\hat{\mathbf{L}}) = \frac{1}{|\mathbf{L}|}\sum_{y\in\mathbf{L}}\min_{\hat{y}\in\hat{\mathbf{L}}}\|y-\hat{y}\|_2+\frac{1}{|\hat{\mathbf{L}}|}\sum_{\hat{y}\in\hat{\mathbf{L}}}\min_{y\in\mathbf{L}}\|\hat{y}-y\|_2, \tag{2}$$

where $\mathbf{L}\subseteq\mathbb{R}^3$ and $\hat{\mathbf{L}}\subseteq\mathbb{R}^3$ are the extracted centerline and centerline annotations, respectively.

Another important quality for the centerline is whether it lies in the mid of the corresponding rib bone, to evaluate such position deviation, we report the Chamfer Distance of the extracted centerline with respect to the annotation of rib labels, denoted as *Line-seg-Chamfer-Distance* (LSCD):

$$LSCD(\mathbf{L},\hat{\mathbf{S}}) = \frac{1}{|\mathbf{L}|}\sum_{y\in\mathbf{L}}\min_{\hat{y}\in\hat{\mathbf{S}}}\|y-\hat{y}\|_2, \tag{3}$$

where $\mathbf{L}\subseteq\mathbb{R}^3$ and $\hat{\mathbf{S}}\subseteq\mathbb{R}^3$ are the extracted centerline and the annotation of rib labels, respectively.

Inspired by normalized surface dice [55], [56], we propose *Normalized-Line-Dice* (NLD) to evaluate the curvature deviation:

$$NLD(\mathbf{L},\hat{\mathbf{L}}) = \frac{|\mathbf{L}\cap B_{\hat{\mathbf{L}}}^{(\tau)}|+|\hat{\mathbf{L}}\cap B_{\mathbf{L}}^{(\tau)}|}{|\mathbf{L}|+|\hat{\mathbf{L}}|}, \tag{4}$$

where $\mathbf{L}\subseteq\mathbb{R}^3$ and $\hat{\mathbf{L}}\subseteq\mathbb{R}^3$ are the extracted centerline and centerline annotation, respectively, and $B_{\mathbf{L}}^{(\tau)}=\{y\in\mathbb{R}^3\mid\exists\,\tilde{y}\in\mathbf{L},\|y-\tilde{y}\|_2\le\tau\}$ denotes the surrounding region of the centerline $\mathbf{L}$ within tolerance distance τ. Here we take $\tau=7$, which is roughly the radius length of the cross surface for an approximate rib cylinder.

2) Quantitative Analysis: We evaluate the method based on LLCD, LSCD, and NLD, and compare the effects of smoothening post-process and CCD in Tab. V. As Chamfer Distance (CD) is not a numerically intuitive metric [57], we focus on NLD for quantitative analysis.

Metrics analysis. As reported in Tab. V, the method with CCD and smoothening performs best. Considering that the extracted centerline could be tortuous in the middle while the annotation is manually refined and smoothened, we also report the effect of smoothening post-process. With smoothening, the extracted centerline fits the annotation better, and has a huge boost on NLD (over 3%), while the performance on LSCD suffers a slight drop. An interpretation is that the smoothening will also deviate the centerline from the middle of the rib, which leads to the performance drop.

Analysis on challenging cases. To evaluate the robustness, we test these settings on all/normal/challenging cases from the test set, respectively. As reported in Tab. V, the methods suffer $7\%\sim29.9\%$ drop of NLD in challenging cases compared to normal cases. While for LLCD and LSCD, the performance degradation in challenging cases is significant. Considering the high clinical importance of rib centerlines, a more robust method is desirable.

3) Qualitative Analysis: As the metrics may not necessarily reflect the prediction quality, We visualize the centerline

TABLE V: **Rib Centerline Extraction Metrics on *RibSeg v2* test set**. The results on all / normal / challenging (A/N/C) cases are reported, respectively ($\frac{A}{N/C}$). The metrics include *Line-line-Chamfer-Distance* (LLCD), *Line-seg-Chamfer-Distance* (LSCD), and *Normalized-Line-Dice* (NLD). Moreover, different settings are also included, where CCD denotes *Connected-Components-Denoising*.

TEASAR Method		LLCD	LSCD	NLD
Smoothen	CCD			
-	-	2709.8 1753.8 / 6174.9	451.6 293.4 / 1020.8	64.9% 67.8% / 54.8%
✓	-	2802.8 1756.5 / 6595.6	466.2 313.1 / 1018.6	68.1% 71.3% / 57.1%
-	✓	892.6 106.9 / **3740.5**	**229.6** **102.2** / **691.8**	78.1% 84.2% / 56.4%
✓	✓	**885.1** **96.1** / 3745.2	238.2 108.3 / 707.5	**81.7%** **88.4%** / **58.5%**

extraction results for more intuitive and detailed analysis.

Visualization analysis. The quality of centerline extraction also heavily depends on the label prediction, i.e., in most cases like Fig. 6 (a), the label prediction with high accuracy can guarantee an accurate centerline with the TEASAR-based method. For some cases like Fig. 6 (b), where the label prediction contains a few cross-labeling regions, the method can easily eliminate the mislabeled regions and obtain a visually satisfying result. However, in some cases where the label prediction suffers a huge accuracy drop, *e.g.*, there are too many mislabeled regions that one centerline may cross several ribs or several centerlines may align to the same rib. As depicted in Fig. 6 (c) where a floating rib is missing, the performance drop on rib labeling heavily affected the TEASAR-based method, and the result contains 3 overlap regions where the centerlines are misaligned. Moreover, as mentioned in Sec. III-C, our method also fails for cases suffering severe rib fracture, where a single rib is broken into several parts, which will affect the result of distance transformation, and resulted in messy curve segments. Such abnormal cases also urge a more robust method for centerline extraction, with centerline annotations provided by *RibSeg v2*.

4) Discussion on annotation and metrics: During the experiment, we noticed that even though the centerline generated by our method perfectly lies in the middle of the rib segmentation by visual assessment, it might not necessarily have a high LSCD value. The interpretation is that the annotations of rib anatomical centerline are not perfect because manual annotation of the centerline is a naturally hard task for humans, and even for well-trained radiologists, it's difficult to locate endpoints that exactly lie in the center of the ribs before connecting them as the centerline curve. However, although being geometrically imperfect, such manually confirmed centerline annotations are still clinically pragmatic and valuable. Moreover, although the proposed Chamfer Distance-based metrics (LSCD and LLCD) could measure the centerline quality given the rib segmentation, they are not numerically intuitive, which also urges more intuitive metrics to evaluate the quality of centerlines.

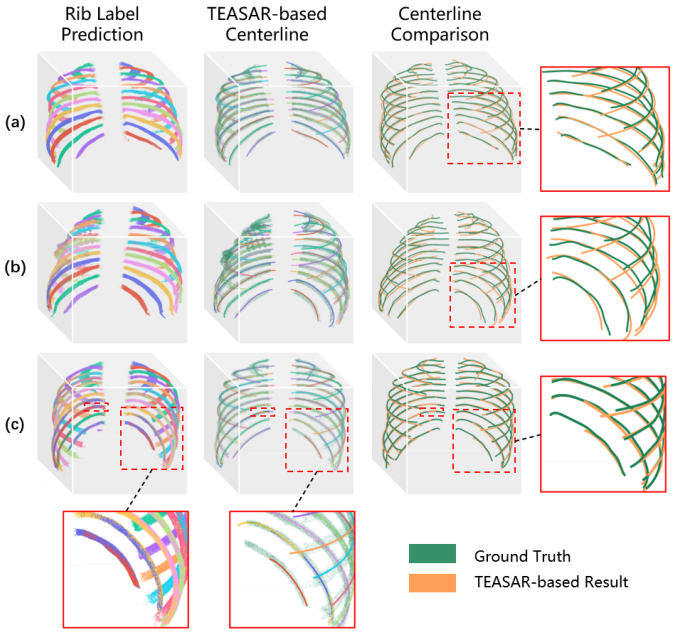

Fig. 6: **Visualization of Centerline Extraction Results**. This figure includes the visualization of rib label prediction, the result of the TEASAR-based method, and a comparison between the extracted centerline and ground truth. To show the relative position of the centerline with respect to rib segmentation, the annotation of rib segmentation is also visualized as background (in the visualization of extracted centerline and centerline ground truth). (a) The case has a flawless label prediction, and its centerline is also flawless. (b) The label prediction of this case has a few cross-labeling, and it is alleviated by the morphological process during the centerline extraction, the resultant centerline is also visually satisfying. (c) The floating ribs of this case are missing, which leads to cross-labeling in the label prediction, and the resultant centerline has 3 misaligned regions.

C. Experiment Settings

In terms of the training settings of rib labeling and rib segmentation, point batches are downsampled to 30,000 points per case in consideration of the trade-off between batch size and input size. The online point augmentation includes scaling, translation, and jittering. The Adam optimizer is adopted to train both models for 250 epochs with a batch size of 4, and a combination of cross-entropy (CE) and Dice loss as the loss function. The initial learning rate was set at 0.001 and decayed by a factor of 0.5 every 20 epochs, with a lower bound of 10^{-5}. In the training stage of the 2-stage method, the input of the segmentation task is point sets of 30,000 downsampled from the binary CT volume, while the input of the labeling task is point sets downsampled from the predictions of segmentation. The inference was conducted with the implementation of PyTorch 1.7.1 and Python 3.9, on a machine with a single NVIDIA Tesla P100 GPU with Intel(R) Xeon(R) CPU @ 2.20 GHz and 150 G memory.

VI. Conclusion

We built the *RibSeg v2* dataset, which is the first open dataset for rib labeling and anatomical centerline extraction. Besides, we explored the challenges of rib labeling and centerline extraction in detail and benchmark *RibSeg v2* with a strong pipeline including a deep learning-based method for rib labeling and a TEASAR-based method for centerline extraction. Then we compared data representations of CT scans as dense voxel grids and sparse point clouds and provided a comprehensive analysis of the abnormal cases where the method might fail. Besides, we also proposed various metrics for each task to provide comprehensive evaluations. Moreover, we also benchmark the rib segmentation task in the ablation study and analyzed all abnormal cases which could lead to performance degradation of the model. Finally, by detailed quantitative and qualitative analysis of the challenging cases, we explored the key challenges of each task, which are valuable to guide future studies on this topic.

The dataset and proposed method show the potential to be clinically applicable, also enhancing the efficiency and performance of downstream tasks, such as the diagnosis of rib fractures and bone lesions. Besides, considering the differences from standard medical image datasets [58], [59] with pixel/voxel grids, the elongated shapes and oblique poses of ribs enable the *RibSeg v2* dataset to serve as a benchmark for curvilinear structures and geometric deep learning. However, there also remains a limitation in this study. For rib anatomical centerline extraction, we apply a TEASAR-based method to extract centerlines from the prediction of rib labels, which is sensitive to rib labeling errors in abnormal cases. Hence, considering the clinical significance of rib anatomical centerline extraction, a more robust method will be favorable.

VII. Acknowledgment

This work was supported by the Science and Technology Planning Project of Shanghai Science and Technology Commission (22Y11910700), Shanghai Key Lab of Forensic Medicine, Key Lab of Forensic Science, Ministry of Justice, China (Academy of Forensic Science) (KF202113), National Natural Science Foundation of China (61976238), and Shanghai "Rising Stars of Medical Talent" Youth Development Program "Outstanding Youth Medical Talents" (SHWJRS [2021]-99).

References

[1] M. Sirmali *et al.*, "A comprehensive analysis of traumatic rib fractures: morbidity, mortality and management," *European Journal of Cardio-Thoracic Surgery*, vol. 24, no. 1, pp. 133–138, 2003.

[2] A. Mansoor, U. Bagci, Z. Xu, B. Foster, K. N. Olivier, J. M. Elinoff, A. F. Suffredini, J. K. Udupa, and D. J. Mollura, "A generic approach to pathological lung segmentation," *IEEE Transactions on Medical Imaging*, vol. 33, pp. 2293–2310, 2014.

[3] A. A. Fokin, J. Wycech, M. Crawford, and I. Puente, "Quantification of rib fractures by different scoring systems." *The Journal of surgical research*, vol. 229, pp. 1–8, 2018.

[4] H. Wang, J. Bai, and Y. Zhang, "A relative thoracic cage coordinate system for localizing the thoracic organs in chest ct volume data," *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 3257–3260, 2005.

[5] H. Shen and M. Shao, "A thoracic cage coordinate system for recording pathologies in lung ct volume data," *2003 IEEE Nuclear Science Symposium. Conference Record (IEEE Cat. No.03CH37515)*, vol. 5, pp. 3029–3031 Vol.5, 2003.

[6] H. Ringl, M. Lazar, M. Töpker, R. Woitek, H. Prosch, U. Asenbaum, C. Balassy, D. Toth, M. Weber, S. Hajdu *et al.*, "The ribs unfolded-a ct visualization algorithm for fast detection of rib fractures: effect on sensitivity and specificity in trauma patients," *European radiology*, vol. 25, no. 7, pp. 1865–1874, 2015.

[7] G. Bier, C. Schabel, A. E. Othman, M. N. Bongers, J. Schmehl, H. Ditt, K. Nikolaou, F. Bamberg, and M. Notohamiprodjo, "Enhanced reading time efficiency by use of automatically unfolded ct rib reformations in acute trauma." *European journal of radiology*, vol. 84 11, pp. 2173–80, 2015.

[8] L. I. Abe, Y. Iwao, T. Gotoh, S. Kagei, R. Y. Takimoto, M. S. G. Tsuzuki, and T. Iwasawa, "High-speed point cloud matching algorithm for medical volume images using 3d voronoi diagram," *2014 7th International Conference on Biomedical Engineering and Informatics*, pp. 205–210, 2014.

[9] L. Jin, J. Yang, K. Kuang, B. Ni, Y. Gao, Y. Sun, P. Gao, W. Ma, M. Tan, H. Kang *et al.*, "Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet," *EBioMedicine*, vol. 62, p. 103106, 2020.

[10] J. Staal, B. van Ginneken, and M. A. Viergever, "Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data," *Medical image analysis*, vol. 11, no. 1, pp. 35–46, 2007.

[11] M. Wu, Z. Chai, G. Qian, H. Lin, Q. Wang, L. Wang, and H. Chen, "Development and evaluation of a deep learning algorithm for rib segmentation and fracture detection from multicenter chest ct images." *Radiology. Artificial intelligence*, vol. 3 5, p. e200248, 2021.

[12] M. Lenga, T. Klinder, C. Bürger, J. von Berg, A. Franz, and C. Lorenz, "Deep learning based rib centerline extraction and labeling," in *MICCAI International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, 2018, pp. 99–113.

[13] H. Shen, L. Liang, M. Shao, and S. Qing, "Tracing based segmentation for the labeling of individual rib structures in chest ct volume data," in *MICCAI*. Springer, 2004, pp. 967–974.

[14] T. Klinder, C. Lorenz, J. Von Berg, S. P. Dries, T. Bülow, and J. Ostermann, "Automated model-based rib cage segmentation and labeling in ct images," in *MICCAI*. Springer, 2007, pp. 195–202.

[15] D. Wu, D. Liu, Z. Puskas, C. Lu, A. Wimmer, C. Tietjen, G. Soza, and S. K. Zhou, "A learning based deformable template matching method for automatic rib centerline extraction and labeling in ct images," in *CVPR*. IEEE, 2012, pp. 980–987.

[16] J. Yang, S. Gu, D. Wei, P. Hanspeter, and B. Ni, "Ribseg dataset and strong point cloud baselines for rib segmentation from ct scans," in *MICCAI*, 2021, pp. 611–621.

[17] M. Sato, I. Bitter, M. Bender, A. Kaufman, and M. Nakajima, "Teasar: tree-structure extraction algorithm for accurate and robust skeletons," in *Proceedings the Eighth Pacific Conference on Computer Graphics and Applications*, 2000, pp. 281–449.

[18] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *MICCAI*. Springer, 2016, pp. 424–432.

[19] W. Wang, H. Feng, Q. Bu, L. Cui, Y. Xie, A. Zhang, J. Feng, Z. Zhu, and Z. Chen, "Mdu-net: A convolutional network for clavicle and rib segmentation from a chest radiograph," *Journal of Healthcare Engineering*, vol. 2020, 2020.

[20] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[21] S. Aylward and E. Bullitt, "Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 61–75, 2002.

[22] S. Ramakrishnan, C. Alvino, L. Grady, and A. Kiraly, "Automatic three-dimensional rib centerline extraction from ct scans for enhanced visualization and anatomical context," in *Medical Imaging 2011: Image Processing*, vol. 7962, 2011, p. 79622X.

[23] Z. Wang and F. Lu, "Voxsegnet: Volumetric cnns for semantic part segmentation of 3d shapes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 2919–2930, 2020.

[24] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3d deep learning," *ArXiv*, vol. abs/1907.03739, 2019.

[25] T. Le and Y. Duan, "Pointgrid: A deep network for 3d shape understanding," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9204–9214, 2018.

[26] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.

[28] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *NIPS*, 2017.

[29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NIPS*, 2017.

[30] J. Liu, B. Ni, C. Li, J. Yang, and Q. Tian, "Dynamic points agglomeration for hierarchical point sets learning," in *ICCV*, 2019, pp. 7546–7555.

[31] J. Yang, Q. Zhang, B. Ni, L. Li, J. Liu, M. Zhou, and Q. Tian, "Modeling point clouds with self-attention and gumbel subset sampling," in *CVPR*, 2019, pp. 3323–3332.

[32] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[33] J. Yu, C. Zhang, H. Wang, D. Zhang, Y. Song, T. Xiang, D. Liu, and W. T. Cai, "3d medical point transformer: Introducing convolution to attention networks for medical point cloud analysis," *ArXiv*, vol. abs/2112.04863, 2021.

[34] I. Drokin and E. Ericheva, "Deep learning on point clouds for false positive reduction at nodule detection in chest ct scans," in *AIST*, 2020.

[35] A. Banerjee, F. Galassi, E. Zacur, G. L. D. Maria, R. P. Choudhury, and V. Grau, "Point-cloud method for automated 3d coronary tree reconstruction from multiple non-simultaneous angiographic projections," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1278–1290, 2020.

[36] T.-R. Liu and T. Stathaki, "Faster r-cnn for robust pedestrian detection using semantic segmentation network," *Frontiers in Neurorobotics*, vol. 12, 2018.

[37] Y. Yu, Y. Makihara, and Y. Yagi, "Pedestrian segmentation based on a spatio-temporally consistent graph-cut with optimal transport," *IPSJ Transactions on Computer Vision and Applications*, vol. 11, pp. 1–17, 2019.

[38] X. Yang, D. Xia, T. Kin, and T. Igarashi, "A two-step surface-based 3d deep learning pipeline for segmentation of intracranial aneurysms," 2020.

[39] Y. Wang, Y. Xu, S. Tsogkas, X. Bai, S. J. Dickinson, and K. Siddiqi, "Deepflux for skeletons in the wild," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5282–5291, 2019.

[40] X. Liu, P. Lyu, X. Bai, and M.-M. Cheng, "Fusing image and segmentation cues for skeleton extraction in the wild," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1744–1748, 2017.

[41] S. M. Yoon, C. Malerczyk, and H. Graf, "3d skeleton extraction from volume data based on normalized gradient vector flow," 2009.

[42] H. Rao, S. Wang, X. Hu, M. Tan, Y. Guo, J. Cheng, B. Hu, and X. Liu, "A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[43] H. Qin, S. Zhang, Q. Liu, L. Chen, and B. Chen, "Pointskelcnn: Deep learning-based 3d human skeleton extraction from point clouds," *Computer Graphics Forum*, vol. 39, 2020.

[44] T. Zhao, D. J. Olbris, Y. Yu, and S. M. Plaza, "Neutu: Software for collaborative, large-scale, segmentation-based connectome reconstruction," *Frontiers in Neural Circuits*, vol. 12, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncir.2018.00101

[45] A. Fornito, A. Zalesky, and M. Breakspear, "Graph analysis of the human connectome: Promise, progress, and pitfalls," *NeuroImage*, vol. 80, 04 2013.

[46] I. Bitter, A. Kaufman, and M. Sato, "Penalized-distance volumetric skeleton algorithm," *IEEE Transactions on Visualization and Computer Graphics*, vol. 7, no. 3, pp. 195–206, 2001.

[47] M. H. Lev and R. G. Gonzalez, "17 – ct angiography and ct perfusion imaging," 2002.

[48] A. Rosenfeld and J. Pfaltz, "Pfalz, j.l.: Sequential operations in digital picture processing. journal of the acm 13(4), 471-494," *J. ACM*, vol. 13, pp. 471–494, 10 1966.

[49] K. Wu, E. Otoo, and K. Suzuki, "Two strategies to speed up connected component labeling algorithms," 01 2005.

[50] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. L. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1979–1986, 2014.

[51] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918, 2019.

[52] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[53] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, 2015.

[54] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471, 2017.

[55] S. Nikolov, S. Blackwell, R. Mendes, J. Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'souza, S. Moinuddin, K. Sullivan, D. Consortium, H. Montgomery, G. Rees, R. Sharma, and O. Ronneberger, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," 09 2018.

[56] K. Zou, S. Warfield, A. Bharatha, C. Tempany, M. Kaus, S. Haker, W. Wells, F. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic radiology*, vol. 11, pp. 178–89, 02 2004.

[57] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin, "Density-aware chamfer distance as a comprehensive metric for point cloud completion," *ArXiv*, vol. abs/2111.12702, 2021.

[58] M. Antonelli, A. Reinke, S. Bakas *et al.*, "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735*, 2021.

[59] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *ISBI*, 2021.