This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313627

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

Liang Jin, Shixuan Gu, Donglai Wei, Jason Ken Adhinarta, Kaiming Kuang, Yongjie Jessica Zhang, Hanspeter Pfister, Bingbing Ni, Jiancheng Yang, and Ming Li

***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.**

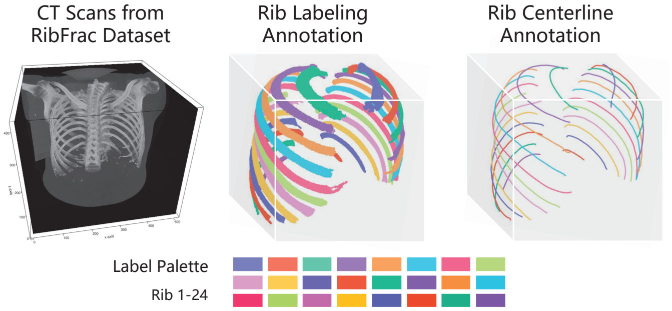

Fig. 1: ***RibSeg v2* Dataset**. *RibSeg v2* extends the annotations of 660 CT scans from the existing *RibFrac* dataset [12], containing labeled rib segmentation and anatomical centerlines. The color palette indicates the label assigned to each rib.

***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 to facilitate 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], [3], bone abnormality quantification [4] and pediatric spinal deformities [5], [6]. Based on rib anatomical centerlines, internal coordinate systems can localize organs for surgery planning and postoperative evaluation [7], as well as registering pathologies such as lung nodules [8]. Moreover, automatic rib labeling and centerline extraction is the key to developing visualization tools of unfolded rib cages [9]–[11], significantly reducing the burden of rib interpretation for clinicians.

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

This work was supported by 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), Shanghai 'Rising Stars of Medical Talent' Youth Development Program 'Outstanding Youth Medical Talents' (SHWJRS [2021]-99), Emerging Talent Program (XXRC2213) and Leading Talent Program (LJRC2202) of Huadong hospital, and Excellent Academic Leaders of Shanghai. This work was supported in part by National Science Foundation of China (U20B2072, 61976137) and Shanghai Jiao Tong University Medical Engineering Cross Research (YG2021ZD18) (Corresponding authors: Ming Li, Jiancheng Yang, and Bingbing Ni.)

Liang Jin and Shixuan Gu contributed equally as co-first authors.

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

Shixuan Gu is with Carnegie Mellon University, PA, USA, and also with Harvard University, MA, USA (shixuangu@g.harvard.edu).

Donglai Wei and Jason Ken Adhinarta are with Boston College, MA, USA (donglai.wei@bc.edu, jason.adhinarta@bc.edu).

Kaiming Kuang is with University of California San Diego, CA, USA, and also with Dianei Technology, Shanghai, China (kakuang@ucsd.edu).

Yongjie Jessica Zhang is with Carnegie Mellon University, PA, USA (jessicaz@andrew.cmu.edu).

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

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

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

Ming Li is with Radiology Department, Huadong Hospital, affiliated to Fudan University, Shanghai, China, and also with Institute of Functional and Molecular Medical Imaging, Shanghai, China (e-mail: minli77@163.com).

metal implant, which is hard to label. Furthermore, extraction of the rib centerline is subject to image noise, artifacts, and the quality of rib labels.

Previous studies on this topic only focus on rib segmentation [13] and trivialize rib labeling as a counting process [14], [15], which underestimates the challenges of false merging and neglects that anatomical labeling of ribs is clinically more desirable. Tracing-based rib segmentation and centerline extraction methods are highly sensitive to initially detected seed points and vulnerable to local ambiguities [16], [17]. Although the deep learning-based method is robust as it learns hierarchical visual features from raw voxels [18], 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 such as image-based modeling and simulations [19], [20].

To tackle these problems, we first develop a benchmark for rib labeling and anatomical centerline extraction, named *RibSeg v2*, including manually inspected annotations of 660 chest-abdomen CT scans (15,466 individual ribs) from *RibFrac* dataset [12]. In addition, we formulate rib labeling as the task of segmenting ribs from CT scans and labeling the binary segmentation into 24 instances, and benchmark the *RibSeg* with a pipeline including a deep learning-based method for rib labeling and skeletonization-based [21] methods for rib anatomical centerline extraction. We further compared the data representations of CT scans as dense voxel grids and sparse point clouds, respectively, and proposed various metrics for each task to perform comprehensive evaluations.

This study is extended from our *RibSeg* (v1) previously presented at MICCAI [22], where we introduced a binary rib segmentation benchmark with 490 CT scans from the RibFrac dataset [12]. In this study, we extend *RibSeg* to a comprehensive benchmark for *rib labeling* and *anatomical centerline extraction* by adding 1) the 170 remaining cases with binary rib segmentation, which are hard to be segmented by the semi-auto method [22], 2) rib labels (1-24) except for the 6 unqualified cases, and 3) annotations of rib anatomical centerlines except for the 6 unqualified cases. All annotations are manually checked, and CT scans that are hard to annotate are categorized into challenging cases by 2 junior radiologists based on visual assessment. Besides, we extended the previous method for binary rib segmentation to rib labeling and anatomical centerline extraction. Finally, by detailed quantitative and qualitative analysis of the challenging cases, we explored the key challenges of each task, which are valuable to facilitate future studies on this topic.

II. Related Works

A. Automatic Rib Analysis

Rib segmentation and labeling. A few studies have addressed rib segmentation and labeling [16], [17] before the era of deep learning, where rib tracing with initial seed point detection is the key method. Supervised deep learning-based segmentation [14] from CT volumes is robust, as it adopts 3D-UNet [23] to learn hierarchical visual features from raw voxels. MDU-Net [24] is proposed to segment clavicles and ribs from CT scans, which combines multiscale feature fusion with the dense connection [25].

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 [13], [26]. Rib tracing-based method is also introduced to centerline extraction [27]. 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 [18] and rib tracing method [15] to the rib cages detected by deep learning method.

B. Deep Learning Models for 3D CT Volumes

Most studies model CT scans as 3D volumes, and work on dense voxel grids, which is computationally expensive. In this study, we represented CT scans as dense voxel grids and sparse point clouds, respectively, for the method comparison.

Voxel grids. 3D-UNet [23] is first introduced to work on sparsely annotated volumetric data. VoxSegNet [28] is further proposed as an effective volumetric method for 3D shape part segmentation, which extracts discriminative features encoding detailed information under limited resolution. PVCNN [29] and PointGrid [30] integrate representations of points and voxels to enhance feature extraction and model efficiency.

Point clouds. Deep learning for point cloud analysis [31] is pioneered by PointNet [32] and DeepSet [33]. Later studies also introduce sophisticated feature aggregation based on spatial graphs [34], [35] or attention [36]. In medical imaging scenarios, point cloud matching has been applied to 3D volumes [11] since 2014. The transformer mechanism [37] is further introduced for medical point cloud analysis [38], and point cloud-based methods are adapted to various medical applications such as nodule detection [39] and vessel reconstruction [40].

C. Semantic Segmentation-guided Methods

The extreme foreground-background imbalance and data sparsity are common challenges of part segmentation tasks in (bio)medical domains. The semantic segmentation-guided method is a common solution, which is also widely used in fine-grain classification tasks such as pedestrian detection [41], [42] where semantic segmentation is first performed to obtain complementary higher-level semantic features. In this study, 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. A similar pipeline is also used in the medical scenario, such as intracranial aneurysm segmentation [43], where vessel segments with aneurysms are detected from the whole CT scan, and segmented by a second model. This pipeline essentially addresses the sparsity issue and eases follow-up tasks.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3313627

D. Skeletonization Methods

In this study, rib anatomical centerline is extracted from well-segmented rib labels, which can be essentially formulated as a skeletonization task for elongated objects.

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 [44], and the skeleton can also be extracted by integrating image and segmentation to obtain complementary information [45]. For 3D skeleton extraction, the previous study utilizes normalized gradient vector flow on volume data [46], and most studies of 3D skeleton focus on human recognition and re-identification [47], *e.g.*, PointSkelCNN is proposed to extract 3D human skeleton from point clouds [48].

Voxel-based TEASAR method. The *Tree-structure Extraction Algorithm for Accurate and Robust Skeletons* (TEASAR) [21], [49] is originally proposed to skeletonize binary discretized 3D occupancy maps of tree-like structures, such as neurons [50], [51]. 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 [52] 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.

Point-based L1-medial skeletonization. L1-medial skeletonization [53] is well-known as a state-of-the-art method to extract curve skeleton for point clouds. It used L1-median as a robust global center of the point cloud, and by adapting L1-medians locally to a point set represeneting a 3D shape, the resultant 1D structure can serve as a localized center of the shape, *i.e.*, the centerline.

III. *RibSeg V2* Dataset

A. Dataset Overview

Most prior studies on rib segmentation or rib centerline extraction use small in-house datasets [24], which makes it inconvenient to conduct comparative studies and develop new methods. To address this issue, we developed the *RibSeg* (v2) dataset containing 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 [12], 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.

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 these 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	0
Development	80	1,780	13	6
Test	160	3,725	32	0

B. Data 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 to obtain rib label segmentation [22]. Based on the high-quality labels, we apply skeletonization methods to extract anatomical centerlines. For abnormal cases where the pipeline fails, we first manually annotate the rib centerlines, and then apply a series of morphological operations to obtain rib label segmentation based on the upsampled centerlines. Each step contains manual checking and refinement, and final annotations are confirmed by 2 junior radiologists and 1 senior radiologist with a human-in-the-loop procedure.

Rib labeling. Rib labeling contains segmenting ribs from CT scans and labeling the segmentation. To generate binary segmentation, we initially adopted the semi-automatic pipeline [22] introduced in *RibSeg* (v1) and successfully produced 519 cases, including 480 cases from *RibSeg* (v1). We describe the primary steps as follows: For each volume, we first filter out non-target voxels by thresholding at 200 HU [54] 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 [55], [56]. The resultant rib segmentation is then labeled from top to bottom and left to right.

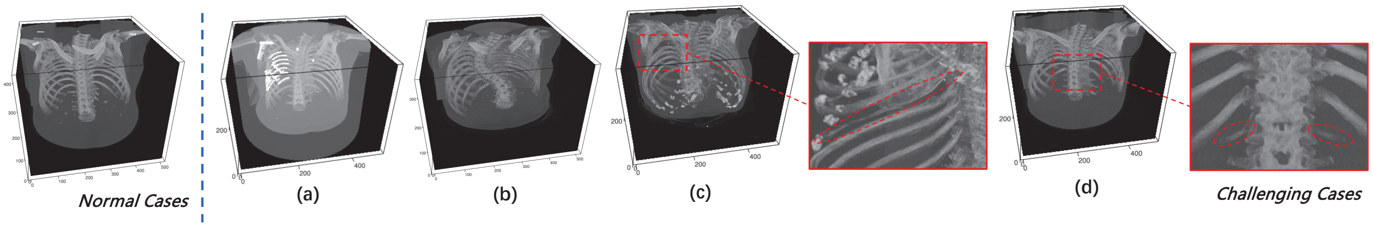

Fig. 2: **Challenging Cases**. 76 cases in *RibSeg v2* are categorized as challenging cases. a) The case suffers HU deviation, and the left 3rd∼9th ribs are connected by metal plates. b) The case contains serious scoliosis. c) The case contains serious fractures in the sternum and the left 5th rib. d)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

For the rest 127 cases where it fails, we turn to a centerline-based pipeline. 1) We first have 1 senior radiologist manually annotate the centerlines. Specifically, the radiologist would annotate 10 points within each rib through meticulous inspection of the CT slices, and the points are further interpolated and smoothened into a 3D curve of 500 points as the centerline. 2) For each rib, we dilate its centerline as a mask and then take the overlap of its mask and CT volume as the rough segmentation. 3) For the resultant rough segmentation, we filter the noise voxels by keeping the largest connected components. In cases where severe fracture breaks the rib into multiple volumes, we adjust the number of the connected components to be kept based on the visual assessment of the resultant segmentation. 4) To ensure the completeness of annotation, we repeatedly dilate the rib segmentation, take the overlap of it and CT volume, and filter the noise voxels.

This pipeline generates high-quality rib labels for 121 cases, and all results are manually checked and refined in a human-in-the-loop procedure to ensure high quality. The rest 6 cases, however, are partly scanned and only cover the middle part of the ribcage, which makes it impossible to label the ribs. Hence, we only provide the annotations of binary segmentation and unlabeled centerline for these cases, and denote them as unqualified in Tab. I, and categorized into a dataset description file, which will be made available together with the *RibSeg v2* dataset. Note that these cases are primarily used for lung inspection, and do not contain rib fractures. We previously incorporated these cases into the evaluation and test sets of the *RibFrac* dataset in order to assess the model's robustness.

Rib anatomical centerline extraction. Annotating a 3D centerline by manually inspecting 2D slices is a naturally challenging task even for well-trained radiologists. This is because it is nearly impossible to locate endpoints that lie precisely in the center of the rib based on 2D perception, as the thickness of the rib is difficult to assess. Even for the 127 cases of centerlines that are manually annotated by radiologists, they might deviate from the center of the ribs. Hence, for all 660 cases, based on the manually confirmed rib segmentation, we extract the rib anatomical centerlines by implementing 1) a variant of the voxel-based TEASAR method, and 2) a point-based L1-medial skeletonization method. For each rib, we compare the results of these 2 skeletonization methods (and manual annotation if contains) based on both visual assessment and numerical analysis.

Specifically, the centerlines generated by skeletonization methods 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. Then we first evaluated all the results based on visual assessment. When two skeletonization-based results (and manual annotation) are visually similar, we calculated the Chamfer distance V-B.1 between the centerlines and rib segmentation as a reference for the deviation of the centerline from the actual centerline of the rib. Then we selected the one with the lowest distance as the finalized centerline. And for the ribs where skeletonization-based methods failed, we used the smoothened manual annotations as the finalized centerline.

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 the segmentation vanished 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 2 junior radiologists and 1 senior radiologist based on visual assessment and consensus review. The total time for segmentation annotation and refinement is about $>$ 400 hours, and the centerline annotation and refinement takes about $>$ 200 hours.

Clinical feasibility evaluation. One of the most important downstream tasks of rib segmentation and centerline extraction is to facilitate fracture diagnosis. Hence, to ensure the clinical feasibility of the annotations, for all cases containing fracutres, we ensure all fractures in the case are 1) covered by the rib segmentation, and 2) passed through by the centerlines.

Specifically, for cases containing rib fractures, we consider the rib segmentation to be clinically feasible if it covers at least 75% of each rib fracture segmentation. The threshold is loosened to be 75% since the fracture segmentation from *RibFrac* Dataset contains the region surrounding the fracture region, which includes the non-rib regions. While the clinical feasibility of centerline annotations are based on radiologists' visual assessment to ensure each centerline line passes through the corresponding rib fracture segmentation.

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

A. Pipeline Overview

Note that *RibSeg v2* contains 2 tasks: 1) rib labeling and 2) anatomical centerline extraction from CT scans. Considering that rib centerlines can be easily obtained from rib labeling segmentation of high quality, we benchmark *RibSeg v2* with a single pipeline that first segments individual ribs from CT scans and then extract centerlines from the segmentation. Specifically, rib labeling is formulated as a multi-class segmentation task and centerlines are extracted using skeletonization-based methods.

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 two-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 skeletonization methods.

B. CT Denoising

Considering the sparsity of ribs in 3D volumes ($< 0.5\%$ voxels) and the high HU value of bones in CT scans ($> 200\,\text{HU}$), 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-Component-Denoising* (CCD). Note 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 [57] and PartNet [58], where the segmentation parts of target objects have a rather balanced distribution.

Frameworks. To alleviate the sparsity issue, we propose a two-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. For comparison, we also test the one-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 voxel grids, which is computationally inefficient [15], [18]. To address the memory issue, we convert the dense 3D volumes to sparse point clouds [22] and adopt point-based networks as the backbone model. Specifically, we tested our benchmark pipeline with PointCNN [59], DGCNN [60], different input settings of PointNet [61] and PointNet++ [34]. Note that PointCNN and DGCNN require a relatively high memory usage and could only afford 2048 points input at most, while PointNet and PointNet++, we also tested 30k points as input. For comparison, we also tested the voxel-based method with nnU-Net [62] 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 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.

D. Skeletonization Methods for Rib Anatomical Centerline Extraction

During the annotation procedure, we found that both TEASAR-based method and L1-medial skeletonization

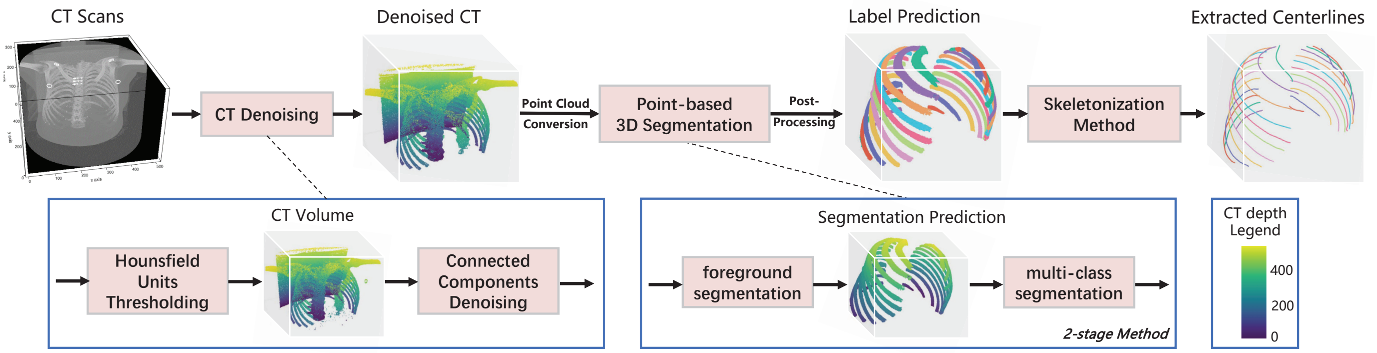

Fig. 3: **The Pipeline of Rib Labeling and Anatomical Centerline Extraction.** The pipeline is divided into 3 steps: 1) **CT Denoising:** CT scans are thresholded by HU value and denoised by *Connected-Component-Denoising* (CCD) to remove most non-bone voxels. 2) **Point-based Rib Labeling:** The denoised CT volume is converted to point clouds for a two-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 skeletonization methods. **Color:** For clear visualization, all binary volumes are colored by axial depth.

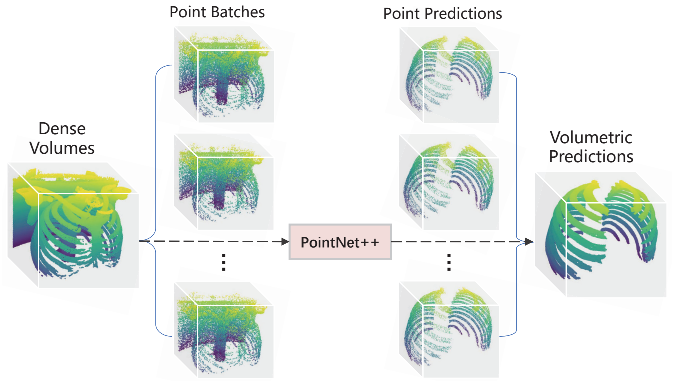

Fig. 4: **Point clouds generation from dense volumes**. The CT volumes are converted to a point cloud and divided into equally-sized point batches as input, and the multi-batch point predictions are concatenated into volumetric predictions.

method can guarantee visually satisfying results as long as the segmentation is correctly labeled. Hence, we simply cascade the skeletonization methods to the learning-based pipeline as an end-to-end baseline method for centerline extraction.

Voxel-based TEASAR method. TEASAR method directly works on volumes. 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 [52] 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.

Point-based L1-medial skeletonization method. L1-medial skeletonization method works on the point clouds. Specifically, in our case, we first 1) convert the dense volume into sparse point clouds. 2) Then we adopted the method in [53] that adds a regularization term to L1-median to prevent the formation of point clusters and uses classical weighted PCA for skeleton branch detection. 3) And the curve skeleton can be generated by iterative contraction where the L1-medians of the local neighborhoods are represented by the point sets. 4) The resultant curves are smoothened and upsampled to 500 points by linear interpolation.

Shortest path-based method. We also include a trivial shortest path-based method for comparison. Specifically, for each rib volume, we first select its endpoints according to coordinates on the transverse plane and perform the shortest path method used in [63]. The resultant point sets are also smoothened and upsampled to 500 points by linear interpolation.

E. Data Eligibility

The 654 of 660 CT volumes from *RibFrac* dataset are included in this study, and the 6 unqualified cases are partly scanned and only cover the middle part of the ribcage, which makes it impossible to label the ribs. All 654 cases are used in the method development.

Note that all CT scans might have different level of HU deviation due to the hardware settings, so the 200 HU thresholding might filter foreground voxels and make the resultant ribs hollow inside. During the dataset development, we added morphological closing to ensure the completeness of the annotation. While for segmentation method, the converted point clouds still preseverve the sufficient geometry of ribs, which indicates that the HU thresholding won't limit the clinical applicability of the pipeline. Hence, all 654 cases are included in the pipeline development.

F. Method Discussion

Clinically, rib anatomical centerlines are more favorable for being used to construct inner coordinate systems for surgery planning, while rib cage labeling is often utilized by visualization tools for structure and morphology analysis. In this study, we implement both two tasks in one pipeline where centerlines are extracted by applying skeletonization methods to the labeled ribs. Although the skeletonization-based method 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.

V. Experiments

A. Experiments on Rib labeling

1) Evaluation Metrics: We first define the Label-Dice [64], [65] of rib i as:

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

where y and $\hat{y}$ indicate the label prediction and ground truth, respectively. For quantitative analysis, we evaluate the performance by reporting the average $\mathrm{Label\text{-}Dice}$ of the 24 ribs, denoted as $\mathrm{Dice}_{\mathrm{avg}}^{(L)}$. While in qualitative analysis, we reflect the performance degradation by reporting the minimal $\mathrm{Label\text{-}Dice}$ amongst the 24 ribs, denoted as $\mathrm{Dice}_{\mathrm{min}}^{(L)}$. Moreover, we report the $\mathrm{Label\text{-}Accuracy}$ of individual ribs to evaluate the method's clinical applicability. Specifically, an individual rib i is counted as correctly labeled if $\mathrm{Recall}_i > 0.7$, 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 $\mathrm{Label\text{-}Accuracy}$ of all / first / intermediate / twelfth rib pairs, respectively.

2) Quantitative Analysis: We first evaluate the methods on *RibSeg v2* test set, comparing the one-stage and two-stage methods with different settings. As depicted in Tab. II, two-stage methods significantly outperform one-stage methods with 0.2% $\sim$ 16.5% higher Label Dice (9.9% on average) and 2.7% $\sim$ 33.6% higher label accuracy (13.7% on average). The interpretation is that the rib in the input volume of the one-stage method is too sparse ($\sim$16%) to provide sufficient features while the two-stage method removes most background noises for the label prediction. Due to the memory requirement, PointCNN and DGCNN could only afford 2048 points input, while for the light models (PointNet and PointNet++), we tested both 2048 and 30k points input for comparison, and the result indicated that increasing input size wouldn't necessarily improve the performance. Besides, the point-based methods outperform the voxel-based method (nnU-Net), as it inputs the geometry of the whole volume instead of voxel patches, which leads to a rich feature representation. Specifically, the best point-based method (DGCNN) enjoys 6.9% higher label Dice values and 4.0% higher label accuracy than nnU-Net.

For inference speed, the point-based method is significantly more efficient for taking sparse point clouds as inputs, as reported in Tab. III, whereas the point-based method is 3 $\sim$ 70× faster than the voxel-based method.

TABLE II: **Rib Labeling Metrics on *RibSeg v2* Test Set**. The metrics include average $\mathrm{Label\text{-}Dice}$ and $\mathrm{Label\text{-}Accuracy}$ of all / first / intermediate / twelfth rib pairs (A/F/I/T). For model comparision, we adopted nnU-Net as the SOTA voxel-based model, and point-based models such as PointCNN, DGCNN, PointNet and PointNet++. Due to the memory requirement of PointCNN and DGCNN, all models take 2048 points as input. For the light PointNet and PointNet++, we further use 30k points input for comparision. Both one-stage and two-stage methods are included.

	Methods	$\mathrm{Dice}_{\mathrm{avg}}^{(L)}$	$\mathrm{Label\text{-}Accuracy}$ (A/F/I/T)
	nnU-Net	83.6%	87.5% / 92.9% / 87.7% / 78.5%
One-stage	PointCNN	67.3%	55.5% / 90.1% / 51.5% / 61.9%
	DGCNN	76.4%	77.4% / 93.6% / 76.8% / 64.9%
	PointNet	70.3%	66.7% / 79.2% / 66.0% / 60.0%
	PointNet (30k)	71.5%	70.1% / 84.0% / 68.9% / 66.8%
	PointNet++	72.0%	70.0% / 89.4% / 68.6% / 62.6%
	PointNet++ (30k)	73.4%	72.0% / 89.4% / 70.9% / 64.9%
Two-stage	PointCNN	83.8%	89.1% / 54.3% / 92.8% / 87.2%
	DGCNN	**90.5%**	**91.5%** / **60.7%** / **96.3%** / **89.4%**
	PointNet	75.1%	73.7% / 50.6% / 75.9% / 74.3%
	PointNet (30k)	71.7%	67.1% / 47.0% / 69.2% / 67.2%
	PointNet++	84.4%	85.4% / 59.1% / 87.9% / 87.9%
	PointNet++ (30k)	84.8%	85.9% / 59.4% / 88.4% / 87.5%

3) Qualitative Analysis: For robustness analysis, we compare the inference results of normal and challenging cases 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: 14.4% lower on average $\mathrm{Label\text{-}Dice}$ and 14.4% lower on $\mathrm{Label\text{-}Accuracy}$. To further investigate the performance degradation, we also analyze the minimal instance-wise $\mathrm{Label\text{-}Dice}$, which is unsatisfying even in normal cases (73.4%) and suffers a huge drop by 11.0% in the challenging cases. Note that the minimal $\mathrm{Label\text{-}Dice}$ occurs on rib 12 or 24 (the 12th pair of ribs). It is exactly these challenging cases that

TABLE III: **Speed Comparison**. The table reports model forward time in seconds (one-stage / two-stage) averaged from 10 sample cases. Post-processing time is not included as it heavily depends on the implementation.

Methods	Forward Time (s)
nnU-Net	72.60 ± 17.19 / -
PointCNN	20.21 ± 5.34 / 24.68 ± 5.67
DGCNN	11.24 ± 3.90 / 13.82 ± 4.11
PointNet	0.34 ± 0.40 / 0.53 ± 0.55
PointNet (30k)	0.71 ± 1.82 / 0.52 ± 0.58
PointNet++	7.66 ± 1.90 / 9.25 ± 1.95
PointNet++ (30k)	0.98 ± 0.50 / 1.40 ± 0.64

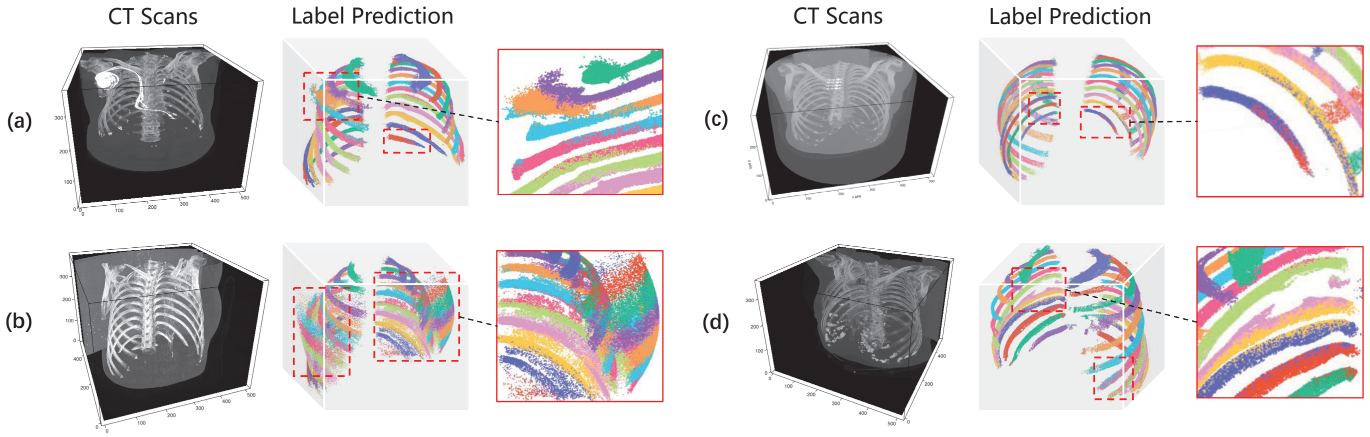

Fig. 5: **Visualization of Label Predictions on Challenging Cases**. (a) The case contains a pacemaker, which partly remains in the label prediction. (b) The case suffers HU deviation, leading to severe cross-labeling. (c) The case misses the 12th pair of floating ribs, causing the cross-labeling. (d) 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	$\text{Dice}^{(L)}_{\text{avg}}$	$\text{Dice}^{(L)}_{\text{min}}$	Label-Accuracy
All	90.5%	72.3%	91.5%
Normal	92.6%	73.4%	94.0%
Challenging	78.2%	62.4%	79.6%

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 (PointNet++). 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 damaged, also causing cross-labeling. In brief, despite the visually satisfying predictions in most cases, the performance degradation in a few challenging cases is significant. 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 1) the similarity/distance between the extracted centerline and annotation, 2) the curvature deviation and 3) the axial deviation of the extracted centerline. We first adopted centerlineDice (clDice) [66] to evaluate the rib-wise distance between extracted centerline and the centerline annotation, and denote the clDice of rib i as:

$$\text{clDice}_i = 2 \times \frac{T_{\text{prec}}\left(L_i, \hat{S}_i\right) \times T_{\text{sens}}\left(\hat{L}_i, S_i\right)}{T_{\text{prec}}\left(L_i, \hat{S}_i\right) + T_{\text{sens}}\left(\hat{L}_i, S_i\right)}, \tag{2}$$

where L_i, $\hat{L}_i$, S_i, and $\hat{S}_i$ indicate the centerline prediction, centerline annotation, label prediction, and label annotation of the rib i, respectively. While T_{prec} and T_{sens} indicate the topology's precision and sensitivity, which are defined as:

$$T_{\text{prec}}\left(L_i, \hat{S}_i\right) = \frac{\left|L_i \cap \hat{S}_i\right|}{\left|L_i\right|}; \quad T_{\text{sens}}\left(\hat{L}_i, S_i\right) = \frac{\left|\hat{L}_i \cap S_i\right|}{\left|\hat{L}_i\right|}. \tag{3}$$

For quantitative analysis, we report the average clDice of the 24 ribs, denoted as $\text{clDice}_{\text{avg}}$.

Inspired by normalized surface dice [67], [68], we propose *Normalized-Line-Dice* (NLD) to evaluate the curvature deviation:

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

where $\mathbf{L} \subseteq \mathbb{R}^3$ and $\hat{\mathbf{L}} \subseteq \mathbb{R}^3$ are the extracted centerline and centerline annotation, respectively, and $B^{(\tau)}_{\mathbf{L}} = \{y \in \mathbb{R}^3 \mid \exists\, \tilde{y} \in \mathbf{L}, \|y - \tilde{y}\|_2 \leq \tau\}$ 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.

Another important measurement is axial deviation, indicating whether the centerline lies in the mid of the corresponding rib bone. Hence, we report the unidirectional Chamfer Distance of the extracted centerline with respect to the annotation of rib segmentation, denoted as *Line-Seg-Chamfer-Distance*

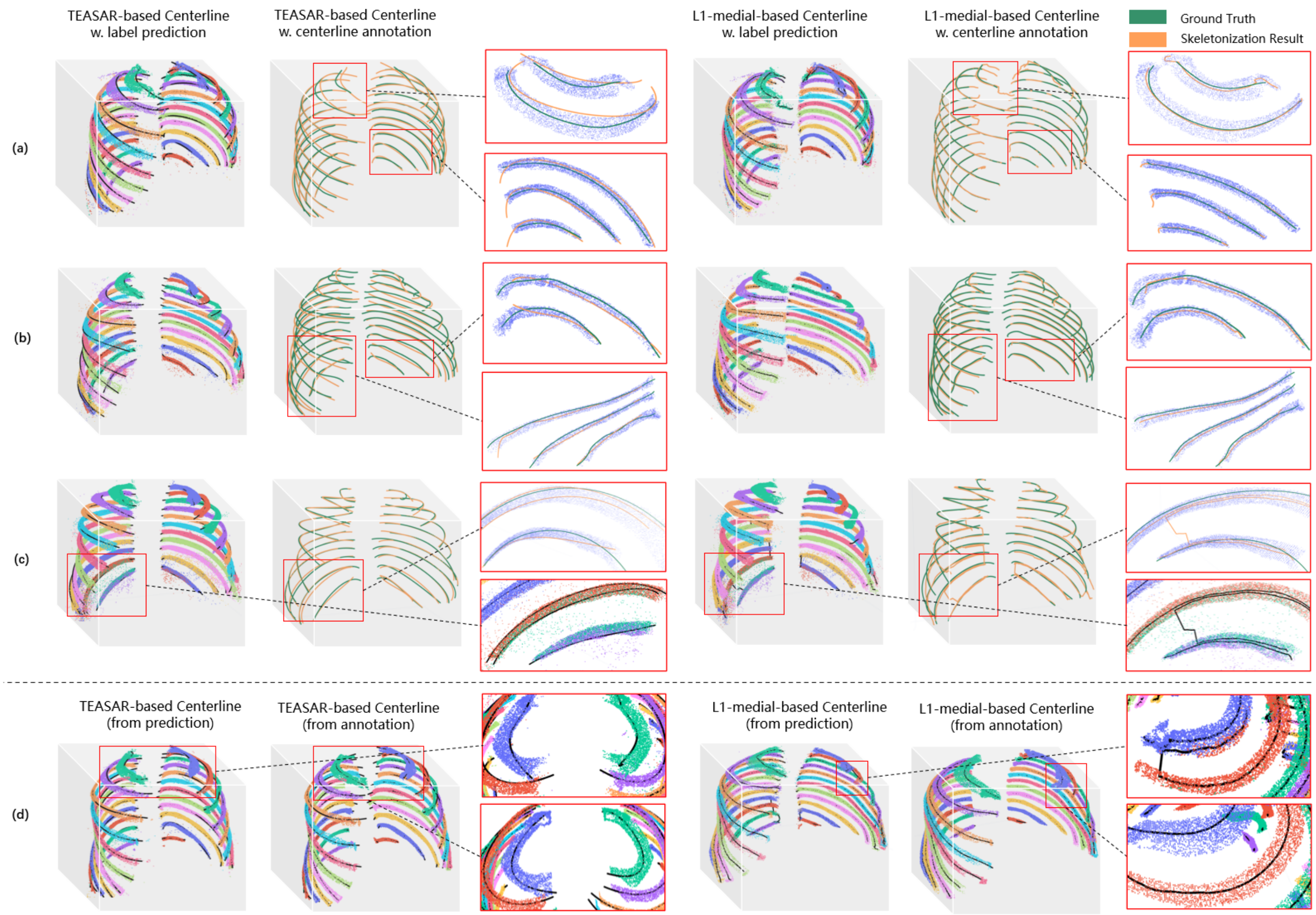
TEASAR-based Centerline w. label prediction
TEASAR-based Centerline w. centerline annotation
L1-medial-based Centerline w. label prediction
L1-medial-based Centerline w. centerline annotation
🟩 Ground Truth
🟧 Skeletonization Result
(a)
(b)
(c)
(d)
TEASAR-based Centerline (from prediction)
TEASAR-based Centerline (from annotation)
L1-medial-based Centerline (from prediction)
L1-medial-based Centerline (from annotation)

Fig. 6: **Visualization of Centerline Extraction Results**. (a∼c) Evaluation on rib centerline extraction pipeline: TEASAR method and L1-medial skeletonization are applied to rib label predictions and the resultant centerlines are compared with the centerline annotation. Shortest path method fails on most cases, which is not included. (d) Ablation study on skeletonization methods: both methods are applied to rib label predictions and annotation, respectively.

(LSCD):

$$\mathrm{LSCD}(\mathbf{L},\hat{\mathbf{S}}) = \frac{1}{|\hat{\mathbf{S}}|}\sum_{\hat{y}\in\hat{\mathbf{S}}}\min_{y\in\mathbf{L}}\|y-\hat{y}\|_2, \tag{5}$$

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. Since Chamfer distance is not numerically intuitive, we report the relative error between LSCD of extracted centerline and LSCD of centerline annotation, denoted as *Line-Seg-Chamfer-Distance-Error* (LSCDError):

$$\mathrm{LSCDError}(\mathbf{L},\hat{\mathbf{L}}) = \frac{\mathrm{LSCD}(\mathbf{L},\hat{\mathbf{S}}) - \mathrm{LSCD}(\hat{\mathbf{L}},\hat{\mathbf{S}})}{\mathrm{LSCD}(\hat{\mathbf{L}},\hat{\mathbf{S}})}. \tag{6}$$

2) Quantitative Analysis: We first applied the skeletonization methods to the rib label prediction with the highest accuracy (two-stage DGCNN) from Sec. V-A. As reported in Tab. V, we evaluated TEASAR method, L1-medial skeletonization and shortest path method based on clDice, NLD, and LSCDError, respectively. Since the skeletonization methods might be sensitive to the rib label predictions from the first stage, we also directly applied these methods to the rib label annotations as an ablation study.

Overall, given the rib label prediction with noises, L1-medial skeletonization performs the best, with a 8.0% higher clDice, a close NLD (0.7% lower), and a 24.3% lower LSCDError than TEASAR method. When applied to the rib label annotations, both L1-medial skeletonization and TEASAR method obtained results with very high accuracy, where L1-medial skeletonization performs slightly better than TEASAR method (2.4% higher clDice, 1.3% higher NLD, and 0.8% LSCDError). And we concluded that such non-learning skeletonization methods can work well on rib centerline extraction task, especially when applied to high quality rib segmentation.

However, the trivial shortest path fails to give a good result even given the rib label annotation, where the resultant path always deviates from the actual centerline. When applied to rib label predictions, it fails on the majority cases and we didn't include its results in Tab. V. The interpretation is that it requires the rib volumes to be one single connected component while the rib label predictions contain too many noises, and the discrete noise points could be detected as ending points, which fails the algorithm.

TABLE V: **Rib Centerline Extraction Metrics on *RibSeg v2* test set**. We include the ablation study of applying skeletonization to rib label predictions (p) and rib annotations (a), respectively (p / a). The methods include l1-medial skeletonization, TEASAR method, and a trivial shortest path. The metrics include *centerlineDice* (clDice), *Normalized-Line-Dice* (NLD), and *Line-Seg-Chamfer-Distance-Error* (LSCDError).

Methods	clDice	NLD	LSCDError
L1-medial	**90.8%** / **97.4%**	82.1% / **97.9%**	**38.8%** / **8.4%**
TEASAR	82.8% / 95.0%	**82.8%** / 96.6%	63.1% / 9.2%
Shortest path	- / 33.5%	- / 55.9%	- / 20.0%

3) Qualitative Analysis: As depicted in Fig. 6, we visualized the centerline extraction results for a more intuitive and detailed analysis. In most cases, with accurate label predictions, both skeletonization methods can guarantee visually satisfying centerlines. As depicted in Fig. 6 (a), the resultant centerline of the voxel-based TEASAR method exceeds the actual rib a little bit and slightly deviates from the annotation, the interpretation is that the TEASAR method includes dilation which dilates the volume making the extracted centerline longer, and also enlarges the noise voxels hence dislocating the centroid. While the point-based L1-medial skeletonization is more shape-sensitive and the resultant centerline aligns the annotation better. Fig. 6 (b) shows a case containing obvious rib fracture, where both methods managed to generate a complete centerline passing through the fracture region. However, the skeletonization methods might suffer a huge performance drop if the prediction contains too much mislabeling. As depicted in Fig. 6 (c), this case misses a floating rib and the label prediction suffers a huge accuracy drop which heavily affected the skeletonization methods. Consequentially, the voxel-based TEASAR result contains 2 centerlines aligning to the same rib, while the point-based L1-medial skeletonization result contains a misaligned centerline crossing 2 ribs. Similarly, the ablation study in Fig. 6 (d) further indicates that the quality of centerline extraction heavily depends on the label prediction. For the voxel-based TEASAR method, the resultant centerline deviates from the rib center. While the point-based L1-medial is much more sensitive to the rib label prediction, as the cross-labeling regions in the prediction will misguide the resultant centerline. Moreover, as mentioned in Sec. III-C, skeletonization methods also fail for cases suffering severe rib fracture, where a single rib is broken into several parts, which will result in messy curve segments. Such abnormal cases also urge a more robust centerline extraction method, 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 center of the rib segmentation by visual assessment, it might not necessarily have the minimal LSCDError value. The interpretation is that annotations of rib anatomical centerlines 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.

C. Experiment Settings

The training of the models was carried out on a cluster with 4 NVIDIA A100. 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. In the training stage of the two-stage method, the input of the segmentation task is point sets of 2,048 / 30,000 downsampled from the binary CT volume, while the input of the labeling task is point sets downsampled from the predictions of segmentation. For point-based models, the Adam optimizer is adopted and a combination of cross-entropy (CE) and Dice loss as the loss function. All experiments settings and trained models are available online at https://github.com/M3DV/RibSeg.

VI. Conclusion

We developed 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 skeletonization-based methods for centerline extraction. We then 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, to evaluate our pipeline, we tested 4 point-based models with different settings and 3 skeletonization methods, and also proposed various metrics for rib segmentation, labeling, and anatomical centerline extraction, providing a comprehensive method evaluation. Finally, by detailed quantitative and qualitative analysis of the categorized challenging cases, we featured the key challenges of each task, which are valuable to guide future studies.

The dataset and proposed method show the potential to be clinically applicable, such as the diagnosis of rib fractures and bone lesions. Besides, considering the differences from standard medical image datasets [69], [70] 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.

There remain limitations in this study. For rib anatomical centerline extraction, we apply a skeletonization-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.

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] Z. Xu, U. Bagci, C. B. Jonsson, S. Jain, and D. J. Mollura, "Efficient ribcage segmentation from ct scans using shape features," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2899–2902, 2014.

[4] 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.

[5] M. Tajdari, A. Pawar, H. Li, F. Tajdari, A. Maqsood, E. Cleary, S. Saha, Y. J. Zhang, J. F. Sarwark, and W. K. Liu, "Image-based modelling for adolescent idiopathic scoliosis: Mechanistic machine learning analysis and prediction," *Computer Methods in Applied Mechanics and Engineering*, vol. 374, p. 113590, 2021.

[6] M. Tajdari, F. Tajdari, P. Shirzadian, A. Pawar, M. Wardak, S. Saha, C. Park, T. Huysmans, Y. Song, Y. J. Zhang, J. F. Sarwark, and W. K. Liu, "Next-generation prognosis framework for pediatric spinal deformities using bio-informed deep learning networks," *Engineering with Computers*, vol. 38, pp. 4061–4084, 2022.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Y. J. Zhang, "Challenges and advances in image-based geometric modeling and mesh generation," *Image-Based Geometric Modeling and Mesh Generation*, 2013.

[20] Y. Zhang, *Geometric Modeling and Mesh Generation from Scanned Images*. CRC Press, Taylor & Francis Group, 2016.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

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

[42] 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.

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

[44] 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.

[45] 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.

[46] S. M. Yoon, C. Malerczyk, and H. Graf, “3D skeleton extraction from volume data based on normalized gradient vector flow,” 2009.

[47] 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.

[48] 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.

[49] 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

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

[51] A. Pawar and Y. J. Zhang, “NeuronSeg_BACH: Automated neuron segmentation using B-spline based active contour and hyperelastic regularization,” *Communications in Computational Physics*, vol. 28, pp. 1219–1244, 2020.

[52] 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.

[53] H. Huang, S. Wu, D. Cohen-Or, M. Gong, H. Zhang, G. Li, and B. Chen, “L1-medial skeleton of point cloud,” *ACM Transactions on Graphics (TOG)*, vol. 32, pp. 1 – 8, 2013.

[54] M. H. Lev and R. G. Gonzalez, “17 – CT angiography and CT perfusion imaging,” 2002.

[55] A. Rosenfeld and J. Pfaltz, “Sequential operations in digital picture processing,” *J. ACM*, vol. 13, pp. 471–494, 10 1966.

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

[57] 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.

[58] 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.

[59] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “Pointcnn: Convolution on x-transformed points,” in *Neural Information Processing Systems*, 2018.

[60] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” *ACM Transactions on Graphics (TOG)*, vol. 38, pp. 1 – 12, 2018.

[61] C. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2016.

[62] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203 – 211, 2020.

[63] P. yu Teng, A. M. Bagci, and N. Alperin, “Automated prescription of an optimal imaging plane for measurement of cerebral blood flow by phase contrast magnetic resonance imaging,” *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 2566–2573, 2011.

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

[65] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,” *BMC Medical Imaging*, vol. 15, 2015.

[66] S. Shit, J. C. Paetzold, A. K. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze, “clDice - a novel topology-preserving loss function for tubular structure segmentation,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 555–16 564, 2020.

[67] 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.

[68] 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.

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

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