A Robust Shape-Aware Rib Fracture Detection and Segmentation Framework with Contrastive Learning

Zheng Cao, Liming Xu, Danny Z. Chen, Honghao Gao, Jian Wu

Abstract—**The rib fracture is a common type of thoracic skeletal trauma, and its inspections using computed tomography (CT) scans are critical for clinical evaluation and treatment planning. However, it is often challenging for radiologists to quickly and accurately detect rib fractures due to tiny objects and blurriness in large 3D CT images. Previous diagnoses for automatic rib fracture mostly relied on deep learning (DL)-based object detection, which highly depends on label quality and quantity. Moreover, general object detection methods did not take into consideration the typically elongated and oblique shapes of ribs in 3D volumes. To address these issues, we propose a shape-aware method based on DL called SA-FracNet for rib fracture detection and segmentation. First, we design a pixel-level pretext task founded on contrastive learning on massive unlabeled CT images. Second, we train the fine-tuned rib fracture detection model based on the pre-trained weights. Third, we develop a fracture shape-aware multi-task segmentation network to delineate the fracture based on the detection result. Experiments demonstrate that our proposed SA-FracNet achieves state-of-the-art rib fracture detection and segmentation performance on the public RibFrac dataset, with a detection sensitivity of 0.926 and segmentation Dice of 0.754. Test on a private dataset also validates the robustness and generalization of our SA-FracNet.**

Index Terms—**Computer-aided Diagnosis, Rib Fracture Detection and Segmentation, Self-supervised Contrastive Learning, Shape-aware Model.**

I. Introduction

RIB fracture is one of the most common thoracic injuries, which is detected in over 10% of all injured patients [1]. The patterns and amount of rib fractures are major indicators for evaluating trauma severity and planning treatments [2]. Chest computed tomography (CT) is a popular medical aid for experienced radiologists to examine chest trauma in the clinic. Lamentably, due to the huge division between tiny and blurry rib fractures and large 3D CT images, manual screening is extremely time-consuming and labor-intensive, which may even result in ignoring small fractures [3] (see supporting document S.I). As a result, the development of automatic and accurate rib fracture recognition methods is highly desired for clinical purposes.

Conventional computer-aided rib fracture detection methods first unfold curved ribs in 3D CT images and then determine whether the ribs are fractured or displaced [4], [5]. These methods rely on handcrafted features that encode unfolded rib features and are not suitable for general applications. Recently, deep learning (DL) methods based on convolutional neural networks (CNNs) achieved remarkable results in various medical image recognition tasks such as organ-at-risk segmentation, fundus disease diagnosis, and COVID-19 detection [6]–[8]. DL-based models also attracted much attention in rib fracture detection due to their high efficiency and accuracy. Zhou *et al.* [9] proposed a rib fracture detection and segmentation method based on multi-modal fusion, combining imaging and other clinical textual information. The model achieved better performance than the basic two-stage target detection Faster RCNN [10]. Founded on 3D-UNnet [11], Jin *et al.* [12] proposed FracNet to improve rib fracture detection/segmentation results. They introduced sliding window sampling to generate regional samples in CT images, reducing the computing complexity of the model in non-rib areas. Liu *et al.* [13] developed a multi-scale segmentation network to enhance the perception of tiny local details. In general, although these known methods were proved to be effective on their datasets, acquiring qualified label data for DL model training still represents a difficult but critical issue.

Contrastive learning is a new DL approach for unsupervised visual representation learning. It can learn from unlabeled samples and reduce the reliance on labeled training data [14]. Wang *et al.* [15] proposed a knowledge distillation method for fracture detection in X-ray images using massive unlabeled data to improve the detection results. Due to the rib's elongated and oblique shape, current contrastive learning methods based on global representations are still limited in achieving good generalized performance. Recently, some pixel-level contrastive learning methods were introduced for semantic segmentation of natural images to attain local representations on specific regions [16]–[18]. However, no known robust contrastive learning method for rib fracture detection can leverage domain-specific and problem-specific cues.

To address the above challenges, in this paper, we propose a new robust shape-aware rib fracture detection and segmentation method based on contrastive learning called SA-FracNet,

Manuscript received 13 July 2022; revised 10 March 2023; accepted 15 March 2023. This research was partially supported by the National Key R&D Program of China under grant No. 2019YFC0118802, the National Natural Science Foundation of China under grant No. 62176231, Zhejiang public welfare technology research project under grant No. NSF Grant CCF-1617735 partly supported LGF20F020013, D. Z. Chen's research. (*Zheng Cao and Liming Xu contributed equally to this work.*) (*Corresponding author: Jian Wu and Honghao Gao.*)

Zheng Cao and Liming Xu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China (e-mail: z.cao@zju.edu.cn; sunrise2019@zju.edu.cn).

Danny Z. Chen is with the Department of Computer Science and Engineering, University of Notre Dame, IN 46556, USA (e-mail: dchen@nd.edu).

Honghao Gao is with the School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China (e-mail: gaohonghao@shu.edu.cn).

Jian Wu is with the Second Affiliated Hospital School of Medicine, and School of Public Health, Zhejiang University, Hangzhou 310058, China (e-mail: wujian2000@zju.edu.cn).

which is a three-stage DL-based approach. Our main idea is to incorporate a more specific geometric representation of ribs and fractures in the contrastive learning stage and fine-tuning stage, respectively. In summary, the main contributions of this work are three-fold: (1) A novel pretext task, pixel-level contrastive learning via massive unlabeled 3D Chest CT images; (2) a fracture shape-aware multi-task segmentation model using signed distance map (SDM); (3) robust segmentation and detection performance on the public RibFrac [12] and private datasets.

II. Related Works

A. Computer-aided Rib Fracture Diagnosis Methods

The rib 3D reconstruction is the most common technology in computer-aided rib fracture diagnoses, which can reconstruct the geometric contour of the rib with the help of the 3D CT image, and then visualize the spatial shape of the rib. General 3D scanning reconstruction technology of ribs mainly includes Multi Planar Reconstruction (MPR) [19], Curved Planar Reconstruction (CPR) [20], Maximum Intensity Projection (MIP) [21], and Volume Rendering Technique (VRT) [22]. Besides, Urbaneja *et al.* [5] proposed a cylindrical projection algorithm for automatic rib expansion.

In recent years, artificial intelligence technology has developed by leaps and bounds. At the same time, deep learning algorithms have achieved leading results in multiple medical image lesion detection tasks [23]–[26], which has also brought new advantages to the computer-aided diagnosis technology of rib fractures. Ibanez *et al.* [27] proposed RiFNet for the rib fracture classification model based on Convolutional Neural Network (CNN). Weikert *et al.* [28] used the deep learning target detection model in multiple natural images to detect rib fractures. The experiments proved that the deep learning model can achieve higher detection sensitivity than clinicians. On this basis, Zhou *et al.* [9] used a two-stage target detection model Faster R-CNN [10] to detect rib fractures in CT images. The detection algorithm first used ResNet [29] to extract features from rib CT images and then used a region proposal network to generate box predictions for rib fractures. However, this method is ineffective for detecting some subtle occult rib fractures, leading to a recall rate of only 86%. Meng *et al.* [30] proposed a multi-stage automatic rib diagnosis system. Yao *et al.* [31] proposed a multi-attention-based rib fracture detection method. However, the multi-stage algorithm prediction process needs to be simplified. Considering that the 2D detection method will lose the spatial context information of rib fractures, Jin *et al.* [12] proposed a deep learning semantic segmentation model FracNet based on 3D-UNet [11], and they reported the detection and segmentation result of rib fractures at the same time.

B. Contrastive Learning and Shape Representation Methods

Contrastive learning aims to learn visual representations by discriminating the instance, which maximizes the similarity of representations from similar sample pairs. InstDisc [32] is the prototype for instance-level contrastive learning for pre-training tasks. On this basis, the contrastive learning model CMC proposed by Van *et al.* [33] takes multiple views of an image as positive sample pairs. MoCo, proposed by He *et al.* [34], uses the momentum of two encoders for contrastive learning, which can avoid the fluctuation of gradient update at the beginning of training. MoCo has two optimized versions, MoCo-v2 [35] and MoCo-v3 [36]. MoCo-v2 utilizes a nonlinear mapping head after the output layer of the encoder, and MoCo-v3 mainly explores the application of MoCo's self-supervised contrastive learning method in the visual Transformer [37] model. Chen *et al.* [14] propose another end-to-end contrastive learning framework SimCLR. The research results of MoCo and SimCLR inspire a series of follow-up studies on self-supervised contrastive learning [38]–[42]. Different from the above instance-level contrastive learning, pixel-level contrastive learning has been introduced for dense feature representations [18]. The pixel-level contrastive learning methods consider the local representation of images to be consistent under different input conditions and use this consistency to construct proxy tasks [43]–[45]. For the task of 3D image segmentation in medical images, Chaitanya *et al.* [16] propose contrastive representation learning of regions, which achieves better segmentation results on three Magnetic Resonance Imaging (MRI) data.

III. Methodology

Fig. 1 shows the architecture of our proposed SA-FracNet, which contains three stages to obtain detection and segmentation results of input 3D thoracic CT images. SA-FracNet aims to learn a pixel-level representation from massive unlabeled 3D thoracic CT images and adopt a fracture shape-aware constraint loss in the downstream segmentation task to improve the perception of rib fracture. It outputs the detection and segmentation results with a sliding-window prediction. In the following part, we present the details of SA-FracNet.

A. pixel-level Contrastive Learning on Unlabelled CT Data

As demonstrated in Fig. 1, SA-FracNet includes a pre-training and fine-tuning process like most self-supervised methods. To attain the pixel-level representations in the pre-training stage, we design the contrastive learning task to learn the spatial consistency of the close pixel in unlabeled CT images. The key idea is to treat each pixel as a single class and distinguish it from the other pixels. We form the positive training pairs by extracting the features of spatially similar pixel pairs via the overlapped sampling and vice versa. Considering the CT image contains simple foreground elements, i.e., ribs, the pixel-level contrastive learning enables the model to discriminate spatially the close pixels, which can be recognized as the rib shape distribution in the CT images.

1) Overlapped Rib Region Sampling Algorithm: Due to the large resolution of the 3D rib CT images and the small proportion of the rib area, we design a rib region overlapping sampling algorithm for sampling a large number of unlabeled CT images in order to improve the training efficiency and avoid the additional computational cost. An illustration of the algorithm is shown in fig. 2, and the detailed process is listed in Algorithm 1.

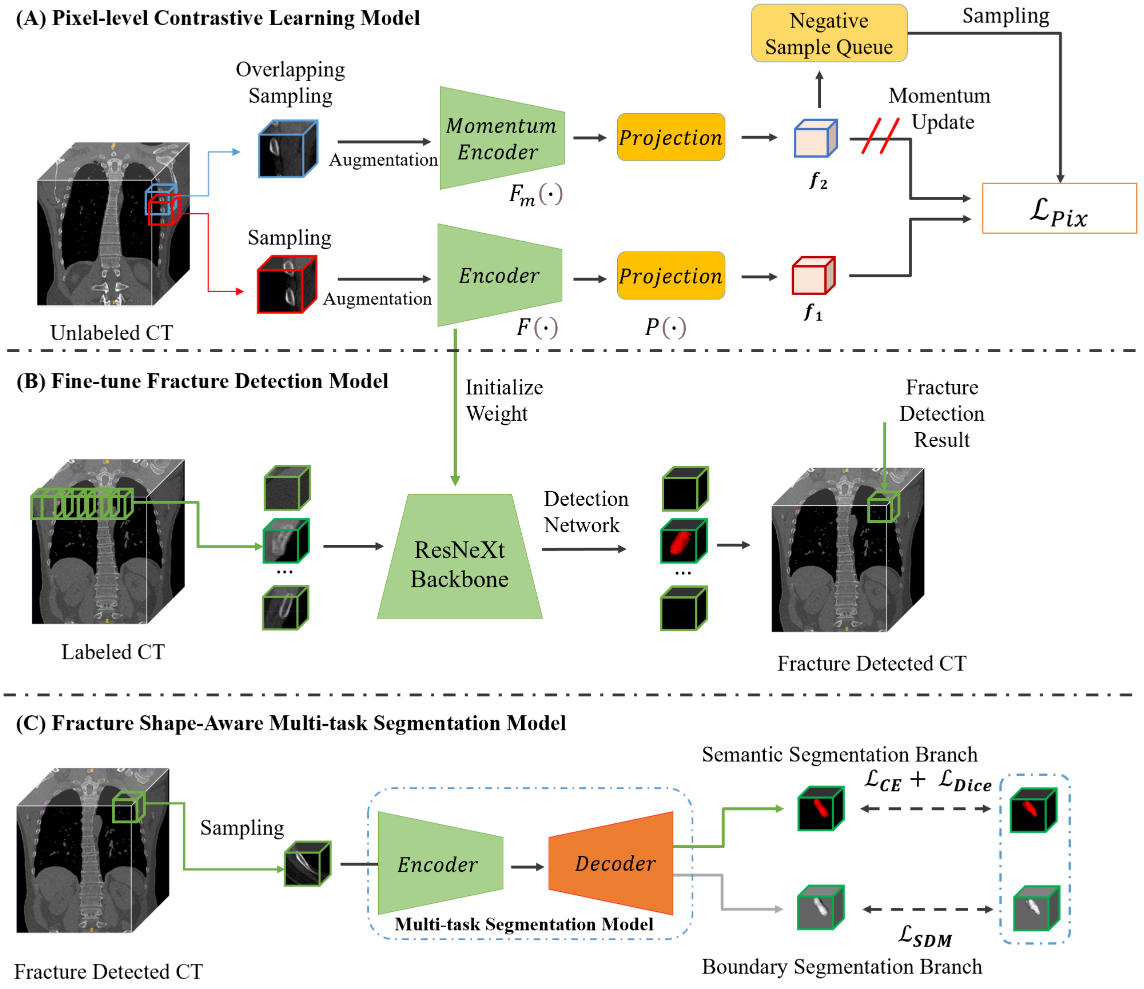

Fig. 1. A schematic view of our proposed SA-FracNet framework, which consists of (A) a pixel-level contrastive learning method, (B) a downstream fracture detection model, and (C) a Fracture Shape-Aware multi-task fracture segmentation network.

The radiodensity of the CT image is defined by the Hounsfield unit (HU). We utilize the prior knowledge of the CT images, i.e., the CT range of ribs is usually 400HU to 1000HU, to control the sampling window includes more rib pixels. As shown in Algorithm 1, the overlapping sampling algorithm first uses threshold sampling to obtain a cuboid sampling area χ_+ (the red sampling area in Fig. 2) with a size of $H \times W \times D$. Then the algorithm adjusts the center (x, y, z) of χ_+ with a small range of random offsets on the three axes (d_x, d_y, d_z) to get another sampling center. We use this coordinate as the center to sample with the same size to get another sampling area χ_- (the blue sampling in Fig. 2 area). $S = \chi_+ \cap \chi_-$ is the overlapping area of the two sampling areas. The ratio of the overlapping area S to the sampling area χ_+ needs to be kept within certain upper and lower bounds. As shown in Fig. 2, the number of effective positive and negative training pairs depends on the overlapping area. Hence, the lower bound of the overlap ratio is to obtain more positive sample pairs, and the upper bound of the overlap ratio is to prevent two samples from being too similar and causing the collapse issue in contrastive learning. In the actual experiment, we set the sampling ratio within the range of $[0.7, 0.9]$. At this time, the proportion of positive and negative sample pairs in the overlapping area is relatively balanced.

The purpose of overlapping sampling is to obtain more effective positive sample pixel pairs, as shown in Fig. 2. After encoding by the encoder, each pixel on the feature map is regarded as a representation of a region corresponding to the original image. Overlapping sampling The two encoded feature maps of an image can also be considered to be overlapping in space. Pixel-level contrastive learning can regard overlapping pixel pairs or similar pixel pairs as positive sample pairs, and others as negative sample pairs.

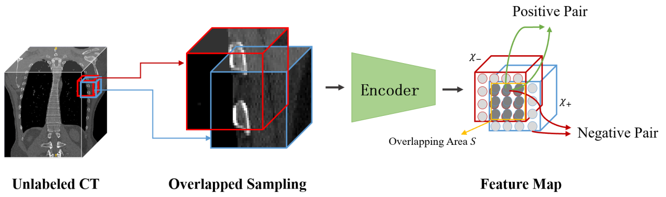

Fig. 2. Diagram of the Overlapped Rib Region Sampling Algorithm

Algorithm 1 Rib Region Coupling Sampling Algorithm.

```
INPUT: Unlabeled Rib CT Dataset $D_u = \{(X_u^i)\}_{i=1}^m$, Number of overlapping samples in single rib CT image $k$, Threshold of the average gray value of pixels in sampling area $\omega$, Upper bound of coordinate offset $d_{max}$, and the sampling couple $\chi_+$ and $\chi_-$ with size $H \times W \times D$
OUTPUT: Collection of sampled overlapping region sample pairs $D_u^s$
For each CT image $X_u^i$ in $D_u$, iter $k$ times, the iteration $iter \leftarrow 0$
While $iter \leq k$:
    Random Select a center point $d = (x, y, z)$ and $d \geq \omega$
    Crop a region $\chi_+ = H \times W \times D$ and calculate the average CT value in this region
    If $\chi_+ \leq \omega$ then go back to select another $\chi_+$
    Generate $(d_x, d_y, d_z)$ from the range $[-d_{max}, d_{max}]$, $\chi_- \leftarrow (x + d_x, y + d_y, z + d_z)$
    If $\chi_+ \cap \chi_- \leq [0.7, 0.9]$ then go back to select another $\chi_-$
    $D_u^s \leftarrow D_u^s + [\chi_+, \chi_-]$
end while
```

2) Architecture of the Pre-training Model: As shown in Fig. 1(A), the model first randomly crops two $64 \times 64 \times 64$ overlapped 3D patches from the same input 3D image to obtain two augmentation views. A regular encoder and a momentum encoder $f(\cdot)$ compute the corresponding features and pass them to the projection head $p(\cdot)$. We use 3D ResNeXt [46] as the regular encoder and set the momentum encoder similar to MoCo-v2 [34]. Since the pixels in the CT images are single-channel, and the semantic features are poorer than natural images, we reduce the output channel of each residual block in ResNeXt to $\{32, 64, 128, 256\}$ in order to prevent the over-fitting issue. The projection head contains 1×1 convolutional layer with 4096 channels, batch normalization layer, ReLu activation layer, and 1×1 convolutional layer with 256 channels, in this order. After obtaining the feature maps of the two views, we warp each pixel to the original 3D space and calculate the corresponding spatial Euclidean distance. The pairs of pixels are considered positive under the condition that the distance is larger than a threshold λ, and negative vice versa:

$$A(i,j) = \begin{cases} 1 & \text{if } d(i,j) \leq \lambda \\ 0 & \text{if } d(i,j) > \lambda \end{cases} \tag{1}$$

where i, and j donate the pixel from the feature map f_1 and f_2 in Fig. 1(A), $d(i,j)$ denotes the Euclidean distance of the two pixels. The loss per pix $\mathcal{L}_{pix}$ can be expressed as:

$$\mathcal{L}_{pix}(i) = -\log \frac{\sum_{j \in \chi_+} e^{\cos(x_i, x'_j)/\tau}}{\sum_{j \in \chi_+} e^{\cos(x_i, x'_j)/\tau} + \sum_{k \in \chi_-} e^{\cos(x_i, x'_k)/\tau}}, \tag{2}$$

where i is a pixel in the overlapped part of the two views, j and k are a positive pixel and a negative pixel, respectively, x_i and x'_j represent features output from the projection head, and τ is a hyper-parameter and equals to 0.3. Then we calculate the overall loss of the overlapping area:

$$\mathcal{L}_{PIX} = \frac{\sum_{i \in S} \mathcal{L}_{pix}(i) + \sum_{j \in S} \mathcal{L}_{pix}(j)}{2\,|S|} \tag{3}$$

Considering the simple foreground of ribs in the CT image, $\mathcal{L}_{PIX}$ can gain a perception of the rib shape from 3D CT images by discriminating the spatially close pixels. It is fundamental to the fracture segmentation downstream task for accurate prediction in boundary areas where labels change.

B. Fine-tune Fracture Detection Model

In the fine-tuning stage, we use a 3D version of RetinaNet [47] for the fracture detection task. The process is shown in Fig. 1(B). The parameters of the encoder $F(\cdot)$ in the pre-training stage are used to initialize the ResNeXt-50 backbone in the detection network. In each training epoch, the detection model selects positive samples with fractures and negative samples without labels as the input. After being encoded by the ResNeXt backbone $F(\cdot)$, the input CT patch is processed by the feature pyramid network for multi-scale feature extraction. Then binary classification of the detection box and regression of the bounding box are performed at each pixel point on the multi-scale feature map. The classification loss and regression loss guide the detection model for gradient update at the same time.

To improve the target detection network's ability to discriminate between normal and fractured ribs and to reduce false positive results in detection, in the training set of the detection model, we not only sampled regions with fractures as the positive sample set but also sampled a certain number of regions without fractures as the negative sample set. Specifically, for each labeled rib fracture, the center of the positive sample is the center of the 3D detection frame labeled with the shape (x_0, y_0, z_0). In contrast, the center of the negative sample is the symmetric coordinate $(512 - x_0, y_0, z_0)$ of the positive sample center about the sagittal plane of the human body. The rib before the fracture is morphologically symmetrical. The positive and negative samples were mixed in a 1:1 ratio, with the positive samples corresponding to the negative samples within the same Batch.

Moreover, due to the large size of a full-scale 3D CT image, training an end-to-end model is limited by GPU memory. Thus, we detect rib fractures in a sliding-window fashion. The specific prediction process is as follows:

- We first slice the single CT image into multiple regions to be detected with a sliding window. The step size of the sliding window is set to half of the window size, with the aim of reducing the impact of prediction boundary problems;
- Detect the rib fracture of each CT patch with the trained model, which outputs the frame prediction as well as its confidence;
- Merge the prediction results of each CT patch. Once the overlapping regions result in multiple detection results, only the larger detection frame is retained when the IoU of each frame is larger than 0.7 or the bigger one can cover the smaller ones.

C. Rib Fracture Shape-aware Multi-task Segmentation Model

As demonstrated in Fig. 1(C), we design the fracture shape-aware multi-task segmentation model to delineate the fracture lesion based on the detection result. We use the dithering sampling strategy in the segmentation model to simulate the positional variation of the detection box and to alleviate the negative impact of inaccurate positioning on the segmentation model. Specifically, for the sampling center of the segmentation, a small offset ($\leq$ 5 pixels) is added to each of the three coordinates (x_0, y_0, z_0) of the detection box center. Then, a cuboid pixel area with a side length of 1.2 times the maximum side length of the detection box is sampled to cover the entire lesion. The lesion area image is then interpolated to a uniform size of $48 \times 48 \times 48$, followed by CT value clipping (-200HU to 1000HU) and normalization to produce the input for the segmentation model. Next, inspired by [48], we use a multi-task segmentation model to jointly segment the fracture region and predict the signed distance map (SDM). SDM is defined by the distance between each pixel and its closest target object boundary, which can provide a shape-aware representation due to the rich features of object shapes and surfaces. We employ 3D ResUnet as the backbone of the segmentation network, which consists of the encoder-decoder module based on ResNet. The output of the segmentation backbone comes from two branches, one for the segmentation map and the other for the signed distance map.

Two kinds of loss functions are designed to fine-tune the segmentation model. For the segmentation map branch, we use a combination of cross-entropy loss and Dice loss for the segmentation map loss, as follows:

$$\begin{aligned}\mathcal{L}_{Seg} =& \mathcal{L}_{CE} + \mathcal{L}_{Dice} = \\ & \sum_t (-G_t \log(P_t) - (1 - G_t)\log(1 - P_t)) + \\ & 1 - \frac{2 \times \sum_t G_t P_t + \epsilon}{\sum_t G_t + \sum_t P_t + \epsilon},\end{aligned} \tag{4}$$

where G_t and P_t denote the ground truth and prediction of the t-th pixel, and ϵ is a small value for avoiding numerical issues.

As for the signed distance map branch, we normalize the SDM ground truth in the range of $[-1, 1]$, and set the output layer with a tanh activation function. We first calculate the SDM ground truth as:

$$G_{SDM} = \begin{cases} -\min_{j \in \partial G} ||i - j||_2 & \text{if } i \in G_{in} \\ 0 & \text{if } i \in \partial G \\ +\min_{j \in \partial G} ||i - j||_2 & \text{if } i \in G_{out} \end{cases} \tag{5}$$

where G denotes the ground truth area. G_{in}, ∂G and G_{out} the inside area, boundary and outside area of the ground truth region respectively. The $||i - j||_2$ represents the Euclidean distance between the pixel i and j. j represents the point on the segmentation boundary that is nearest to the pixel i. The loss of SDM is computed as:

$$\mathcal{L}_{SDM} = \text{MSE}(P_{SDM}, G_{SDM}). \tag{6}$$

where the MSE denotes the mean squared error function. G_{SDM} and P_{SDM} are the ground truth of signed distance map and the prediction of the signed distance map respectively. When jointly fine-tuning the segmentation model with the segmentation map and signed distance map, the final fracture shape-aware (FSA) loss is defined as a combination of the segmentation map loss and signed distance map loss as:

$$\mathcal{L}_{FSA} = \mathcal{L}_{Seg} + \alpha \times \mathcal{L}_{SDM}. \tag{7}$$

In order to reduce the training difficulty, we set a hyper-parameter α to adjust the influence of SDM. In the first 50 epochs, α is set to 0, increasing to 0.5 linearly during 50-100 epochs. α equals 0.5 eventually until the convergence of the training process.

Besides, false positive predictions in object detection may cause false positive predictions in segmentation. Therefore, we employ the following two false positive suppression strategies in its segmentation prediction process: (1) removal of rib fracture predictions with low confidence, and (2) removal of rib fracture predictions within segmented regions that are too small.

The confidence prediction of rib fracture is generated jointly by the detection model and the segmentation model. Specifically, in the detection model, it generates the confidence p_{det} of the detection box prediction. In the segmentation model, the confidence p_{seg} of the segmentation prediction is represented by the average predicted confidence of all pixels within the segmentation area. Specifically, it can be expressed as:

$$p_{seg} = \frac{1}{|\Omega|} \sum_{i \in \Omega} p_i \tag{8}$$

where Ω represents the set of all pixels within the predicted segmentation of the fracture lesion, p_i denotes the predicted confidence of the segmentation model at pixel i with a range of $[p_0, 1]$. p_0 is the confidence threshold of the segmentation model, and the pixel region with confidence greater than p_0 is considered to be the fracture lesion area. The ultimate confidence prediction p of rib fracture can be expressed as:

$$p = 0.5 \times p_{det} + 0.5 \times p_{seg} \tag{9}$$

In the segmentation model's prediction process, if the segmentation's confidence prediction is less than 0.5, the fracture lesion's segmented result is considered a false positive prediction and is removed from the final result. Moreover, we also remove fracture segmentation predictions with small areas. This strategy mainly relies on empirical judgment, i.e., the total number of false positive predicted pixels is no more than 100. These false positive predictions are mainly caused by noise and artifacts in the rib CT images, which interfere with the model. In the inference process, we remove the segmentation prediction area with a total of less than 100 pixels as false positive predictions.

IV. Experiments and Results

In this section, we will report the experimental results of rib fracture detection and rib fracture segmentation tasks separately. Please refer to S.II - S.IV for more details about the dataset, the evaluation metrics and the implementation in the supporting document. S.V and S.VI report the visualization results. Meanwhile, we provide the ablation study results and analysis in S.VII.

A. Rib Fracture Detection Results

To compare the detection performance of the SA-FracNet, we reproduce two state-of-the-art fracture detection methods, including the 2D target detection network based on Faster R-CNN (Zhou *et al.* [9]), and the 3D detection method based on 3D UNet (Jin *et al.* [12]). We also test some general detection models like Mask R-CNN [49], FCOS [50], and RetinaNet [47]. In the comparative experiments, we use the same ResNeXt backbone and post-processing strategies to attain fair results. Table I and Table II report the fracture detection performance on the public rib fracture dataset and private dataset, respectively.

The experimental results on the public dataset show that SA-FracNet achieves optimal results in both the maximum recall and FROC metrics for rib fracture detection. It can achieve a maximum recall of 92.64% and an average recall of 83.51% with a maximum false positive prediction of less than 6, both of which are higher than all other detection methods, which indicates that SA-FracNet can detect rib fractures missed in other models. Although the single-stage detection model RetinaNet has many more false positive predictions, its recall rate is higher, with a maximum recall rate of 86.80%. The recall rate is higher than other two-stage detection models and the Anchor-free detection model FCOS under the same conditions. Besides, RetinaNet has high prediction efficiency and flexible deployment and is used as the base detection model in SA-FracNet.

In addition, 3D RetinaNet is higher than the 2D detection model in all detection metrics of the detection task, which indicates that the 2D detection method lacks spatial modeling ability and is prone to miss some rib fractures. SA-FracNet, based on its baseline model 3D RetinaNet, improves the detection recall of public and private datasets by 3.4% and 7.6%, respectively, which indicates the efficiency of the self-supervised pre-training. The experimental results on the private dataset demonstrate that all the detection models except SA-FracNet show a significant decrease in detection metrics on the private dataset, ranging from a 6% to a 13% decrease in average recall. Notably, the 3D UNet-based detection method [12] has a 13.62% decrease in recall compared with the performance on the public dataset. In contrast, compared with the baseline model 3D RetinaNet, the performance of SA-FracNet does not degrade significantly because the large-scale self-supervised pre-training enhances the robustness of the model. It still achieves a maximum recall of 91.24% on the private dataset.

B. Rib Fracture Segmentation Results

In order to demonstrate the effectiveness of the multitask segmentation model SA-FracNet proposed in this paper, we select four state-of-the-art semantic segmentation models in the comparative experiments, including 3D Unet [11], 3D DUnet [51] and 3D Attention Unet [11]. Moreover, the fracture detection model proposed by Jin *et al.* [12] and Mask R-CNN [49] can also generate segmentation results. Each segmentation model's experiments are based on the detection result of the SA-FracNet. Besides, we use the same hyper-parameter configurations as the SA-FracNet in the training phases.

The segmentation results of the fracture lesion region of the model on both public and private datasets are demonstrated in Table III. Experimental results show that the multitask segmentation model SA-FracNet achieves the highest segmentation Dice score and the smallest 95HD distance on both public and private datasets, and its average Dice score on both datasets reaches more than 75%, and the average 95HD is both less than 7.5, which indicates that the multitask segmentation model has good robustness.

The deformable convolutional segmentation method improves fracture lesion segmentation, but the SA-FracNet method, which uses segmentation boundary constraints as an auxiliary task, performs better. Specifically, SA-FracNet has a 2.6% higher Dice metric and reduces the 95% Hausdorff distance by 13.7% compared to 3D-DUnet on a private dataset. The deformable convolution method is relatively complex as it models rib morphology by convolutional offsets. Conversely, the segmentation boundary constraint auxiliary task only needs to focus on accurate boundary prediction, simplifying the task's complexity. Hence, this method is more suitable for segmenting rib fracture lesions. Compared to the 3D-Unet segmentation model, the 3D Attention Unet with an attention mechanism does not improve segmentation results in the fracture region. It suggests that the attention mechanism focuses too much on the lesion region and not enough on the surrounding rib region, preventing effective segmentation of the lesion boundary.

It is worth mentioning that the method proposed by Jin *et al.* [12] and Mask R-CNN [49] generated significant differences in segmentation metrics on both test sets. Considering the differences in detection results, it suggests that the generalizability problem of the model is mainly reflected in the detection task. The focus of attention of the segmentation task is the accurate prediction of complex morphological fracture lesions when the detection algorithm detects fractures. The segmentation algorithm is relatively robust, indicating that as a downstream task of the detection model, the self-supervised pre-training approach also indirectly improves the generalizability of the segmentation model.

V. Conclusions

In this paper, we propose SA-FracNet based on deep learning (DL) for rib fracture detection and segmentation in computed tomography (CT) scans. The proposed model consists of a pre-text task based on contrastive learning, a fine-tuned rib fracture detection model, and a fracture shape-aware multi-task segmentation network. The results of the

TABLE I

Comparison of our proposed method with known detection methods on the public test set (80 samples). The best values of each category are marked in **bold**.

Model	Detection (%)						Detection Metric	
	0.5	1	2	4	8	FROC	Recall (%)	Avg FP
Zhou *et al.* [9]	67.32	74.89	81.54	84.20	84.85	78.56	84.85	6.73
Jin *et al.* [12]	66.02	75.11	81.82	87.66	90.26	80.17	90.69	5.28
Mask R-CNN [49]	68.18	77.27	80.52	83.77	84.63	78.87	84.63	6.58
FCOS [50]	67.10	76.19	80.74	84.41	85.50	78.79	85.93	8.25
RetinaNet [47]	67.96	76.62	80.95	85.28	86.36	79.43	86.80	9.85
3D RetinaNet [47]	69.05	77.92	82.24	86.58	88.96	80.95	89.61	8.43
SA-FracNet	**71.21**	**79.65**	**84.42**	**89.83**	**92.42**	**83.51**	**92.64**	**5.98**

TABLE II

Comparison of our proposed method with known detection methods on the private test set (40 samples). The best values of each category are marked in **bold**.

Model	Detection (%)						Detection Metric	
	0.5	1	2	4	8	FROC	Recall (%)	Avg FP
Zhou et al [9]	55.76	62.21	69.12	76.96	79.26	68.66	79.26	7.65
Jin *et al.* [12]	58.06	63.59	70.96	76.50	78.34	69.49	78.34	8.83
Mask R-CNN [49]	56.68	66.36	72.35	78.80	81.11	71.06	81.56	7.18
FCOS [50]	58.99	67.74	74.19	79.72	81.57	72.44	82.03	8.55
RetinaNet [47]	59.91	68.20	74.65	80.65	82.49	73.18	82.95	9.25
3D RetinaNet [47]	61.29	70.05	76.50	82.48	83.87	74.83	84.79	7.98
SA-FracNet	**67.59**	**78.80**	**83.41**	**88.94**	**90.32**	**81.81**	**91.24**	**6.53**

TABLE III

Comparison of our proposed method with known segmentation methods. Separate results obtained on the public test set (80 samples) and private test set (40 samples) are reported in each method's top and bottom rows, respectively. The best values of each category are marked in **bold**.

Model	Public Dataset Results		Private Dataset Results	
	Dice(%)	95HD	Dice(%)	95HD
Jin *et al.* [12]	71.51	13.16	62.39	15.33
Mask R-CNN [49]	65.36	17.98	60.72	18.83
3D AttUnet [11]	70.27	13.84	71.93	14.26
3D Unet [11]	71.13	12.52	73.86	12.07
3D DUnet [51]	74.08	8.33	75.21	7.96
SA-FracNet	**75.46**	**7.16**	**77.14**	**6.88**

experiments conducted on both public and private datasets have demonstrated the robustness and generalization of the proposed model. SA-FracNet has achieved state-of-the-art performance with a detection sensitivity of 0.926 and segmentation Dice of 0.754, surpassing existing methods of automatic rib fracture detection in terms of precision, sensitivity, and robustness.

The proposed method has the potential to significantly improve the clinical evaluation and treatment planning for patients with rib fractures. Our approach provides a promising tool for automatic rib fracture diagnosis by utilizing potentially unlimited clinical data. This paper may also inspire future research in the development of more shape-aware models and enhancing the accuracy and speed of rib fracture diagnosis via deep learning methods.

References

[1] D. W. Ziegler and N. N. Agarwal, "The morbidity and mortality of rib fractures," *The Journal of Trauma*, vol. 37, no. 6, pp. 975–979, 1994.

[2] L. May, C. Hillermann, and S. Patil, "Rib fracture management," *BJA Education*, vol. 16, no. 1, pp. 26–32, 2016.

[3] B. Zhang, C. Jia, R. Wu, B. Lv, B. Li, F. Li, G. Du, Z. Sun, and X. Li, "Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: A clinical evaluation," *The British Journal of Radiology*, vol. 94, no. 1118, p. 20200870, 2021.

[4] 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.

[5] A. Urbaneja, J. De Verbizier, A.-S. Formery, C. Tobon-Gomez, L. Nace, A. Blum, and P. A. G. Teixeira, "Automatic rib cage unfolding with ct cylindrical projection reformat in polytraumatized patients for rib fracture detection and characterization: feasibility and clinical application," *European Journal of Radiology*, vol. 110, pp. 121–127, 2019.

[6] Z. Cao, C. Mu, H. Ying, and J. Wu, "Full scale attention for automated COVID-19 diagnosis from CT images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3213–3216.

[7] Z. Cao, C. Sun, W. Wang, X. Zheng, J. Wu, and H. Gao, "Multi-modality fusion learning for the automatic diagnosis of optic neuropathy," *Pattern Recognition Letters*, vol. 142, pp. 58–64, 2021.

[8] Z. Cao, B. Yu, B. Lei, H. Ying, X. Zhang, D. Z. Chen, and J. Wu, "Cascaded SE-ResUnet for segmentation of thoracic organs at risk," *Neurocomputing*, vol. 453, pp. 357–368, 2021.

[9] Q.-Q. Zhou, W. Tang, J. Wang, Z.-C. Hu, Z.-Y. Xia, R. Zhang, X. Fan, W. Yong, X. Yin, B. Zhang *et al.*, "Automatic detection and classification of rib fractures based on patients' ct images and clinical information via convolutional neural network," *European Radiology*, vol. 31, no. 6, pp. 3815–3825, 2021.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[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. Liu, Z. Cui, Y. Sun, C. Jiang, Z. Chen, H. Yang, Y. Zhang, D. Wu, and D. Shen, "Multi-scale segmentation network for rib fracture

classification from CT images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2021, pp. 546–554.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[15] Y. Wang, K. Zheng, C.-T. Cheng, X.-Y. Zhou, Z. Zheng, J. Xiao, L. Lu, C.-H. Liao, and S. Miao, "Knowledge distillation with adaptive asymmetric label sharpening for semi-supervised fracture detection in chest X-rays," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 599–610.

[16] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 546–12 558, 2020.

[17] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7303–7313.

[18] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 684–16 693.

[19] G. Bier, C. Schabel, A. 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, no. 11, pp. 2173–2180, 2015.

[20] I. Franke, A. Pingen, H. Schiffmann, M. Vogel, D. Vlajnic, R. Ganschow, M. Born *et al.*, "Cardiopulmonary resuscitation (cpr)-related posterior rib fractures in neonates and infants following recommended changes in cpr techniques," *Child abuse & neglect*, vol. 38, no. 7, pp. 1267–1274, 2014.

[21] R. C. Mackersie, T. G. Karagianes, D. B. Hoyt, and J. W. Davis, "Prospective evaluation of epidural and intravenous administration of fentanyl for pain control and restoration of ventilatory function following multiple rib fractures." *The Journal of trauma*, vol. 31, no. 4, pp. 443–9, 1991.

[22] T. Yuan, X. MI, and B. MA, "Clinical research on diagnosis of traumatic rib fractures by multislice ct vrt and dr plain film," *China Medical Equipment*, pp. 92–93, 2013.

[23] A. Mobiny and H. V. Nguyen, "Fast capsnet for lung cancer screening," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 741–749.

[24] M. D. Levine and A. M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 155–164, 1985.

[25] H.-t. Zhang, J.-s. Zhang, H.-h. Zhang, Y.-d. Nan, Y. Zhao, E.-q. Fu, Y.-h. Xie, W. Liu, W.-p. Li, H.-j. Zhang *et al.*, "Automated detection and quantification of covid-19 pneumonia: Ct imaging analysis by a deep learning-based software," *European journal of nuclear medicine and molecular imaging*, vol. 47, no. 11, pp. 2525–2532, 2020.

[26] S. Gunz, S. Erne, E. J. Rawdon, G. Ampanozi, T. Sieberth, R. Affolter, L. C. Ebert, and A. Dobay, "Automated rib fracture detection of postmortem computed tomography images using machine learning techniques," *arXiv preprint arXiv:1908.05467*, 2019.

[27] V. Ibanez, S. Gunz, S. Erne, E. J. Rawdon, G. Ampanozi, S. Franckenberg, T. Sieberth, R. Affolter, L. C. Ebert, and A. Dobay, "Rifnet: Automated rib fracture detection in postmortem computed tomography," *Forensic Science, Medicine and Pathology*, vol. 18, no. 1, pp. 20–29, 2022.

[28] T. Weikert, L. A. Noordtzij, J. Bremerich, B. Stieltjes, V. Parmar, J. Cyriac, G. Sommer, and A. W. Sauter, "Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography," *Korean Journal of Radiology*, vol. 21, no. 7, p. 891, 2020.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] X. H. Meng, D. J. Wu, Z. Wang, X. L. Ma, X. M. Dong, A. E. Liu, and L. Chen, "A fully automated rib fracture detection system on chest ct images and its impact on radiologist performance," *Skeletal Radiology*, vol. 50, no. 9, pp. 1821–1828, 2021.

[31] L. Yao, X. Guan, X. Song, Y. Tan, C. Wang, C. Jin, M. Chen, H. Wang, and M. Zhang, "Rib fracture detection system based on deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[32] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[33] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[34] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[35] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[36] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[38] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[39] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi, "Boosting contrastive self-supervised learning with false negative cancellation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2785–2795.

[40] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

[41] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[42] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[43] P. O. O Pinheiro, A. Almahairi, R. Benmalek, F. Golemo, and A. C. Courville, "Unsupervised learning of dense visual representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4489–4500, 2020.

[44] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 052–10 062.

[45] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatially consistent representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1144–1153.

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[48] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 552–561.

[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[50] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[51] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "Dunet: A deformable network for retinal vessel segmentation," *Knowledge-Based Systems*, vol. 178, pp. 149–162, 2019.