Rectifying Supporting Regions with Mixed and Active Supervision for Rib Fracture Recognition

Yi-Jie Huang, Weiping Liu, Xiuying Wang, Qu Fang, Renzhen Wang ,Yi Wang, Huai Chen, Hao Chen, Deyu Meng, Lisheng Wang

***Abstract*— Automatic rib fracture recognition from chest X-ray images is clinically important yet challenging due to weak saliency of fractures. Weakly Supervised Learning (WSL) models recognize fractures by learning from large-scale image-level labels. In WSL, Class Activation Maps (CAMs) are considered to provide spatial interpretations on classification decisions. However, the high-responding regions, namely *Supporting Regions* of CAMs may erroneously lock to regions irrelevant to fractures, which thereby raises concerns on the reliability of WSL models for clinical applications. Currently available Mixed Supervised Learning (MSL) models utilize object-level labels to assist fitting WSL-derived CAMs. However, as a prerequisite of MSL, the large quantity of precisely delineated labels is rarely available for rib fracture tasks. To address these problems, this paper proposes a novel MSL framework. Firstly, by embedding the adversarial classification learning into WSL frameworks, the proposed Biased Correlation Decoupling and Instance Separation Enhancing strategies guide CAMs to true fractures indirectly. The CAM guidance is insensitive to shape and size variations of object descriptions, thereby enables robust learning from bounding boxes. Secondly, to further minimize annotation cost in MSL, a CAM-based Active Learning strategy is proposed to recognize and annotate samples whose *Supporting Regions* cannot be confidently localized. Consequently, the quantity demand of object-level labels can be reduced without compromising the performance. Over a chest X-ray rib-fracture dataset of 10966 images, the experimental results show that our method produces rational *Supporting Regions* to interpret its classification decisions and outperforms competing methods at an expense of annotating 20% of the positive samples with bounding boxes.**

***Index Terms*— Convolutional Neural Network (CNN), Class Activation Map (CAM), Supporting Region, Mixed Supervised Learning (MSL), Active Learning (AL)**

This work was supported in part by Shanghai Intelligent Medicine Project (2018ZHYL0217) and SJTU Translational Medicine Cross Research Fund (YG2019ZDA26, ZH2018QNA05).*(Corresponding author: Lisheng Wang.)*

Y. Huang, Huai Chen and L. Wang are with Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education (email: huangyj_wuhan@sjtu.edu.cn; chenhuai@sjtu.edu.cn; lswang@sjtu.edu.cn)

W. Liu, Q. Fang and Y. Wang are with Department of Algorithm and Research, Shanghai Aitrox Technology Co.Ltd (email: liuwp@fosun.com; fangqu@fosun.com; wangyi@fosun.com)

X. Wang is with School of Computer Science, The University of Sydney (email: xiu.wang@sydney.edu.au)

R. Wang and D. Meng are with School of Mathematics and Statistics, Xi'an Jiaotong University (email: wrzhen@stu.xjtu.edu.cn; dymeng@mail.xjtu.edu.cn)

Hao Chen is with Department of Computer Science and Engineering, The Chinese University of Hong Kong (email: jackie.haochen@gmail.com)

I. Introduction

RIB fracture is the most common thoracic injury, often resulted from chest trauma, blunt forces as well as medical conditions such as cancer and obesity. Incurrence of rib fracture impairs life quality of the patients, and it is associated with morbidity and mortality [1]. Chest X-ray remains the most effective diagnostic imaging for recognizing rib fracture [2]. However, in projectional 2D X-ray images, apart from multiple overlapping structures, the characteristics of fractures including small sizes and various shapes, arbitrary locations, and normally with blur boundary and weak saliencies make manual rib fracture recognition highly expertise-dependent. Hence, it is in an urgent demand on automatically recognizing rib fractures from chest X-ray images to assist clinicians.

Nowadays, fully supervised learning (FSL) has become benchmarks in lesion localization based on large-scale training datasets with object-level labels, such as pixel-level masks or bounding boxes. FSL is majorly formulated as CNN-based detection or segmentation frameworks. However, collecting object-level labels (referred to as fine-grained labels) by large quantities is of high cost in terms of time and human resources. In addition, collecting exact delineation on hard-to-describe fractures suffers from inter-operator and intra-operator style variation, which cast significant burden on label quality control. Consequently, these drawbacks limit FSL models' clinical practice on chest X-ray images.

On the contrary, weakly supervised learning (WSL) [3]–[5] is majorly formulated as CNN-based classifiers that learn from image-level labels (namely "weak" labels) on disease existence, which can be collected from large-scale diagnosis reports at acceptable cost. Despite the data diversity from brute-force data accumulation, in many chest X-ray tasks, clinical satisfactory performance is yet to meet largely because that these needle-in-haystack tasks lead to WSL's unsatisfactory spatial attention. Fundamentally, to assist clinicians to perceive the rationale for WSL decision-making, classifiers predict object localization from high responding regions of Class

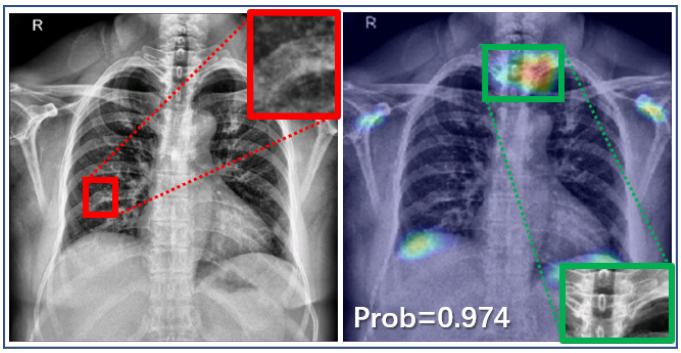

Fig. 1: An example of correctly classifying a sample based on incorrect *Supporting Regions*, where red block indicates a rib-fracture, and green block indicates classification-driven CNN's *Supporting Regions* to its 97.4% confidence of fracture existence.

Activation Maps (CAMs) [6], [7], namely *Supporting Regions*. However, as illustrated in Fig. 1, for fracture localization from chest X-ray, *Supporting Regions* are erroneously located to non-fracture contents though an eventual true positive decision on fracture recognition is produced. The problem of Mis-located *Supporting Regions* limits generalizability of WSL-based computer-aided diagnosis systems and raises clinicians' concerns about the systems' decisions' reliability and indicates WSL's clinically unacceptable biased understanding of the disease, which is mainly caused by two reasons: (1) Firstly, the co-existence of targets (fractures) and background contents (ribs) in the training data formulates biased correlation, as in the water-boat problem [8] where CNN identifies "boat" from seeing "water" because of such co-existence in the training stage. (2) Secondly, WSL models are implicitly multiple-instance learning (MIL) frameworks with their Global Pooling layers assigning spatial pseudo-labels [9], which propagates early mistakes to following epochs and further magnifies the correlation bias. Researchers propose to solve the Mis-located *Supporting Regions* problem in natural images by rebuilding unbiased dataset [7]. However, tackling this problem from dataset perspective is not feasible for rib fracture recognition as the distraction in biased correlation are not understandable and not isolatable. There are also WSL-based researches aiming to improve attention maps. Zhu et al. [10] improved attention maps as well as classification performance by incorporating size priors as a sparsity constraint to CAMs. In addition, [11]–[13] fuse *Supporting Regions* from multiple cascaded models to improve the completeness of targets by erasing prior models' *Supporting Regions* from training images and mining complementary *Supporting Regions* with posterior models. However, they are still challenged by the Mis-located *Supporting Regions* problem as the problem is not pointedly tackled.

To tackle the Mis-located *Supporting Regions* problem, mixed supervised learning (MSL) strategies are proposed in [8], [14], [15] that learn from mixture of large-scale weak labels and extra fine-grained labels. Detection-based MSL methods start from well-trained detection models and learn from weak labels by Expectation-Maximization [16] and Multiple-Instance Learning [17]. In case well-trained detection models are not available, multiple researches [8], [14], [15], [18] proposed classification-based frameworks with CAM-fitting mechanisms which pixel-wisely fit WSL-driven CAMs to object regions. An inspiring approach is to fit CAMs to pixel-level masks [8], [15] or saliency maps inside bounding boxes [8]. However, in chest X-ray images, due to unclear boundaries and overlapped contents, pixel-level masks are extremely challenging to acquire. On the other hand, even in-box saliency maps are difficult to acquire mainly due to high similarity between the fracture and its surroundings in chest X-ray images. In addition, annotating masks is on average $8\times$ slower and thereby more expensive than annotating bounding boxes [19]. With only bounding boxes available as fine-grained labels for disease localization from chest X-ray images, Li et al. [14] proposed a unified loss to learn both localization and classification without compulsory requirement on saliency in the bounding boxes. Instead, it is required that the bounding boxes should enclose the objects as tightly as possible to act as pseudo-masks of rib fractures. However, learning of pseudo-masks can be distracted by nearby non-fracture regions contained in the pseudo-masks, whose shape and size descriptions are significantly affected by inter-operator and intra-operator style variance due to unclear fracture boundaries.

In addition, reducing the quantity demand on fine-grained labels for MSL training remains an open question and yet to investigate. In the literature, Active Learning (AL) models are proposed to selectively query labels from external oracles. Fundamentally, in AL methods [20]–[22], the trained image-level classifiers produce probabilities at image-level to evaluate necessity of annotation. Since classification decision is substantially isolated from *Supporting Regions* correctness, the Mis-located *Supporting Regions* problem remains in these models. Alternatively, Tang et al. [18] presents to mine severity of disease from clinical reports as an indicator of learning difficulty, yet it is challenging to guarantee the availability of the description. Consequently, typical AL methods are not feasible to effectively recommend samples for fine-annotation.

To tackle the Mis-located *Supporting Regions* problem in computer aided diagnosis systems for fracture recognition in chest X-ray images, involving informative fine-grained annotations is necessary but often demands massive extra cost. In this paper, our proposed holistic approach named Supporting Region Rectified Network (SR^2-Net) focuses on mitigating the dependence on annotation quantity and quality by synergistically incorporating novel methods of MSL and Active Learning. Firstly, the proposed MSL method named Fitting-by-not-Fitting alleviates the reliance on high-quality annotations by reducing requirement on annotations from precise pixel-level to bounding-box-level. Specifically, it guides CAMs to target regions ("fitting") by not fitting CAM elements to pixel-wise masks ("not fitting") as in [8], [14], [15], but instead learns from bounding boxes while discriminating genuine targets from in-box background. Secondly, by designing a novel CAM-based uncertainty metric, the proposed Active Learning criterion Query-by-bagging *w.r.t.* CAMs (QBag-CAMs) reduces the demand on the quantity of fine-grained annotations

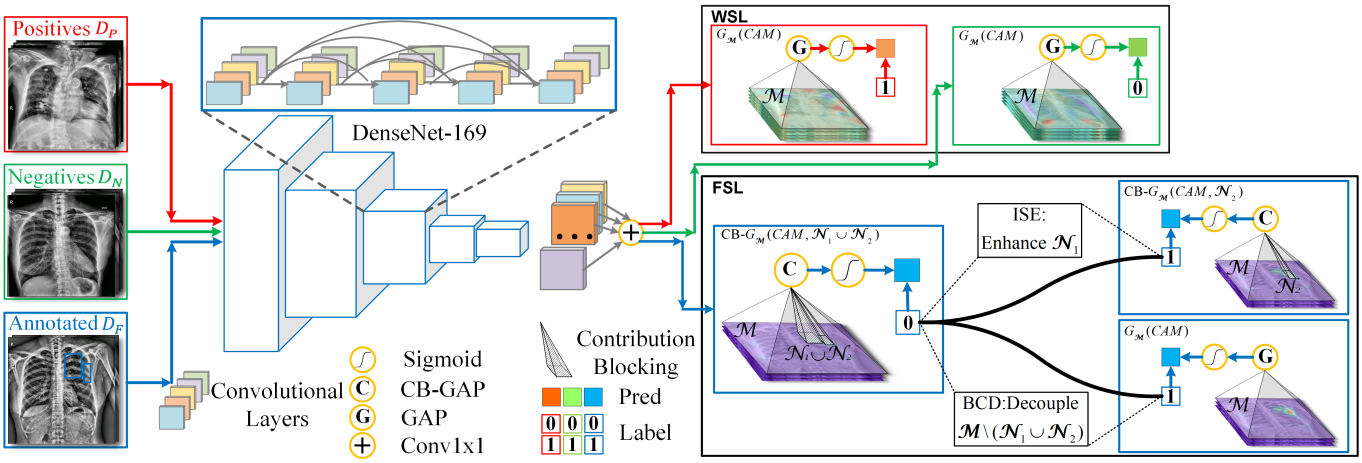

Fig. 2: The MSL training scheme has two branches: the WSL and the FSL branch. Training samples with weak labels pass through the red and green route for WSL, while samples with fine-grained labels pass through the blue route for FSL. In the FSL branch, CB-GAP excludes in-box CAMs from contributing to classification scores. The FSL branch performs Fitting-by-not-Fitting, which is composed of two types of adversarial classification learning: the Biased Correlation Decoupling (BCD) task that learns from pseudo-negative labels by blocking all in-box CAM, and the Instance Separation Enhancing (ISE) task that learns from positive labels by preserving only one in-box CAM while excluding other boxes (when multiple boxes exist).

by identifying samples of high priority for querying localization annotations. Specifically, calculating uncertainty over CAMs effectively evaluates localization difficulty as desired, which can be substantially isolated from commonly used classification uncertainty as in [22]. In summary, the major achievements of this paper are as follows:

1) The proposed Fitting-by-not-Fitting guides WSL-derived CAMs to rib fractures using bounding boxes without being affected by background contained in them. Experiment results show that Fitting-by-not-Fitting produces interpretable localization results to support its classification decisions and outperforms state-of-the-art by a significant margin.
2) Compared to typical Active Learning strategies, QBag-CAMs selects samples prone to Mis-located *Supporting Regions* problem and annotate them with bounding boxes. QBag-CAMs effectively reduces the quantity demand on fine-grained labels without compromising FSL-comparable performance.

The remainder of this paper is organized as follows. We describe our method in Section II and report the experimental results in Section III. Section IV further discusses some insights as well as issues of the proposed method. The conclusions are drawn in Section V.

II. Method

A novel framework Supporting Region Rectified Network (SR^2-Net) is proposed in this paper. Firstly, a Mixed Supervised Learning method named Fitting-by-not-Fitting is introduced to relax the quality requirement on fine-grained labels. Secondly, to minimize annotation cost, an Active Learning strategy named QBag-CAMs is proposed to evaluate potential performance gain from annotating candidate samples with bounding boxes. Initially, we adopt the similar architectures as in the recent methods [3], [13], [14], [23] that reverse the order of Global Pooling layers (generally denoted as G) and Fully Connected layers (FC), so that CAMs are generated before performing Global Pooling. In the training stage, the training samples are divided into three groups: positive dataset D_P, negative dataset D_N and positive samples with fine-grained labels D_F which satisfies $D_F \subset D_P$. As is illustrated in Fig. 2, samples from D_P and D_N pass through the WSL route with equal probability and are learned following a traditional classification scheme, while samples from D_F pass through the FSL route with sampling frequency p_F and are learned following the Fitting-by-not-Fitting routine. In the test stage, all samples pass through the WSL route.

A. Fitting-by-not-Fitting: Adversarial Classification Learning for Supporting Region Rectification

Fitting-by-not-Fitting is a Mixed Supervised Learning method that learns from fine-grained annotations as bounding boxes instead of pixel-wise masks as in [8], [14], [15]. This method address the performance challenge posed by the lack of clear fracture boundaries in chest X-ray images.

By revisiting the bounding box information, image contents are classified into two major categories according to confidence or certainty levels: (1) information with high certainty, namely ascertained information, including distracting contents in background regions as excluded from bounding boxes, and object instances as separately annotated by the bounding boxes, and (2) information of uncertainty, namely uncertain information, including the size and shape of the fractures in chest X-ray due to operator-dependence. Correspondingly, our hypothesis is that ascertained information can be safely learned with full trust, while uncertain information should be handled with spatially-restricted attention mining.

To this end, by embracing biased correlation decoupling

and instance separation enhancing, our proposed Fitting-by-not-Fitting enables spatially-restricted attention mining. As illustrated in Fig. 2, over the samples with fine-grained labels, the proposed method guides classifiers attention mining via two-fold adversarial classification learning:

1) Biased Correlation Decoupling (BCD): To tackle the Mislocated *Supporting Regions* problem caused by biased correlation between distracting background contents and positive classification decisions, we decouple the correlation using the ascertained information of background from bounding box annotation. In particular, as background does not support a positive decision, a classifier with unbiased correlation should recognize a positive sample as negative only if in-box image contents' contribution to the classification probability is blocked. To this end, we design an adversarial classification learning strategy that learns contrastively from positive samples with both positive (following standard classification criterion) and pseudo-negative labels (with all fractures' contribution blocked). Specifically, the contribution blocking is embedded into a Contribution-Blocked (CB) extension of Global Pooling layer G:

$$CB\text{-}G_{\mathcal{M}}(CAM, \mathcal{N}) = G_{\mathcal{M}\setminus\mathcal{N}}(CAM), \tag{1}$$

where $\mathcal{M}$ indicates the whole space of CAM, $\mathcal{N}$ indicates the mask covering all bounding box regions and $G_{\mathcal{M}\setminus\mathcal{N}}(CAM)$ indicates a Global Pooling layer computing over out-of-box regions.

Next, given a positive sample $I_k \in D_F$ with bounding box label and its $CAM(I_k)$, we acquire two classification scores:

$$\begin{aligned} P_{pos} &= Sigmoid(G_{\mathcal{M}}(CAM(I_k))), \\ P_{p\text{-}neg} &= Sigmoid(CB\text{-}G_{\mathcal{M}}(CAM(I_k), \mathcal{N})), \end{aligned} \tag{2}$$

where the positive probability P_{pos} is acquired by typical Pooling layer computed over $\mathcal{M}$, while the pseudo-negative probability $P_{p\text{-}neg}$ is acquired by Contribution-Blocked Global Pooling layer $CB\text{-}G$. In the training stage, the loss function for decoupling is defined as below:

$$\mathcal{L}_{Decouple}(I_k) = -\log P_{pos} - \log(1 - P_{p\text{-}neg}), \tag{3}$$

where by minimizing the loss function $\mathcal{L}_{Decouple}$, P_{pos} approaches 1 and $P_{p\text{-}neg}$ approaches 0, thereby $\mathcal{M}\setminus\mathcal{N}$ is decoupled from the decisive contribution of in-box region $\mathcal{N}$.

2) Instance Separation Enhancing (ISE): The objective of BCD is confined to restricting high CAM responses to a union of sub-regions of $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2 ... \cup \mathcal{N}_K$ given multiple instances $K > 1$. However, it's widely reported [11]–[13] that CAMs tend to highlight only the most discriminative region of target objects and can fail to recall all objects. To tackle this problem, we fully utilize the ascertained information from separate delineation of multiple instances which is complementary to the background to enhance instance separation and to further improve localization performance.

To this end, we propose an extra adversarial classification task to highlight that each instance is important enough to support a positive decision. Given a sample $I_k \in D_F$ with multiple bounding boxes $\mathcal{N}$, by recovering one random instance's CAM contribution while keeping others blocked, a pseudo-positive sample is derived from a pseudo-negative sample. The corresponding $CB\text{-}G$ is defined as below:

$$\begin{aligned} &CB\text{-}G_{\mathcal{M}}(CAM, \mathcal{N} - \mathcal{N}_r) = G_{\mathcal{M}\setminus\mathcal{P}}(CAM), \\ &\mathcal{P} = \mathcal{N} - \mathcal{N}_r, \end{aligned} \tag{4}$$

where $\mathcal{N}_r (1 \le r \le K)$ is a randomly selected bounding box out of $\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_K$.

Here we acquire two classification scores:

$$\begin{aligned} P_{p\text{-}pos} &= Sigmoid(CB\text{-}G_{\mathcal{M}}(CAM(I_k), \mathcal{P})), \\ P_{p\text{-}neg} &= Sigmoid(CB\text{-}G_{\mathcal{M}}(CAM(I_k), \mathcal{N})), \end{aligned} \tag{5}$$

where both the pseudo-positive probability $P_{p\text{-}pos}$ and pseudo-negative probability $P_{p\text{-}neg}$ are acquired by a Contribution-Blocked Global Pooling layer, yet they block different regions of the same image. Training of ISE is driven by the loss function defined as below:

$$\mathcal{L}_{Instance}(I_k) = -\log P_{p\text{-}pos} - \log(1 - P_{p\text{-}neg}), \tag{6}$$

where by minimizing $L_{Instance}$, $P_{p\text{-}pos}$ approaches 1 and $P_{p\text{-}neg}$ approaches 0, thereby each $\mathcal{N}_r$ where $1 \le r \le K$ from $\mathcal{N}$ forms the decisive contribution of a positive classification decision.

3) Spatially-restricted Attention Mining: With the proposed adversarial classification learning strategy, typical WSL's self-exploration attention mining is spatially-restricted from the full image space $\mathcal{M}$ to some sub-regions of each separately annotated instance $\mathcal{N}_r \subseteq \mathcal{N}$ where $1 \le r \le K$. With spatial range significantly narrowed down, the attention mining can be improved by a large extent.

4) Contribution-Blocked Global Pooling Layers: With the targets of above-mentioned adversarial classification learning, we formally define the computing criterion of $G_{\mathcal{M}\setminus\mathcal{N}}$ or $G_{\mathcal{M}\setminus\mathcal{P}}$. As the Contribution-Blocking operation does not violate the computing criterion of existing Global Pooling layers, formulating Contribution-Blocked G is natural and clean:

CB-GAP: Instead of averaging the whole feature maps, the Contribution-Blocked version of Global Average Pooling performs average computing over regions out of bounding boxes:

$$CB\text{-}GAP(CAM, \mathcal{N}) = \frac{\sum_{(h,w)\in\mathcal{M}} (CAM \odot (1 - \mathcal{N}))}{HW - \sum_{(h,w)\in\mathcal{M}} \mathcal{N}}, \tag{7}$$

where $\odot$ denotes element-wise multiplication, H and W denote height and width of the CAM grid, respectively, and $1 \le h \le H, 1 \le w \le W$.

CB-GMP: Instead of picking the maximum element from the whole feature maps, the Contribution-Blocked version of Global Max Pooling selects maximum from elements out of bounding boxes:

$$CB\text{-}GMP(CAM, \mathcal{N}) = \max_{\mathcal{M}\setminus\mathcal{N}} (CAM). \tag{8}$$

CB-LSE: Instead of summarizing the whole feature maps, Contribution-Blocked version of $log\text{-}sum\text{-}exp$ (LSE) performs its computing over regions out of bounding boxes:

$$\begin{aligned} &CB\text{-}LSE(CAM, \mathcal{N}) = \\ &\frac{1}{r} \log\{\frac{\sum_{(h,w)\in\mathcal{M}} \exp(r \odot CAM) \odot (1 - \mathcal{N})}{HW - \sum_{(h,w)\in\mathcal{M}} \mathcal{N}}\}, \end{aligned} \tag{9}$$

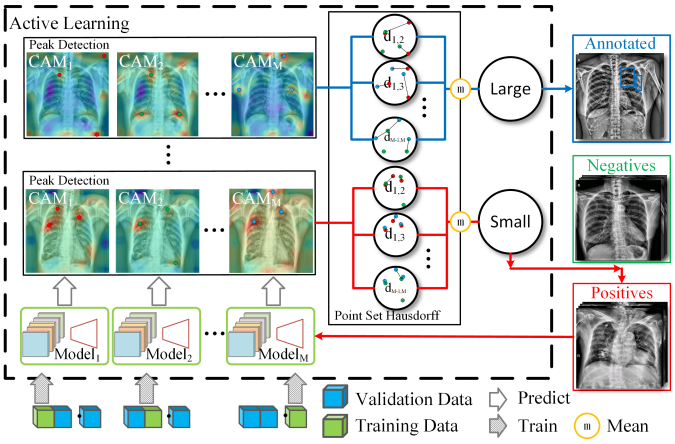

Fig. 3: Illustration of the Active Learning module's acquisition function $a(D_P, Models)$. Committee $Models$ are trained with different non-overlapped subsets of the training set. Then positive samples are fed to committee $Models$ and produced CAMs other than classification probabilities are used to evaluate disagreement. The disagreement is computed as point set distances over local peak sets. Larger distance indicates a samples' being more prone to Mis-located *Supporting Regions* problem, and therefore the samples with large distances are chosen for bounding box labeling.

where r is a hyper-parameter.

5) Loss Function: For samples with weak labels only, the typical classification loss is used. For samples with fine-grained annotations, adversarial losses are included. Formally, the overall loss function is denoted as:

$$\begin{aligned}\mathcal{L}_{all}(I_k) =& \rho(I_k \in D_P)\log p_{I_k} + \\ & \rho(I_k \in D_N)\log(1-p_{I_k}) + \\ & \rho(I_k \in D_F)\mathcal{L}_{Decouple}(I_k) + \\ & \rho(I_k \in D_F)\rho(K>1)\mathcal{L}_{Instance}(I_k),\end{aligned} \tag{10}$$

where $\rho(c) = 1$ if the condition c is satisfied, else $\rho(c) = 0$.

B. Active Learning for Rectifying Supporting Regions

As fine-grained labels are acquired to guide *Supporting Regions*, the proposed Active Learning method, namely Query-by-Bagging $w.r.t.$ CAMs (QBag-CAMs), adopts a novel query strategy to effectively recognize and query annotation for candidates prone to Mis-located *Supporting Regions* problem, thereby saves the annotation expense over candidates whose rib fractures can be localized with confidence by WSL.

To this end, the query strategy probes CAMs instead of image-level probabilities to bridge the gap between classifiers and evaluation of localization difficulty. The query strategy follows a general Query-by-Bagging [21] cycle of committee training, evaluating committee disagreement and querying labels for samples on which the committee members disagree most. Specifically, in each Active Learning iteration, we first train committee $Models = (Model_1, ..., Model_M)$ of M identical models (same as SR^2-Net) with training set (D_P, D_N, D_F)'s non-overlapped subsets $Ds = [(D_{P_1}, D_{N_1}, D_{F_1}), ..., (D_{P_M}, D_{N_M}, D_{F_M})]$, where for $\forall i, j \in [1, M]$, the subset size $|D_{P_i}| = |D_{P_j}|$, $|D_{N_i}| = |D_{N_j}|$ and $|D_{F_i}| = |D_{F_j}|$. Next, we compute their disagreement over CAMs instead of classification probabilities and regard large disagreement as high annotation priority. The Active Learning workflow is described in Algorithm 1.

Algorithm 1 Solving Active Learning Problem by QBag-CAMs

Input:
SR^2-Net: ImageNet pretrained main network
$Models$: Committee networks $(Model_1, ..., Model_M)$
(D_P, D_N, D_F): Positive training dataset, negative training dataset and positive training dataset with fine-grained annotations, respectively
D_f: Positive dataset to be annotated with bounding boxes each AL iteration
s: Sample number of D_f to select each AL iteration
Output:
SR^2-Net_t: Fully trained main network
Functions:
$CAMs \leftarrow P(I_k, Models)$ {CAM of $Models$ given $I_k \in D_P$}
$Model_{t+1} \leftarrow T(D, Model_t)$ {train $Model_t$ with dataset D}
$\boldsymbol{S} \leftarrow F(CAMs)$ {peak feature extraction from $CAMs$}
$\mu_k \leftarrow Avg(\mathcal{H}(\boldsymbol{S}))$ {mean pairwise Hausdorff Distance}
Initialize:
$t \leftarrow 1$, $D_F \leftarrow \emptyset$
$Ds \leftarrow [(D_{P_1}, D_{N_1}, D_{F_1}), ..., (D_{P_M}, D_{N_M}, D_{F_M})]$
SR^2-$Net_t \leftarrow T((D_P, D_N, D_F), SR^2\text{-}Net)$
$Models_t \leftarrow T(Ds, Models)$

while performance of SR^2-Net_t keeps improving **do**
  for $I_k \in D_P$ **do**
    $CAMs \leftarrow P(I_k, Models_t)$
    $\boldsymbol{S} \leftarrow F(CAMs)$
    $\mu_k \leftarrow Avg(\mathcal{H}(\boldsymbol{S}))$
  Rank samples in D_P according to μ_k
  $D_f \leftarrow$ top s of samples of D_P, annotate D_f
  $D_F \leftarrow D_F \cup D_f$
  for each $(D_{P_i}, D_{N_i}, D_{F_i})$ in Ds **do**
    $D_{f_i} \leftarrow D_{F_i} \cap D_f$
    $D_{F_i} \leftarrow D_{F_i} \cup D_{f_i}$
    $Model_{i_{t+1}} \leftarrow T((D_{P_i}, D_{N_i}, D_{F_i}), Model_{i_t})$
  SR^2-$Net_{t+1} \leftarrow T((D_P, D_N, D_F), SR^2\text{-}Net_t)$
  $Models_{t+1} \leftarrow (Model_{1_{t+1}}, ..., Model_{M_{t+1}})$
  $t \leftarrow t+1$
return SR^2-Net_t

Formally, as illustrated in Fig. 3, the query strategy is formulated as an acquisition function $a(D_P, Models)$ that selects s samples to be annotated with bounding boxes by external experts and added to D_F from D_P. Specifically, given a k-indexed candidate sample I_k, we calculate the disagreement over $(CAM_1, CAM_2, ..., CAM_M)$ predicted by $Models$. However, as CAMs are soft heatmaps, point-wise disagreement metric [24] that summarizes the disagreement over each CAM element can be misleading due to (1) class imbalance between rib fractures and chest background; (2) different response magnitudes and sizes among the overlapped *Supporting Regions* of the same target; (3) large disagreement

produced by closely located but not overlapped *Supporting Regions*, while they can all be part of the target object. Therefore, the disagreement is to be evaluated over features extracted from CAMs.

1) Feature Extraction from CAMs: To alleviate the effect of class imbalance, size inconsistency and close non-overlapped responses, we use local peaks to better describe *Supporting Regions*. Fundamentally, *Supporting Regions* of a given $Model_m$ $(m < M)$ are defined as local peak responses of the CAM_m. Since that CAMs are soft heatmaps without value range limitation, local peaks are detected using local maximum filters [25]. Next, we assign a tun-able parameter N to restrict detected peaks to a limited number. Then the acquired *Supporting Regions* are described as a point set $S_m = [(h_1, w_1, p_1), (h_2, w_2, p_2), ..., (h_N, w_N, p_N)]$, where (h_i, w_i) indicates the peak coordinate with top ith response p_i. Also, peaks with magnitude smaller than threshold $Th(CAM)$ are abandoned, therefore the size of S_m is reduced to N_-. For all committee $Models$, we acquire a set $\boldsymbol{S} = (S_1, S_2, ..., S_M)$ of *Supporting Region* sets.

2) Disagreement Evaluation from Features: With *Supporting Regions* peak point sets $\boldsymbol{S}$ from different models, a proper disagreement metric is their distances, where larger distance indicates stronger disagreement. To evaluate point set distance, it's a common practice to utilize Hausdorff Distance [26], [27]. Also, partial matchness should be rewarded since that it's considered also a good though not perfect case where n $(0 < n < N_-)$ peaks are closely located, because the number of fractures in one sample can be less than N_-. Therefore, we use a variant of typical max-based Hausdorff distance, the mean Hausdorff distance as follows:

$$d_{i,j} = \frac{1}{|S_i| + |S_j|}\left(\sum_{a \in S_i} \min_{b \in S_j} \|a - b\| + \sum_{b \in S_j} \min_{a \in S_i} \|b - a\|\right), \quad (11)$$

where $|S_i|$ indicates the number of peaks N_- of S_i, $\|a - b\|$ denotes Euclidean distance between a and b, $i \neq j$ and $S_i, S_j \in \boldsymbol{S}$. Hausdorff distance between each pair of point set denotes the disagreement of each pair of committee models. The overall uncertainty over a sample can be calculated as the mean value of the point set distances as below:

$$\mathcal{H}(\boldsymbol{S}) = (d_{1,2}, ..., d_{M-1,M}), \mu_k = Avg(\mathcal{H}(\boldsymbol{S})), \quad (12)$$

where larger μ_k indicates that the corresponding sample is more likely to introduce Mis-located *Supporting Regions* problem to the training stage and is therefore necessary to be annotated with bounding boxes.

III. Experiments

A. Dataset and Preprocessing

In total 10966 pre-operative chest X-ray images are collected from Foshan Chancheng Central Hospital. The images are of high resolution ($> 2000 \times 2000$ pixels) and in Dicom format, and multiple scans from the same patient are excluded. We divided them into positive subset (including 2434 positive studies with rib fractures) and negative subset (with 8532 negative studies from either healthy persons or patients free of rib fractures). The positive subset is further divided into five non-overlapping subsets that are respectively annotated with bounding boxes by five experienced radiologists. It is worth noting that these pre-operative X-ray images do not contain traits of steel plates. And thereafter, it is more challenging to recognize fractures from our dataset in comparison to MURA [5] where CNNs can identify fractures by referring the implanted steel plates. The positive and negative samples are randomly divided into training, validation and test sets following the 7:1:2 criterion respectively. The test set is composed of 2188 (484 positive and 1704 negative) samples.

Initial experiments showed that lower resolution degrades the performance (896×896 degrades AUPRC by 2%), while higher resolution severely restricts trainable mini-batch size and also limits the performance. Therefore, the images are firstly resized to 1024×1024, and then we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [28] and normalize the images to have zero mean and unit standard deviation before passing the images to CNNs.

B. Implementation Details

The tested methods are all performed on a workstation platform with $2\times$ Xeon E5 CPU (8C16T) @ 2.4 Ghz, 128GB RAM and $4\times$ NVIDIA GTX 1080 Ti GPU with 11GB GPU memory using Ubuntu 16.04 system. The code is implemented with Python 3.6 and PyTorch 1.0.

1) Training Process: The backbone network DenseNet-169 [29] is pretrained by ImageNet [30]. We used Adam [31] optimizer at a learning rate of 10^{-4}. The weights of convolution kernels were penalized with 10^{-4} L2 norm for better generalization capability. We assign 8 samples per mini-batch (2 samples per GPU) since that larger batch size given high resolution samples exceeds GPU memory limitation. Samples from D_F are sampled by sampling frequency p_F and samples from D_P and D_N are sampled by equal probability $(1 - p_F)/2$. For each round of experiment, we select models with best validation classification score (AUPRC) for the result report.

2) Hyper-Parameters: Due to GAP's superior performance in initial experiments, GAP module is chosen as Global Pooling module G in our experiments over GMP and LSE. Also, we did grid search to M, the size of committee among (3,6,9), N, the number of supporting region peaks per sample among (3,6,9) and p_F, the sampling frequency of set D_F among (0.1,0.2,0.3,0.4,0.5), and assigned $M = 3, N = 3$ and $p_F = 0.3$ according to validation performance. The number s of fine-grained labeling each round is assigned as 5% (85) total number of the training positive samples. The threshold $Th(CAM)$ for peak detection is set as $0.2 \times \max_{(h,w) \in \mathcal{M}}(CAM)$.

3) Data Augmentation: We include random horizontal flipping, random rotation between $(-30°, 30°)$ and gray-scale jitterring between $(-0.1\delta, 0.1\delta)$ where δ denotes standard deviation of image intensities.

TABLE I: Comparing the results of self-attention to the proposed method

Method \ Metrics	AUPRC	AUROC	FROC
DenseNet-169 [29]	0.714	0.877	0.463
DenseNet-169-CBAM [32]	0.714	0.886	0.421
ACoL [13]	0.708	0.884	0.467
SR^2-Net + 20% bbox	**0.773**	**0.918**	**0.749**

C. Evaluation Metrics

We evaluate the fracture recognition performance from different methods with three common evaluation metrics including classification metrics Area Under Receiver Operating Characteristic Curve (AUROC) and Area Under Precision-Recall Curve (AUPRC), and a localization metric Free-Response ROC (FROC).

1) AUROC: The Receiver Operating Characteristic (ROC) Curve is widely used to reflect classification performance. It measures the true positive rate ($\frac{TP}{TP+FN}$) against the false positive rate ($\frac{FP}{FP+TN}$) at various threshold settings. To quantify the performance to a scalar, Area Under ROC Curve (AUROC) is calculated. It's worth noting AUROC is sensitive to class-imbalance issue. We use this metric to keep up with the X-ray disease classification community [3], [4], [14].

2) AUPRC: The Precision-Recall (PR) Curve measures the precision ($\frac{TP}{TP+FP}$) rate against the recall rate ($\frac{TP}{TP+FN}$) at various threshold settings. To quantify the performance to a scalar, Area Under PR Curve (AUPRC) is calculated. It is worth noting that as reported in [4], given imbalanced test set (positives fewer than negatives), depending on the imbalance extent, the calculated AUPRC can be usually significantly lower than AUROC.

3) FROC: The Free-Response ROC Curve (FROC) is widely used in object detection tasks and is defined as the plot of sensitivity versus the average number of false-positives per image. To quantify FROC to a scalar, area under FROC is defined as the average sensitivity at five predefined false positive rates: (1, 2, 4, 8, 16) false positives per image. The FROC metric is calculated over positive validation / test samples. For heatmap-based methods, we firstly normalize CAMs to the range of [0,1], and then the top 3% response is selected for generating bounding boxes covering isolated regions, and finally the probability assigned to each generated bounding box is derived from max response in the box. The threshold $Th(IoU)$ is assigned as 0.1 for both detection-based and heatmap-based frameworks to exclude the effect of imperfect shape and size description.

D. Results

The training set is composed of 1704 positive samples and 5972 negative samples. By annotating all positive samples with bounding boxes, a FSL RetinaNet [33], achieved 0.755 localization FROC. When taking all these samples for WSL training, the baseline DenseNet-169 resulted in unsatisfactory FROC of 0.463 while achieving AUROC of 0.877 and AUPRC of 0.714. In comparison, our proposed model reached 0.749 localization FROC and in turn enjoyed 4% higher AUROC (0.918), 6% higher AUPRC (0.773) and 87.8% overall accuracy at the expense of only 20% (340) of the positive samples being annotated with bounding boxes.

Over a few typical hard samples with rib fractures of (a) weak saliency and (b) multiple objects, visual comparison of localization results between different methods is illustrated in Fig. 4. As is shown, our method correctly predicted rib fractures with few false positives, while WSL-based methods displayed severe Mis-located *Supporting Regions* and competing MSL-based methods had significantly more failures. Also, due to enhanced object separation awareness, our method correctly localized multiple instances like FSL did, while WSL methods had only one correct detection. As illustrated in Fig. 5, the three samples were selected based on their corresponding Pointset Distances. The smaller the Pointset Distance is, the easier the sample is to locate fractures from. For instance, the leftmost sample in Fig. 5 is considered as easy sample with small Pointset Distance corresponding to blue color in the color bar. In this easy sample, the fractures are clear from occlusion or overlapping with other structures and the CAMs generated by committee models consistently focus on the true fracture regions. In the intermediate sample, fractures overlap with bronchus, and while the generated CAMs diffusing with the background, there is one true positive CAM for the fracture region. The rightmost sample shows a hard case that the fracture occurs in overlapped rib bones. The CAMs generated are soft and none of the highlighted regions focus to the fracture. Our experimental analyses showed that when the fracture regions overlapped with other contents, it was more likely to result in higher committee disagreements, and therefore harder to be recognized; and in contrast, in the easy samples, the target fracture regions were often clear from occlusion or overlapping with other structures. However, the locations of the fractures have not shown obvious impact on hard samples or easy samples.

Next, we compare our method to related works by evaluating AUROC, AUPRC and FROC as follows:

1) ***Comparison with weakly supervised models****:* To justify the necessity of extra fine-grained labels provided as bounding boxes in MSL-based models, we compared our proposed MSL model with the other three weakly supervised models using the scores shown in TABLE. I: (1) **DenseNet-169 [29]**: Baseline, a state-of-the-art classification network trained with weak labels only. It had a slightly higher performance compared to [3]. There are two possible reasons for this less satisfactory FROC: The Mis-located *Supporting Region* problem, and the problem that CAMs can focus only to the most discriminative part of targets [11]–[13], which results in bad shape descriptions and failing to recall multiple objects. However, the latter problem would merely contribute to the loss of FROC: On the one hand, the IoU threshold $Th(IoU) = 0.1$ is low, thereby correctly detected targets with bad shape descriptions does not get lower scores. On the other hand, the FROC score on the test samples with multiple objects is not significantly lower. Therefore, the major problem is from the Mis-located *Supporting Regions*. (2) **Convolutional Block Attention Module (CBAM) [32]**: CBAM forms self-supervised

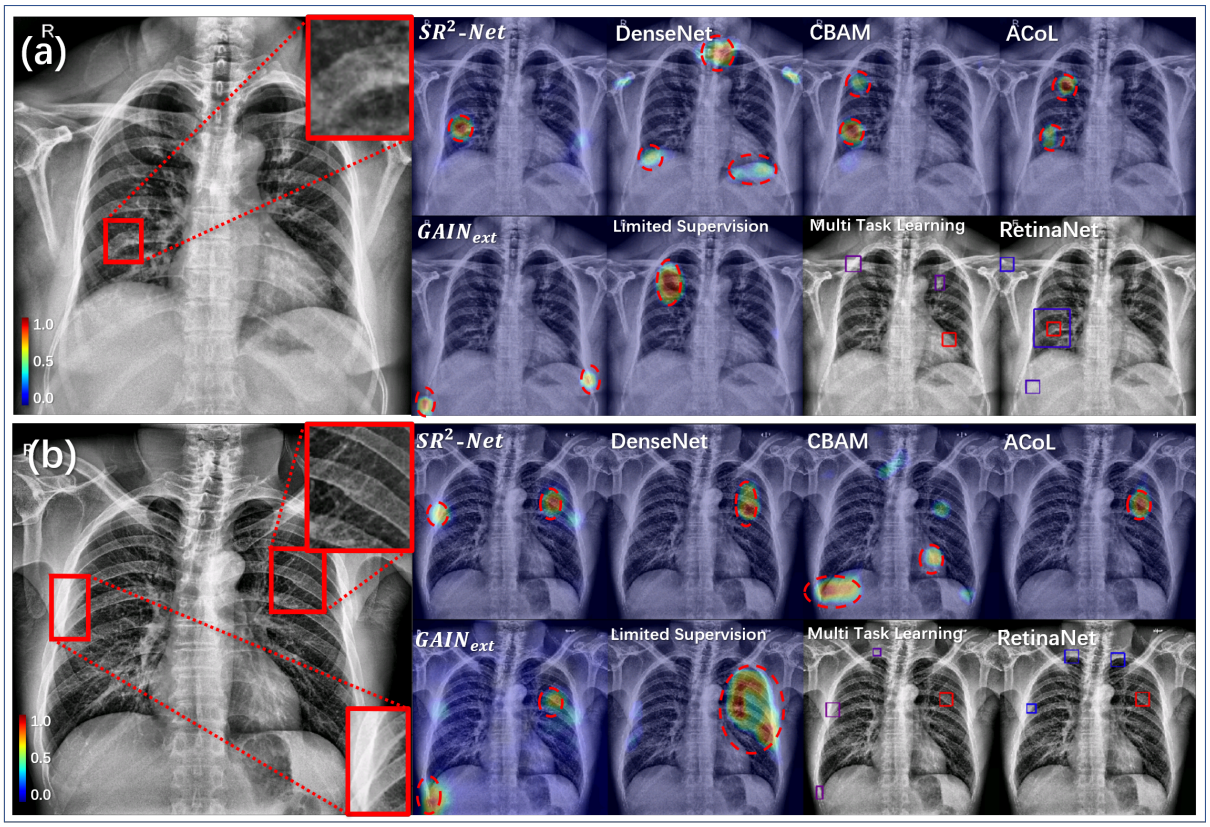

Fig. 4: Typical rib fracture samples with (a) weak saliency and (b) multiple objects. Large images in the left are annotated with ground truth bounding boxes. Smaller ones show localization results of different methods. For both heatmaps and bounding boxes, red denotes high responses while blue denotes low responses.

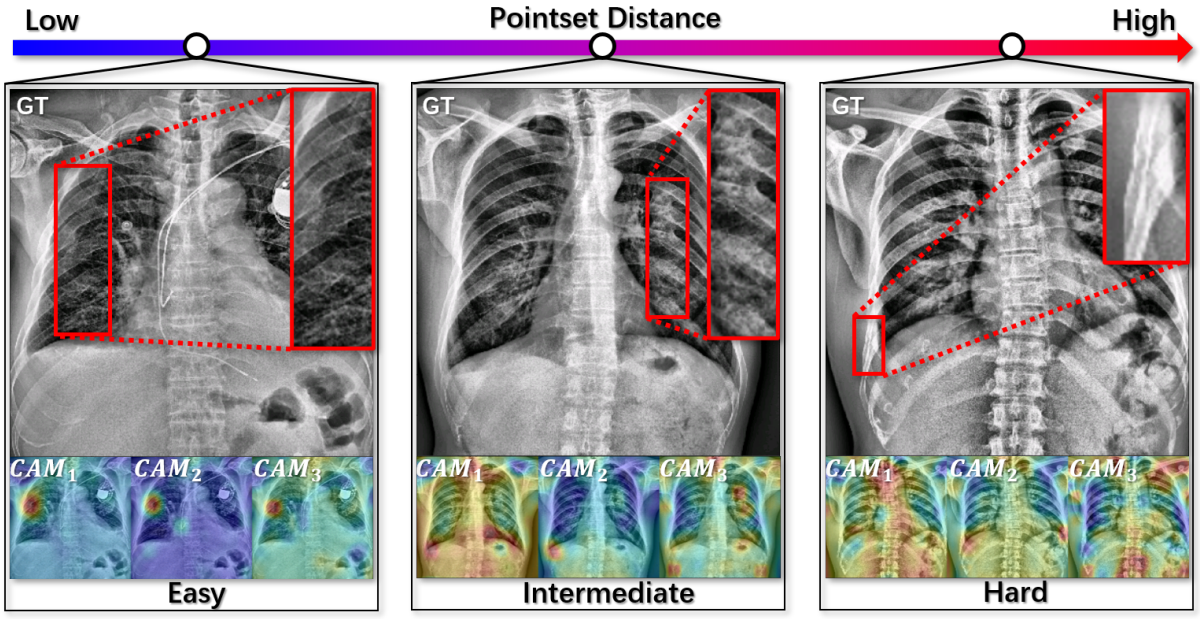

Fig. 5: Illustration of easy-to-locate, intermediate, and hard-to-locate samples selected by QBag-CAMs Active Learning strategy. The color bar, ranging from blue to red, corresponds to increasing Hausdorff Distances of the samples, which accordingly indicate easy to hard samples.

spatial and channel-wise attention modules to arbitrary backbones (DenseNet-169 in this paper). However, its achieved performance was not significantly different from DenseNet-169. (3) **Adversarial Complementary Learning (ACoL)**

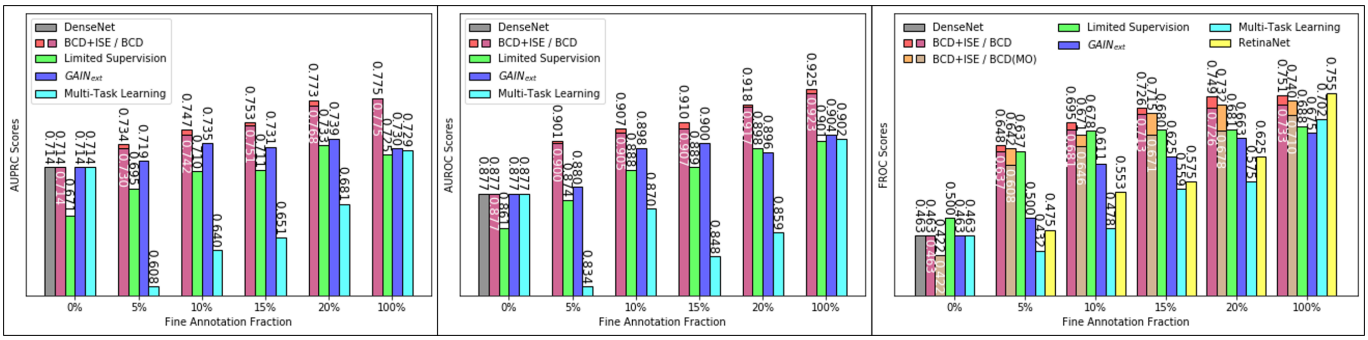

Fig. 6: Performance comparison between MSL methods given 100% weak labels and accumulating fine-grained labels by QBag-CAMs. The proposed method (BCD + ISE) is compared with standalone BCD in a bar painted with two colors. In the FROC figure, the BCD+ISE/BCD(MO) bar indicates the scores on the multiple-object test subset.

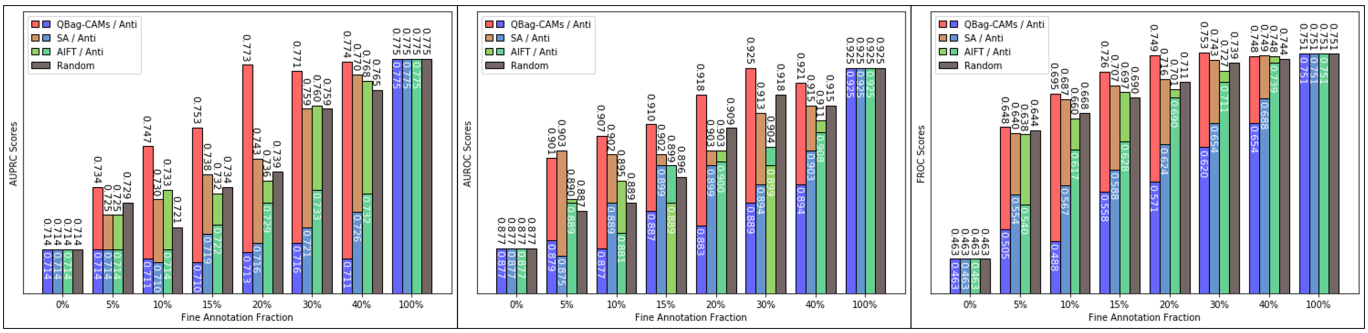

Fig. 7: Comparison between different Active Learning methods given 100% weak labels and accumulating fine-grained labels. Each Active Learning method is also compared with its Anti-Active Learning counterpart in a bar painted with two colors.

[13]: As an improvement of Adversarial Erasing [11], [12], [34], it employs two classifiers A and B and fuse their attention maps to acquire the final localization results. Despite some of its operations being similar to our method in appearance, the experiment suggests that self-driven Adversarial Erasing did not help the network focus to true rib fractures because complementary regions were not unconditionally bias-free as Mis-located *Supporting Regions* problem was not pointedly tackled. (4) SR^2-Net: Apart from significantly improved localization performance (which is natural as fine-grained labels were used), the proposed method boosted classification performance by 6%, which is a large margin. This suggests that extra supervision provided better understanding of target objects and thereby improved the generalization capability. The Discussion section further analyzes the underlying merits.

2) ***Comparison with mixed supervised models****:* We compared our method to state-of-the-art MSL methods by adding different amounts of fine-grained labels to a fixed number of weak labels, as is shown in Fig. 6. Note that since compared methods did not include any Active Learning strategy, all evaluated MSL methods utilized our proposed QBag-CAMs in this experiment to exclude its effect, and all reference methods were DenseNet-169 [29] backboned to exclude the effect of different backbones. (1) **Multi-Task Learning (MTL)**: We trained a RetinaNet [33] with its encoder injected with an extra head for classification training using D_P and D_N, and samples from both D_N and D_F were used to train the detection head. The experimental results showed lowered classification performance compared to raw DenseNet-169 and lowered localization performance compared to a standalone RetinaNet. This result suggests that MTL is not suitable for this task. (2) $GAIN_{ext}$ **[8]**: We compared SR^2-Net to its most direct competitor $GAIN_{ext}$ [8], which is a mask-based MSL method. With only bounding boxes as object-level labels and saliency being impractical to use, we had to transform the bounding box regions to bounding box-shaped pseudo-masks. The results suggest that $GAIN_{ext}$ enjoyed certain performance gains from fine-grained labels but the overall performance was inferior. (3) **Limited Supervision [14]**: Li *et al.* [14] fits patch score heatmaps to bounding boxes in a strict manner, which can also be affected by background contained in bounding boxes. The results suggest that this method also enjoyed certain performance gains but the overall performance was inferior. A common practice of (2) and (3) is that they fit bounding boxes in a strict manner, and we assume that this operation is sub-optimal. To further validate this assumption, see the Discussion section. (4) Ablation Study: We compare Fitting-by-not-Fitting consisting of both BCD and ISE to its standalone BCD counterpart. As ISE is proposed to capture multiple targets, the boosts over classification metrics (AUPRC and AUROC) are reasonably limited. On the other hand, Fitting-by-not-Fitting without ISE has a lower upper-bound in FROC localization score and therefore does not match the fully-supervised RetinaNet. Over a subset of the test set where each sample contains multiple fractures, the FROC gaps are significantly larger, as illustrated in the BCD+ISE/BCD(MO)

bars in the FROC part of Fig. 6.

*3) **Comparison with fully supervised models**:* We also compared our proposed method to a FSL model, RetinaNet [33], to explore the upper-bound of localization performance. A major finding is that our SR^2-Net trained with only 20% fine-grained labels produced comparable FROC to a RetinaNet trained with 100% fine-grained labels, while other MSL methods failed to meet this upper-bound and were outperformed by a large gap. The classifier-based methods' FSL-comparable localization capability suggests that the training scheme is proper and seamless. Furthermore, our method significantly outperformed RetinaNet given 5%-20% fine-grained labels, which suggests that large-scale weak labels were also contributing to localization performances. To further validate this finding, see the Discussion section.

*4) **Comparison with Active Learning methods**:* In this section, we combined different Active Learning methods to our Fitting-by-not-Fitting module, whose scores are shown in Fig. 7: (1) Random Selection: Without any active learning scheme, images for fine-grained labeling were sampled in a random manner. With the increment of D_F's size, each evaluation metric gains certain growth but in a slower manner compared to QBag-CAMs. (2) Active and Incremental Fine-Tuning (AIFT) [22]: A recent classification probabilities-based AL method evaluates "worthines" by combination of patch scores' uncertainty and disagreement. Though it's localization-aware, the experiments showed results without significant difference compared to the random sampling strategy. A major cause is that rib fractures are small local diseases so that both easy samples and hard samples displayed strong patch disagreement. (3) Suggestive Annotation (SA) [24]: This method proposes a point-wise Active Learning method associated with representative sampling, and is an ablation counterpart of our method by excluding peak extraction and peak distance metric. As CAMs are soft heatmaps, by summarizing point-wise disagreements, the large number of weak disagreements overwhelms strong disagreements in small target areas and the uncertainty metric is therefore less discriminative. (4) QBag-CAMs: By the proposed method, images for fine-grained labeling were sampled by evaluating their distance over CAM peak point sets. The proposed method not only achieved best overall performance given half number of fine-grained labels compared to (1), (2) and (3), but performed equally well given only 20% selected positive training samples (340) annotated with bounding boxes compared to using 100% (1704) fine-grained labels. Each Active Learning method was also compared with its Anti-Active Learning counterpart [35]. As is discussed in [35], a better AL metric should produce worse performance when used in anti-AL form. As is shown in Fig. 7, the anti-AL counterpart of the proposed QBag-CAMs showed mostly weakened performance regarding both classification and localization.

IV. Discussion

The experimental validations demonstrated that our SR^2-Net improved the performance of rib fracture recognition and produced accurate *Supporting Regions* to assist spatial interpretation on classification decision. Our major findings are: our method is able to (1) rectify classifiers' biased understanding of target objects; (2) relax the requirement on precisely-delineated object-level labels; (3) more effectively utilize large-scale weak labels.

Firstly, the rectification of biased understanding of rib fractures was achieved by improvement of the training process. By taking a training sample $I_k \in D_P$ as example, Fig. 8 compares I_k's CAMs to illustrate the learning process of our method and WSL method at their increasing epochs. While the probabilities approached the label gradually, the high responding regions in CAMs generated by DenseNet were erroneously locked to distracting image contents, which means, biased understanding of target objects remained in the network and got strengthened over time. In comparison, with the help of a small subset D_F annotated with fine-grained labels, SR^2-Net's *Supporting Regions* improved over time on a weakly-labeled sample from D_P. Consequently, our method enabled rectified attention mining from weakly-labeled samples to achieve better convergence quality.

Secondly, we relaxed the strict requirement [8], [14], [15] on annotation quality, which is mainly attributed to our Fitting-by-not-Fitting strategy. The further experiments on exaggerating label inconsistency validated our method's robustness: By enlarging the bounding boxes by different ratios (also enlarged validation/test bounding boxes), we evaluated the performance degradation, as is listed in Fig. 9. While [14] and [8] witnessed significantly lowered performances, our proposed method remained nearly consistent performances in terms of both classification and localization. In particular, the advantage of robustness was derived from Fitting-by-not-Fitting's careful separation of ascertained background-foreground decoupling, instance separation and uncertain object description within bounding boxes.

Thirdly, SR^2-Net's bias-free convergence made it more effectively and efficiently learns from weak labels. To validate the contribution of weak labels, an additional experiment is conducted: Given 40% (680) randomly selected positive training samples with fine-grained labels and 40% negative samples (2388), we progressively added remaining positive and negative training samples with weak labels to 100% (1704 positive:5968 negative) once by 20% (340 positive:1194 negative). The samples with fine-grained labels are randomly selected because Active Learning cannot be applied without the remaining samples. Interestingly, as observed from Fig. 10, while the baseline DenseNet encountered performance saturation as weak labels accumulate (60%-100%), the performance gains of the proposed method remained near-linear. The different patterns of performance gains from weak labels indicate that our method may have potential in acquiring further performance gains from data accumulation. In the future, we will collect more data to further validate these patterns.

Finally, as we effectively evaluated the certainty of localization via QBag-CAMs, a major future work is to combine semi-supervised learning to our framework to further reduce the need for fine-grained labels. Samples with certain localization can be used as pseudo-labeled samples. Our method is robust

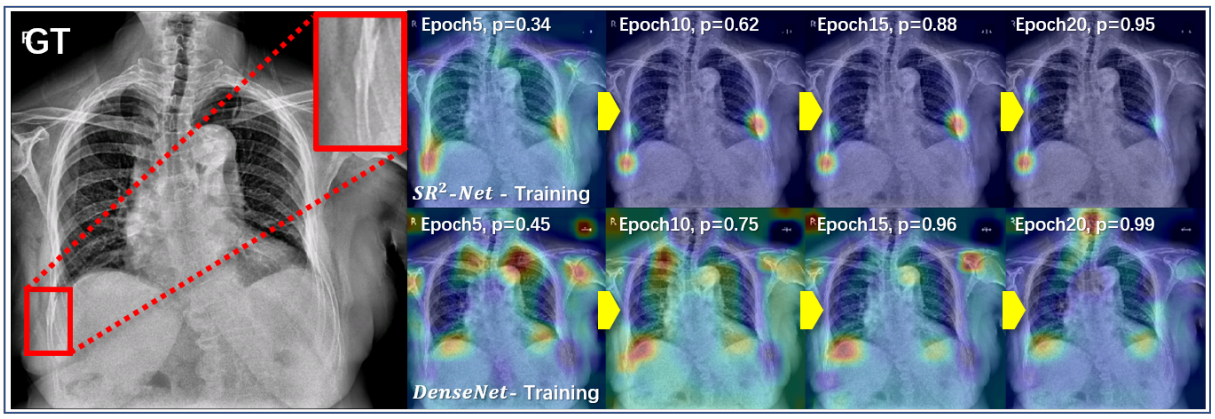

Fig. 8: CAM comparison between our method and WSL-based DenseNet at their different training epochs. As the classification probabilities p approach the label, DenseNet's CAMs are erroneously locked to distracting image contents while our method gets improved over time.

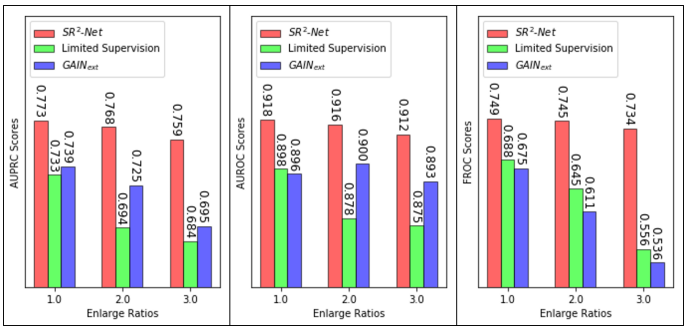

Fig. 9: The results of different methods regarding different ratios of enlarged bounding boxes.

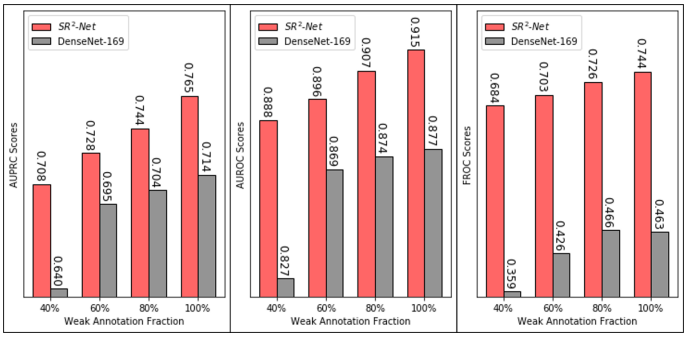

Fig. 10: The results of SR^2-Net and WSL given 40% randomly selected fine-grained labels and accumulating weak labels.

to coarse pseudo-labels and has the potential to effectively utilize them.

V. Conclusion

In this paper, aiming to address the Mis-located *Supporting Regions* problem in a rib fracture recognition task, we proposed a novel method named SR^2-Net. Firstly, other than adopting hard-to-acquire masks, we guided WSL-derived CAMs with bounding box labels by a novel Mixed Supervised Learning method, which is robust to background contained in bounding boxes. In addition, we proposed an Active Learning strategy to select samples prone to Mis-located *Supporting Regions* problem to effectively reduce the quantity need for fine-annotation. Our experiments on a large chest X-ray rib-fracture dataset with 10,966 images demonstrate that our method can produce rational *Supporting Regions* for more accurate spatial interpretations on classification decisions. Our experimental comparisons validate that our method outperforms state-of-the-art methods; in particular, with only 20% of the positive samples with bounding boxes, our method achieves comparable performance as expensive fully supervised learning model.

References

[1] B. S. Talbot, C. P. Gange, Jr., A. Chaturvedi, N. Klionsky, S. K. Hobbs, and A. Chaturvedi, "Traumatic rib injury: patterns, imaging pitfalls, complications, and treatment," *Radiographics*, vol. 37, no. 2, pp. 628–651, 2017.

[2] A. Assi and Y. Nazal, "Rib fracture: Different radiographic projections," *Polish Journal of Radiology*, vol. 77, no. 4, p. 13, 2012.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2097–2106.

[4] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intel.*, 2019, pp. 590–597.

[5] P. Rajpurkar *et al.*, "Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs," in *Conf. Med. Imag. Deep Learn.*, 2018, openReview.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[7] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization." in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[8] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu, "Guided attention inference network," *IEEE Trans. Pattern Anal. Mach. Intel.*, 2019, early Access.

[9] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2132–2141.

[10] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 603–611.

[11] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1568–1576.

[12] Q. Hou, P. Jiang, Y. Wei, and M. Cheng, "Self-erasing network for integral object attention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 549–559.

[13] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1325–1334.

[14] Z. Li *et al.*, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8290–8299.

[15] H. Yang, J. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Trans. Med. Imag.*, 2019, early Access.

[16] W. Zhu, Y. S. Vang, Y. Huang, and X. Xie, "Deepem: Deep 3d convnets with em for weakly supervised pulmonary nodule detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2018, pp. 812–820.

[17] Zhang *et al.*, "Mixed supervised object detection with robust objectness transfer," *IEEE Trans. Pattern An Mach. Intel.*, vol. 41, no. 3, pp. 639–653, 2018.

[18] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2018, pp. 249–258.

[19] A. Bearman, O. Russakovsky, V. Ferrari, and F. Li, "Whats the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 549–565.

[20] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Annual Workshop, Comput. Learn. Theory*. ACM, 1992, pp. 287–294.

[21] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 1998, pp. 1–9.

[22] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4761–4772.

[23] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1713–1721.

[24] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 399–407.

[25] S. van der Walt *et al.*, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[26] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 15, no. 9, pp. 850–863, 1993.

[27] M. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Proc. Int. Conf. Pattern Recognit.*, vol. 1. IEEE, 1994, pp. 566–568.

[28] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," in *Proc. Conf. Vis. Biomed. Comput.* IEEE, 1990, pp. 337–345.

[29] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Int. Conf. Neural Inf. Process. Sys.*, 2012, pp. 1097–1105.

[31] D. P. Kingma and J. Ba., "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent.*, 2015.

[32] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[33] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. PP, no. 99, pp. 2999–3007, 2017.

[34] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3544–3553.

[35] A. Jiménez-Sánchez *et al.*, "Medical-based deep curriculum learning for improved fracture classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 694–702.