FRACTURE DETECTION AND LOCALIZATION IN CHEST X-RAYS USING SEMI-SUPERVISED LEARNING WITH DYNAMIC SHARPENING

Lijuan Lu[⋆] *Shun Miao*[†] *Ling Ye*[‡]

[⋆] Tsinghua University, Beijing, China
[†] PAII Inc., Bethesda, MD, USA
[‡] Ping An Technology, Shenzhen, China

ABSTRACT

In this work, we present a low-cost and efficient method for training a rib and clavicle fracture detection model for chest X-ray (CXR) in a semi-supervised setting where only a small portion of training data with location annotation. Our method leverages the teacher-student model paradigm which forms a consensus prediction of unknown labels using the output under different input augmentation conditions. And most importantly, we develop a dynamic sharpening method to make the pseudo label generated by the teacher model approximate to the true label with low entropy. This dynamic sharpening method adaptively adjusts the sharpening effect according to the performance of the model during the training process, which can effectively cope with the label imbalance problem in the real world, and improve the model sensitivity. The experiment results demonstrate that our method achieves the state-of-the-art fracture detection performance, i.e., an area under receiver operating characteristic curve (AUROC) of 0.9767 and a free-response receiver operating characteristic (FROC) score of 0.9300, significantly outperforming previous approaches by a gap of 1.00% and 3.68% respectively.

***Index Terms*—** Dynamic Sharpening, Semi-supervised Learning, Fracture Detection, Chest X-ray

1. INTRODUCTION

Deep learning (DL) has gained remarkable success in computer vision tasks in recent decades which motivated a range of applications in computer-aided diagnosis (CAD) based on DL, e.g., a CAD system for digital X-ray mammograms via deep learning detection and classification [1, 2]. Convolutional Neural Networks (CNNs) excel at image analysis tasks in supervised learning, they can learn effective representations of images from a large amount of labeled training data. However, annotating training data is a particular bottleneck in the case of medical image object detection task, which requires the participation of medical experts and cost a lot of labor and time. Therefore, reducing annotation costs can be critical for real-world applications of DL, including medical imaging and beyond.

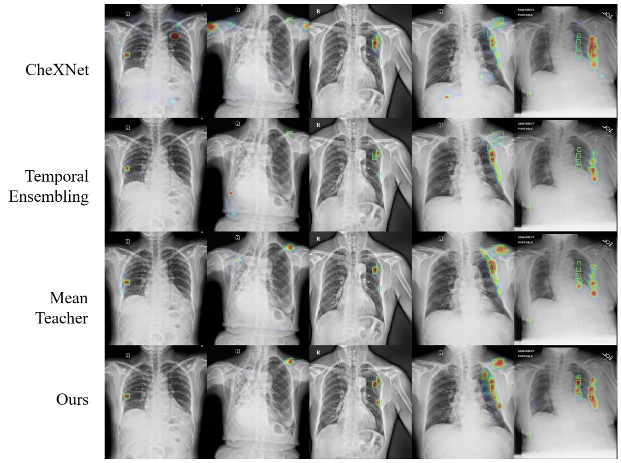


Fig. 1. Visualization of fracture localization results. GT bounding-boxes are shown in green.

In this work, we obtain an accurate rib and clavicle fracture detection and localization model for CXR images with a small number of annotations. Since the abundant public datasets with weakly labels are accessible, e.g., CheXpert [3], PadChest [4], MIMIC [5], semi-supervised learning (SSL) methods are feasible to apply for training a target model, which were already widely used for various tasks in medical image analysis, e.g., skin lesion segmentation [6], thoracic disease identification and localization [7], lung nodule detection [8].

Recently, the majority of SSL methods that achieve state-of-the-art performance are focusing on a consistency-based approach that constrains the model to output invariant results over different perturbations. For example, [9] forms pseudo targets through an ensemble of the model's current predictions and those earlier predictions which also ensures the timing consistency of the model output results, or [10] averages model weights instead of label predictions. However, using only consistency constraints for unlabeled samples without

Thanks to Ping An Technology for funding.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

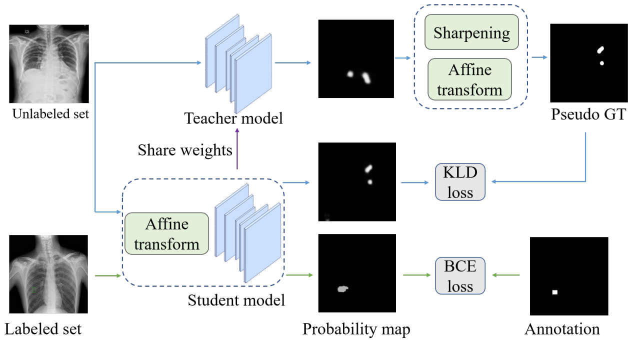


Fig. 2. An overview of the proposed method. The student model is trained via back-propagation. The prediction of the teacher model should be sharpened and transformed.

encouraging the model to predict low-entropy results makes it difficult to mine potential information from unlabeled samples quickly and efficiently when labeled images are limited, especially in the case of object class imbalance (e.g., the normal region is much larger than the fracture region in a fracture CXR.).

For handling this problem, Berthelot et al [11] and Wang et al [12] propose holistic methods using sharpened pseudo-labels to achieve both minimum entropy regularization and consistency regularization. But the sharpening operation with constant hyperparameters they applied could not effectively match the changes of the model (e.g., the sharpening operation proposed by Wang et al [12] could inhibit the activation of the fracture region when the model's sensitivity is low in the early stage of training.). We believe that sharpening operations need to be adjusted in time to suit variant model characteristics, so we propose a dynamic sharpening operation that can automatically adjust the hyperparameters according to the performance of the model.

Overall, we make the following contributions: 1) We propose a cost-effective knowledge distillation-based SSL approach to train an accurate rib and clavicle fracture detection model for CXRs using limited region-level annotations. 2) The dynamic sharpening operation we presented is proved to be effective in generating pseudo Ground-Truth (GT). 3) Compare the classification and localization performance with state-of-the-art methods, our method significantly improves the AUROC by 1.00% and the FROC by 3.68%.

2. METHOD

Different fracture types usually cause significant differences in the scale and shape of fracture characteristics, e.g., the scale of acute fractures is often much larger than chronic fractures. Feature pyramid network (FPN) [13] and ResNet-50 [14] backbone are employed to deal with this challenge. The model finally outputs the probability map for locating the fracture, and the maximum value in the probability map is used as the classification score of the fracture.

Dataset We total collected 32184 CXR images for model training and evaluation. 10438 are collected from a hospital with 9858 negative are denoted by N_{hos} and 580 positive CXRs which are annotated to provide region-level labels are denoted by P_{hos}. 21746 CXRs with image-level labels are collected from public dataset CheXpert [3], PadChest [4] and MIMIC [5] with 16921 negative CXRs are denoted by N_{pub} and 4825 positive CXRs are denoted by P_{pub}.

2.1. Training with fully supervision

In the supervised pre-training setting, CXRs in the negative set from the hospital (N_{hos}) and region-level labeled positive set (P_{hos}) are used to train the network. GT masks are generated by assigning one to the pixels in the fracture region and $zero$ elsewhere. Since a low-sensitivity model always appears with the sample class imbalance (e.g., 580 vs. 9858) and object class imbalance, we use a weighted binary cross-entropy (BCE) loss to improve performance. For CXR I in P_{hos} and N_{hos}, the weighted BCE loss between the GT mask y and the probability map p can be written as

$$L_{sup} = -\alpha_l(\sum_{i \in I}(\alpha_r y_i \log p_i + (1 - y_i)\log(1 - p_i))$$

Where y_i and p_i denote value at pixel i of y and p. α_l and α_r denote weight for handling sample class imbalance and object class imbalance respectively. α_l is set by inverse sample class frequency. α_r is set to 10 in all our experiments following.

2.2. Training with weakly supervision

In the SSL scenarios, we aim to improve the performance of the model by all data which include abundant images without reliable location annotations. As shown in Fig.2, similar to Π-model [9], our network acts as both the student network and the teacher network. The difference is that we do not directly use the output of the teacher network as a pseudo GT but combine a priori image-level annotation to sharpen the output, and the sharpened probability map serves as a pseudo GT to supervise the training of the student network.

For approximating GT from image-level annotations, following recent work [12, 15, 11], we obtain a dynamic sharpening operation $DS(\cdot)$, as described in section 2.3, that can adapt to changes in the model training process. For CXR I in P_{pub}, the predictions of the teacher model are sharpened as $p' = f^{\text{Affi}}(DS(p))$, where $f^{\text{Affi}}(\cdot)$ denotes the affine transformation operation in data augmentation. The KLD loss is employed to constrain the semi-supervised training with the following equation

$$L_{sem} = \sum_{i \in I} (p_i' \log \frac{p_i'}{p_{si}} + (1 - p_i') \log \frac{1 - p_i'}{1 - p_{si}})$$

Where p' denotes the generated pseudo GT. p_s denotes the output of the student model. The total loss used to train the student model is the sum of L_{semi} and L_{sup} which can be expressed as

$$L = L_{semi} + \ L_{sup}$$

2.3. Approximate Ground-truth from Image-level label

Since a class imbalance causes a low sensitivity model, the results obtained in the early training are too different from the GT. To obtain a pseudo GT, we sharpen the output of the teacher network by performing a power function-based operation which can be expressed as

$$DS(y) = \begin{cases} \frac{y^a}{t^{a-1}}, & 0 \leq y \leq t \\ \frac{(y-t)^{1/a}}{(1-t)^{1/a-1}} + t, & t < y \leq 1 \end{cases}$$

As shown in Fig. 3(a), the parameters a and t directly affect the degree and effect of sharpening. t is the center of sharpening. Intuitively, t is a threshold, the function is used to amplify y larger than t and suppress y smaller than t. a represents the degree of sharpening, the greater it is, the greater the sharpening intensity.

The value of t should be set below 0.5 (in the public perception, 0.5 is the threshold for judging whether there is a fracture). Also, its value should be related to the performance of the current model. Therefore, we set the value of t according to m, the median of the model's predicted scores for the positive CXRs, and employ a variable sharpening intensity a as

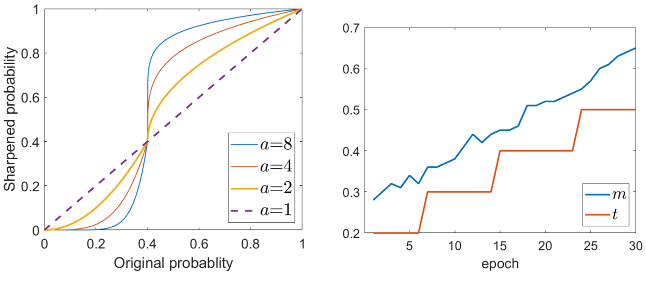
$$t = \min(0.5, \ \max(m - 0.1, \ 0.2))$$

$$a = \begin{cases} 4, & t > 0.35 \\ 6, & t \leq 0.35 \end{cases}$$

Fig. 3(b) shows the value of t and m changes with epoch during SSL training. In the SSL training, the method of knowledge distillation is used to extract fracture information in the positive dataset to improve the performance of the model. Therefore, the sensitivity of the model is gradually increasing and the parameter t increases accordingly.

2.4. Implementation Details

We trained our model on a workstation with a single Intel Xeon E5-2650 v4 CPU@ 2.2 GHz, 128 GB RAM, 4 NVIDIA Quadro RTX 8000 GPUs. All methods are implemented in Python 3.6 and PyTorch v1.6. We use the ImageNet pretrained weights to initialize the backbone network. Adam optimizer is employed in all methods. A learning rate of 4e-5, a weight decay of 0.0001, and a batch size of 32 are used to train the model for 30 epochs. All images are padded to square and resized to 1024 × 1024 for network training and inference. We randomly perform rotation, horizontal flipping, intensity and contrast jittering to augment the training data. 344 CXRs are randomly selected as a hold-out testing set. All experiments are conducted using five-fold cross-validation with an 80%/20% training and validation split, respectively. The trained model is evaluated on the validation set after every training epoch, and the one with the highest validation AUROC is selected as the best model for inference.



(a) (b)

Fig. 3. (a) Dynamic sharpening function at t = 0.4. (b) The values of t (the sharpening center) and m (the median of predict scores of positive CXRs) in each epoch.

3. EXPERIMENTS

Evaluation Metrics We adopt two metrics, AUROC and FROC, to evaluate both fracture classification and localization performances. The maximum response of the probability map is taken as the classification score. FROC is applied to evaluate the fracture localization performance, since our method predicting probability maps without bounding-box, the standard FROC metric based on bounding-box predictions is inapplicable. To quantify FROC, we calculate an FROC score as the average recall at five false positive (FP) rates: (0.1, 0.2, 0.3, 0.4, 0.5) FPs per image. We also report the recall at FP=0.1 (Recall@FP=0.1), the most clinically relevant setting [16]. In detail, we generate the mask of the probability map under different thresholds and calculate the recall according to whether the midpoint of the bounding box in GT is activated in the mask or not. We also find the smallest bounding rectangle of the fracture area in the mask. A detected fracture area is considered as false positive (FP) if the center of its bounding rectangle is not in the range of any bounding box in the GT.

Table 1. Fracture classification and localization performance comparison with state-of-the-art models. Our method achieves the top performance, outperforming all baseline methods by significant margins.

Method	**AUROC**	**FROC**	**Recall @FP=0.1**
CheXNet	0.9667	0.6555	0.5520
TE	0.9343	0.8636	0.7478
MT	0.9467	0.8932	0.7860
Supervised pre-training	0.8906	0.7823	0.6726
Ours	0.9767 **(+1.00%)**	0.9300 **(+3.68%)**	0.8283 **(+5.22%)**

3.1. Comparison with baseline methods

All the comparison methods are implemented with the same network backbone ResNet50 [14] for a fair comparison.

SSL methods We implement current state-of-the-art semi-supervised methods for comparison, including mean teacher (MT) [10] and Temporal Ensemble (TE) [9]. These consistency-based methods applied the same perturbations to the inputs, as elaborated in the implementation details of our method. Especially, when applying the TE method, the network is evaluated twice per input per epoch, the undisturbed output is accumulated into ensemble target for eliminating the effect of affine transformation on the pseudo label.

Image classification methods We implement a state-of-the-art X-ray CAD method based on image classification, CheXNet [17]. A fracture localization map is generated using CAM [18].

Table 1 and Fig. 4 quantitatively summarizes the results of these experiments. As can be seen, Our proposed method measures an AUROC of 0.9767 and an FROC score of 0.9300, which significantly outperforms the most performance baseline method by a 1.00% gap on the AUROC, and a 3.68% gap on the FROC score. CheXNet achieves the superior classification performance among baseline methods, the AUROC achieved by it is 0.9667 which outperforms the SSL methods by 2.00%. CheXNet outperforms the SSL methods in classification performance because it uses image-level labels for fully supervised training. While SSL methods fail to utilize the labels of image-level labeled positive images, resulting in some hard-to-classify images failing to be classified correctly. However, the SSL methods use location information thus the FROC scores are significantly better than CheXNet. Our method utilizes the prior information of image-level labels which baseline methods do not use and yields low entropy pseudo-labels that are closer to GT in the training process. As a result, our student model can distillate more accurate knowledge from the teacher model and achieve state-of-the-art performance in both classification and localization.

Table 2. Ablation study on the sharpening hyperparameters.

a	t	**AUROC**	**FROC**	**Recall @FP=0.1**
4	0.2	0.9536	0.8933	0.8225
	0.3	0.9560	0.9103	0.8115
	0.4	0.9660	0.9123	0.8080
6	0.2	0.9542	0.9018	0.8322
	0.3	0.9588	0.9225	0.8151
	0.4	0.9537	0.9024	0.8180
	Ours	**0.9767**	**0.9300**	**0.8283**

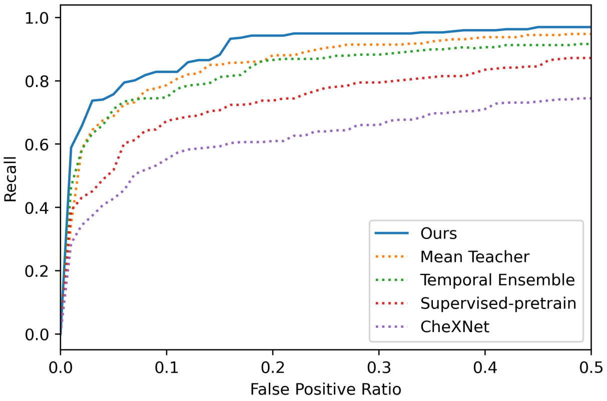


Fig. 4. FROC curves of fracture detection results using different methods.

3.2. Ablation experiments

The ablation studies (see Table 2) are conducted to show the effectiveness of using dynamic parameters in sharpening operations. Our dynamic sharpening achieves the best performance on all metrics. This is because a large constant t (e.g., t=0.4) could suppress the fracture regions that are not yet activated at the beginning of the model training, and a small constant t (e.g., t=0.2) could cause more false positives. Therefore, the value of t needs to be changed during the training process in order to gradually improve the model performance. As it can be seen, when $t = 0.4$, the model works better at $a = 4$ than at $a = 6$, we think that a too-large a would expand the activation area, so we adjust the sharpening intensity a with the change of parameter t.

4. CONCLUSION

We propose a simple and efficient SSL method base on the knowledge distillation method of dynamic sharpening. The proposed sharpening solution improves model performance by generating more correct pseudo-labels and effectively extracting information from unlabeled images. Our method reports the state-of-the-art performance on a hold-out testing set.

5. REFERENCES

[1] Mugahed A Al-Antari, Mohammed A Al-Masni, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim, "A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification," *International journal of medical informatics*, vol. 117, pp. 44–54, 2018.

[2] Mohammed A Al-Masni, Mugahed A Al-Antari, Jeong-Min Park, Geon Gi, Tae-Yeon Kim, Patricio Rivera, Edwin Valarezo, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system," *Computer methods and programs in biomedicine*, vol. 157, pp. 85–94, 2018.

[3] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," vol. 33, no. 01, pp. 590–597, 2019.

[4] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical image analysis*, vol. 66, pp. 101797, 2020.

[5] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[6] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," *arXiv preprint arXiv:1808.03887*, 2018.

[7] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei, "Thoracic disease identification and localization with limited supervision," pp. 8290–8299, 2018.

[8] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang, "Focalmix: Semi-supervised learning for 3d medical image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3951–3960.

[9] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[10] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

[11] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[12] Yirui Wang, Kang Zheng, Chi-Tung Cheng, Xiao-Yun Zhou, Zhilin Zheng, Jing Xiao, Le Lu, Chien-Hung Liao, and Shun Miao, "Knowledge distillation with adaptive asymmetric label sharpening for semi-supervised fracture detection in chest x-rays," pp. 599–610, 2021.

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," pp. 2117–2125, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] Xinyu Zhang, Yirui Wang, Chi-Tung Cheng, Le Lu, Adam P Harrison, Jing Xiao, Chien-Hung Liao, and Shun Miao, "A new window loss function for bone fracture detection and localization in x-ray images with point-based annotation," 2021.

[17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," pp. 2921–2929, 2016.