You've Got Two Teachers: Co-evolutionary Image and Report Distillation for Semi-supervised Anatomical Abnormality Detection in Chest X-Ray

Jinghan Sun[1,2], Dong Wei[2], Zhe Xu[2,3], Donghuan Lu[2], Hong Liu[1,2], Liansheng Wang[1(✉)], and Yefeng Zheng[2]

[1] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
{jhsun,liuhong}@stu.xmu.edu.cn, lswang@xmu.edu.cn
[2] Tencent Healthcare (Shenzhen) Co., LTD, Tencent Jarvis Lab, Shenzhen, China
{donwei,caleblu,yefengzheng}@tencent.com
[3] The Chinese University of Hong Kong, Hong Kong, China
jackxz@link.cuhk.edu.hk

Abstract. Chest X-ray (CXR) anatomical abnormality detection aims at localizing and characterising cardiopulmonary radiological findings in the radiographs, which can expedite clinical workflow and reduce observational oversights. Most existing methods attempted this task in either fully supervised settings which demanded costly mass per-abnormality annotations, or weakly supervised settings which still lagged badly behind fully supervised methods in performance. In this work, we propose a co-evolutionary image and report distillation (CEIRD) framework, which approaches semi-supervised abnormality detection in CXR by grounding the visual detection results with text-classified abnormalities from paired radiology reports, and vice versa. Concretely, based on the classical teacher-student pseudo label distillation (TSD) paradigm, we additionally introduce an auxiliary report classification model, whose prediction is used for report-guided pseudo detection label refinement (RPDLR) in the primary vision detection task. Inversely, we also use the prediction of the vision detection model for abnormality-guided pseudo classification label refinement (APCLR) in the auxiliary report classification task, and propose a co-evolution strategy where the vision and report models mutually promote each other with RPDLR and APCLR performed alternatively. To this end, we effectively incorporate the weak supervision by reports into the semi-supervised TSD pipeline. Besides

J. Sun and D. Wei—Contributed equally; J. Sun contributed to this work during an internship at Tencent.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_35.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 363–373, 2023.
https://doi.org/10.1007/978-3-031-43907-0_35

the cross-modal pseudo label refinement, we further propose an intra-image-modal self-adaptive non-maximum suppression, where the pseudo detection labels generated by the teacher vision model are dynamically rectified by high-confidence predictions by the student. Experimental results on the public MIMIC-CXR benchmark demonstrate CEIRD's superior performance to several up-to-date weakly and semi-supervised methods.

Keywords: Anatomical abnormality detection · Semi-supervised learning · Co-evolution · Visual and textual grounding · Chest X-ray

1 Introduction

Chest X-ray (CXR) is the most commonly performed diagnostic radiograph in medicine, which helps spot abnormalities or diseases of the airways, blood vessels, bones, heart, and lungs. Given the complexity and workload of clinical CXR reading, there is a growing interest in developing automated methods for anatomical abnormality detection in CXR [19]—especially using deep neural networks (DNNs) [14,17,20], which are expected to expedite clinical workflow and reduce observational oversights. Here, the detection task involves both localization (*e.g.*, with bounding boxes) and characterization (*e.g.*, cardiomegaly) of the abnormalities. However, training accurate DNN-based detection models usually requires large-scale datasets with high-quality per-abnormality annotations, which is costly in time, effort, and expense.

To completely relieve the burden of annotation, a few works [2,22,25] resorted to the radiology reports as a form of weak supervision for localization of pneumonia and pneumothorax in CXR. The text report describes important findings in each CXR and is available for most archive radiographs, thus is a valuable source of image-level supervision signal unique to medical image data. However, studies have shown that there are still apparent gaps in performance between image-level weakly supervised and bounding-box-level fully supervised detection [1,11]. Alternatively, seeking for a trade-off between annotation effort and model performance, semi-supervised learning aims to achieve reasonable performance with an acceptable quantity of manual annotations. Semi-supervised object detection methods have achieved noteworthy advances in the natural image domain [4,21,23]. Most of these methods were built on the teacher-student distillation (TSD) paradigm [10], where a teacher model is firstly trained on the labeled data, and then a student model is trained on both the labeled data with real annotations and the unlabeled data with pseudo labels generated (predicted) by the teacher. However, compared with objects in natural images, the abnormalities in CXR can be subtle and less well-defined with ambiguous boundaries, thus likely to introduce great noise to the pseudo labels and eventually lead to suboptimal performance of semi-supervised learning with TSD.

In this paper, we present a co-evolutionary image and report distillation (CEIRD) framework for semi-supervised anatomical abnormality detection in CXR, incorporating the weak supervision by radiology reports. Above all, on

the basis of TSD [10], CEIRD introduces an auxiliary, also semi-supervised, multi-label report classification natural language processing task, whose prediction is used for noise reduction in the pseudo labels of the primary vision detection task, *i.e.*, report-guided pseudo detection label refinement (RPDLR). Then, noting that the performance of the auxiliary language task is crucial to RPDLR, we inversely use the abnormalities detected in the vision task to filter the pseudo labels in the language task, for abnormality-guided pseudo classification label refinement (APCLR). In addition, we implement an iterative co-evolution strategy where RPDLR and APCLR are performed alternatively in a loop, where either model is trained while fixing the other and using the other's prediction for pseudo label refinement. To the best of our knowledge, this is the first work that approaches semi-supervised abnormality detection in CXR by grounding report-classified abnormalities with the visual detection results in the paired radiograph [5,24], and vice versa. Besides the cross-modal pseudo label refinement, we additionally propose self-adaptive non-maximum suppression (SA-NMS) for intra-(image-)modal refinement, too, where the predictions by both the teacher and student vision models go through NMS together to produce new pseudo detection labels for training. In this way, the pseudo labels generated by the teacher are dynamically rectified by high-confidence predictions of the student who is getting better as training goes. Experimental results on the MIMIC-CXR [12,13] public benchmark show that our CEIRD outperforms various up-to-date weakly and semi-supervised methods, and that its building elements are effective.

To summarize, our contributions include: (1) the complementary RPDLR and APCLR for noise reduction in both vision and language pseudo labels for improved semi-supervised training via mutual grounding, (2) the co-evolution strategy for joint optimization of the primary and auxiliary tasks, and (3) the SA-NMS for dynamic intra-image-modal pseudo label refinement, all contributing to the superior performance of the proposed CEIRD framework.

2 Method

Problem Setting. In semi-supervised anatomical abnormality localization, a data set comprising both unlabeled samples $D_u = \{(x_i^u, r_i^u)\}_{i=1}^{N_u}$ and labeled ones $D_l = \{(x_i^l, r_i^l, A_i)\}_{i=1}^{N_l}$ is provided for training, where x and r are a CXR and accompanying report, respectively, $A_i = \{(y^l, B^l)\}$ is the annotation for a labeled sample including both bounding boxes $\{B^l\}$ and corresponding categories $\{y^l\}$, and $N^l \ll N^u$ for practical use scenario. It is worth noting that $\{y^l\}$ can also be considered as classification labels for the report. The objective is to obtain a detection model that can accurately localize and classify the abnormalities in any testing CXR (without report in practice), by making good use of both the labeled and unlabeled CXRs plus the accompanying reports in the training set.

Method Overview. Figure 1 provides an overview of our framework. Suppose a pretrained teacher vision model F_t^I (*e.g.*, on labeled data) for abnormality

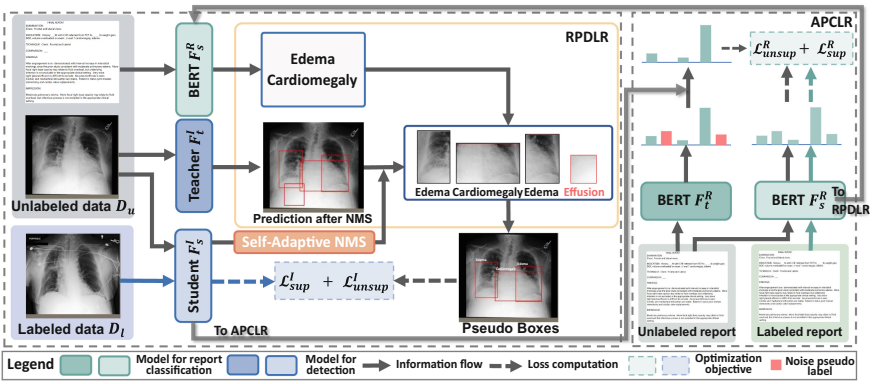

Fig. 1. Overview of the proposed framework. RPDLR: report-guided pseudo detection label refinement; APCLR: abnormality-guided pseudo classification label refinement.

detection in CXR is given, together with a pretrained language model F_s^R for multi-label abnormality classification of reports. On the one hand, we generate for an unlabeled image x_i^u pseudo detection labels with F_t^I and filter the pseudo labels by self-adaptive non-maximum suppression (NMS). Meanwhile, we feed the corresponding report r_i^u into F_s^R and use the prediction for report-guided pseudo detection label refinement (RPDLR). To this end, we obtain refined pseudo labels to supervise the student vision model F_s^I toward better anatomical abnormality localization. On the other hand, we also pass the detection predictions by F_s^I to a teacher language model F_t^R for abnormality-guided pseudo classification label refinement (APCLR), to better supervise the student language model F_s^R on unlabeled data for report-based abnormality classification. In turn, the better language model F_s^R helps train better vision models via RPDLR, thus both types of models co-evolve during training. Note that the real labels are used to train both student models along with the pseudo ones. After training, we only need the student vision model F_s^I for abnormality localization in testing CXRs.

Preliminary Pseudo Label Distillation for Semi-supervised Learning. Both of our baseline semi-supervised vision and language models follow the teacher-student knowledge distillation (TSD) procedure [10]. For report classification, we first train a teacher model F_t^R on labeled reports, and then train a student model F_s^R to predict real labels on labeled reports and pseudo labels produced by F_t^R on unlabeled ones with the loss function:

$$\mathcal{L}^R = \mathcal{L}_{\text{sup}}^R + \mathcal{L}_{\text{unsup}}^R = \sum\nolimits_{D_l} \mathcal{L}_{\text{cls}}^R \left(\hat{y}^l, y^l\right) + \sum\nolimits_{D_u} \mathcal{L}_{\text{cls}}^R \left(\hat{y}, y^{pr}\right), \tag{1}$$

where $\mathcal{L}_{\text{cls}}^R$ is the cross-entropy loss, $\{\hat{y}\} = F_s^R(r)$ is the prediction by the student model, $\{y^{pr}\} = F_t^R(r^u)$ is the set of pseudo labels generated by the teacher model. In each batch, labeled and unlabeled instances are sampled according to a controlled ratio. The resulting report classification model F_s^R will be utilized

later to help with the primary task of abnormality detection in CXR. Similarly, a student vision model F_s^I for abnormality detection in CXR is trained in semi-supervised setting by distilling from a teacher vision model F_t^I trained on labeled CXRs, with the loss function:

$$\begin{aligned}\mathcal{L}^I = \mathcal{L}^I_{\text{sup}} + \mathcal{L}^I_{\text{unsup}} = &\sum\nolimits_{D_l} \left[\mathcal{L}^I_{\text{cls}}\left(\hat{y}^l, y^l\right) + \mathcal{L}^I_{\text{reg}}\left(\hat{B}^l, B^l\right)\right] \\ &+\sum\nolimits_{D_u} \left[\mathcal{L}^I_{\text{cls}}\left(\hat{y}, y^{pv}\right) + \mathcal{L}^I_{\text{reg}}\left(\hat{B}, B^{pv}\right)\right],\end{aligned} \tag{2}$$

where $\{(\hat{y}, \hat{B})\} = F_s^I(x)$ are the predictions by the student model, $\{(y^{pv}, B^{pv})\} = F_t^I(x^u)$ are the pseudo class and bounding box labels generated by the teacher model, $\mathcal{L}^I_{\text{cls}}$ is the focal loss [16] for abnormality classification, and $\mathcal{L}^I_{\text{reg}}$ is the smooth L1 loss for bounding box regression.

Self-adaptive Non-maximum Suppression. During the TSD, the teacher vision model F_t^I is kept fixed. While its knowledge suffices for guiding the student vision model F_s^I in the early stage of TSD, it may somehow impede the learning of F_s^I when F_s^I gradually improves by also learning from the large amount of unlabeled data. Therefore, to gradually improve quality and robustness of the pseudo detection labels as F_s^I learns, we propose to perform self-adaptive non-maximum suppression (SA-NMS) to combine the pseudo labels $\{(y^{pv}, B^{pv})\}$ output by F_t^I and the predictions $\{(\hat{y}, \hat{B})\}$ by F_s^I in each mini batch. Specifically, we perform NMS on the combined set of the pseudo labels and predictions: $\{(y^{cv}, B^{cv})\} = \text{NMS}\big(\{(y^{pv}, B^{pv})\} \bigcup \{(\hat{y}, \hat{B})\}\big)$, and replace $\{(y^{pv}, B^{pv})\}$ in Eq. (2) with $\{(y^{cv}, B^{cv})\}$ for supervision by unlabeled CXRs. In this way, highly confident predictions by the maturing student can rectify imprecise ones by the teacher, leading to better supervision signals stemming from unlabeled data.

Report-Guided Pseudo Label Refinement. In routine clinics, almost every radiograph in archive is accompanied by a report describing findings, abnormalities (if any), and diagnosis. Compared with the captions of natural images, the report texts constitute a unique (to medical image analysis) and rich source of extra information in addition to the image modality. To this end, we propose report-guided pseudo detection label refinement (RPDLR) to make use of this cross-modal information for semi-supervised anatomical abnormality detection in CXR. Specifically, we use the student language model F_s^R (trained with Eq. (1)) to refine the pseudo detection labels. Given a pair of unlabeled image x^u and report r^u, we obtain the set of abnormalities $\{(y^{cv}, B^{cv})\}$ detected in x^u after SA-NMS, and the set of abnormalities $\{\hat{y}\}$ classified in r^u by F_s^R. Then, we only keep the pseudo detection labels whose categories are in the report-classified abnormalities:

$$\{(y^v, B^v)\} = \left\{\left(y_j^{cv}, B_j^{cv}\right) \middle| \, y_j^{cv} \in \{\hat{y}\}\right\}. \tag{3}$$

Eventually, we train the student vision model F_s^I using $\{(y^v, B^v)\}$ in Eq. (2).

Co-evolutionary Semi-supervised Learning with Cycle Pseudo Label Refinement. As the auxiliary student language model F_s^R plays an important role in RPDLR, it is reasonable to optimize its performance which in turn would

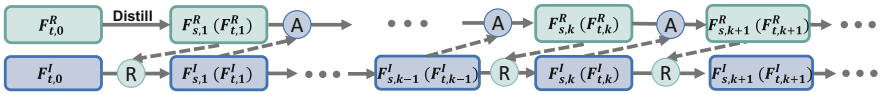

Fig. 2. Illustration of the co-evolution strategy. "R" and "A" represent report- and abnormality-guided pseudo label refinements (RPDLR and APCLR), respectively.

benefit the primary task of abnormality detection. Therefore, we further propose an inverse, abnormality-guided pseudo classification labels refinement (APCLR) to help with semi-supervised training of the report classification model. Similarly in concept to the RPDLP, given a pair of unlabeled image x^u and report r^u, we obtain the set of abnormalities $\{(\hat{y}, \hat{B})\}$ detected in x^u by the student vision model F_s^I, and the set of classification pseudo labels $\{y^{pr}\}$ generated for r^u by the teacher language model F_t^R. We retain only the pseudo labels $\{y_j^{pr} | y_j^{pr} \in \{\hat{y}\}\}$, by excluding the report-classified abnormalities not detected in the paired CXR.

Ideally, one should use an optimal report classification model for refinement of the abnormality detection pseudo labels, and vice versa. However, the two models are mutually dependent on each other in a circle. To solve this dilemma, we implement an alternative co-evolution strategy to refine the abnormality detection and report classification pseudo labels iteratively, in *generations*. As shown in Fig. 2, the k^{th} generation student vision model $F_{s,k}^I$ is distilled from the teacher $F_{t,k-1}^I$, whose pseudo labels are refined by the prediction of the frozen student language model $F_{s,k}^R$ via RPDLR. Subsequently, $F_{s,k}^I$ is frozen and used to 1) help train the $(k+1)^{\text{th}}$ student language model $F_{s,k+1}^R$ via APCLR, and 2) serve as the teacher vision model in next generation: $F_{s,k}^I \rightarrow F_{t,k}^I$.[1] Note that in each generation the students are reborn with random initialization [8]. Thus the co-evolution continues to optimize the vision and report models cyclically with cross-modal mutual promotion. After training, we only need the K^{th} generation student vision model $F_{s,K}^I$ for abnormality detection in upcoming test CXRs.

3 Experiments

Dataset and Evaluation Metrics. We conduct experiments on the chest radiography dataset MIMIC-CXR [12,13], with the detection annotations provided by MS-CXR [3]. MIMIC-CXR is a large publicly available dataset of CXR and free-text radiology reports. MS-CXR provides bounding box annotations for part of the CXRs in MIMIC-CXR (1,026 samples). MIMIC-CXR includes 14 categories of anatomical abnormalities for multi-label classification of the reports, while there are only eight categories in the bounding box annotations of MS-CXR. For consistency, we exclude samples in MIMIC-CXR that have abnormality labels outside the eight categories of MS-CXR, leaving 112,425 samples. Thus, in our semi-supervised setting, 1,026 samples are labeled and the rest are

[1] The initial teachers $F_{t,0}^I$ and $F_{t,0}^R$ are obtained by training on the labeled data only.

Table 1. Abnormality detection results on the test data, using mAP (%) with the IoU thresholds of 0.25, 0.5, and 0.75. TSD: teacher-student distillation.

mAP	CAM [26]	AGXNet [25]	Sup.	TSD [10]	STAC [21]	LabelMatch [4]	Soft Teacher [23]	Ours	Semi-oracle
@0.25	20.47	29.96	37.91	38.29	39.26	39.92	40.17	**41.93**	42.61
@0.5	11.20	15.62	32.84	33.95	35.01	36.40	36.59	**37.20**	37.39
@0.75	3.05	7.44	19.21	19.51	23.90	24.06	24.78	**25.12**	25.66

not.[2] We split the labeled samples for training, validation, and testing according to the ratio of 7:1:2, and use the remaining samples as our unlabeled training data. We focus on the frontal views in this work. The mean average precision (mAP) [7] with the intersection over union (IoU) threshold of 0.25, 0.5, and 0.75 is employed to evaluate the performance of abnormality detection in CXR.

Implementation. The PyTorch [18] framework (1.4.0) is used for experiments. For report classification, we employ the BERT-base uncased model [6] with eight linear heads. Stochastic gradient descent with the momentum of 0.9 and learning rate of 10^{-4} is used for optimization. The batch size is set to 16 reports. For abnormality detection, we employ RetinaNet [16] with FPN [15]+ResNet-101 [9] as backbone. We resize all images to 512×512 pixels and use a batch size of 16. Data augmentation including random cropping and flipping is performed. Our implementation and hyper-parameters follow the official settings [16]. Unless otherwise stated, we evolve the vision and language models for two generations, and train both models for 2000 iterations in each generation (including initial training of the teacher models). The ratio of labeled to unlabeled samples in each mini batch during the semi-supervised training is empirically set to 1:1 and 2:1 for the language and vision models, respectively. The source code is available at: https://github.com/jinghanSunn/CEIRD.

Comparison with State-of-the-Art (SOTA) Methods. We compare our proposed co-evolution image and report distillation (CEIRD) framework with several up-to-date detection methods, including weakly supervised: CAM [26] (locating objects based on class activation maps), AGXNet [25] (aiding CAM-based localization with report representations), fully supervised on labeled training data only (Sup.), baseline semi-supervised via teacher-student pseudo-label distillation (TSD; see Eq. (2)) [10], and three SOTA semi-supervised (STAC [21], LabelMatch [4], and Soft Teacher [23]) object detection methods.

The results are shown in Table 1, from which we have the following observations. First, both fully supervised (by the labeled data only) and semi-supervised methods outperform the weakly-supervised by large margins, proving the efficacy of using limited annotations. Second, all semi-supervised methods outperform the fully supervised (by the labeled data only) by various margins, demonstrating

[2] In this work, we deliberately construct the semi-supervised setting by ignoring the report labels provided in MIMIC-CXR for methodology development. When it comes to a practical new application with no such label available, *e.g.*, semi-supervised lesion detection in color fundus photography, our method can be readily applied.

Table 2. Ablation study results on the validation data.

	Baseline	(a)	(b)	(c)
RPDLR	-	✓	✓	✓
CoE+APCLR	-	-	✓	✓
SA-NMS	-	-	-	✓
mAP@0.25	37.31	39.10	39.72	**41.86**
mAP@0.5	32.76	35.87	35.90	**36.18**
mAP@0.75	17.96	22.02	23.55	**24.49**

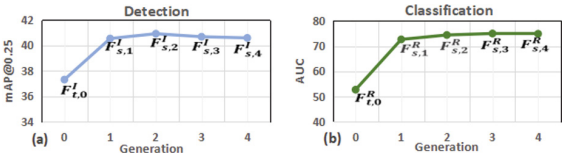

Fig. 3. Performance of (a) vision and (b) report models as a function of generations. AUC: area under the receiver operating characteristic curve.

apparent benefit of making use of the unlabeled data, too. Third, our CEIRD achieves the best performance of all the semi-supervised methods for the mAPs evaluated at three different IoU thresholds, outperforming the second best (Soft Teacher) by up to 1.76%. These results clearly demonstrate the advantage of our method which innovatively integrates the semi-supervision by unlabeled images and the weak supervision by texts. In addition, we evaluate a semi-oracle of our method, where the ground truth report labels provided in MIMIC-CXR are used for RPDLR, instead of the auxiliary model's prediction. As we can see, our method is marginally short of the semi-oracle, *e.g.*, 37.20% versus 37.39% for mAP@0.5, suggesting that our co-evolution strategy can effectively mine the relevant information from the reports. We provide in the supplementary material visualizations of example detection results by Soft Teacher and our method, where ours are visually superior with fewer misses (left), fewer false positives (middle), and better localization (right). Lastly, we also provide performance evaluation of the auxiliary report classification task in the supplement.

Ablation Study. We conduct ablation studies on the validation data to investigate efficacy of the novel building elements of our CEIRD framework, including: report-guided pseudo detection label refinement (RPDLR), co-evolution strategy (CoE) with abnormality-guided pseudo classification label refinement (APCLR), and self-adaptive non-maximum suppression (SA-NMS). We use the preliminary teacher-to-student pseudo label distillation as baseline (Eq. (2)). As shown in Table 2, RPDLR (column (a)) substantially boosts performance upon the baseline by 1.79–4.06% in mAPs thanks to the refined pseudo detection labels. By adopting CoE+APCLR (column (b)), we achieve further performance improvements up to 1.53% as the auxiliary report classification model gets better together. Last but not the least, the introduction of SA-NMS (column (c)) also brings improvements up to 2.14%. These results validate the novel design of our framework. In addition, we conduct experiments to empirically determine the optimal number of generations for the co-evolution. The results are shown in Fig. 3, where the 0^{th} generation means fully supervised models trained on the labeled data only (*i.e.*, the initial teacher models $F^I_{t,0}$ and $F^R_{t,0}$). As we can see, the vision and report models improve in the first two and three generations, respectively, and then remain stable in the following ones, confirming that both models promote each other with the co-evolution strategy. Since our ulti-

mate objective is abnormality detection in CXR, we select two generations for comparison with other methods.

4 Conclusion

In this work, we proposed a new co-evolutionary image and report distillation (CEIRD) framework for semi-supervised anatomical abnormality detection in chest X-ray. On the basis of a preliminary teacher-student pseudo label distillation, we first presented self-adaptive NMS to mingle highly confident predictions by both the teacher and student for improved pseudo labels. We then proposed report-guided pseudo detection label refinement (RPDLR) that used abnormalities classified from the accompanying radiology reports by an auxiliary language model to eliminate unmatched pseudo labels. Meanwhile, we further proposed an inverse, abnormality-guided pseudo classification label refinement (APCLR) making use of the abnormalities detected in X-ray images for better language model training. In addition, we implemented a co-evolution strategy that looped the RPDLR and APCLR to iteratively optimize the main vision detection model and auxiliary report classification model in an alternative manner. Experimental results showed that our CEIRD framework achieved superior performance to up-to-date semi-/weakly-supervised methods.

Acknowledgement. This work was supported by the National Key R&D Program of China under Grant 2020AAA0109500/2020AAA0109501 and the National Key Research and Development Program of China (2019YFE0113900).

References

1. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_34
2. Bhalodia, R., et al.: Improving pneumonia localization via cross-attention on medical images and reports. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 571–581. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_53
3. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision-language processing. arxiv Preprint: arxiv:2204.09817 (2022)
4. Chen, B., et al.: Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14381–14390 (2022)
5. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2Ground: weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2601–2610 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arxiv Preprint: arxiv:1810.04805 (2018)

7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int. J. Comput. Vision **88**, 303–308 (2009)
8. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning, pp. 1607–1616. PMLR (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arxiv Preprint: arxiv:1503.02531 (2015)
11. Ji, H., et al.: A benchmark for weakly semi-supervised abnormality localization in chest X-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13433, pp. 249–260. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_24
12. Johnson, A., et al.: MIMIC-CXR-JPG-chest radiographs with structured labels. PhysioNet (2019)
13. Johnson, A.E., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arxiv Preprint: arxiv:1901.07042 (2019)
14. Lakhani, P., Sundaram, B.: Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology **284**(2), 574–582 (2017)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
17. Oğul, B.B., Koşucu, P., Özçam, A., Kanik, S.D.: Lung nodule detection in X-ray images: a new feature set. In: Lacković, I., Vasic, D. (eds.) 6th European Conference of the International Federation for Medical and Biological Engineering. IP, vol. 45, pp. 150–155. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11128-5_38
18. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
19. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed. Eng. Online **17**(1), 1–23 (2018)
20. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arxiv preprint: arxiv:1711.05225 (2017)
21. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arxiv Preprint: arxiv:2005.04757 (2020)
22. Tam, L.K., Wang, X., Turkbey, E., Lu, K., Wen, Y., Xu, D.: Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 45–55. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_5
23. Xu, M., et al.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3060–3069 (2021)

24. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4683–4693 (2019)
25. Yu, K., Ghosh, S., Liu, Z., Deible, C., Batmanghelich, K.: Anatomy-guided weakly-supervised abnormality localization in chest X-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13435, pp. 658–668. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_63
26. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)