Anatomy-Driven Pathology Detection on Chest X-rays

Philip Müller[1](✉), Felix Meissen[1], Johannes Brandt[1], Georgios Kaissis[1,2], and Daniel Rueckert[1,3]

[1] Institute for AI in Medicine, Technical University of Munich, Munich, Germany
philip.j.mueller@tum.de
[2] Helmholtz Zentrum Munich, Munich, Germany
[3] Department of Computing, Imperial College London, London, UK

Abstract. Pathology detection and delineation enables the automatic interpretation of medical scans such as chest X-rays while providing a high level of explainability to support radiologists in making informed decisions. However, annotating pathology bounding boxes is a time-consuming task such that large public datasets for this purpose are scarce. Current approaches thus use weakly supervised object detection to learn the (rough) localization of pathologies from image-level annotations, which is however limited in performance due to the lack of bounding box supervision. We therefore propose anatomy-driven pathology detection (ADPD), which uses easy-to-annotate bounding boxes of anatomical regions as proxies for pathologies. We study two training approaches: supervised training using anatomy-level pathology labels and multiple instance learning (MIL) with image-level pathology labels. Our results show that our anatomy-level training approach outperforms weakly supervised methods and fully supervised detection with limited training samples, and our MIL approach is competitive with both baseline approaches, therefore demonstrating the potential of our approach.

Keywords: Pathology detection · Anatomical regions · Chest X-rays

1 Introduction

Chest radiographs (chest X-rays) represent the most widely utilized type of medical imaging examination globally and hold immense significance in the detection of prevalent thoracic diseases, including pneumonia and lung cancer, making them a crucial tool in clinical care [10,15]. Pathology detection and localization – for brevity we will use the term *pathology detection* throughout this work – enables the automatic interpretation of medical scans such as chest X-rays by predicting bounding boxes for detected pathologies. Unlike classification, which only predicts the presence of pathologies, it provides a high level of explainability supporting radiologists in making informed decisions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_6.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 57–66, 2023.
https://doi.org/10.1007/978-3-031-43907-0_6

However, while image classification labels can be automatically extracted from electronic health records or radiology reports [7,20], this is typically not possible for bounding boxes, thus limiting the availability of large datasets for pathology detection. Additionally, manually annotating pathology bounding boxes is a time-consuming task, further exacerbating the issue. The resulting scarcity of large, publicly available datasets with pathology bounding boxes limits the use of supervised methods for pathology detection, such that current approaches typically follow weakly supervised object detection approaches, where only classification labels are required for training. However, as these methods are not guided by any form of bounding boxes, their performance is limited.

We, therefore, propose a novel approach towards pathology detection that uses anatomical region bounding boxes, solely defined on anatomical structures, as proxies for pathology bounding boxes. These region boxes are easier to annotate – the physiological shape of a healthy subject's thorax can be learned relatively easily by medical students – and generalize better than those of pathologies, such that huge labeled datasets are available [21]. In summary:

- We propose anatomy-driven pathology detection (ADPD), a pathology detection approach for chest X-rays, trained with pathology classification labels together with anatomical region bounding boxes as proxies for pathologies.
- We study two training approaches: using localized (anatomy-level) pathology labels for our model *Loc-ADPD* and using image-level labels with multiple instance learning (MIL) for our model *MIL-ADPD*.
- We train our models on the Chest ImaGenome [21] dataset and evaluate on NIH ChestX-ray 8 [20], where we found that our Loc-ADPD model outperforms both, weakly supervised methods and fully supervised detection with a small training set, while our MIL-ADPD model is competitive with supervised detection and slightly outperforms weakly supervised approaches.

2 Related Work

Weakly Supervised Pathology Detection. Due to the scarcity of bounding box annotations, pathology detection on chest X-rays is often tackled using weakly supervised object detection with Class Activation Mapping (CAM) [25], which only requires image-level classification labels. After training a classification model with global average pooling (GAP), an activation heatmap is computed by classifying each individual patch (extracted before pooling) with the trained classifier, before thresholding this heatmap for predicting bounding boxes. Inspired by this approach, several methods have been developed for chest X-rays [6,14,20,23]. While CheXNet [14] follows the original approach, the method provided with the NIH ChestX-ray 8 dataset [20] and the STL method [6] use Logsumexp (LSE) pooling [13], while the MultiMap model [23] uses max-min pooling as first proposed for the WELDON [3] method. Unlike our method, none of these methods utilize anatomical regions as proxies for predicting pathology bounding boxes, therefore leading to inferior performance.

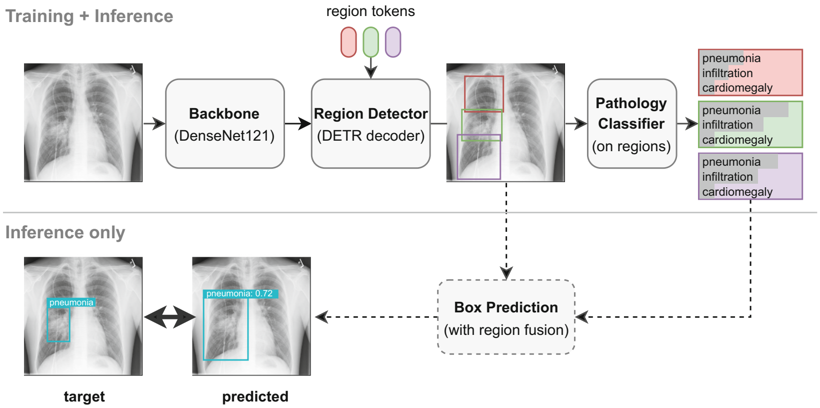

Fig. 1. Overview of our method. Anatomical regions are first detected using a CNN backbone and a shallow detector. For each region, observed pathologies are predicted using a shared classifier. Bounding boxes for each pathology are then predicted by considering regions with positive predictions and fusing overlapping boxes.

Localized Pathology Classification. Anatomy-level pathology labels have been utilized before to train localized pathology classifiers [1,21] or to improve weakly supervised pathology detection [24]. Along with the Chest ImaGenome dataset [21] several localized pathology classification models have been proposed which use a Faster R-CNN [16] to extract anatomical region features before predicting observed pathologies for each region using either a linear model or a GCN model based on pathology co-occurrences. This approach has been further extended to use GCNs on anatomical region relationships [1]. While utilizing the same form of supervision as our method, these methods do not tackle pathology detection.

In AGXNet [24], anatomy-level pathology classification labels are used to train a weakly-supervised pathology detection model. Unlike our and the other described methods, it does however not use anatomical region bounding boxes.

3 Method

3.1 Model

Figure 1 provides an overview of our method. Given a chest X-ray, we apply a DenseNet121 [5] backbone and extract patch-wise features by using the feature map after the last convolutional layer (before GAP). We then apply a lightweight object detection model consisting of a single DETR [2] decoder layer to detect anatomical regions. Following [2], we use learned query tokens attending to patch features in the decoder layer, where each token corresponds to one predicted bounding box. As no anatomical region can occur more than once in each chest X-ray, each query token is assigned to exactly one pre-defined anatomical region, such that the number of tokens equals the number of anatomical regions. This one-to-one assignment of tokens and regions allows us to remove the Hungarian

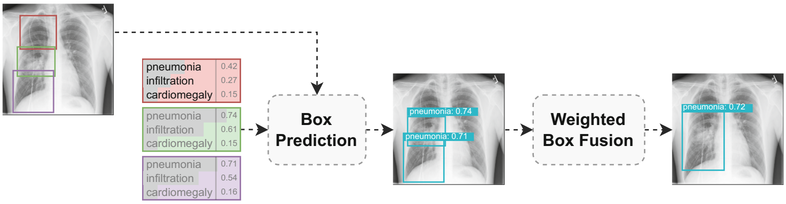

Fig. 2. Inference. For each pathology, the regions with pathology probability above a threshold are predicted as bounding boxes, which are then fused if overlapping.

matching used in [2]. As described next, the resulting per-region features from the output of the decoder layer will be used for predictions on each region.

For predicting whether the associated region is present, we use a binary classifier with a single linear layer, for bounding box prediction we use a three-layer MLP followed by sigmoid. We consider the prediction of observed pathologies as a multi-label binary classification task and use a single linear layer (followed by sigmoid) to predict the probabilities of all pathologies. Each of these predictors is applied independently to each region with their weights shared across regions.

We experimented with more complex pathology predictors like an MLP or a transformer layer but did not observe any benefits. We also did not observe improvements when using several decoder layers and observed degrading performance when using ROI pooling to compute region features.

3.2 Inference

During inference, the trained model predicts anatomical region bounding boxes and per-region pathology probabilities, which are then used to predict pathology bounding boxes in two steps, as shown in Fig. 2. In step (i), pathology probabilities are first thresholded and for each positive pathology (with probability larger than the threshold) the bounding box of the corresponding anatomical region is predicted as its pathology box, using the pathology probability as box score. This means, if a region contains several predicted pathologies, then all of its predicted pathologies share the same bounding box during step (i). In step (ii), weighted box fusion (WBF) [19] merges bounding boxes of the same pathology with IoU-overlaps above 0.03 and computes weighted averages (using box scores as weights) of their box coordinates. As many anatomical regions are at least partially overlapping, and we use a small IoU-overlap threshold, this allows the model to either pull the predicted boxes to relevant subparts of an anatomical region or to predict that pathologies stretch over several regions.

3.3 Training

The anatomical region detector is trained using the DETR loss [2] with fixed one-to-one matching (i.e. without Hungarian matching). For training the pathology classifier, we experiment with two different levels of supervision (Fig. 3).

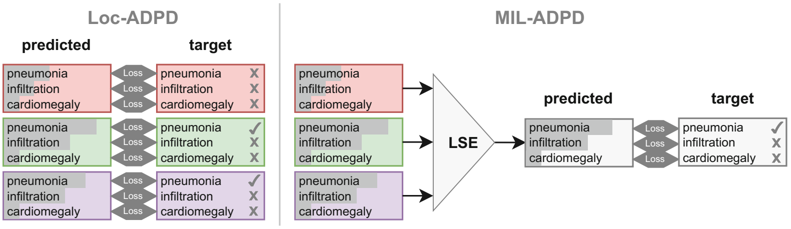

Fig. 3. Training. **Loc-ADPD:** Pathology predictions of regions are directly trained using anatomy-level supervision. **MIL-ADPD:** Region predictions are first aggregated using LSE pooling and then trained using image-level supervision.

For our *Loc-ADPD* model, we utilize anatomy-level pathology classification labels. Here, the target set of observed pathologies is available for each anatomical region individually such that the pathology observation prediction can directly be trained for each anatomical region. We apply the ASL [17] loss function independently on each region-pathology pair and average the results over all regions and pathologies. The decoder feature dimension is set to 512.

For our *MIL-ADPD* model, we experiment with a weaker form of supervision, where pathology classification labels are only available on the per-image level. We utilize multiple instance learning (MIL), where an image is considered a bag of individual instances (i.e. the anatomical regions), and only a single label (per pathology) is provided for the whole bag, which is positive if any of its instances is positive. To train using MIL, we first aggregate the predicted pathology probabilities of each region over all detected regions in the image using LSE pooling [13], acting as a smooth approximation of max pooling. The resulting per-image probability for each pathology is then trained using the ASL [17] loss. In this model, the decoder feature dimension is set to 256.

In both models, the ASL loss is weighted by a factor of 0.01 before adding it to the DETR loss. We train using AdamW [12] with a learning rate of 3e−5 (Loc-ADPD) or 1e−4 (MIL-ADPD) and weight decay 1e−5 (Loc-ADPD) or 1e−4 (MIL-ADPD) in batches of 128 samples with early stopping (with 20 000 steps patience) for roughly 7 h on a single Nvidia RTX A6000.

3.4 Dataset

Training Dataset. We train on the Chest ImaGenome dataset [4,21,22][1], consisting of roughly 240 000 frontal chest X-ray images with corresponding scene graphs automatically constructed from free-text radiology reports. It is derived from the MIMIC-CXR dataset [9,10], which is based on imaging studies from 65 079 patients performed at Beth Israel Deaconess Medical Center in Boston, US. Amongst other information, each scene graph contains bounding boxes for 29

[1] https://physionet.org/content/chest-imagenome/1.0.0 (PhysioNet Credentialed Health Data License 1.5.0).

unique anatomical regions with annotated attributes, where we consider positive `anatomical finding` and `disease` attributes as positive labels for pathologies, leading to binary anatomy-level annotations for 55 unique pathologies. We consider the image-level label for a pathology to be positive if any region is positively labeled with that pathology.

We use the provided jpg-images [11][2] and follow the official MIMIC-CXR training split but only keep samples containing a scene graph with at least five valid region bounding boxes, resulting in a total of 234 307 training samples. During training, we use random resized cropping with size 224 × 224, apply contrast and brightness jittering, random affine augmentations, and Gaussian blurring.

Evaluation Dataset and Class Mapping. We evaluate our method on the subset of 882 chest X-ray images with pathology bounding boxes, annotated by radiologists, from the NIH ChestXray-8 (CXR8) dataset [20][3] from the National Institutes of Health Clinical Center in the US. We use 50% for validation and keep the other 50% as a held-out test set. Note that for evaluation only pathology bounding boxes are required (to compute the metrics), while during training only anatomical region bounding boxes (without considering pathologies) are required. All images are center-cropped and resized to 224 × 224.

The dataset contains bounding boxes for 8 unique pathologies. While partly overlapping with the training classes, a one-to-one correspondence is not possible for all classes. For some evaluation classes, we therefore use a many-to-one mapping where the class probability is computed as the mean over several training classes. We refer to the supp. material for a detailed study on class mappings.

4 Experiments and Results

4.1 Experimental Setup and Baselines

We compare our method against several weakly supervised object detection methods (CheXNet [14], STL [6], GradCAM [18], CXR [20], WELDON [3], MultiMap Model [23], LSE Model [13]), trained on the CXR8 training set using only image-level pathology labels. Note that some of these methods focus on (image-level) classification and do not report quantitative localization results. Nevertheless, we compare their localization approaches quantitatively with our method. We also use AGXNet [24] for comparison, a weakly supervised method trained using anatomy-level pathology labels but without any bounding box supervision. It was trained on MIMIC-CXR (sharing the images with our method) with labels from RadGraph [8] and finetuned on the CXR8 training set with image-level labels. Additionally, we also compare with a Faster-RCNN [16] trained on a small subset of roughly 500 samples from the CXR8 training set that have been

[2] https://physionet.org/content/mimic-cxr-jpg/2.0.0/ (PhysioNet Credentialed Health Data License 1.5.0).

[3] https://www.kaggle.com/datasets/nih-chest-xrays/data (CC0: Public Domain).

annotated with pathology bounding boxes by two medical experts, including one board-certified radiologist.

Table 1. Results on the NIH ChestX-ray 8 dataset [20]. Our models Loc-ADPD and MIL-ADPD, trained using anatomy (An) bounding boxes, both outperform all weakly supervised methods trained with image-level pathology (Pa) and anatomy-level pathology (An-Pa) labels by a large margin. MIL-ADPD is competitive with the supervised baseline trained with pathology (Pa) bounding boxes, while Loc-ADPD outperforms it by a large margin.

Method	Supervision		IoU@10-70	IoU@10		IoU@30		IoU@50	
	Box	Class	mAP	AP	loc-acc	AP	loc-acc	AP	loc-acc
MIL-ADPD (ours)	An	Pa	7.84	14.01	0.68	8.85	0.65	7.03	0.65
w/o WBF			5.42	11.05	0.67	7.97	0.65	3.44	0.64
Loc-ADPD (ours)	An	An-Pa	**10.89**	**19.99**	**0.85**	**12.43**	**0.84**	**8.72**	**0.83**
w/o WBF			8.88	17.02	0.84	9.65	0.83	7.36	0.83
w/ MIL			10.29	19.16	0.84	10.95	0.83	8.00	0.82
CheXNet [14]	–	Pa	5.80	12.87	0.58	8.23	0.55	3.12	0.52
STL [6]	–	Pa	5.61	12.76	0.57	7.94	0.54	2.45	0.50
GradCAM [18]	–	Pa	4.43	12.53	0.58	6.67	0.54	0.13	0.51
CXR [20]	–	Pa	5.61	13.91	0.59	8.01	0.55	1.24	0.51
WELDON [3]	–	Pa	4.76	14.57	0.61	6.18	0.56	0.34	0.51
MultiMap [23]	–	Pa	4.91	12.36	0.61	7.13	0.57	1.35	0.53
LSE Model [13]	–	Pa	3.77	14.49	0.61	2.62	0.56	0.42	0.54
AGXNet [24]	–	An-Pa	5.30	11.39	0.59	6.58	0.56	4.14	0.54
Faster R-CNN	Pa	–	7.36	9.11	0.79	7.62	0.79	7.26	0.78

For all models, we only consider the predicted boxes with the highest box score per pathology, as the CXR8 dataset never contains more than one box per pathology. We report the standard object detection metrics *average precision (AP)* at different IoU-thresholds and the *mean AP (mAP)* over thresholds $(0.1, 0.2, \ldots, 0.7)$, commonly used thresholds on this dataset [20]. Additionally, we report the localization accuracy (loc-acc) [20], a common localization metric on this dataset, where we use a box score threshold of 0.7 for our method.

4.2 Pathology Detection Results

Comparison with Baselines. Table 1 shows the results of our MIL-ADPD and Loc-ADPD models and all baselines on the CXR8 test set. Compared to the best weakly supervised method with image-level supervision (CheXNet) our methods improve by large margins (MIL-ADPD by Δ+35.2%, Loc-ADPD by Δ+87.8% in mAP). Improvements are especially high when considering larger IoU-thresholds and huge improvements are also achieved in loc-acc at all thresholds. Both models also outperform AGXNet (which uses anatomy-level supervision) by large

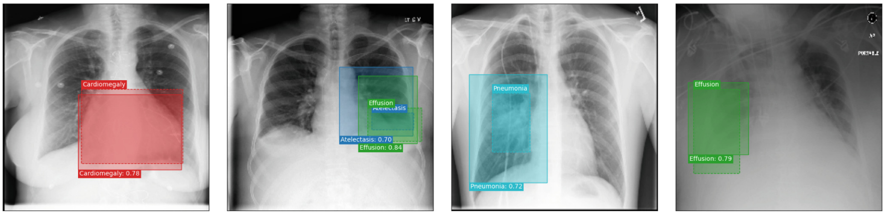

Fig. 4. Qualitative results of Loc-ADPD, with predicted (solid) and target (dashed) boxes. Cardiomegaly (red) is detected almost perfectly, as it is always exactly localized at one anatomical region. Other pathologies like atelectasis (blue), effusion (green), or pneumonia (cyan) are detected but often with non-perfect overlapping boxes. Detection also works well for predicting several overlapping pathologies (second from left). (Color figure online)

margins (MIL-ADPD by $\varDelta + 47.9\%$ and Loc-ADPD by $\varDelta + 105.5\%$ mAP), while improvements on larger thresholds are smaller here. Even when compared to Faster R-CNN trained on a small set of fully supervised samples, MIL-ADPD is competitive ($\varDelta + 6.5\%$), while Loc-ADPD improves by $\varDelta + 48.0\%$. However, on larger thresholds (IoU@50) the supervised baseline slightly outperforms MIL-ADPD, while Loc-ADPD is still superior. This shows that using anatomical regions as proxies is an effective approach to tackle pathology detection. While using image-level annotations (MIL-ADPD) already gives promising results, the full potential is only achieved using anatomy-level supervision (Loc-ADPD). Unlike Loc-ADPD and MIL-ADPD, all baselines were either trained or fine-tuned on the CXR8 dataset, showing that our method generalizes well to unseen datasets and that our class mapping is effective.

For detailed results per pathology we refer to the supp. material. We found that the improvements of MIL-ADPD are mainly due to improved performance on Cardiomegaly and Mass detection, while Loc-ADPD consistently outperforms all baselines on all classes except Nodule, often by a large margin.

Ablation Study. In Table 1 we also show the results of different ablation studies. Without WBF, results degrade for both of our models, highlighting the importance of merging region boxes. Combining the training strategies of Loc-ADPD and MIL-ADPD does not lead to an improved performance. Different class mappings between training and evaluation set are studied in the supp. material.

Qualitative Results. As shown in Fig. 4 Loc-ADPD detects cardiomegaly almost perfectly, as it is always exactly localized at one anatomical region. Other pathologies are detected but often with too large or too small boxes as they only cover parts of anatomical regions or stretch over several of them, which cannot be completely corrected using WBF. Detection also works well for predicting several overlapping pathologies. For qualitative comparisons between Loc-ADPD and MIL-ADPD, we refer to the supp. material.

5 Discussion and Conclusion

Limitations. While our proposed ADPD method outperforms all competing models, it is still subject to limitations. First, due to the dependence on region proxies, for pathologies covering only a small part of a region, our models predict the whole region, as highlighted by their incapability to detect nodules. We however note that in clinical practice, chest X-rays are not used for the final diagnosis of such pathologies and even rough localization can be beneficial. Additionally, while not requiring pathology bounding boxes, our models still require supervision in the form of anatomical region bounding boxes, and Loc-ADPD requires anatomy-level labels. However, anatomical bounding boxes are easier to annotate and predict than pathology bounding boxes, and the used anatomy-level labels were extracted automatically from radiology reports [21]. While our work is currently limited to chest X-rays, we see huge potential for modalities where abnormalities can be assigned to meaningful regions.

Conclusion. We proposed a novel approach tackling pathology detection on chest X-rays using anatomical region bounding boxes. We studied two training approaches, using anatomy-level pathology labels and using image-level labels with MIL. Our experiments demonstrate that using anatomical regions as proxies improves results compared weakly supervised methods and supervised training on little data, thus providing a promising direction for future research.

References

1. Agu, N.N., et al.: AnaXNet: anatomy aware multi-label finding classification in chest X-ray. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 804–813. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_77
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
3. Durand, T., Thome, N., Cord, M.: Weldon: weakly supervised learning of deep convolutional neural networks. In: CVPR, pp. 4743–4752 (2016). https://doi.org/10.1109/CVPR.2016.513
4. Goldberger, A.L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
6. Hwang, S., Kim, H.-E.: Self-transfer learning for weakly supervised lesion localization. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 239–246. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_28
7. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI, pp. 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590

8. Jain, S., et al.: Radgraph: extracting clinical entities and relations from radiology reports. In: NeurIPS (2021)
9. Johnson, A.E.W., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
10. Johnson, A.E.W., et al.: Mimic-cxr database (version 2.0.0). PhysioNet (2019)
11. Johnson, A.E.W., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
13. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR, pp. 1713–1721 (2015). https://doi.org/10.1109/CVPR.2015.7298780
14. Rajpurkar, P., et al.: CheXnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017). https://doi.org/10.48550/arXiv.1711.05225
15. Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., Rosenow, E.C., III.: Interpretation of plain chest roentgenogram. Chest **141**(2), 545–558 (2012)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, vol. 28 (2015)
17. Ridnik, T., et al.: Asymmetric loss for multi-label classification. In: ICCV, pp. 82–91 (2021). https://doi.org/10.1109/ICCV48922.2021.00015
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: CVPR, pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
19. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis. Comput. **107**, 104117 (2021). https://doi.org/10.1016/j.imavis.2021.104117
20. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR, pp. 2097–2106 (2017). https://doi.org/10.1109/CVPR.2017.369
21. Wu, J., et al.: Chest imagenome dataset for clinical reasoning. In: NIPS (2021)
22. Wu, J.T., et al.: Chest imagenome dataset (version 1.0.0). PhysioNet (2021)
23. Yan, C., Yao, J., Li, R., Xu, Z., Huang, J.: Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In: ACM BCB, pp. 103–110 (2018)
24. Yu, K., Ghosh, S., Liu, Z., Deible, C., Batmanghelich, K.: Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In: Wang, L., et al. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 658–668. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_63
25. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319