Fine-Grained Localization, Classification and Segmentation of Lungs with Various Diseases

Julian Berger* Tibor Bleidt* Martin Büßemeyer* Marcus Ding* Moritz Feldmann*
Moritz Feuerpfeil* Janusch Jacoby* Valentin Schröter* Bjarne Sievers*
Moritz Spranger* Simon Stadlinger* Paul Wullenweber* Sarel Cohen† Vanja Doskoč†
Tobias Friedrich†
Hasso Plattner Institute, Germany
*{firstname.lastname}@student.hpi.uni-potsdam.de, †{firstname.lastname}@hpi.de

Abstract

The fine-grained localization and classification of various lung abnormalities is a challenging yet important task for combating diseases and, also, pandemics. In this paper, we present one way to detect and classify abnormalities within chest X-ray scans. In particular, we investigate the use of binary image classification (to distinguish between healthy and infected chests) and the weighted box fusion (which constructs a detection box using the proposed boxes within range). We observe that both methods increase the performance of a base model significantly.

Furthermore, we improve state of the art on lung segmentation, even in the presence of abnormalities. We do so using transfer learning to fine-tune a UNet model on the Montgomery and Shenzhen datasets. In our experiments, we compare standard augmentations (like crop, pad, rotate, warp, zoom, brightness, and contrast variations) to more complex ones (for example, block masking and diffused noise augmentations). This way, we obtain a state-of-the-art model with a dice score of 97.9%. In particular, we show that simple augmentations outperform complex ones in our setting.

1. Introduction

Preface Investigating chest X-rays (CXR) is an important and challenging task. Both the check for the presence of abnormalities and the classification thereof are crucial. Automating this fine-grained localization and classification task with the help of deep learning would result in better patient outcomes, lift up health care quality worldwide, and save many lives by making leading expert-level diagnoses scalable and thus widely accessible. Since most datasets provide no or only partial information on the location and annotations of the abnormalities, many existing approaches focus on classification without detection or implicit localization [32, 4, 13, 15]. The novel VinDr-CXR dataset [22] provides the much-needed locations of the abnormalities, making a fine-grained classification and localization possible.

In order to facilitate the recognition of abnormalities, medical images usually have to be segmented first. The segmented areas can then be used to calculate more complicated metrics. Thus, the classification and localization of lung abnormalities are tightly bound with the segmentation. We focus on the segmentation of lungs in standard posteroanterior chest X-ray (CXR) scans, which can be used to diagnose many diseases and abnormalities.

Related Work Many studies aim to classify and detect various abnormalities such as pneumonia and cancer in CXRs for clinical use in computer-aided detection (CAD) systems. Several methods have been developed in the last years to generate localized predictions for CXRs without available localized training data due to the lack of sufficiently large datasets. Class activation maps (CAM) are used to obtain a localization of abnormalities from a classification task [33]. Similarly, in the Unified DCNN framework, the weights and activations extracted from the network can be used to detect if abnormalities are present and then locate them [18]. Another approach exploits the structured property of CXR images and locates abnormalities via contrastive learning with a learnable alignment module to align input images geometrically [19]. Besides using standard image classification techniques and verifying them against activation heatmaps [2] or predicting regions without having them in the training data, few studies approach the CXR diagnosis using object detection architecture due to the lack of respective datasets. Most recently, the feasibility of a two-stage classification and detection using YOLOv2 with DenseNet on a small proprietary dataset with 3 500 images and 5 class labels is investigated [6]. At the time of writing, the VinDr-CXR dataset is only used as one of five CXR datasets to evaluate Federated Learning [36].

Also, the segmentation of lungs in chest X-rays is still a pending problem in computer science, especially in the presence of abnormalities. Transfer learning approaches for medical tasks have been investigated in the literature, for example, by pre-training a model on a large dataset containing one disease and applying this knowledge to a small dataset with a different disease [16] or by using a model pre-trained on ImageNet [25]. Another approach is to add an attention mechanism to the UNet architecture [26] in order to achieve improved results in lung segmentation [8]. Another way uses the same approach to determine the usefulness of fine-tuning [21] and manages to achieve nearly state-of-the-art results with only a few image samples for training. Also, focusing specifically on biological abnormalities that result in highly obfuscated X-ray images, one can train a model on an X-ray dataset for lung segmentation that does not include these abnormalities and transfer it to one that did [27].

Our Contribution We suggest improvements for the fine-grained localization, classification, and segmentation tasks. We use the *weighted boxes fusion* (WBF) method to improve the non-maximum suppression (NMS) for the fine-grained localization task. On top of that, we use a *2-class-classification* (2CC) for the fine-grained localization and classification task to distinguish between healthy chests and those which are not. This helps remove false positives when the chest X-ray does not contain any abnormality. We observe that the proposed approaches (WBF + 2CC) significantly improve the base model's performance, leading to a mAP-score of 0.242.

We advance the state of the art for lung segmentation, even in the presence of diseases, using a transfer learning approach with a UNet [26]. We compare both complex and straightforward augmentations and evaluate their impact on the model. This way, we reduce the margin of error of the state-of-the-art model by over 40% with a resulting dice score of 97.9%. Most notably, we observe that simple augmentations increase the model's performance more than complex ones in our setting (compare Section 3.1), reinforcing prior findings in the literature [8]. Contrary to the previous literature [8], we observe that the Adam optimizer [17] does converge as expected.

2. Data Set

For the **disease classification and localization**, we use the VinDr-CXR dataset [22]. It is provided as part of the *VinBigDa Chest X-Ray Abnormalities Detection* Kaggle challenge [23] by the Vingroup Big Data Institute and includes $18\,000$ labeled images. There are 14 different lung condition classes and a "No Finding" label, indicating the absence of an abnormality. The images are split up into $15\,000$ images belonging to the training set and $3\,000$ unlabeled images (of which $2\,700$ (90%) are available as the test set). To monitor our training, we use 750 images (5% of the training set) as a validation set. Out of the $15\,000$ images in the training set, $10\,606$ do not contain any abnormalities. We remark that the class distribution for the local labels is heavily imbalanced.

For the **lung segmentation** task, we use the Pulmonary Chest X-Ray Abnormalities dataset [14, 5], a copy of which can be found on Kaggle [20] for example. This dataset consists of X-ray images of patients with and without tuberculosis. It was created by the Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China, and the National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. Furthermore, the X-ray images from the Montgomery hospital contain lung segmentations. The missing lung segmentations for the X-rays obtained from the Shenzen hospital can be found as separate dataset [14, 5, 31]. It was manually annotated by the students and teachers of the Computer Engineering Department, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine ”Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. It can be downloaded from Kaggle [1] as well. The Shenzhen dataset contains 336 images of patients with tuberculosis and 326 images of patients without tuberculosis, while the Montgomery set contains 80 images of patients with tuberculosis and 58 of patients without tuberculosis. The set contains images of different abnormalities, such as effusions and miliary patterns, but misses other possible abnormalities, such as severe opacification. We split it into training (80%), validation (10%) and test (10%) set.

3. Architectures

Next we discuss the models we use for the fine-grained classification and localization task (Section 3.1) and for the segmentation task (Section 3.3). In Section 3.2, we discuss possible improvements of the base model from Section 3.1.

3.1. Fine-Grained Classification and Localization

We use the Detectron2 framework [34], which provides a variety of easy-to-use, pre-trained models for object detection. The models were pre-trained on the COCO dataset. For object detection, Detectron2 provides three main model architectures in a variety of nuances: Faster R-CNN, RetinaNet, and RPN & Fast-RCNN. The model we choose is the `faster_rcnn_R_101_FPN_3x` from the Detectron2 model zoo [35]. This model has the second-highest baseline score for object detection among the models available in the library while using significantly less memory than better models [34]. We train it with a batch size of 4, a cosine learning rate scheduler with a base learning rate of 0.001, and a Non-Maximum Suppression (NMS) threshold of 0.8. To evaluate our models during training, we use Detectron2's `COCOEvaluator` to compute the mean average precision (mAP) scores on the validation set as this is an established metric for evaluating object detection performance [24]. The evaluation of the test set is done on the Kaggle server using the *standard PASCAL VOC 2010 mAP* [7] with an IoU threshold of 0.4. Since Detectron2 does not have this specific metric built-in, we use its existing, more general mAP implementation. To distinguish between the two metrics, we refer to the first as the *test mAP* (implementation in Kaggle) and the *validation mAP* (implementation in Detectron2 side).

To improve our model's generalization capabilities, we use several types of image augmentation. We start by applying augmentations built into Detectron2, such as random saturation (0.7-1.3), contrast (0.7-1.3), brightness (0.7-1.3), rotation ($-10°$-$10°$), horizontal flips and cropping (0.8 relative range). Each augmentation has a probability of 0.2 except the horizontal flips, which have a probability of 0.5. The results of $10\,000$ iterations with evaluation every $1\,000$ steps on a `faster_rcnn_R_101_FPN_3x` model shows no improvement, the AP score even decreases. Additionally we use Gaussian blur with a kernel size of 5. We experiment with diffused noise augmentation [27], wherea random disk sets of varying radii, smoothed with a Gaussian kernel, is appended and is used to scan for Pulmonary Opacification in X-ray images [27], which can also be caused by COVID-19.

The results of our augmentation experiments are displayed in Figure 1. Overall using only the built-in augmentations worsened the mAP on the validation set. Using diffused noise instead improved the validation mAP slightly. Adding all built-in augmentations except the rotation increased the score further. One phenomenon attracted our attention, as it seemed that certain augmentations improved the score for certain types of diseases. For example, the diffused noise augmentation improved the detection of the lung infiltration condition, while training without the rotation augmentation improved the detection of Pneumothorax. In future experiments, it might be advantageous to explore selective augmentations further.

3.2. Improvements on Top of the Base Model

To further improve our fine-tuned model from Section 3.1, we apply **weighted box fusion** (WBF) [29]. It constructs average boxes by utilizing the confidence scores of all proposed boxes. The method shows promising results on different datasets. Apart from functioning in place of non-maximum suppression (NMS), it can also fuse the boxes of multiple trained models. This effectively results in creating a model ensemble. We use an implementation available on GitHub [28]. Our models (that is, the base model with various parameter and augmentation settings) make 100 predictions per image and $300\,000$ predictions for all images. Using only the boxes of a single model, employing WBF lowers the test mAP compared to our baseline mAP of 0.165. For different IoU thresholds, the test mAP varies between 0.087 and 0.165. The IoU threshold and the resulting accuracy seem to be correlated: the lower the IoU threshold, the worse the accuracy. When using the boxes of our second-best model (0.163 test mAP), the results were only slightly worse. The mAP varied between 0.127 and 0.163 depending on different IoUs and model weights. Using the best six models (all > 0.150 mAP), weighting the best model by factor five, the second-best by factor two, and using an IoU threshold of 0.4, we were able to improve the initial mAP of 0.165 to 0.186. The initial number of $1\,800\,000$ boxes decreased to $613\,000$ boxes due to applying WBF.

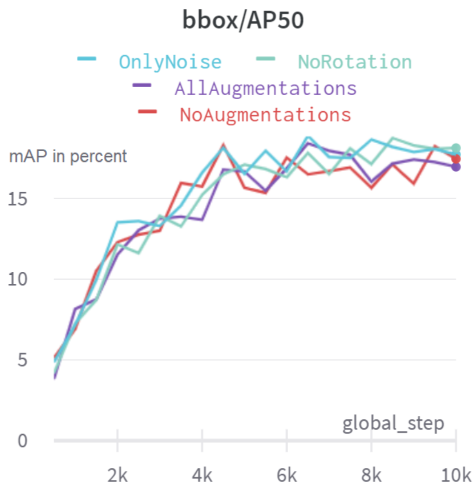

Figure 1: The validation mAP on the validation set with built-in augmentations.

Our base model from Section 3.1 is only trained with abnormal chests to ensure good disambiguation of the different diseases because healthy chests dominate the training set. Hence the model cannot distinguish between healthy and abnormal chests. This leads to many false positives if it is shown healthy chests. Since over 70 percent of the training set are healthy chests, it is crucial to reduce the number of false positives returned by the model. Therefore, we apply a 2CC model on top of our base model's output. Here, healthy chests are identified by a "healthy chest" 1-pixel bounding box and contain no other bounding boxes. Now, the 2CC replaces our base model's prediction with the "healthy chest" 1-pixel box if its confidence for a healthy chest exceeds a certain threshold α. Additionally, we use a second, lower threshold β at which the "healthy chest" coding is added to but does not replace the other predicted boxes to account for cases where the 2CC is not as confident that it has found a completely healthy chest. We train a `resnet18` backbone model with 5-StratifiedKFold cross-validation (CV) for 15 epochs on a downscaled 256x256 pixel version of the dataset. Its validation accuracy is around 0.93. Deeper models and higher resolution images are tested but resulted in overfitting.

We find that an α of 0.997 and β of 0.3 yielded the best results. 919 images are above the upper threshold, and $2\,314$ images are above the lower threshold. Using 2CC, we can improve the test mAP significantly to 0.242. An ensemble of the three best 2CC models (all > 0.9 validation accuracy) results in a test mAP of 0.237, and we see room for improvement with better models.

3.3. Segmentation

We use the ResNet-34 architecture [10] organized in the way of a UNet [26]. We choose fastai [12] as our deep learning library. It already provides pretrained weights (on ImageNet) for our ResNet-34 UNet architecture, and it enables us to fine-tune the parameters. Given the dataset's manageable size, using pretrained models instead of training from scratch is essential to obtain good performance.

In our training runs, we use the following hyperparameters. The batch size is set to 4. Moreover, we apply a weight decay of 0.01. Since the raw images are pretty large, we downsample every image to be one-tenth of its original size to fit our model with the given images onto one GPU. As for the learning rate, we use a learning rate finder to determine a reasonable learning rate. Next, we train the model in two-phase fine-tuning. First, we fit the last layers of the pretrained model with higher learning rates for about 5 epochs. Then, we unfreeze all layers and train for another 5 epochs with a lower learning rate. We evaluate our model's performance using various standard metrics for evaluating the segmentation task. This way, we can observe the model's strengths and weaknesses. We use the *dice metric* [30], which captures the ratio of how many pixels were correctly classified with respect to the size of both the ground truth and the predicted segmentation masks. Furthermore, we use the *accuracy metric*, which intuitively and classically describes the ratio of how many pixels were classified correctly overall. Lastly, as both the dice and accuracy score do not reveal any information about the error types and their relation to one another, we also consider the standard metrics *precision* and *recall* which are usually used in classification tasks. The model is trained until convergence with a learning rate of 0.001 and then fine-tuned for five epochs with a lower learning rate. We are using the Adam optimizer [17] which converged as expected. This differs from the observation that Adam does not converge as well as plain SDG [8].

4. Results

We compare three different approaches for the fine-grained localization task, as displayed in Table 1. We observe that WBF and 2CC together significantly improve our model's performance, leading to a test mAP of 0.242.

For the classification and detection tasks, we use various augmentations during the training process. We compare the influence of standard augmentations (transformations, brightness, contrast variations, etc.) to more complex ones, such as *block masking* [27], where half of the image, horizontally or vertically, gets replaced by gray pixels and *diffused noise* [27], where the brightness of circular areas is increased. Interestingly, the more complex augmentations ultimately result in slightly inferior performance in our setting, see Table 2, endorsing similar observations [8].

Approach	Test mAP
Basic Training	0.165
Basic Training + WBF	0.186
Basic Training + WBF + 2CC	0.242

Table 1: The resulting scores of the approaches discussed in Sections 3.1 and 3.2.

Model	dice	precision	recall	accuracy
Our Model (OM)	**97.9%**	97.0%	97.0%	98.4%
OM+ExAugm	96.2%	96.7%	95.6%	98.0%
Adv. ATTN [8]	96.2%	-	-	-
ATTN [8]	95.8%	-	-	-

Table 2: We compare our model (OM) and our model with complex augmentations (OM+ExAugm) from Section 3.3 to the state of the art [8].

5. Future Work

In order to obtain even better results, we propose to merge the tasks of localization and classification of lung abnormalities with the segmentation of lungs containing abnormalities in the following ways. First, one can use the localization of diseases to improve the segmentation by, for example, removing the diseased tissue before training or by providing the location of the diseases as additional input to the neural network. Another way is to use the lungs' segmentation to improve the localization of the diseases by, for example, providing the lung mask as additional information.

Due to time and space constraints, we leave a comparison to a randomly initialized neural network which would demonstrate the benefits of the ImageNet pretraining [9], a comparison to other possible benchmarks, the consideration of various other metrics, detailed experiments on the impact of the augmentations, and a hyperparameter optimization to future work. Furthermore, studying how 2CC helps eliminating false positives [3, 11] and the effects of applying 2CC before the detection, is left to future work.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] Kaggle competition: Lung masks for Shenzhen Hospital chest X-ray set. https://www.kaggle.com/yoctoman/shcxr-lung-mask, 2018. 2

[2] Ivo Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*, 2019. 1

[3] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision (ECCV)*, volume 12348, pages 558–573, 2020. 4

[4] Aurelia Bustos, Antonio Pertusa, José María Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest X-ray image dataset with multi-label annotated reports. *Medical Image Anal.*, 66:101797, 2020. 1

[5] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Medical Imaging*, 33(2):577–590, 2014. 2

[6] Yongwon Cho, Sang Min Lee, Young-Hoon Cho, June-Goo Lee, Beomhee Park, Gaeun Lee, Namkug Kim, and Joon Beom Seo. Deep chest x-ray : Detection and classification of lesions based on deep convolutional neural networks. *International Journal of Imaging Systems and Technology*, 31, 2020. 1

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 3

[8] Gusztáv Gaál, Balázs Maga, and András Lukács. Attention U-Net based adversarial architectures for chest X-ray lung segmentation. In *Workshop on Applied Deep Generative Networks co-located with 24th European Conference on Artificial Intelligence (ADGN@ECAI)*, volume 2692, 2020. 2, 4

[9] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *IEEE/CVF International Conference on Computer Vision, (ICCV)*, pages 4917–4926, 2019. 4

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[11] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European Conference on Computer Vision (ECCV)*, volume 7574, pages 340–353, 2012. 4

[12] Jeremy Howard and Sylvain Gugger. Fastai: A layered API for deep learning. *Information*, 11(2):108, 2020. 4

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Conference on Artificial Intelligence (AAAI)*, pages 590–597, 2019. 1

[14] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y. Wang, P. Lu, and C. J. McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014. 2

[15] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, 2019. 1

[16] Barleen Kaur, Paul Lemaître, Raghav Mehta, Nazanin Mohammadi Sepahvand, Doina Precup, Douglas L. Arnold, and Tal Arbel. Improving pathological structure segmentation via transfer learning across diseases. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data (DART)*, volume 11795, pages 90–98, 2019. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2015. 2, 4

[18] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Fei-Fei Li. Thoracic disease identification and localization with limited supervision. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, pages 139–161. 2019. 1

[19] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest X-Ray diagnosis via contrast induced attention network with limited supervision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10631–10640, 2019. 1

[20] K. Scott Mader. Kaggle competition: Pulmonary chest X-ray abnormalities. https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities, 2018. 2

[21] B. Maga. Chest X-ray lung and heart segmentation based on minimal training sets. *CoRR*, 2021. 2

[22] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. *CoRR*, 2021. 1, 2

[23] Ha Q. Nguyen, Hieu H. Pham, Nhan T. Nguyen, Dung B. Nguyen, Minh Dao, Van Vu, Khanh Lam, and Linh T. Le. Kaggle competition: VinBigData chest X-ray abnormalities detection. https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection, 2021. 2

[24] R. Padilla, S. L. Netto, and E. A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020. 3

[25] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for

medical imaging. In *Neural Information Processing Systems (NeurIPS)*, pages 3342–3352, 2019. 2

[26] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015. 2, 4

[27] Raghavendra Selvan, E. Dam, Sofus Rischel, Kaining Sheng, Mads Nielsen, and A. Pai. Lung segmentation from Chest X-rays using variational data imputation. *CoRR*, 2020. 2, 3, 4

[28] Roman Solovyev. Github: Weighted boxes fusion. https://github.com/ZFTurbo/Weighted-Boxes-Fusion, 2019. 3

[29] Roman Solovyev and Weimin Wang. Weighted boxes fusion: ensembling boxes for object detection models. *CoRR*, 2019. 3

[30] T. Sørensen, T. Biering-Sørensen, and J. T. Sørensen. A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on danish commons. 1948. 4

[31] Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Peng Gang, Wei Zeng, and Yuri Gordienko. Chest X-Ray analysis of tuberculosis by deep learning with segmentation and augmentation. *CoRR*, 2018. 2

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition, (CVPR)*, pages 3462–3471, 2017. 1

[33] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. 1

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[35] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2 model zoo. https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md, 2019. 2

[36] Zhuoning Yuan, Zhishuai Guo, Yi Xu, Yiming Ying, and Tianbao Yang. Federated deep AUC maximization for heterogeneous data with a constant communication complexity. *CoRR*, 2021. 2