TOWARDS THE AUTOMATIC GENERATION OF MEDICAL REPORTS IN LOW SUPERVISION SCENARIOS

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)
genehmigte

Dissertation von
CONSTANTIN SEIBOLD

geboren am 18.10.1995 in Stuttgart, Deutschland

Tag der mündlichen Prüfung: 20. April 2023

Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen
Korreferent: Prof. Dr. med. Dr. rer. nat. Jens Kleesiek

Constantin Seibold: *Towards the Automatic Generation of Medical Reports in Low Supervision Scenarios*

Dedicated to my parents
Margit and Gerhard Seibold

Nanos gigantum humeris insidentes

— Bernard of Chartres

ABSTRACT

Automated medical report generation with the basis of deep learning has the potential to be a technology that can severely improve working conditions in radiology departments all over the world. However, a significant hindrance to the development of such systems is the requirement of enough training data for the explicit modeling of visible pathological and anatomical structures that allows for meaningful processing in subsequent clinical steps. As gathering such densely structured data is difficult and time-consuming primarily due to the need for medical experts during annotation, we develop methods that can leverage low supervision scenarios to enable the accurate localization of human anatomy and pathology in chest radiographs, the most common imaging procedure worldwide. Specifically, we investigate three low-data scenarios in order to provide a basis of tools for the use of medical reporting methods.

First, we look at the plain available data, images, and associated medical reports. As medical reports are typically unstructured data, these tend to be parsed to essential few-dimensional binary vectors indicating the presence and absence of a pathology. These are utilized to train models for the recognition of diseases. However, as proper reporting requires not just recognizing an anomaly but its position, we develop methods that can learn from medical reports in a parsed or free-text manner to not only accurately recognize but also localize pathologies. Furthermore, by breaking free from the parsed report, we can avoid label noise and enable practitioners to query at will for findings of interest, leading towards simplified integration of deep learning models into personalized structured reporting.

While identifying visually anomalous patterns is essential for medical reports, only the combination with anatomical structures lets a doctor put everything into perspective to develop strategies to treat a patient. Because all anatomical structures tend to always occur in a patient, they are hardly learnable through strategies commonly employed in weakly supervised learning. As such, we have to rely on some form of manual annotation. To lessen the burden of annotation, we, in the next step, investigate scenarios in which we have a few manually annotated images while leaving significantly more data unlabeled, namely semi-supervised learning. We notice that in scenarios with overlapping label structures, commonly used semi-supervised approaches tend to fail, especially when moving towards having nearly no annotated data. As such, we developed a reference-guided pseudo-labeling scheme that allows us to utilize the co-occurrence of existing classes in unlabeled data. As a result, we manage to surpass existing semi-supervised methods for anatomy segmentation of chest radiographs.

While our approach shows immense potential, it still requires manual annotations at a scale that is not humanly possible to perform usable anatomical segmentation. To overcome this hurdle, we look at other medical domains. Due to their shared radiation type, both computer tomographies and X-ray images look similar when projected onto a two-dimensional plane. Following this train of thought, we aggregate any available anatomical annotation for computer tomography images and, by projecting the volumes with their annotations back to 2D, develop the PAX-Ray(++) dataset. This dataset enables the first-ever fine-grained segmentation of human anatomy in chest radiographs, which we validate via human annotation and downstream tasks.

This thesis has built a vital basis for automated medical report generation as we provided methods for the open recognition of pathological and anatomical structures and has advanced medical image analysis by identifying novel tasks and developing algorithms and datasets. Our experiment results show great promise for real-life applications of such algorithms.

ZUSAMMENFASSUNG

Automatisierte medizinische Berichterstellung auf der Grundlage von Deep Learning hat das Potenzial, die Arbeitsbedingungen in der Radiologie weltweit erheblich zu verbessern. Ein wesentliches Hindernis für die Entwicklung solcher Systeme ist jedoch die Notwendigkeit einer ausreichenden Anzahl von Trainingsdaten für die explizite Modellierung sichtbarer pathologischer und anatomischer Strukturen, die eine sinnvolle Verarbeitung in nachfolgenden klinischen Schritten ermöglicht. Da das Sammeln solcher feinen Annotationen schwierig und zeitaufwändig ist, unter Anderem durch den Mangel an medizinische Experten, entwickeln wir Methoden, die in Szenarien mit geringer Überwachung agieren können, um die genaue Lokalisierung der menschlichen Anatomie und Pathologie in Thoraxröntgenbildern, dem weltweit am häufigsten verwendeten bildgebenden Verfahren, zu ermöglichen. Konkret untersuchen wir drei Szenarien mit wenigen Daten, um eine Basis von Werkzeugen für den Einsatz von medizinischen Berichterstattungsmethoden zu schaffen.

Zunächst betrachten wir die verfügbaren Daten, Bilder und zugehörigen medizinischen Berichte. Da es sich bei medizinischen Berichten um unstrukturierte Daten handelt, werden diese typischerweise in wenig dimensionale binäre Vektoren zerlegt, die das Vorhandensein von bestimmten Pathologien beschreiben. Diese werden genutzt um Modelle zur Krankheitserkennung zu trainieren. Da jedoch eine ordnungsgemäße Berichterstattung nicht nur das Erkennen einer Anomalie, sondern deren Position erfordert, entwickeln wir Methoden, die aus medizinischen Berichten in einer geparsten oder Freitextform lernen können, um Pathologien nicht nur genau zu erkennen, sondern auch zu lokalisieren. Indem wir uns von den geparsten Berichten lösen, können wir außerdem Rauschen bei der Labelextraktion vermeiden und es Ärzten ermöglichen, beliebig nach Befunden von Interesse suchen zu lassen, was zu einer erleichterten Integration von Deep Learning Modellen in strukturierte Befundungsmethoden führen kann.

Die Identifizierung von visuell anomalen Mustern ist für medizinische Berichte zwar unerlässlich, aber erst die Kombination mit anatomischen Strukturen ermöglicht es dem Arzt, alles in die richtige Perspektive zu rücken und Strategien für die Behandlung eines Patienten zu entwickeln. Da alle anatomischen Strukturen in der Regel immer bei einem Patienten vorkommen, lassen sie sich kaum durch Strategien erlernen, die beim schwach überwachten Lernen üblich sind. Daher müssen wir auf eine Form der manuellen Annotation zurückgreifen. Um die Last der Annotation zu verringern, untersuchen wir im nächsten Schritt Szenarien, in denen wir einige wenige manuell gelabelte Bilder haben, aber wesentlich mehr Daten unannotiert lassen, nämlich das semi-überwachte Lernen. Wir stellen fest, dass in Szenarien mit sich überschneidenden Beschriftungsstrukturen gängige semi-überwachte Ansätze scheitern, insbesondere dann, wenn fast keine annotierten Daten vorhanden sind. Aus diesem Grund haben wir ein referenzgeleitetes Pseudo-Labeling-Schema entwickelt, das es uns ermöglicht, das gemeinsame Auftreten bestehender Klassen in nicht beschrifteten Daten zu nutzen. Dadurch ereichen wir es die Schwachstellen anderer semi-überwachte Methoden für die Segmentierung von anatomischen Strukturen in Thoraxröntgenbildern zu umgehen.

Obwohl unser Ansatz ein immenses Potenzial aufweist, benötigen wir immer noch manuelle Annotationen in einem Umfang, der für den Menschen nicht sammelbar ist. Um diese Hürde zu überwinden, schauen wir uns andere medizinische Bereiche an. Aufgrund ihrer gemeinsamen Strahlungsart sehen sowohl Computertomografien als auch Röntgenbilder ähnlich aus, wenn sie auf eine zweidimensionale Ebene projiziert werden. Diesem Gedankengang folgend, fassen

wir alle verfügbaren anatomischen Annotationen für Computertomografiebilder zusammen und entwickeln den PAX-Ray(++)-Datensatz, indem wir die Volumina mit ihren Beschriftungen auf eine 2D-Ebene zurückprojizieren. Dieser Datensatz ermöglicht die allererste feingranulare Segmentierung der menschlichen Anatomie in Röntgenbildern des Brustkorbs, die wir durch menschliche Annotation und Downstreamtasks validieren.

Diese Arbeit hat eine wichtige Grundlage für die automatisierte Erstellung medizinischer Berichte geschaffen, da wir Methoden für die offene Erkennung pathologischer und anatomischer Strukturen bereitgestellt und die medizinische Bildanalyse durch die Identifizierung neuer Aufgaben und die Entwicklung von Algorithmen und Datensätzen vorangetrieben haben. Die Ergebnisse unserer Experimente sind sehr vielversprechend für die Anwendung solcher Algorithmen in der Praxis.

ACKNOWLEDGEMENTS

This journey over the last three and a half years was filled with quite a few ups and downs, and I am thankful for the many people I have managed to cross paths with. First of all, a thank you to Prof. Dr.-Ing. Rainer Stiefelhagen, with whom my research was only possible. The combination of challenge and trust he brought towards me allowed me to massively grow, not only as a researcher but also as a person during this project. I wouldn't be where I am today without his perspective on research. In a similar vein, I would like to thank Prof. Dr. med. Dr. rer. nat. Jens Kleesiek for the continuous assistance from a more medical perspective during this thesis. He was always willing to provide medical resources to assist in generating studies, datasets, or the like, which massively helped in the planning and execution of various subprojects.

During my stay at the CV:HCI, I have made many great memories that I hope to never forget, be that the joined long productive nights just before a big deadline, the exciting Werewolf game nights during our retreats in Bad Herrenalb, or just the simple brainstorming in front of a whiteboard or while drinking some coffee (or some weird Cola+juice mixture...). I will miss the time spent inspecting Schmidhuber-memes, discussing (sometimes more, sometimes less) genius ideas or playing table tennis with one of my best friends and colleagues, Simon (who might be one of the most adventurous non-alcoholic drinkers I've ever witnessed) the most. His permanent optimism turned too many adverse situations upside down and this research experience into an (albeit sometimes frustrating) entertaining ride. I am incredibly thankful to Saquib, who has been a continuous inspiration since my master's thesis due to his pragmatic yet extraordinary approaches toward various research topics. Furthermore, his attitude towards ideas made it a joy to "disrupt" his chess games with great discussions. Vivek was sometimes quite a controversial person within our lab. However, I always enjoyed our conversations (and late-night coding sessions!). His somewhat easygoing view on life was great to learn from, and I loved our random trips to the lakesides. Alina always appeared as the person to look up to in our lab. She was friendly and helpful regarding any Ph.D.-related thing but also hit a perfect balance between delivering excellent research and maintaining a healthy work-life balance. As such, she provided ongoing motivation regarding what can be achieved. Despite just joining recently, Alexander has become quite a dear friend with whom it was great fun to discuss the wonders of the nnUNet or go bouldering. Obviously, I also owe a massive "Thank You!" to Corinna, who always was kind enough to assist with any organizational question, no matter how stupid it might have been (bringing Luna to the office also always brought a big smile to my face!). Overall, I loved the atmosphere at our lab, despite it sometimes being reeaaally empty during COVID times and for this, I am also thankful to our other members over the years, namely, in no particular order Manel (also thanks for helping me with setting up my workstation), Monika (I still got your bouncy ball!), Tobias (pls no *sudo rm -rf /home/**), David (you sometimes appear like a monkey while bouldering, absolute powerhouse!), Marios (a mathematicians mind truly works wonders!), Jiaming, Zdravko (supervising practical courses with you and Jiaming was always a pleasure), Kunyu, Omar, Thorsten, Karin and Kailun.

Also, a big thanks to all the doctors who helped me grasp the practical applications of our work (and for dragging me back down to the ground). A special big thanks to Matthias, who is obviously the "nicest and most handsome" attending to ever exist. It was always great fun to banter with you about various things and plan and design projects and papers (although most of our work will not be featured in this thesis). Similarly, thanks to Moon-Song Kim, whose kind and

go-getting attitude made it a joy to work with! Also, thanks to Victoria Mayer and Jacob Murray for assisting in user studies!

It would not have come to this research without the funding of the HIDSS4Health graduate school. A big thanks to Dr. Ines Reinarzt, who managed so many needs of us students in a seemingly near self-sacrificing manner for most of my time there. Thanks to Dr. Kathrin Brunk, Jasmin Thiel, and Dr. Nicole Merkle for organizational matters. Thank you to the additional members of my thesis advisory committee: Prof. Dr. Klaus Maier-Hein and Prof. Dr. Heinz-Peter Schlemmer, who both continued to give meaningful feedback throughout this thesis. Obviously, also "Thank You"s go out to several other students who made this "school"-life much more enjoyable. I wish I could name everyone but so many great people have joined the graduate school in the meantime that this becomes quite difficult. Also, thanks to the SCC and the HOREKA cluster, which enabled a parts of the experiments mentioned in this thesis.

I would also like to thank my family and friends for our great moments together. I am deeply grateful to my parents, Gerhard and Margit, who have always provided me with enough freedom and assistance to pursue what I want to my heart's content. Similarly, thanks to my aunt and uncle, Doris and Robert, who always welcomed me with hospitality during my time in Karlsruhe. Thanks to my original friend group of Vincent, Alex, Maxim, Maxi, and David. Even though our lives have become relatively spread out, it is always great to know to be able to fall back on you and have someone to talk to (obligatory, "we really should organize something more explicit this summer"...). A massive thanks to Caro and Joana (plus their inventory of Alex and David), who made life in Karlsruhe an overall joy. Your presence in our flat provided much-needed relief from everyday research, either through Joana's weekend expeditions or tellings of Caro's DnD adventures. I will miss our doorframe conversations. Tim deserved a big thanks for not only providing close (heh, as if) competition in the Jigglypuff-Pokemon Trainer matchup but also for staying an amazing friend during my time in Karlsruhe, who significantly helped keep my motivation high during the writing phase of the thesis. Thanks to Remi for great conversations about topics ranging about various topics from math, life, and video games (also for teaching me some chess basics)! I want to give an immense thanks to Joan, Julien, Alex, and Patri for introducing me back to social life after the global break that was COVID. I enjoyed all our trips, parties and alternative activities I managed to participate in! Also major thanks to Alex and Julie for keeping me company during various sports activities, whether jogging, calisthenics or swimming! I may be able to learn a few more words of French or Spanish in the future (sorry for my incapability...). Finally, a massive thanks to Nadja for keeping me sane on the finishing straight!

CONTENTS

I BACKGROUND 1
1 INTRODUCTION 3
  1.1 Medical Reports in the Clinical Setting 3
  1.2 On the Design of Report Generation Systems 4
  1.3 Thesis Roadmap and Contributions 6
2 RELATED WORK 9
  2.1 Automating Medical Report Generation 9
  2.2 Understanding of Vision and Language 11
  2.3 Learning from Limited Annotations 13
  2.4 Chest Radiograph Datasets 15
  2.5 Multiple Instance Learning 16

II LEARNING FROM LIMITED SUPERVISION 19
3 LEARNING FROM RADIOLOGIC REPORTS 21
  3.1 Pathology Recognition from Report-Level Labels 22
    3.1.1 Online Instance-Level Pseudo-Labeling via Self-Guidance 22
    3.1.2 Multiple Instance Learning for Chest Radiograph Diagnosis 25
    3.1.3 Experimental Setup 26
    3.1.4 Results 28
  3.2 Breaking with Fixed Set Pathology Recognition 31
    3.2.1 Global-Local Contrastive Learning 32
    3.2.2 Model Inference 34
    3.2.3 Localization with Vision-Language Models 36
    3.2.4 Prompt-based Dataset Extension 36
    3.2.5 Experimental Setup 36
    3.2.6 Results 37
  3.3 Chapter Conclusion 40
4 LEARNING FROM LITTLE AMOUNTS OF ANNOTATIONS 43
  4.1 Understanding CXRs: Multi-Label Segmentation 43
  4.2 Reference-guided Pseudo-Label Generation 45
    4.2.1 Preliminary Definitions 45
    4.2.2 Nearest-Neighbors from References for Pseudo-Label Generation 45
    4.2.3 Consistency Regularization 48
  4.3 Experimental Setup 48
    4.3.1 Datasets 49
    4.3.2 Implementation Details 49
    4.3.3 Ablation Studies 50
    4.3.4 Quantitative Results 51
  4.4 Chapter Conclusion 54
5 LEARNING FROM PARTIALLY ANNOTATED DOMAINS 57
  5.1 Annotations of Chest X-Rays via Computer Tomography Projection 58
    5.1.1 Automated Generation of Projected CXR Datasets 58
    5.1.2 Experimental Setup 62

5.1.3 Results 63
5.2 Extending the PAXRay dataset 64
5.2.1 Adjusted Label Accumulation 66
5.2.2 Dataset Analysis 70
5.2.3 Experimental Setup 74
5.2.4 Results 75
5.3 Grounding Medical Reports via Anatomy Descriptors 84
5.3.1 Results for Medical Phrase Grounding 86
5.4 Chapter Conclusion 87

III INSIGHTS 89

6 IMPACT ON THE FIELD 91
6.1 New Research Directions 91
6.2 New Datasets 92
6.3 New Methods 92
7 FUTURE WORK 93
7.1 Automated Reporting 93
7.1.1 Dense Predictions for Report Generation 93
7.1.2 Automated Extraction of Visual Biomarkers 93
7.2 Modelling of Structures for Computer Tomography Back-Projection 94

IV APPENDIX 95

A AUTHORED PUBLICATIONS 97

BIBLIOGRAPHY 99

LIST OF FIGURES

Figure 1 Overview of the conceptual design of holistic automated report generation. We follow general structure defined in the reporting consensus of the RSNA Radiology Reporting Committee [127]. The automatic generation of the "Observations" section is an accumulation of (a) identifying relevant findings, (b) retrieving corresponding anchors based on established priors, and (c) quantification of findings based on anchors. 5
Figure 2 Overview of the contribution areas and the underlying research questions of this thesis. We develop methods for (1) learning from available medical reports, (2) learning scraps of manually annotated data, and (3) the utilization of partially annotated domains for dense predictions. 6
Figure 4 Excerpt of a structured report for chest radiograph analysis as endorsed by the TLAP [238, 341]. 9
Figure 3 An exemplary workflow for the creation of datasets based on medical reports for the classification of pre-defined findings. Based on the manual annotation of selected samples a text classifier is trained. The text classifier is then applied on either a subset or the remaining amount of the raw dataset. This process can be enhanced through the integration of visual classifiers trained on labels from the text predictions. 10
Figure 5 In our framework, the network trained on report-based information reads chest X-ray images and produces overall image-level pathology prediction scores and their corresponding locations. 21
Figure 6 Illustration of supervision for different loss function concepts for MIL. Strict bag-level supervision (left) provided, Zhou *et al.* 's BIL [338] (center left), Morfi *et al.* 's MMM [183] (center right) and on the right our proposed SGL. 23
Figure 7 Overview of our framework for thoracic disease identification and localization. A chest X-ray is passed through an FCN producing a prediction map. This map is used to compute the instance- and bag-level losses. 26
Figure 8 We display exemplary positive bags of size 10 for MNIST-bags (top) and CIFAR10-bags (bottom). Positive instances denoted in green, negatives in red. 26
Figure 9 Ablation study on different aspects of our SGL on MNIST-bags. 28
Figure 10 Test AUC and IoU for MNIST-Bags for differing avg. instances per bag. . . . 28
Figure 11 AUC and IoU for CIFAR10-Bags for differing number of positive instances per bag with size of 50 and 100. 29
Figure 12 We compare the patch-wise predictions between a mean-pooling trained baseline to our proposed method for different diseases. The value ranges from 0 (blue) to 1 (red). We show prediction boxes (orange) around the connected component of the maximum prediction and Ground-Truth bounding boxes (green). 31

Figure 13 Illustration of our proposed report-guided training scheme. A medical report is split into sentences and processed by a text encoder. The image is passed through an image encoder with augmented views. We apply self-supervised SimSiam-losses (SSV), sentence-level local contrast between sentences and an associated image embedding (L), as well as report-level global contrast between report embeddings and image (G). Furthermore, we apply the losses building on local and global contrast on the augmented views used in SSV. . . . 33
Figure 14 Illustration of inference schemes for open set multi-label classification. (a) *Cosine* computes just the similarity between image features and the class prompt. (b) *Basic* uses prompts for both presence and absence of the class and normalizes the similarity to either. (c) *Detailed* uses a set of positive and negative prompts for the class. . . . 34
Figure 15 Illustration of the QKV-attention pooling (left) and the restructuring for localization (right) [335]. . . . 36
Figure 16 Performance changes based on differences in prompt generation. Class wise performance on the left. Mean performance to the right. Models trained on MIMIC. . . . 37
Figure 17 Contributions of data scaling for chest radiograph dataset. Performance change of adding additional chest X-ray datasets with prompt-based captions. . . . 38
Figure 18 We compare the patch-wise predictions of the considered inference schemes for different diseases. The value ranges from 0 (blue) to 1 (red). We show prediction boxes (orange) around the connected component of the prediction over a 0.8 threshold and Ground-Truth bounding boxes (green). . . . 39
Figure 19 Illustration of the QKV-attention pooling (left) and the restructuring for localization (right) [335]. . . . 40
Figure 20 Comparison of pseudo-labeling by examples. On the left, we display confidence-based pseudo-labeling for multi-class settings. In the center, we show confidence-based pseudo-labeling for multi-label settings. On the right, we show our proposed nearest-neighbor-based pseudo-labeling for multi-label segmentation. We note that an imbalance in confidence can lead to several ignored classes. . . . 44
Figure 21 Overview of the proposed training step for unlabeled images. For each unlabeled image, we extract its features in addition to a pool of sampled annotated images to generate pseudo-labels. In parallel, we use the predictions of a weakly augmented sample to act as supervision for a strongly augmented version of the image. On the right, we illustrate our reference-based pseudo-label generation for k = 5. . . . 45
Figure 22 Example of equidistant pixel-sampling via Nearest-Neighbor interpolation for the generation of our reference pool. . . . 46
Figure 23 Display of super-classes of our annotation scheme of anatomy in chest radiographs. . . . 49
Figure 24 Qualitative Segmentation Results on the JSRT [257] dataset for $N_l = 6$. . . . 51
Figure 25 Qualitative Segmentation Results on our extended anatomical X-Ray annotations. We display segmentations for RPG and RPG^+ in the center, while we show one of the reference images used for training on the right. . . . 53

Figure 26 Qualitative Results for $MLDS$,RPG, and RPG^{+} on RETOUCH for different amounts of training examples. 53
Figure 27 Dataset creation protocol [243]. We apply 3D segmentation methods to generate comprehensive annotations of a CT dataset. Afterwards, the CTs and their labels are projected to 2D, emulating X-rays. 57
Figure 28 Comparison of different real and projected samples for the frontal and lateral view 59
Figure 29 Compressed diagram of our initial proposed label structure for PAX-Ray. . . 60
Figure 30 Examples of overlapping annotations in both lateral and frontal view of PAX-ray 61
Figure 31 Axial slice of two lung halves overlapping sagittally. 62
Figure 32 Qualitative results of a UNet on the test set of the PAXRay dataset 63
Figure 33 Qualitative results of a UNet trained on our PAX-ray dataset for a patient in OpenI 64
Figure 34 Segmentation performance in IoU on the test split for all classes of our proposed PAXRay dataset 65
Figure 35 The workflow for scaling up PAX-Ray. We collect labeled datasets describing different categories, such as bones or abdominal regions. These datasets are used to train nnUNet ensembles, which are then applied to AutoPET to generate a shared label space. This shared label space is then used to retrain the model and generate pseudo-labels for several thorax CT datasets. The CT volumes and pseudo-labels are projected to a two-dimensional plane and post-processed using anatomical priors and morphological operations, resulting in PAX-Ray++. Finally, 2D segmentation models such as UNet are trained on the projected data to predict anatomical structures in CXR accurately. 67
Figure 36 Illustration of projections of PAX-Ray 5.1.1.2 (left) and PAX-Ray++5.2.1.3 (center). We display the difference image on the right. 69
Figure 37 Compressed diagram of our proposed label structure for frontal and lateral images in PAX-Ray. 70
Figure 38 Sample of frontal projected x-rays from the RibFrac, COVID-19-AR, COVID-19-1, COVID-19-2, RSNA PE, and LIDC-IDRI dataset. We show labels belonging to the categories *Airways*, *Vascular System*, *Bones*, and *Abdomen*. . . 71
Figure 39 Sample of lateral projected x-rays from the RibFrac, COVID-19-AR, COVID-19-1, COVID-19-2, RSNA PE, and LIDC-IDRI dataset. We show labels belonging to the categories *Airways*, *Vascular System*, *Bones*, and *Abdomen*. . . 72
Figure 40 (a) #instances per category for frontal images. (b) #instances per category for lateral images. 73
Figure 41 (a) Number of instances per category vs. number of categories. Position indicates mean, size indicates standard deviation. (b) Plot indicating distribution of instance sizes. We compare X-Ray and real world datasets. PAX-Ray++ shows a large amount of labeled instances which tend to be smaller than labeled instances in natural image datasets like COCO or PASCAL VOC. 74
Figure 42 FID scores and qualitative examples of MIP and Matsubara *et al.* 's method with and without histogram equalization (HE) for frontal and lateral views. 76

Figure 43 FID scores and qualitative examples of Matsubara *et al.* 's method applied on CT's with different axial spacings for frontal and lateral views. We can see that for CTs with higher spacing the FID noticably increases. 76
Figure 44 (a) Segmentation performance in IoU on the validation split for all classes of our proposed PAXRay++ dataset for the settings using ViT-B and ResNet-50 backbones, and the impact data augmentations. (b) Segmentation performance in IoU on real CXR from the PadChest dataset for all classes compared to medical expert annotations. 78
Figure 45 Qualitative segmentation results for lateral projections for the class categories of lungs, vascular systems, bones and abdomen. 79
Figure 46 Qualitative segmentation results for lateral projections for the class categories of lungs, vascular systems, bones and abdomen. 80
Figure 47 (a) Segmentation performance in IoU on real CXR from the PadChest dataset for frontal classes. (b) Segmentation performance in IoU on real CXR from the PadChest dataset for lateral classes. 81
Figure 48 Qualitative segmentation results for frontal CXR for the class categories of lungs, vascular systems, bones and abdomen. We compare against the corrections of Annotator 1 and 2. 82
Figure 49 Qualitative segmentation results for frontal CXR for the class categories of lungs, vascular systems, bones and abdomen. We compare against the corrections of Annotator 1 and 2. 83
Figure 50 Chord diagramm of the co-occurences of *Anatomies* and *Findings* in medical reports for patients in the OpenI dataset 84
Figure 51 Anatomy Grounding Baseline. Using NER divide phrases into *Anatomy*, *Anatomy-Modifier*, *Observation*, *Observation-Modifier*. For the images, we generate proposals using our anatomy segmentation model. We extract word-wise embeddings for the phrase/anatomy labels, aggregate them and retrieve the most similar region for each phrase 85
Figure 52 We show ground truth (green), retrievals (orange) and expected retrieval (yellow). If anatomy phrases are identified a result is provided. Otherwise, the segmentation, albeit accurate, is not retrieved. 87

LIST OF TABLES

Table 1 Comparison of related datasets for Chest X-Rays with our PAXRay variants. R, C, B, E, L, M, P stand for Report, Class Labels, Bounding Boxes, Ellipsoids, Lines, Masks, Polygons respectively. 15
Table 2 Comparison of classification performance for pathologies on the ChestX-Ray14 dataset. Here, 70% of all images were used for training with no bounding box annotations available. Evaluations were performed on the official test split containing 20% of all images. "*" denotes usage of additional bounding box supervision. At., Card., Cons., Ed., Eff., Emph., Fib., Inf., Nod. , Pl. Th. , Pn., Pt. stand for Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Pleural Thickening, Pneumonia, and Pneumothorax respectively. 29
Table 3 Disease localization accuracy are evaluated with a classification threshold of 0.5. "*" denotes additional bounding box supervision. Abbreviations follow Table 2. 30
Table 4 We evaluate the impact of the scores from the model heads and their combinations. 37
Table 5 Ablation on the impact of the different loss parts described in Section 3.2.1.2. 37
Table 6 Classification performance on MIMIC, CheXpert and Chest-XRay14. * indicates that the model was trained with additional PadChest and ChestX-Ray14 data. 38
Table 7 Comparison of Nearest-Neighbor performance for different image pool sizes. 50
Table 8 Comparison of RPG for different k and image pool sizes. 'OOM' denotes '**O**ut **O**f **M**emory'. 50
Table 9 Performance comparison of our work to related work on the datasets JSRT. * denotes that due to lacking convergence the model was trained twice the iterations. Bold and underlines denote best and second best performance. . 51
Table 10 Performance comparison on JSRT and our extended annotations (Custom). * denotes a training of twice the iterations. Bold and underlined denote best and second best performance respectively. 52
Table 11 Performance comparison on Retouch. Bold and underlined denote best and second best performance. 54
Table 12 Category-wise segmentation performance in IoU on the test split of our proposed PAXRay dataset 66
Table 13 Comparison of considered CT datasets for PAX-Ray++ regarding size, number of labels, label domain, volume spacing, number of slices and segmentation performance of a nnUNet [112] in 5-fold cross validation. Top half displays datasets used for label aggregation, while the bottom part was used as imaging source for PAX-Ray++. 68
Table 14 Hit rates of region proposals for different IoU thresholds (denoted by subscript). 86
Table 15 Medical phrase grounding performance on OpenI showing Top-k region retrieval performance at different IoU thresholds (denoted by the subscript). 86

ACRONYMS

CT	Computed Tomography
CXR	Chest Radiograph
MRI	Magnetic Resonance Imaging
TLAP	Template Advisory Panel
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GPT	Generative Pre-trained Transformers
MIL	Multiple-Instance Learning
SSL	Semi-Supervised Learning
CAM	Class Activation Mapping
SGL	Self-Guiding Loss
FCN	Fully Convolutional Network
RPG	Reference-Guided Pseudo-Label Generation
PAX-Ray	Projected Anatomy for X-Ray dataset
OOM	Out-of-Memory
IoU	Intersection over Union
AuROC	Area under the Receiver-Operator-Characteristic

Part I

BACKGROUND

1

INTRODUCTION

In this thesis, we aim to develop methods enabling quantifiable automated reporting for radiological cases. Our goal is to create a comprehensive process that captures relevant observations, such as pathologies or foreign objects, as well as their location and severity. The development of such models demands fine-grained annotations. However, in the medical field, the amount of trained medical personnel required for annotation purposes is unavailable, resulting in the challenge of gathering sufficient training data. Therefore, we propose methods that bring us towards automated reporting in scenarios with limited supervision. Namely, we investigate (1) weakly supervised scenarios given available medical reports, (2) semi-supervised scenarios where scraps of data are manually annotated, and (3) the utilization of partially annotated domains for dense analysis. We display a high-level summary of the addressed research questions in Fig 2. Our findings have the potential to generate clear, accurate medical reports and improve the overall insight in the field of medical reporting.

Are there any abnormalities or areas of concern that could indicate a cardiac or pulmonary issue?

What is the concrete positioning of any medical implants or devices in the patients body?

Are there potential complications or risks for a surgery based on the imaging findings?

Such questions and many more build the foundation for treating patients. Radiologists interpret images of the human body and provide answers to these questions about a patient's health based on various modalities such as chest X-rays (CXR), computed tomography (CT) scans, and magnetic resonance imaging (MRI). This process involves understanding the anatomy and physiology of the human body and the various imaging modalities used to produce medical images. This holistic understanding of a given image enables them to answer questions from departments about their daily cases.

In this thesis, we will investigate the essential steps that bring us towards the development of automated systems that can mimic a radiologist's level of understanding, accurately interpret medical images and correctly convey visual information.

1.1 MEDICAL REPORTS IN THE CLINICAL SETTING

Radiologists play a crucial role in the diagnostic process by utilizing medical imaging techniques to help diagnose patients and either avoid or even assist in the planning of potentially risky invasive procedures. After selecting the appropriate imaging modality, the radiologist interprets the resulting images and conveys their findings in a human-understand-able format, either verbally or in a written report [3]. These reports typically describe all of the relevant information in the images in a way easily understandable by humans [65, 121, 215]. What is considered relevant information often depends on the patient's indication. For example, a patient experiencing chest pain may have their lung and heart examined [131], while a patient post-surgery may have their medical devices and potential complications evaluated [69, 80, 214].

Every year, millions of medical imaging exams are performed in Germany alone, and interpreting these images takes time. For example, a standard chest radiograph may only take a minute to interpret and write the report, but more difficult exams like CT scans can take up to more than 15 minutes.[68, 131, 142]. With hundreds of millions of exams performed each year in Germany alone [229], this amounts to a minimum of 247 accumulative years spent annually on the analysis of radiologic imaging[1]. As requirements for detailed reports increase continually, many radiology departments struggle to keep up within the contracted hours [43, 207]. These demands pose a severe issue for the healthcare system, particularly in the face of an ongoing shortage of medical personnel [9, 10, 56], leading to avoidable errors due to rushed examination [21, 261] or even burnout due to straining doctors [7, 22, 84].

Artificial intelligence has the potential to help radiologists by reducing their workload in the diagnostic process for both image analysis and report generation [129, 219]. Deep learning, in particular, has shown great success in many vision and language tasks [42, 141, 221, 286] due to the availability of large-scale datasets. Automated systems have used this technology to assist humans in various fields such as science [27, 62, 126], content generation [210, 211, 226], and information retrieval [224, 333]. It has also been used in radiological imaging for the automated recognition of specific structures, such as lung nodules [144] or abdominal organs [185]. However, to be of use in this medical setting, automated reporting systems, instead of focusing on specific tasks, have to concisely capture all relevant behavior of interest, provide an accurate descriptor and give a comprehensible assessment of the presented state [93, 127].

In this thesis, we aim to identify and clear some of the notable obstacles that hinder the application of deep learning methods for generating medical reports.

1.2 ON THE DESIGN OF REPORT GENERATION SYSTEMS

Radiology reports provide crucial information about a patient's medical imaging exams. These reports can take on different forms, but they tend to include metadata about the patient, information about why the exam was performed, the findings from the exam, and the radiologist's conclusions and recommendations. This work focuses on the findings section, which is essential for any conclusions and recommendations. Depending on the task, this follows either a structured report or a free text. Structured reporting is becoming increasingly important in radiology, as it allows for uniform and precise descriptions of patients, regardless of the radiologist's experience [241]. Task-specific experts often design a structured report to provide all the necessary information for downstream analysis [65, 175]. For example, Fink et al. [65] developed a structured report template to describe pulmonary embolisms, including information about their presence, location, and severity, as well as related quantitative information for further diagnosis. This approach allows for more streamlined reporting and more straightforward analysis of patient cohorts. However, developing and implementing a structured reporting system for numerous diseases can require significant time and rapid environmental changes, such as those seen during the COVID-19 pandemic [137, 304, 314], may require noticeable changes structured reporting formats to new findings.

On the other hand, free text narratives are known methods as they are well-established and do not require an initial introduction to the format. Furthermore, the free text allows for more flexibility and personalization in reporting [90]. It can dynamically adjust to a given image and incorporate a wide range of reporting styles. However, due to its nature, it may omit certain in-

---

[1] We assume a lower bound of one minute per exam.

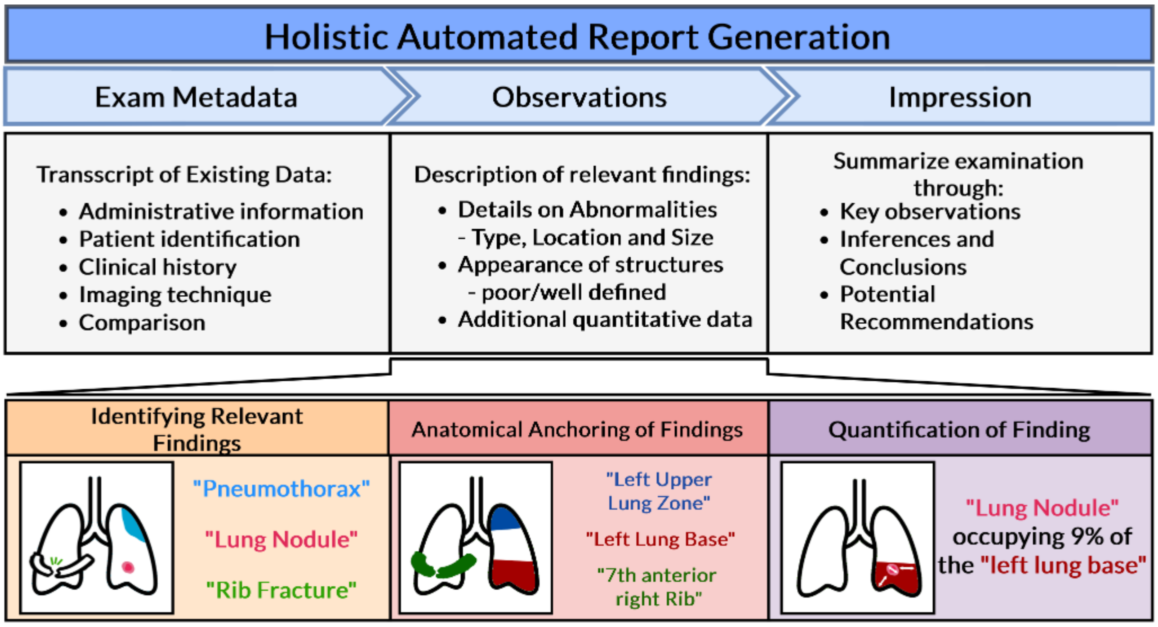

Holistic Automated Report Generation
Exam Metadata
Observations
Impression
Transscript of Existing Data:
• Administrative information
• Patient identification
• Clinical history
• Imaging technique
• Comparison
Description of relevant findings:
• Details on Abnormalities
- Type, Location and Size
• Appearance of structures
- poor/well defined
• Additional quantitative data
Summarize examination through:
• Key observations
• Inferences and Conclusions
• Potential Recommendations
Identifying Relevant Findings
"Pneumothorax"
"Lung Nodule"
"Rib Fracture"
Anatomical Anchoring of Findings
"Left Upper Lung Zone"
"Left Lung Base"
"7th anterior right Rib"
Quantification of Finding
"Lung Nodule" occupying 9% of the "left lung base"

Figure 1: Overview of the conceptual design of holistic automated report generation. We follow general structure defined in the reporting consensus of the RSNA Radiology Reporting Committee [127]. The automatic generation of the "Observations" section is an accumulation of (a) identifying relevant findings, (b) retrieving corresponding anchors based on established priors, and (c) quantification of findings based on anchors.

formation or become less parsable compared to a structured report, potentially hindering retrospective studies [93] and requiring more time during generation [90].

Given recent widespread developments in radiology and pandemic events, radiologic reports need to adapt based on expert knowledge, follow a structured format based on precise information, and allow the extraction of quantitative diagnostic data to categorize patients properly. Keeping these developments in mind, we follow the concept of requirements for what we define as holistic, automated report generation, which we illustrate in Fig. 1. While the overall reporting structure is similar to standardized manual reporting, with sections for metadata of the exam such as administrative information or patient identifiers, observations that report on visible findings, and conclusions, we see the findings section as consisting of three major components.

1. Identifying relevant information on the desired indication
2. Grounding the information based on known semantic priors such as anatomical concepts
3. Quantifying the severity of the identified finding

This conceptual design allows a radiologist to build a specified reporting system by defining findings of interest. In addition, the modular design enables highly flexible adaptability to the desired task, and underlying models provide interpretable information through dense predictions and extraction of quantitative features.

The challenge of implementing a holistic reporting system is the need for dense predictions, such as bounding box detections or mask segmentations, which require expert annotation of thousands of images. This tedious task is even more difficult in the medical field, where pathologies are hard to identify, and correct delineation is difficult due to imaging techniques. As a result,

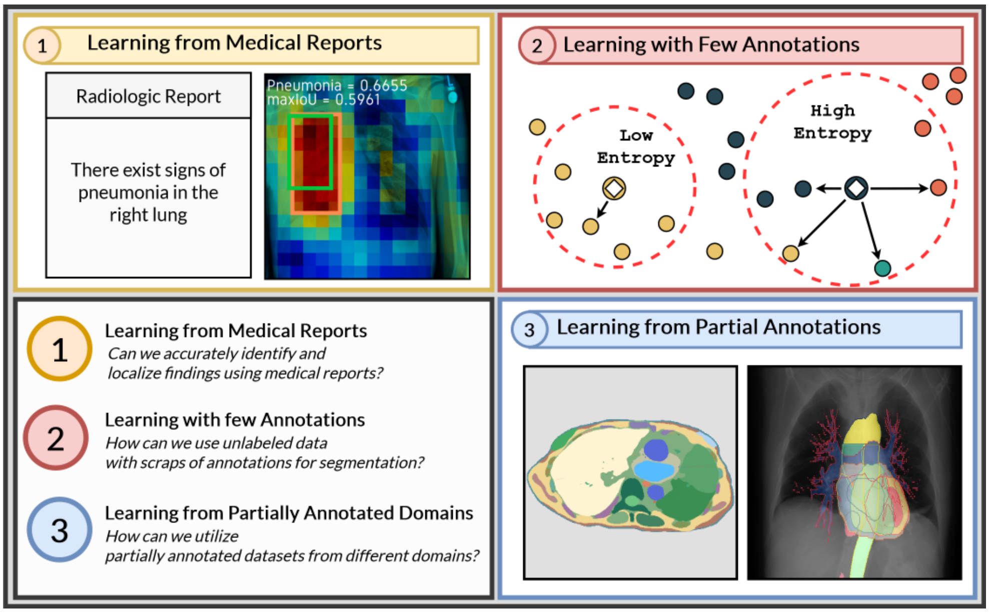

1 Learning from Medical Reports
Radiologic Report
There exist signs of pneumonia in the right lung
Pneumonia = 0.6655
maxIoU = 0.5961
2 Learning with Few Annotations
Low Entropy
High Entropy
1 Learning from Medical Reports
Can we accurately identify and localize findings using medical reports?
2 Learning with few Annotations
How can we use unlabeled data with scraps of annotations for segmentation?
3 Learning from Partially Annotated Domains
How can we utilize partially annotated datasets from different domains?
3 Learning from Partial Annotations

Figure 2: Overview of the contribution areas and the underlying research questions of this thesis. We develop methods for (1) learning from available medical reports, (2) learning scraps of manually annotated data, and (3) the utilization of partially annotated domains for dense predictions.

the manual annotation of a single medical image can take hours, making the dense annotation of a whole dataset nearly impossible.

So, what can we do to achieve a holistic reporting system? While directly learning to produce an entire report is complex, using the limited amount of available data to learn to solve partial tasks seems reasonable. We explore methods for learning from medical reports, using small amounts of dense supervision, and demonstrate how we can leverage easier-to-annotate domains to build the basis for holistic reporting systems. Thus, we focus on developing algorithms for identifying arbitrary medical findings and anatomical structures.

1.3 THESIS ROADMAP AND CONTRIBUTIONS

This dissertation focuses on developing algorithms that bring us closer towards holistic radiologic reporting systems in limited data scenarios. Initially, we discuss and delineate us from related work in Chapter 2. Afterwards, we investigate three low supervision settings, which we illustrate in Fig. 2. This includes utilizing implicit knowledge from medical reports (Chapter 3), leveraging small amounts of densely annotated labels (Chapter 4), and making use of multiple partially annotated domains (Chapter 5). Finally, we review our overall contributions to the field (Chapter 6) and discuss potential directions of future work (Chapter 7).

LEARNING FROM MEDICAL REPORTS

Can we accurately identify and localize medical findings using training data directly from medical reports? In Chapter 3, we focus on scenarios of language-based supervision signals. By developing deep neural networks using automatically generated labels from medical reports, we demonstrate how modeling the problem as multiple instance learning can improve the localization of pathologies while improving classification performance (based on our *ACCV 2020* publication [245]). Subsequently, we expand this setting by abandoning the concept of image-level labels and adapting straight text supervision. Conditioning models directly on text allows us to break away from predefined definitions and manual annotation as we present the first foundation model that enables open set finding identification and localization in medical images(as published in *MICCAI 2022* [246]).

Learning from Limited Amounts of Annotation

How can we utilize abundant unlabeled data in combination with limited annotations to perform semantic segmentation? Unlabeled data is often abundant in medical imaging applications, but creating densely annotated data can be a tedious and challenging task. In Chapter 4, we explore how to use small amounts of annotations as a reference for unlabeled samples to perform semantic segmentation. Our focus is on anatomy segmentation in chest x-rays, where the overlap of different structures makes typical semi-supervised approaches unsuitable. We propose a novel pseudo-labeling mechanism designed explicitly for multi-label segmentation. We argue that visually similar regions between labeled and unlabeled images likely contain the same semantics. We use this information to match pixels in an unlabeled image to the semantics of the best-fitting pixel in a reference set. This approach avoids pitfalls such as confirmation bias, which is common in purely prediction-based pseudo-labeling (published in *AAAI 2022* [242]).

Learning from Partially Annotated Domains

How can we utilize various partially annotated datasets? In Chapter 5, we investigate the idea of a specific case of transfer learning following some elementary assumptions. The human body remains unchanged regardless of the imaging modality, but some modalities allow for more accurate visual feature delineation than others. Similarly, a CT allows for more accurate delineation of visual features than an X-Ray and many models suffice with a handful of training examples for many tasks. As such, there exist several CT datasets specific for a particular task, such as organs at risk segmentation. We show how we can leverage different annotations made in the CT domain to accurately segment Chest X-Rays (published in *BMVC 2022* [243]). We also demonstrate how to scale the size of these projected datasets and how resulting models can be used to identify anatomies at a level close to the inter-annotator agreement.

A Note on implementation

Constantin Seibold is solely responsible for the implementation of *all sections* of this work.

2

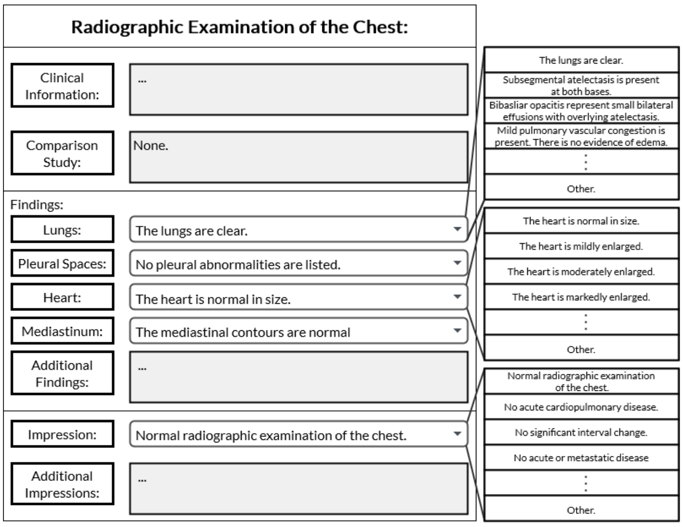
RELATED WORK

The methods proposed in this thesis fall into the intersection between computer vision and natural language processing, radiology, medical image analysis, and learning with limited data. While several of our results affect different research areas, we see our most significant impact on methods for developing densely annotated medical datasets. We provide significant contributions in areas such as multi-label open-set inference and semi-supervised segmentation and propose conceptual changes to the ideas of medical reporting, including the integration of dense annotations for the automatic generation of quantifiable reports. This chapter provides a comprehensive overview of the literature influencing our work.

2.1 AUTOMATING MEDICAL REPORT GENERATION

Written medical reports serve as a crucial bridge between clinical departments and between doctors and patients, providing crucial information about a patient's current state. [90, 127]. The reports contain, apart from metadata such as the sex or age of the patient, relevant information about its current state based on an underlying indication and presented data. In radiology, reports focus on identifying anomalies and comprehensibly transferring that knowledge to others. This transfer can be done in a standardized or free-text manner. While standardized reports provide clear and consistent results due to their enforced structure, free-text reports allow for dynamic descriptions of particular cases and require a smaller training period. For a system to effectively reconstruct this process, it must first learn to analyze an image accurately and then clearly convey complex medical information in a way that medical professionals and non-medical readers can easily understand.

One approach to medical report generation is to use template generation, where the system fills in the blanks in a predefined template with the appropriate medical information [200]. As structured reports focus on certain data elements in a specific format, many settings can be formulated as a classification task. This allows for streamlining models to predict predefined classes similar to options in a doctor's evaluation [127].


Radiographic Examination of the Chest:

Clinical Information: ...
Comparison Study: None.
Findings:
Lungs: The lungs are clear.
Pleural Spaces: No pleural abnormalities are listed.
Heart: The heart is normal in size.
Mediastinum: The mediastinal contours are normal
Additional Findings: ...
Impression: Normal radiographic examination of the chest.
Additional Impressions: ...

The lungs are clear.
Subsegmental atelectasis is present at both bases.
Bibasliar opacitis represent small bilateral effusions with overlying atelectasis.
Mild pulmonary vascular congestion is present. There is no evidence of edema.
⋮
Other.
The heart is normal in size.
The heart is mildly enlarged.
The heart is moderately enlarged.
The heart is markedly enlarged.
⋮
Other.
Normal radiographic examination of the chest.
No acute cardiopulmonary disease.
No significant interval change.
No acute or metastatic disease
⋮
Other.

Figure 4: Excerpt of a structured report for chest radiograph analysis as endorsed by the TLAP [238, 341].

We display an example of a typical structured reporting template for standard chest radiograph analysis endorsed by the Template Library Advisory Panel (TLAP) consisting of members of the RSNA and ESR [238, 341] in Fig. 4. This template consists

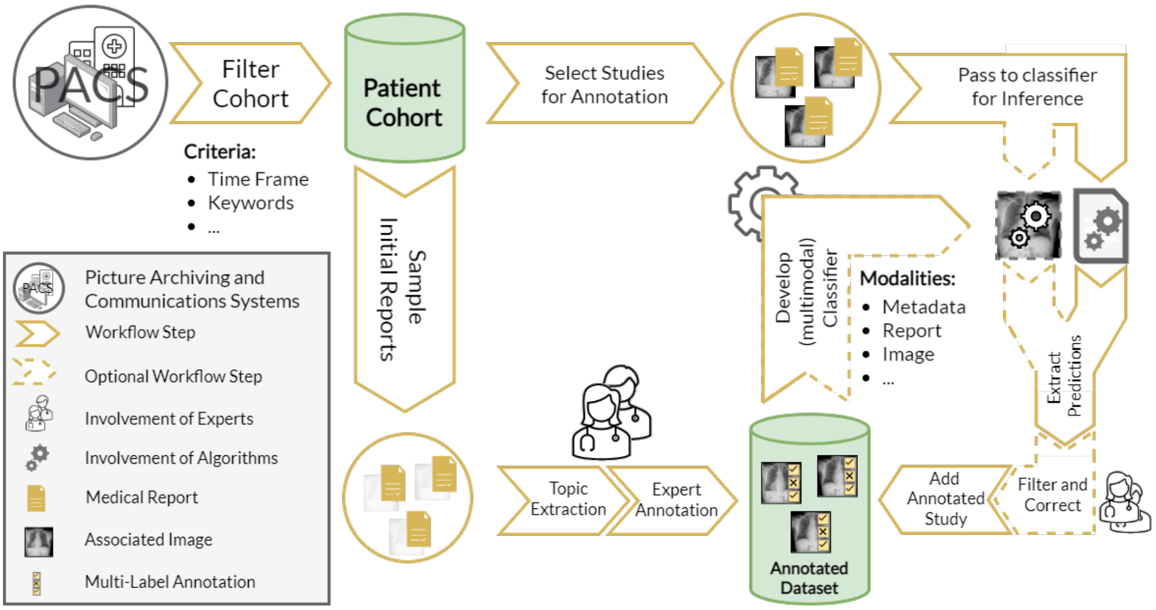


Figure 3: An exemplary workflow for the creation of datasets based on medical reports for the classification of pre-defined findings. Based on the manual annotation of selected samples a text classifier is trained. The text classifier is then applied on either a subset or the remaining amount of the raw dataset. This process can be enhanced through the integration of visual classifiers trained on labels from the text predictions.

of a mix of free-text fields for metadata, such as clinical information and comparisons and drop-down menus for commonly affected areas in chest radiographs. The drop-down menus contain options to describe the state of an anatomy. I.e., the varying deviations of heart size or differently affected areas of the pleura by effusions. In cases that stray away from the most common findings, the radiologist can mark the option “other” and elaborate via the free-text area “Additional Findings”. This template structure enables concise and standardized processing of typical cases while also enabling documentation of outliers.

Recent datasets [24, 52, 110, 292] have enabled the development of deep learning approaches for the classification of findings [17, 102, 187, 199, 209, 245, 293, 322]. These datasets are typically annotated for a fixed set of findings [110, 259, 292], with a manual report-level annotation to denote the presence or absence of a particular class. We display an exemplary workflow in Fig. 3. One starts by selecting a set of cases with clean visual data for manual report-level annotation to denote the presence or absence of a class. Then, using these text annotations, one can apply a text classifier for multi-label classification of medical reports for the remaining dataset [110, 292]. Some approaches refine the annotation process by taking iterative training of visual multi-label classification models using these annotations into account [23]. The ensemble of text and visual classifiers has led to a more accurate classification of findings [23, 293]. Many works build on these datasets to identify diseases, using different kinds of convolutional neural networks or transformers for multi-label classification [12, 25, 30, 85, 86, 149, 155, 157, 159, 167, 193, 208, 209, 231, 251, 271, 287, 290, 292, 293, 311, 320]. These models can complete simple template-based reports, but are unable to adapt to provide detailed information such as measurements or deviate from the predefined setting. As a side note, although the template-based reporting approach is straightforward and natural to apply, it is not well investigated in current literature and clinical settings.

Another direction is the automatic generation of free text reports through the use of visual encoders and language decoders in an end-to-end manner [19, 103, 119, 151, 272, 293, 297]. Some models incorporate comparisons to prior knowledge through retrieval of similar cases in the training set [60, 158] or learned prototypes [288], as well as external knowledge from knowledge graphs [204, 262, 329]. However, the underlying structure often relies on standard visual encoders, such as DenseNet, [103, 105]. Language models, ranging from Recurrent Neural Networks (RNN) [303] using Long Short-Term Memory (LSTM) [100, 293] to recent transformer-based methods like Generative Pre-trained Transformers (GPT) [11, 19, 206]. While these methods produce human-readable text, they can be unreliable as they suffer from flaws in the training data such as data imbalance [235] or model design as they are unable to extract quantitative measurements due to the lack of dense predictions. Moreover, it has been shown that in some cases, even simple baselines not relying on image features can beat these methods in commonly used metrics [11], indicating a disconnect from models that provide a human-like comprehension of images.

Lastly, we want to mention the importance of exact measurements and localizations in radiology. Several studies have quantified the severity of diseases based on the size and position of anomalies such as clots in the vasculatory system [65]. Other research has examined the relationship between anatomy volume and the occurrence of pathologies [6, 220, 302]. The measurement of tumor size is also crucial in determining the malignancy and treatment of cancer or its correlation to other pathologies [16, 89, 192]. While detailed measurements like those used in research can provide a more comprehensive view of a patient's condition [276], they are often not performed in clinical practice due to the time-consuming nature of manual measurements [98]. While machines built on dense predictions can automate these tedious tasks, they are often not used in automatic reporting because of a lack of available training material, particularly in chest radiographs. They are, therefore, mainly limited to specific use cases.

Our contribution: One of the significant challenges in current medical reporting methods is their limited reliability of predictions, inability to provide quantifiable data, and reliance on restricted training data. The concept of holistic automatic reporting addresses these issues by utilizing visually interpretable predictions, which can be challenging to obtain in the medical field through heavily annotated training data. In Section 3.2, we abandon the concept of label supervision for free-text reports in pathology classification models leading to open-set classification, which, in turn, allows for a greater range of potential subtleties through template-based reporting. In addition, we improve disease localization through our self-guiding loss in Section 3.1.

The reliability and interpretability of a report are often restricted by the network's ability to recognize structures in medical images. However, as this data type is not always available for training, we have developed methods that enable deep learning methods to perform these tasks for anatomical and pathological information in Section 5.2. This allows for the automated generation of medical reports incorporating a wide range of information such as concrete sizes of anatomical structures or lung nodules.

2.2 UNDERSTANDING OF VISION AND LANGUAGE

Human comprehension of every day life situations depends on their cognition of various signals. As we form semantic concepts from gathered information, we can convey this information to others through common signals such as images or language. As we intend to approach more human-like models, it is essential for models to not only grasp visual aspects but also link these to corresponding semantic concepts as found to allow for a common ground of communication between humans and machines.

With the emergence of deep learning, there have been significant strides in the juncture of semantic concepts of both computer vision and natural language processing, thanks to advances in CNNs [96, 141, 258] and word embeddings [181, 196]. Early efforts in this field used two-stage training pipelines that were specific to particular downstream tasks, such as Visual Question Answering [5] or Image Captioning [283] using datasets like COCO Captions [36, 163] or Flickr30k [202]. These models typically combined trained visual models, like plain convolutional networks [96, 141, 258], with textual encoders, using late fusion methods like concatenation or feature refinement to merge the two feature spaces [5, 100, 113, 174, 283]. More recently, multi-modal attention approaches have been proposed to better capture the alignment between image and text regions. These models adapt the influence of correspondences between the two modalities based on their weight [132, 309, 318]. While these models allowed to tap into the potential of multi-modal understanding, they have been limited by the size of the training datasets, which limits their ability to understand semantic concepts and go beyond what they have seen before.

To overcome this limitation and to create universal vision-language models, recent approaches follow the concept of large-scale multi-modal pre-training utilizing datasets with millions [190, 205, 240, 250] or even billions of image-text pairs [34, 116, 197, 239]. To use this large amount of data, most methods typically employ fusion-based or dual encoder-based approaches [38, 153, 205, 330]. Dual encoder architectures use separate encoders for each modality and only integrate the interaction between the modalities through similarity measures between the image and text representations [116, 205, 330]. On the other hand, fusion encoders use additional self-attention layers to model a shared latent space [38, 152, 326]. While fusion-based architectures tend to perform better on visual reasoning tasks, dual-encoder architectures are more flexible and excel in many standard vision tasks. [1]

Zhang *et al.* [330] lead the way with the first dual encoder architecture for vision-language pre-training. They proposed a contrastive objective between an image and a sampled sentence from a medical report, pulling co-occurring image-sentence pairs closer in latent space and pushing non-co-occurring pairs further apart. This was applied for finetuning classification models and retrieving Chest X-Rays and corresponding reports. Similarly, Radford *et al.* [205] and Jia *et al.* [116] apply the same objective on the natural image domain and, apart from potential finetuning, show the effectiveness of this approach for open-set [2] recognition by treating image classification as an image-to-text retrieval problem, where the goal is to return the most similar text class concept (also known as a prompt) for a given image. It is worth noting that the choice of prompts can have a significant impact on recognition ability, with swings in performance of, i.e., more than 40% for satellite images [205, 336, 337], indicating that in addition to model quality, domain knowledge may be necessary to make full use of pre-trained models.

For tasks such as image-based reasoning or visual grounding, many models rely on dense predictions in the form of bounding boxes or masks [4, 38, 78, 267, 317, 323]. These models often use pre-trained models such as Faster R-CNN [218] or panoptic feature pyramid networks [134] trained on a high-quality dataset such as MS COCO [163] or Visual Genome [139], to generate high-quality proposals [78, 164, 316]. These proposals can then be used to explicitly model intra-

---

[1] Fusion-based architectures can be less efficient due to their need to encode all possible image-text pairs to compute similarity scores, which can hinder their use for open-set recognition or explicit retrieval [73].

[2] The used definition of "open-set" or "zero-shot" recognition becomes somewhat fuzzy in literature. While initially, one considers recognition of classes unseen during training, several models see many of the concepts during training, however, are not explicitly trained for classification.

image associations [124, 164, 316], which in turn allows for the solution of complex tasks such as high-level reasoning [298].

Our contribution: Whereas most recent methods stated above rely on the abundance of available web data, the medical domain often faces difficulties in obtaining sufficient data and imbalanced data distribution for disease patterns. This thesis aims to address these challenges and facilitate a holistic understanding of medical images. Thus, we investigate, on the one hand, how we can leverage report supervision to enable dynamic pathology recognition on chest X-rays. In Section 3.2, we establish the first multi-label inference scheme for open-set recognition with vision-language models [3]and investigate the impact of prompting in the medical domain. Furthermore, we show how we can effectively scale datasets for vision-language pre-training through prompt-based supervision. Additionally, we address the limited availability of dense annotations in the medical field by investigating methods for learning with limited data (Chapter 4) and introducing a new, densely annotated fine-grained dataset for the segmentation of anatomical structures, enabling anatomy-based image analysis (Chapter 5).

2.3 LEARNING FROM LIMITED ANNOTATIONS

The availability of densely annotated datasets has helped to improve understanding of complex scenarios in fields like autonomous driving [83] and smart living [49, 203]. However, the high annotation cost can be a major barrier to deep learning in many applications. For example, it can take around an hour to annotate a single mask image in the Cityscapes dataset [46]. While it is relatively easy for most people to understand and segment a street scene and thus can be outsourced to laypeople, this may not be the case for more specialized domains like medicine, which often require input from domain experts for proper annotations [216]. As domain experts are often hard to come by, it is essential to consider methods that can reduce the annotation burden and help to build datasets for comprehensive image understanding.

One approach that has gained popularity is using weak supervision, which relies on the co-occurrence of class labels and inherent image features. By using image-label pairs, networks learn to recognize discriminative regions of objects, which can be visualized using saliency methods like class activation mappings (CAMs) [29, 248, 334] or guided backpropagation [263]. For visualization, CAMs adapt the commonly used concept of the average pooling of the feature maps before a final classification layer by removing the pooling layer and replacing the linear classification layer with a 1×1 convolution. This way, one applies the classification weights to the individual elements of the feature map followed, which can be normalized to generate a heatmap to indicate the most relevant regions within the image. By using saliency maps as hints, several methods have been able to retrieve bounding boxes [55, 74, 154, 301] or segmentation masks [256, 300] using strictly image-level supervision. Some combine these methods to merge network knowledge with low-level image features resulting from region proposal methods such as EdgeBoxes [339], Selective Search [278] or BING [39] to generate more accurate regions of interest.

A different tactic is Semi-supervised learning (SSL) which combines few densely labeled samples with a larger number of unlabeled samples to improve the performance of machine learning models. This field has seen significant progress in recent years through different conceptual di-

---

[3] Tiu *et al.* [274] independently introduced a similar framework for open-set pathology recognition around the same time as our MICCAI 2022 [246] publication.

rections [14, 15, 28, 33, 191, 198, 260] as most methods follow one or combinations of directions such as consistency regularization [70, 115, 260, 273] or entropy minimization [82].

Consistency regularization is based on the idea that different views of the same data describe the same semantics. Thus the encodings made by a model for these different views should also be similar, usually enforced with a loss term. In many cases, a simple augmented version (i.e. slight cropping) is used as a supervision signal for more strongly augmented views of a sample (multiple iterations of RandAug [48]) as in FixMatch [260]. For example, French *et al.* [70] use CutMix [324] to enforce consistency between mixed outputs and the predictions from corresponding mixed inputs for semi-supervised segmentation. Ouali *et al.* [191] aligns the outputs of the main segmentation decoder module and auxiliary decoders trained on different perturbations to enforce consistent feature representations. Chen *et al.* [35] use two independent networks with the same structure and enforce consistency between their predictions.

Entropy minimization aims to maximize a classifier's confidence by approaching the optimal decision boundary for an underlying system by extrapolating information from annotated to unlabeled ones [81]. One way to do this is through pseudolabeling [28, 111, 147]. Approaches typically train a classifier with unlabeled data using targets derived from a model's predictions with a confidence value above a certain threshold, assuming that predictions with high confidence are more likely to be correct. [147]. The threshold can be set manually or dynamically depending on the prediction [150, 310]. However, pseudo-labeling can be problematic if the model is poorly calibrated or makes incorrect high-confidence predictions [87, 223], leading to noisy training data. PseudoSeg [340] adapts FixMatch [260] and thus enforces consistency between segmentations of weakly and strongly augmented images employing GradCAM [248]. Contrary to score-based pseudo-labeling, one can follow a transductive setting and perform label propagation. This generation process of pseudo-labels is often not feasible in an online setting as the requirement of a nearest-neighbor graph to the high demand on run-time and memory consumption for label-propagation. It is performed after a set amount of iterations [111, 165, 254]. Even when not considering all data simultaneously, Taherkani *et al.* [266] match clusters of unlabeled data to their most similar classes in an offline procedure. In this fashion, pseudo-labeling literature can be divided into online variants, which build pseudo-labels for unlabeled data directly during forward pass [147, 260], and offline variants, which generate new targets for the dataset in greater intervals [28, 33, 111, 198, 308].

Another recently growing idea for segmentation with limited resources is the utilization of synthetic training data [222, 230, 236]. Instead of using manually annotated data, which can be costly and time-consuming, methods utilizing simulation platforms such as video game engines can mass-produce synthetic training data [222, 225]. The challenge lies in matching the concepts of the synthetic source domain with the ones of the desired target application. Several methods attempt to solve this via latent space or distribution alignment of the source and target domains [282, 305], variants of pseudolabeling [128], and consistency regularization [237]. Synthetic data has been utilized in games like Sims, GTA, and Witcher 3 [138, 222, 225], the concept of utilizing an easy to annotate source domain as a basis for the segmentation of others has also been applied in the field of radiology. For example, Jiang *et al.* [117] used annotations of CT scans to train models for the segmentation of lung tumors in MRI images through image translation methods based on CycleGAN. Similarly, Dou *et al.* [59] utilized adversarial feature alignment between CT and MRI for whole heart segmentation.

Our contribution: The concept of holistic automatic reporting heavily depends on visually interpretable predictions, which are difficult to obtain in the medical field through densely

	Dataset	**Year**	**Source**	**Frontal**[4]	**Lateral**	**Annotations**	**Findings**	**Anat.**	**Instances**
Original	JSRT[257]	2000		247	-	M	1	5	1.2K
	Montgomery [114]	2013		138	-	L	1	-	-
	Shenzen [114]	2013		615	-	L	1	-	-
	OpenI [52]	2016		3.8K	3.6K	R	-	-	-
	CXR14 [292]	2017		112K	-	L, B	14	-	1.6K
	CheXpert [110]	2019		191K	32K	L	14	-	-
	PadChest [23]	2019		108K	49K	L	170	104	-
	MIMIC-CXR [120, 121]	2019		224K	133K	R,C	14	-	-
	Object-CXR[97]*	2020		9K	-	B, E, P	1	-	18K
	VinDr-CXR [188]	2020		18K	-	B	22	-	9K
Subset	RSNA Pneumonia[255]	2019	CXR14	26K	-	M	1	-	6K
	SIIM-ACR[64]	2020	CXR14	12K	-	M	1	-	3.5K
	ReflacX[18]	2021	MIMIC-CXR	3K	-	ET	N/A	N/A	N/A
	ChestX-Det[160]	2021	CXR14	3,5K	-	P	14	-	9.6K
	CLiP[268]	2021	CXR14	30K	-	L	3	-	17K
Ours	PAXRay[243]	2022	RibFrac[118]	440	440	M	-	166	145K
	PAXRay++	2023	Misc[118]	7.2K	7.2K	M	-	158	2M

Table 1: Comparison of related datasets for Chest X-Rays with our PAXRay variants. R, C, B, E, L, M, P stand for Report, Class Labels, Bounding Boxes, Ellipsoids, Lines, Masks, Polygons respectively.

annotated training data. To address this challenge, we propose several strategies to alleviate the annotation burden. We propose a loss function based on dynamic entropy minimization for classification models, which allows for more interpretable class activation mappings (Section 3.1). We also demonstrate how to extract multi-label open-set localizations(Section 3.2). Furthermore, we intertwine online-generated pseudo-labels with consistency regularization to alleviate drawbacks in either of the two for multi-label segmentation(Chapter 4). This task is especially relevant in domains with many overlapping visual structures, as in chest X-Rays. Additionally, in Chapter 5, we use computer tomography to generate large-scale, densely annotated datasets of pseudo chest x-rays, enabling the segmentation of real-world x-ray data, a crucial step in adequately analyzing these images.

2.4 CHEST RADIOGRAPH DATASETS

A deep learning-based system can only be as good as the data it is built on. While some datasets aim to provide as much data as possible [239, 240], others prioritize detailed annotations to enable a wide range of downstream tasks [88, 163]. In the medical domain, however, gathering annotations can be difficult due to the need for medical experts and the challenges of sharing highly personal data. We display a list of available datasets in this domain in Table 1.

Initial public datasets were relatively small-scale (less than 1K images). They focused on particular classes such as JSRT [257] for nodule detection and the Shenzhen as well as the Montgomery County dataset [114] on tuberculosis. JSRT also provides anatomical masks for the heart, lungs, and clavicles. The IU OpenI dataset has been the first dataset containing image-report pairs with studies containing both lateral and frontal views, but, as it contains less than 10K weakly labeled images, training deep networks with it can lead to non-generalizable results [11].

Recently, several chest x-ray datasets have been released that build on available image-report pairs. Datasets such as ChestX-Ray14 [292], CheXpert [110], and PadChest [23] automatically

---

[4] Frontal refers to AP/PA Chest X-Rays alike.

parse pathology labels from medical reports using workflows similar to what is displayed in Fig. 3 and provide them with corresponding images. This narrows the focus and protects personal information contained in medical reports. In addition, MIMIC-CXR [121] provides the image-report pairs openly, allowing for a broader range of information. While these datasets have made significant progress in automating x-ray processing as they provide vast amounts of annotated images (>100K images), they suffer from the lack of dense mask annotations, which are essential for many computer vision approaches [253, 321].

There have also been efforts to expand these larger datasets through additional dense annotations. For example, the SIIM-ACR dataset [64] provides 3576 mask annotations for pneumothorax in 12,954 frontal images, while the RSNA Pneumonia Challenge [64] contains 6012 masks for 26,684 cases of pneumonia. The VinDr-CXR [188] provides box annotations for 18K images. However, they also display a very long-tail distribution. Bigolin et al. [18] provide roughly 3,000 sets of eye-tracking data from radiologists corresponding to medical reports and images. Lian et al. [160] provide 9.6K instance masks for 3578 images for diseases in thirteen common categories of the ChestX-Ray14 dataset. The CLiP dataset contains 18K line annotations for catheter and tube types.

Our contribution: Overall, the lack of anatomical annotations limits their utility for reliable reasoning (e.g., correct alignment of an endotracheal tube within the trachea) and hinders our ability to understand and interpret the images in their entirety. Therefore, to further advance the field of automated x-ray processing, it will be essential to prioritize the inclusion of detailed anatomical annotations in future datasets.

In Chapter 5, we introduce the PAX-Ray dataset, a comprehensive resource for anatomy segmentation in chest radiographs. With over one million mask annotations for various anatomical structures in both frontal and lateral views, PAX-Ray is a unique and valuable resource for researchers and practitioners. Our fine-grained label structure includes 160 labels covering the thoracic, abdominal, vasculatory, and bone systems, allowing for a thorough analysis of chest radiographs. Given the labor-intensive nature of manually annotating such a dataset, the release of PAX-Ray significantly contributes to automated x-ray processing. It will facilitate further research on anatomical understanding and automatic report generation.

2.5 MULTIPLE INSTANCE LEARNING

Multiple Instance Learning (MIL) has become a widely adopted assumption within weakly supervised learning [26, 171, 234, 285], often used in weakly supervised detection [41, 63, 269] and segmentation [148, 252, 270]. MIL works by assuming that a sample consists of a “bag” of instances, and the bag is considered positive for a certain class if there is at least one positive instance within the bag. However, during training, only the bag-level label is available. While initially proposed for drug activity prediction [57], it has since found a use for fields such as common signal processing [136, 176, 296, 338] and the analysis of large, difficult-to-annotate datasets, which often face a small amount of positive and large amounts of negative elements, like histopathology images [47, 108, 148, 264].

While the choice of pooling function is often a max-operator or an approximation for deep MIL networks [159, 176, 292], recent research has explored the use of learnable pooling functions that combine embeddings or predictions of instances to make bag-level predictions [107, 161, 176, 296, 320, 338] but also on how to leverage the use of artificial supervision within a MIL setting to train the network additionally through instance-level losses [183, 249, 295, 338]. MIL has also been shown to improve by introducing artificial instance labels for prediction scores above a certain threshold. However, this method has limited supervision as it ignores the association between positive and negative instances. [249, 295, 338].

Our contribution: Our research addresses the challenges faced by traditional MIL methods for data with imbalanced class distributions. In these scenarios, the minimization of entropy leads to clumped value distributions within a bag, making it difficult to apply strict thresholding for pseudo-label generation. We propose a novel approach that uses instance-level scores to adapt a threshold to create artificial instance-level supervision dynamically. This allows the network to learn implicitly from the bag label while also considering explicit instance-level predictions in training. Furthermore, by dividing the prediction maps into the foreground, background, and ambiguous regions, the network can provide instance-wise targets of varying levels of certainty.

Part II

LEARNING FROM LIMITED SUPERVISION

3

LEARNING FROM RADIOLOGIC REPORTS

Automated medical report generation has the potential to become a streamlining tool in the the healthcare process. We start of with one of the pivotal problems of automated reporting, namely, the identification and localization of relevant findings. Since such information is contained within a medical report, the main goal of this chapter is to investigate methods that are able to extract the required knowledge to teach neural networks to identify findings. We, first, take a look at learning from parsed labels and how we can make use of implicit information to localize findings in Section 3.1. As parsed labels restrict a networks potential, we, in Section 3.2, show subsequently how report-level supervision enables neural networks to identify findings in an open-set manner.

The ability to accurately identify abnormalities in medical images is crucial for effective diagnosis and treatment. Deep learning techniques have made great strides in this area, enabling the reliable recognition of a wide range of categories in images, as demonstrated on numerous large datasets [141]. This same approach can be applied to the analysis of medical observations, such as the millions of chest radiographs (CXR) captured annually [186]. The availability of large CXR datasets like Open-I and ChestX-ray14 [23, 52, 110, 121, 292] has allowed for the use of deep neural networks to assist in the detection of pulmonary abnormalities [12, 25, 85, 149, 155, 157, 159, 167, 193, 208, 209, 231, 251, 271, 287, 290, 292, 293, 311]. The underlying concept is displayed in Fig. 5. The CXR of a patient is passed through a network. The network processes the visual features via a series of convolutions [146] or attention blocks [281] to identify the occurence of a pathology. The location of a finding can then inferred based on the activations of the network.

But how are these networks trained? While in the natural image domain most commonly there exists enough labelled data to train sophisticated models for detection [168, 213, 218] or instance segmentation [95], most existing large-scale CXR datasets rely on the usage of medical reports as manual annotation is hard to perform in the clinical setting on a large scale due to the unattainability of medical staff.

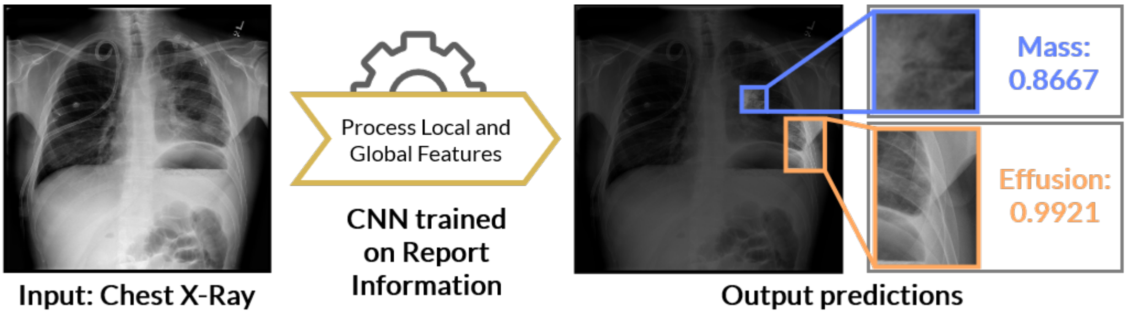

Process Local and Global Features
CNN trained on Report Information
Input: Chest X-Ray
Mass: 0.8667
Effusion: 0.9921
Output predictions

Figure 5: In our framework, the network trained on report-based information reads chest X-ray images and produces overall image-level pathology prediction scores and their corresponding locations.

In the following sections, we develop methods that make use of different training setup involving reports and show not only how to improve localization through our self-guided loss but also how we enable open set finding recognition through report-guided contrastive training and subsequent prompting.

3.1 PATHOLOGY RECOGNITION FROM REPORT-LEVEL LABELS

The following section is based on our publication in *ACCV 2020* [244].

One common approach for training neural networks for finding recognition often involves the use of labels from medical reports as targets for image-based training [110, 292]. These labels are generated in a similar manner as displayed in Fig. 3. While this workflow saves time, it can lead to inaccurate labels and hinder the training of advanced models. As a result, the identification and localization of findings is at best considered weakly supervised due to the lack of dense annotations and inherent label noise stemming from the annotation process.

There are two main approaches to weakly-supervised pathology localization in CXRs: network saliency and Multiple-Instance Learning (MIL). Saliency-based methods [12, 25, 193, 209, 251, 271, 287, 292] use visualization techniques like class activation mappings (CAM), GradCAM, or excitation backpropagation to implicitly predict locations [248, 325, 334]. However, these methods rely on global average pooling, which can lead to less indicative decisions due to healthy regions outweighing the few regions of interest containing abnormalities. MIL-based methods, on the other hand, use Fully Convolutional Networks (FCN) to implicitly learn patch-level predictions for localization [159, 167, 231, 311]. In MIL, the input data is treated as a bag of instances where the label is only available at the bag level. This approach works well for medical images, as small regions may define the presence of a pathology within the overall image.

Our focus is on MIL-based approaches for diagnosing and localizing pulmonary abnormalities in CXRs. Previous MIL-related work has explored the use of different pooling functions to aggregate predictions or embeddings [107, 161, 176, 292, 296, 312, 319, 338]. However, we argue that this approach ignores the potential of using instance-level predictions in training. We present a novel loss formulation split into two stages. The first stage leads to conventional bag-level classification, while the second stage leads to more precise predictions by generating auxiliary supervision from instance-level predictions. By separating the prediction maps into foreground, background, and ambiguous regions, the network can provide itself with instance-wise targets at different levels of certainty.

3.1.1 *Online Instance-Level Pseudo-Labeling via Self-Guidance*

We will first define multiple-instance learning. We then introduce our proposed Self-Guiding Loss (SGL) and explain how it differs from existing MIL-loss formulations. Finally, we will discuss the use of SGL for classification and weakly supervised localization of CXR pathologies in a MIL setting.

3.1.1.1 *Preliminaries of Multiple-Instance Learning*

Consider a set of bag-of-instances, $\mathscr{B}$, with size N, and their associated labels, $\mathscr{B} = \{(B_1, y_1), \ldots, (B_N, y_N)\}$. Each bag-of-instances, B_i, has N_i instances, $B_{i,j} \in B_i$, where $j \in \{1, \ldots, N_i\}$. The labels, $y_i \in \{0,1\}^C$, describe the presence or absence of C classes, which can occur independently of each other. The label for a specific class $c \in \{1, \ldots, C\}$, for a bag and an instance is denoted by $y_i^c \in \{0,1\}$ and $y_{i,j}^c \in \{0,1\}$, respectively. A label of 1 is referred to as positive, while a label of 0 is referred to as

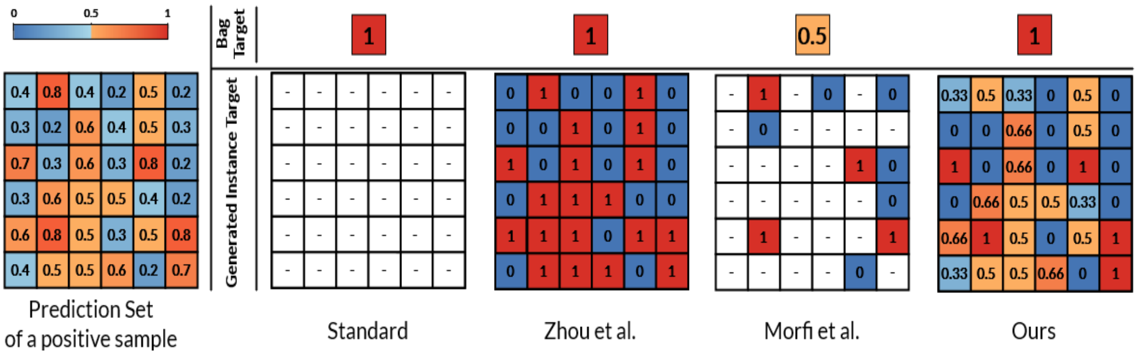


Figure 6: Illustration of supervision for different loss function concepts for MIL. Strict bag-level supervision (left) provided, Zhou *et al.*'s BIL [338] (center left), Morfi *et al.*'s MMM [183] (center right) and on the right our proposed SGL.

negative. The MIL assumption requires that $y_i^c = 1$ if and only if there exists at least one positive instance, which can be defined as:

$$y_i^c = \max_j y_{i,j}^c. \tag{1}$$

Note that while the bag-level annotation, y_i^c, is available in the training data, the instance-level annotation, $y_{i,j}^c$, is unknown.

Our goal is to learn a classifier that can predict the likelihood of each instance within a bag belonging to each class. In many deep MIL approaches, this classifier may consist of a convolutional backbone, Ψ, linked with a pooling layer, Φ, to combine predictions or features. The class-wise likelihood of a single instance is denoted by $p_{i,j}^c(B_{i,j}) \in [0,1]$, with:

$$\begin{split}\mathbf{p}_i^c(B_i) = \{p_{i,1}^c(B_{i,1}), p_{i,2}^c(B_{i,2}), \dots, \\ p_{i,N_i}^c(B_{i,N_i})\} = \Psi_c(B_i)\end{split} \tag{2}$$

representing the set of all instance-level predictions for class c of the i-th bag. These instance-level predictions are then aggregated using a pooling layer to obtain bag-level predictions:

$$p_i^c(B_i) = \Phi_c(\mathbf{p}_i^c(B_i)) \tag{3}$$

where $p_i^c(B_i) \in [0,1]$. For simplicity, we omit the arguments of the presented functions from this point on.

3.1.1.2 *Defining the Self-Guiding Loss*

The Self-Guiding Loss (SGL) is designed to address the challenges of the MIL setting, where there is a lack of knowledge about the correct instance labels and an imbalance between positive and negative instances. In MIL, it is common to merge instance predictions and train the model by optimizing any loss function using the bag's label, y, and the bag prediction, $\mathbf{p}$. This level of supervision is shown on the left in Fig 6. The bag label is depicted in the top row, while the types of instance supervision are displayed in the bottom. Numbers represent the target label, while a dash ("-") denotes that there is no supervision for that particular instance. While this level of supervision can lead to accurate bag-level predictions, it does not ensure that the determining instances are correctly inferred.

Our loss formulation is split into two parts to address this issue. The first part defines the bag-level loss, while the second part describes how the network's predictions can be used to create artificial supervision for each instance.

Bag-Level Loss. The bag-level loss functions in the same way as classic MIL approaches. It generates bag-level predictions by aggregating the network's instance-level predictions. We calculate the loss for this stage using common loss functions, $\mathscr{L}$, such as the binary cross-entropy, by passing the prediction and target for all classes and bags as follows:

$$\mathscr{L}_{Bag}(\mathscr{B}, y) = \frac{1}{C \cdot N} \sum c \sum_{i} \mathscr{L}(p_i^c, y_i) \tag{4}$$

where $i \in 1, \ldots, N$ and $c \in 1, \ldots, C$. This loss depends on the choice of the pooling function, Φ, and allows the instance-level loss to step in to provide more distinct supervision.

Instance-Level Loss. To define the instance-level loss, we start by assuming that a network trained solely on bag labels will inevitably assign some positive instances a higher prediction score than most negative instances. Based on this, we identify three types of instance predictions. Instances with a high score are likely to be considered positive, while those with a low score are likely to be considered negative. Instances with scores close to the decision boundary are ambiguous, as they may easily be swayed during training and do not provide a clear indication of the actual class of the instance. We therefore establish three types of supervision based on the certainty level of each prediction within a bag.

Our first step is to normalize the prediction set using min-max feature scaling to avoid biases resulting from algorithmic decisions or data imbalance. The resulting rescaled bag of predictions, θ, is defined as:

$$\theta_{ij}^c = \frac{p_{ij}^c - \min(\mathbf{p}_i^c)}{\max(\mathbf{p}_i^c) - \min(\mathbf{p}_i^c)} \tag{5}$$

where min and max return the minimum and maximum values within a set, respectively. We then use the normalized predictions to create a ternary mask, M, depicting targets based on the previously mentioned cases, similar to Hou *et al.* [104] and Zhang *et al.* [327]. To do this, we define upper and lower thresholds, δ_h and δ_l, to partition the prediction set, with $\delta_h + \delta_l = 1$ and $\delta_h \geq \delta_l \geq 0$. Anything larger than the upper threshold, δ_h, is considered a positive instance, while everything lower than δ_l is considered negative. The target mask, M, is then defined for each instance, j, in the bag, i, for class c as:

$$M_{i,j}^c = \begin{cases} 0 & \text{, if } \theta_{i,j}^c < \delta_l \text{ or } y_i^c = 0 \\ \theta_{i,j}^c & \text{, if } \delta_l \leq \theta_{i,j}^c \leq \delta_h \\ 1 & \text{, if } \delta_h < \theta_{i,j}^c \end{cases} . \tag{6}$$

For distinct positive and negative predictions, we can provide instance-wise supervision with target values of 1 and 0, respectively. We can also assume, based on Eq.1, that all instances within negative bags are negative, and set their masks to 0. However, for uncertain regions, it is more difficult to assign explicit labels. While we want to guide the network's decision process, we also need to account for potential misassignments. To address this, we set the target value to θ rather than a fixed value. This approach is similar to the label smoothing technique [265]. Instead of using maximal target values, we insert the adjusted value into the loss function as the target value. This slightly pushes the loss towards the most extreme predictions within the uncertain instance set, steadily increasing the number of distinct positive and negative predictions over the course of training.

We can construct the loss using a basic loss function, $\mathscr{L}$, such as binary cross entropy, by using M as the target. The instance-level loss is then defined as:

$$\mathscr{L}Inst(\mathscr{B}, M) = \sum_i \sum_c \sum_j 2^{\alpha_i^c - 1} \cdot \mathscr{L}(p^c i, j, M_{i,j}^c), \tag{7}$$

where each part is normalized by the number of pixels with the respective supervision types. This strengthens the network's decision process for its more certain instances. We also introduce a weighting factor, α, to influence the impact of each bag based on the overall certainty of its prediction. We make the assumption that a positive bag will always also contain a majority of negative instances, thus, the network should be able to seggregate between the two camps. We define α as:

$$\alpha_i^c = \max(\max(\mathbf{p}_i^c) - \mathrm{median}(\mathbf{p}_i^c)), 1 - y_i) \tag{8}$$

Since a positive bag in a common MIL setting should have a low median value due to a limited number of positive instances, it is given a high weight if the network is able to clearly separate positive from negative predictions. Therefore, for positive bags, $\alpha = 0$ if all predictions are the same value, and $\alpha = 1$ if the network is able to clearly separate positive from negative instances, under the assumption that the number of positive instances is much smaller than the number of negative ones. For negative bags, $\alpha = 1$ holds due to the given supervision.

The complete loss is now defined as

$$\mathscr{L}_{SGL}(p_i, y_i) = \mathscr{L}_{Bag} + \lambda \cdot \mathscr{L}_{Inst}, \tag{9}$$

where λ is the hyperparameter that weighs the instance-level loss. An example of the final supervision for our loss can be seen in Fig 6. The standard approach on the left does not use instance-level supervision. In the center left, Zhou *et al.* 's BIL provides a positive label for instances above the 0.5 threshold and a negative label for others, while maintaining the bag supervision. The MMM loss by Morfi *et al.* in the center right considers positive labels for the maximum instances and negative labels for the minimum instances. It also uses a target of 0.5 for a mean-pooled prediction. In contrast, our loss adapts its assumed supervision to the produced predictions. Instead of simply using the maximum or applying a fixed threshold, we threshold on a rescaled set of predictions, which helps to avoid a common issue with imbalanced data. Our formulation incorporates all instance predictions while providing a margin of error based on the network's certainty over the smoothed targets, θ, and the weighing factor, α.

3.1.2 *Multiple Instance Learning for Chest Radiograph Diagnosis*

We study the MIL scenario for CXR diagnosis where the goal is to predict the presence or absence of specific pathologies based on small, individual patches of the image. These patches, which can be seen as the instance, are grouped into a larger collection, or bag, and the label for the bag is determined by whether any of the instances suggest the presence of the pathology. For example, the presence of a nodule, which may be small and hard to see in the overall image, can be identified through the analysis of individual instances. While we are provided with image-level labels for the pathologies, we do not have access to more detailed information such as bounding boxes or pixel-level annotations. The goal of our model is to be able to classify the bag of instances and provide insights into which regions of the image are affected by the pathology.

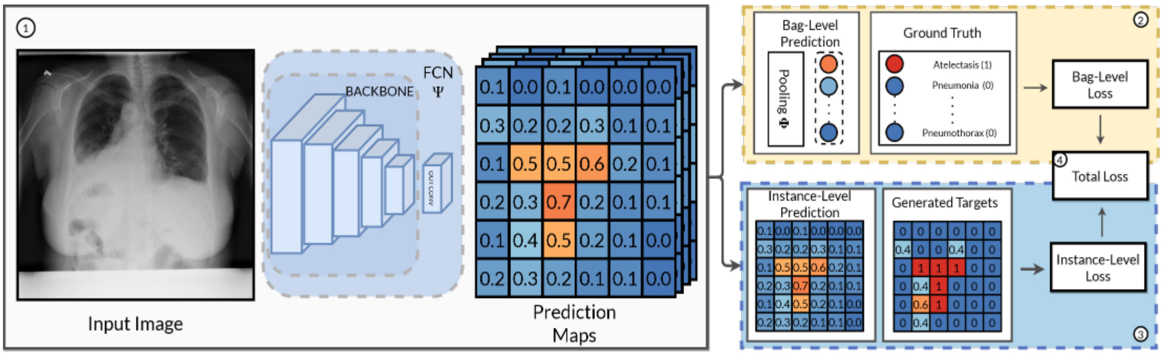


Figure 7: Overview of our framework for thoracic disease identification and localization. A chest X-ray is passed through an FCN producing a prediction map. This map is used to compute the instance- and bag-level losses.

Overview. Figure 7 illustrates our scheme for CXR diagnosis. The process begins with an FCN that processes the CXR images, resulting in patch-wise classification scores for each abnormality. The number of patches is determined by the size of the perceptual field, which is related to the backbone architecture of the network. Each patch is processed independently through a 1 × 1 convolutional classification layer. Here, we do not add any additional modules to the backbone network. The patch-level predictions are then aggregated using a pooling layer to produce a bag-level prediction, which is used to calculate the bag-level loss. In the next step, an instance-level loss function is applied based on the patch-level predictions. Finally, both the instance-level and bag-level losses are combined and optimized, with an additional penalty applied to the occurrence of non-zero elements in M using an L_2-norm, which maximizes the patch-wise confidence scores.

Choice of Pooling Function. The selection of an appropriate pooling function is crucial for accurate bag-level prediction in a MIL-setting. Max and mean pooling can result in imprecise decisions. In the context of MIL for CXR diagnosis, the Noisy-OR function has been used, but it can suffer from numerical instability due to the product of multiple instances. Instead, we have chosen to use Softmax pooling, which has been successful in audio event detection and allows each instance to influence the bag-level loss based on its intensity, providing a balance between instance-level predictions [176, 296].

3.1.3 *Experimental Setup*

3.1.3.1 *Datasets*

MNIST-Bags. In a similar manner to Ilse *et al.* [107], we use the MNIST-bags [107, 146] dataset to test our method in a MIL setting. The dataset consists of grayscale MNIST images of size 28 × 28 that have been resized to 32 × 32. A bag is considered positive if it contains at least one image with the label "9". The number of images in a bag follows a Gaussian distribution based on a fixed bag size. We study the effect of different average bag sizes and training set sizes on the performance of the model. During evaluation, we use 1000

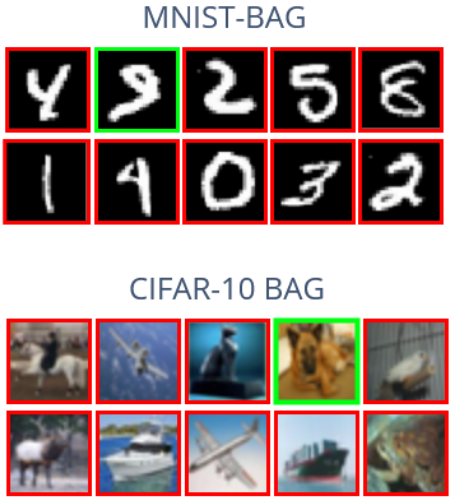


Figure 8: We display exemplary positive bags of size 10 for MNIST-bags (top) and CIFAR10-bags (bottom). Positive instances denoted in green, negatives in red.

bags created from the MNIST test set with the same bag size as used in training and average the results of ten training runs.

CIFAR10-Bags. We are creating collections of image patches called bags from the CIFAR10 [140], similar to those we previously created from the MNIST dataset. We have a total of 2500 training bags and 5000 test bags, each with a fixed number of instances. A bag is considered to be positive if it contains at least one instance with the label "dog". We display visual examples in Fig. 8. We are interested in exploring the impact of having different numbers of positive instances per bag on the performance of our model. To do this, we will average the results of five training runs.

ChestX-ray14. To demonstrate the effectiveness of our loss function for medical diagnosis, we conduct experiments using the ChestX-ray14 dataset [292]. The images are originally 1024×1024 in size, which we resize them to 512×512 for our experiments. We use the standard train/val/test split provided by Wang *et al.* [292], which gives us a split of 70%/10%/20%. Additionally, we evaluate our localization performance on the subset of 880 images with bounding boxes for 8 of the 14 pathologies.

3.1.3.2 *Implementation Details*

For the experiments on MNIST-Bags, we use the LeNet5 model [146] as described in Ilse *et al.* [107] and apply max-pooling with $\delta_l = 0.3$ and $\lambda = 1$ unless otherwise specified. We train the BIL model [338] using mean-pooling because we found that it was unable to be trained with max-pooling.

For the experiments on CIFAR10-bags, we train a ResNet-18 model [96] with the same optimizer hyperparameters and a batch size of 64 for 50 epochs. We also apply max-pooling with $\delta_l = 0.3$ and $\lambda = 1$.

For the ChestX-ray14 experiments, we use a ResNet-50 model [96] initialized with ImageNet pretraining, following the base model of Wang *et al.* [292]. We modify the final fully connected and pooling layers with a convolutional layer of kernel size 1×1, resulting in the same number of parameters as in the work of Wang *et al.* . Standard image normalization techniques [233] are applied. During training, we randomly crop the images to size 7/8 of the input image size, and use the full image size during test time. The model is trained for 20 epochs using Adam optimization [133] with a learning rate of 10^{-4}, weight decay of 10^{-4}, and $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is decayed by a factor of 0.1 every 10 epochs, and we set $\delta = 0.3$ and $\lambda = 20$. We increase the value of λ to keep the magnitudes of the two losses similar. The model is implemented using PyTorch [194].

3.1.3.3 *Evaluation Metrics*

To evaluate the performance of our network, we use the area under the receiver operating characteristic curve (AUC) as a measure of classification ability. To assess the network's localization ability, we follow the approach of Russakovsky et al. [233] and calculate the average intersection-over-union (IoU) score for each class. We threshold the probability map at a scalar value T_p to obtain the predicted area and then calculate the intersection between the predicted and ground truth areas. For MNIST and CIFAR10 bags, we use a threshold of $T_p = 0.5$ for the positive instances. The localization accuracy is calculated as the number of correct predictions (hit) divided by the total number of predictions $\frac{\#hit}{\#hit+\#miss}$, where a correct prediction is defined as a correct class prediction with an IoU greater than a predefined threshold T_{IoU}.

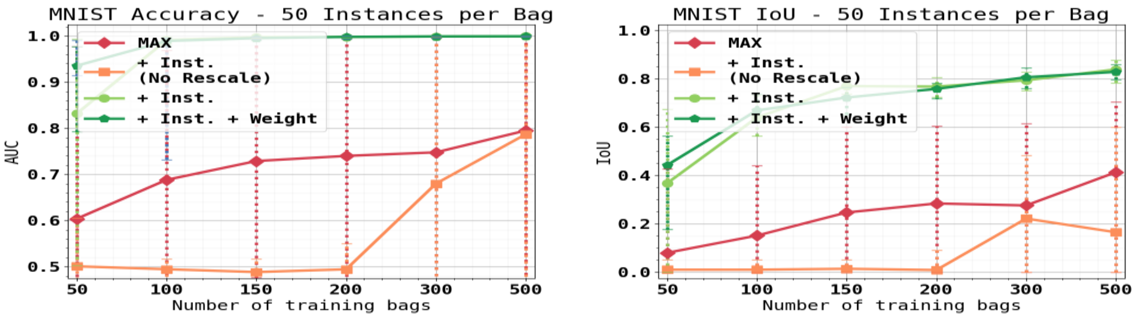


Figure 9: Ablation study on different aspects of our SGL on MNIST-bags.

3.1.4 *Results*

Ablations on MNIST-Bags. In Fig. 9 (a), we present ablation studies involving different combinations of the loss components. We start with the max-pooling baseline and successively add parts of the proposed self-guidance loss (SGL). On its own, max-pooling struggles with identifying positive and negative bags, but it improves slowly in terms of IoU and AUC with increasing numbers of training bags. When we add the proposed loss without the rescaling mentioned in Eq. 3.1.1.2 and the weighting component (shown as "Inst. (No Rescale)"), the method becomes unable to learn, as even random initializations may lead the network to incorrect conclusions. When we add the rescaling component (shown as "Inst."), the model significantly outperforms the previous versions in both metrics. This achieves higher maximums than using the applied weighting factor α, as shown by "Inst. + Weight." However, the addition of the weighting factor provides a more stable training, especially for smaller amounts of training data.

MNIST-Bags. The results of the AUC and IoU evaluations for mean bag sizes of 10, 50, and 100 with a varying number of training bags are shown in Fig. 10. The top and bottom rows display the results from left to right, respectively. The mean, best, and worst results for each method are presented. For small bags, our method performs similarly to the simple max-pooling baseline in both AUC and IoU. We attribute this average performance to the small number of instances in

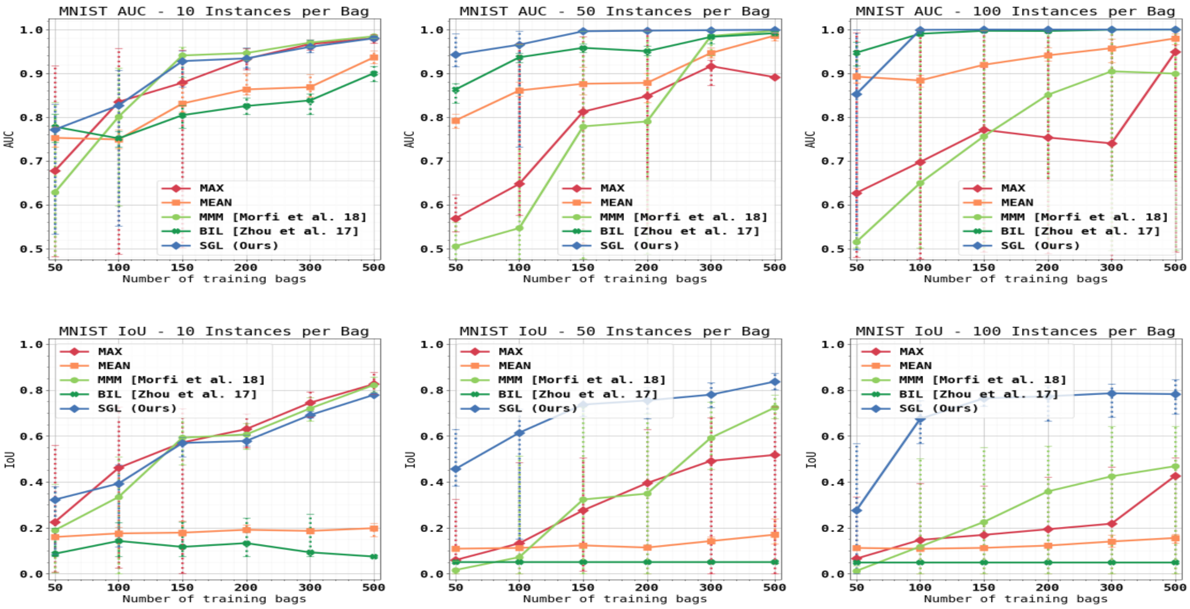


Figure 10: Test AUC and IoU for MNIST-Bags for differing avg. instances per bag.

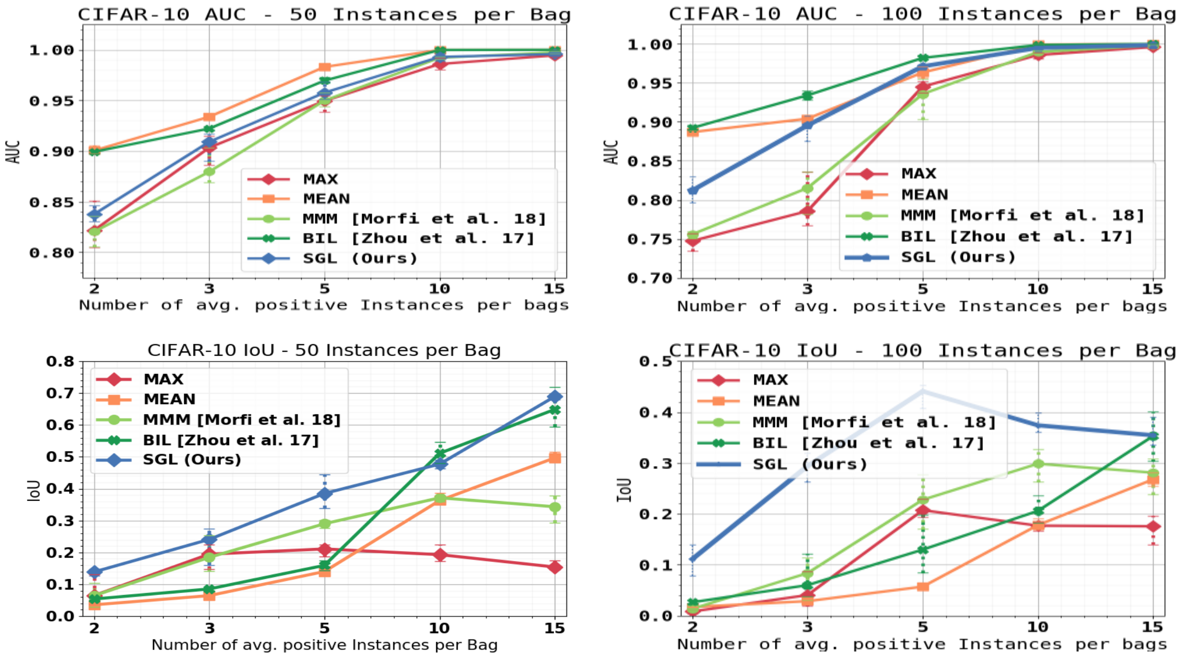


Figure 11: AUC and IoU for CIFAR10-Bags for differing number of positive instances per bag with size of 50 and 100.

a bag, which does not allow for effective use of our ternary training approach. As the bag size is increased to 50 and 100, our proposed loss performs better than the max-pooling baseline, as well as the other methods, for both metrics. This is particularly evident in the IoU, where our method achieves nearly double the performance of the next best method for almost all amounts of training bags. We find that our approach does not trade off between confident predictions and overall AUC, but rather facilitates a training environment that improves both metrics. It is worth noting that while increasing the number of training bags improves performance for any bag size, our method achieves exceptional performance for both AUC and IoU with a small number of examples for larger bags. This suggests that self-guidance can improve a method regardless of dataset size.

CIFAR10-Bags. The results for the mean bag sizes of 50 and 100 with varying numbers of positive instances per bag are shown in the top and bottom rows of Figure 11. We see that for smaller bag sizes, straightforward mean-pooling achieves the highest AUC scores for CIFAR10-Bags. In

	At.	Card.	Cons.	Ed.	Eff.	Emph.	Fib.	Hernia	Inf.	Mass	Nod.	Pl. Th.	Pn.	Pt.	Mean
Wang *et al.*	0.70	0.81	0.70	0.81	0.76	0.83	0.79	0.87	0.66	0.69	0.67	0.68	0.66	0.80	0.75
Li *et al.* *	**0.80**	0.87	**0.80**	0.88	0.87	0.91	0.78	0.70	0.70	0.83	0.75	0.79	0.67	0.87	0.81
Liu *et al.* *	0.79	0.87	0.79	**0.91**	**0.88**	0.93	0.80	0.92	0.69	0.81	0.73	0.80	**0.75**	0.89	**0.83**
ResNet-50+SGL	0.78	**0.88**	0.75	0.86	0.84	**0.95**	**0.85**	**0.94**	**0.71**	**0.84**	**0.81**	**0.81**	0.74	**0.90**	**0.83**

Table 2: Comparison of classification performance for pathologies on the ChestX-Ray14 dataset. Here, 70% of all images were used for training with no bounding box annotations available. Evaluations were performed on the official test split containing 20% of all images. “*” denotes usage of additional bounding box supervision. At., Card., Cons., Ed., Eff., Emph., Fib., Inf., Nod. , Pl. Th. , Pn., Pt. stand for Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Pleural Thickening, Pneumonia, and Pneumothorax respectively.

T_{IoU}	Model	At.	Card.	Eff.	Inf.	Mass	Nod.	Pn.	Pt.	Mean
0.1	Wang *et al.* [292]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	Li *et al.* [159]*	**0.71**	**0.98**	**0.87**	**0.92**	**0.71**	0.40	0.60	**0.63**	**0.73**
	Liu *et al.* [167]	0.39	0.90	0.65	0.85	0.69	0.38	0.30	0.39	0.60
	SGL (Ours)	0.67	0.94	0.67	0.81	**0.71**	**0.41**	**0.66**	0.43	0.66
0.3	Wang *et al.* [292]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	Li *et al.* [159]*	**0.36**	**0.94**	**0.56**	**0.66**	0.45	**0.17**	**0.39**	**0.44**	**0.50**
	Liu *et al.* [167]	0.34	0.71	0.39	0.65	**0.48**	0.09	0.16	0.20	0.38
	SGL (Ours)	0.31	0.76	0.30	0.43	0.34	0.13	**0.39**	0.18	0.36
0.5	Wang *et al.* [292]	0.05	0.18	0.11	0.07	0.01	0.01	0.01	0.03	0.06
	Li *et al.* [159]*	0.14	**0.84**	**0.22**	0.30	0.22	0.07	**0.17**	**0.19**	**0.27**
	Liu *et al.* [167]	**0.19**	0.53	0.19	**0.47**	**0.33**	0.03	0.08	0.11	0.24
	SGL (Ours)	0.07	0.32	0.08	0.19	0.18	**0.10**	0.12	0.04	0.13
0.7	Wang *et al.* [292]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	Li *et al.* [159]*	0.04	**0.52**	0.07	0.09	0.11	**0.01**	**0.05**	0.05	0.12
	Liu *et al.* [167]	**0.08**	0.30	**0.09**	**0.25**	**0.19**	**0.01**	0.04	**0.07**	**0.13**
	SGL (Ours)	0.02	0.01	0.1	0.00	0.04	0.00	0.03	0.01	0.01

Table 3: Disease localization accuracy are evaluated with a classification threshold of 0.5. "*" denotes additional bounding box supervision. Abbreviations follow Table 2.

general, our method (SGL) performs better than straightforward max-pooling for all numbers of instances. In terms of IoU, our method outperforms other methods for a wide range of numbers of positive instances per bag. For larger bag sizes, SGL achieves similar AUC performance to BIL, which was trained using mean-pooling, while outperforming it in terms of IoU for all numbers of positive instances per bag. The self-guidance of our method addresses the limitations of max-pooling and improves accuracy for any bag size or number of positive instances.

ChestX-Ray 14: Multi-Label Pathology Classification. Table 2 shows the AUC scores for different disease classes, comparing the results of our loss function with those of other classification approaches: Wang et al. [292], Li et al. [159], and Liu et al. [167]. All of these methods use a ResNet-50 network as the backbone and the latter two also make use of bounding box supervision and architectural adaptations. Our approach outperforms the baseline ResNet-50 of Wang et al. in all categories, and achieves better classification performance than the other methods in 9 out of 14 classes in total. We also achieve a higher mean performance than the other methods, which use additional bounding box annotations and architectural modifications such as additional networks [167] or additional convolutional layers [159, 167].

ChestX-Ray 14: Pathology Localization. To evaluate the localization accuracy of different methods, we measure the accuracy over a range of intersection-over-union (IoU) thresholds. For our method, we upsample the prediction maps using nearest-neighbor interpolation, and then construct bounding boxes around the connected components of the maximum prediction. The results are shown in Table 3 for IoU thresholds of $T_{IoU} \in 0.1, 0.3, 0.5, 0.7$. In Figure 12, we also provide qualitative examples for each pathology, where the expert annotation is shown in green and the predicted bounding box is in orange.

Our method performs well across all pathologies at an IoU threshold of $T_{IoU} = 0.1$, and generally outperforms the baseline method of Wang et al. [292]. However, for higher thresholds, our model performs worse than the more specialized approaches of Li et al. [159] and Liu et al. [167]. We attribute this suboptimal quantitative performance to two factors: the low spatial resolution of our output, which can make it difficult to achieve high IoU scores, especially for naturally small classes like nodules; and the overall coarse and partially faulty annotation for example in the case of infiltrates, which affect the lung area but may also include the cardiac area, reducing the IoU. Despite these limitations, our proposed method still produces more precise predictions

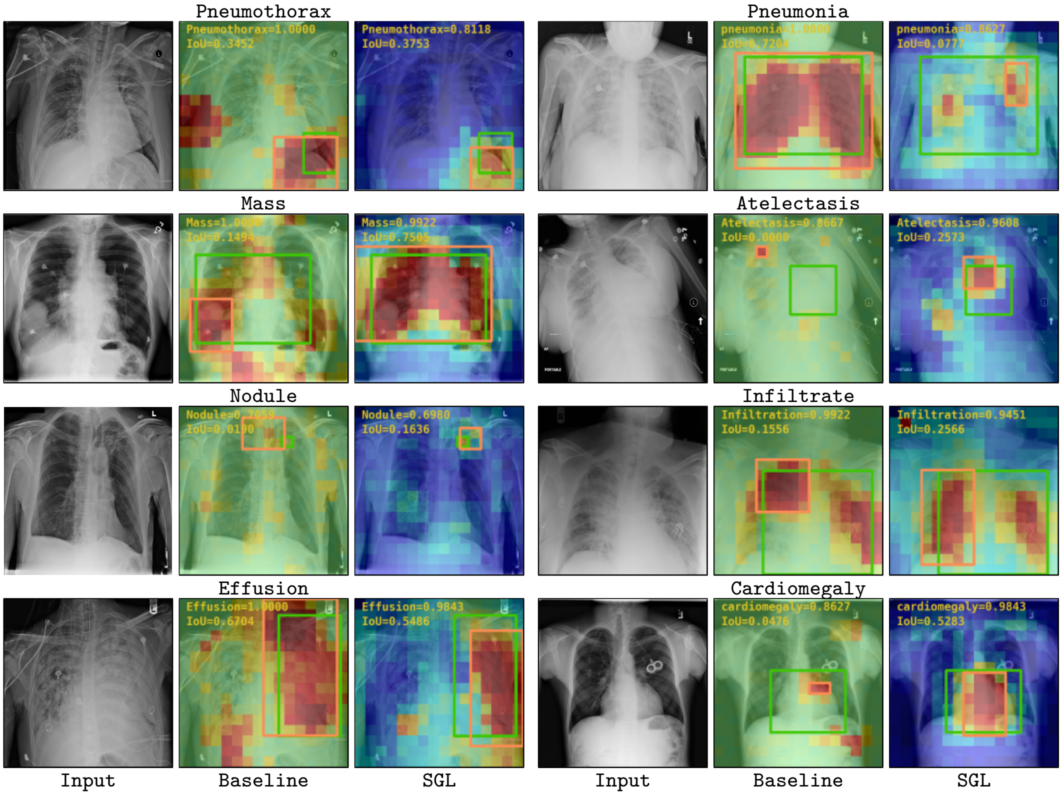


Figure 12: We compare the patch-wise predictions between a mean-pooling trained baseline to our proposed method for different diseases. The value ranges from 0 (blue) to 1 (red). We show prediction boxes (orange) around the connected component of the maximum prediction and Ground-Truth bounding boxes (green).

compared to the baseline model and is able to more clearly distinguish between healthy and abnormal tissue. These results suggest that our loss function leads to refined predictions.

3.2 BREAKING WITH FIXED SET PATHOLOGY RECOGNITION

The following section is based on our publication in *MICCAI 2022* [246].

While our Self-Guiding Loss improves the identification of findings, we still rely on the diversity and quality of the label extraction process. When we now want to be able to identify pathologies that are not part of the training set, the label extraction would have to be adapted accordingly. Adding new disease classes requires a significant amount of time and effort to annotate the data, and subsequently retrain the system. To avoid this hurdle inherent in label supervision, several works train models directly using text supervision [330], but to access the information encoded within the model weights to properly perform finding inference, models still need to finetuned on pre-defined labels.

Recent advances in contrastive language-image pre-training have shown that it is possible to achieve object recognition without relying on predetermined class definitions [116, 197, 205, 294, 330]. These models learn to represent images and textual descriptions in a joint feature space and use text prompts to transform the recognition task into a matching task between text and image embeddings. However, applying them to radiological reports is non-trivial because reports have a different structure compared to natural image captions. Radiological reports consist of multiple distinct sentences that together describe all relevant information, whereas natural image captions are typically defined by a single object.

In the following, we present a method for report-guided training of vision-language models and show how we can query these models to perform open set pathology recognition. This enables not only a more flexible training process but also more dynamic identification and localization of different pathologies. We, thereby, investigate the factors affecting the performance of multi-modal training and inference.

3.2.1 *Global-Local Contrastive Learning*

We display the training process of our vision-language model for disease recognition in Figure 13. To handle the complexity of medical reports, we split the representations into sentence-level and report-level ones, which are derived from a shared visual and language encoder. Our model utilizes the shared feature space to account to predict the presence or absence of a pathology by providing respective prompts.

3.2.1.1 *Model Overview*

Contrastive language and image pretraining (CLIP) has demonstrated effectiveness in object recognition in natural images [205] by learning from image-caption pairs where a caption often corresponds to a short sentence. However, medical reports typically consist of multiple sentences describing different aspects of the image (i.e. state of the pleural space, size of the heart, ...). To consider not only the global report context but also the individual subsets, our model captures text information on a local sentence- (s) and global report-level (R) level. As such, we separate our model into an image encoder ξ, which embed an image I as $z_I = \xi(I) = \phi_\xi(\psi(I))$, and a text encoder θ, which encodes sentences $z_s = \theta(s)$ and reports $z_R = \phi_\theta(\theta(R))$.

During training, for a given report R, we capture the local context by splitting the report into its individual sentences $R = \{s_1, \dots, s_n\}$ and extracting sentence-level embeddings $z_{s_i} = \theta(s_i), s_i \in R$. To create global embeddings that contain the full information of the report, we use attention pooling ϕ to aggregate the sentence-level embeddings, which corresponds to a single layer QKV-attention block with the Query conditioned on the average of the sentence embeddings [280]: $z_R = \theta(R) = \phi_\theta([z_{s_1}, \dots, z_{s_{|R|}}])$. To incorporate the embeddings z_I, z_s, and z_R into shared multi-modal representations, we project the sentences and reports through linear transformations pr^S and pr^R into separate feature spaces. We also project the image embedding z_I twice, once into a global representation $pr^G(z_I)$ to align with $pr^R(z_R)$ and once into a local representation $pr^L(z_I)$ for alignment with $pr^S(z_s)$.

For clarity in subsequent formulas, for a dataset of image-report pairs $(I_i, R_i) \in \{(I_1, R_1), \dots, (I_N, R_N)\}$ we omit inputs for our projections, *e.g.* the global projection of an image I_i as pr_i^G instead of $pr_i^G(\theta(I_i))$ or for the projection of the k^{th} sentence from report R_i, we write pr_{ik}^S.

3.2.1.2 *Training Objectives*

Local Contrast: While radiological reports provide information about a patient's health, not every sentence is directly related to specific findings. Some sentences may mention clinical procedures or required follow-up examinations. However, we can assume that all clinically relevant information is present in a subset of sentences within the report due to doctors' obligations to document their findings. This property parallels the previously mentioned MIL assumption.

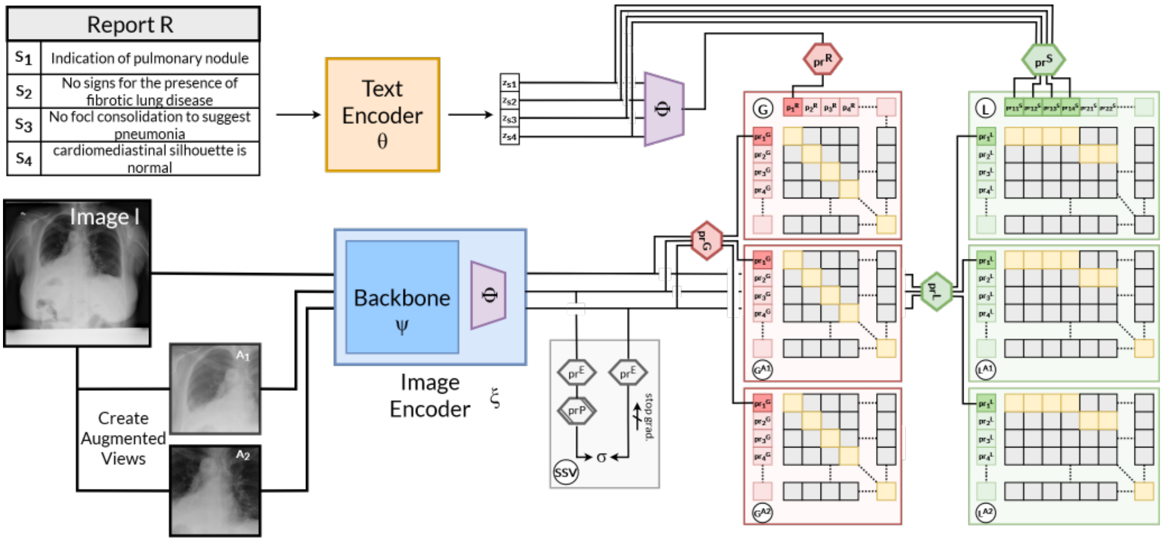

Report R
S1 Indication of pulmonary nodule
S2 No signs for the presence of fibrotic lung disease
S3 No focl consolidation to suggest pneumonia
S4 cardiomediastinal silhouette is normal
Text Encoder θ
Image I
Backbone ψ
Φ
Image Encoder ξ
Create Augmented Views
A1
A2
prR
prS
prG
prL
prE
prP
σ
stop grad.
SSV
G
L
GA1
GA2
LA1
LA2

Figure 13: Illustration of our proposed report-guided training scheme. A medical report is split into sentences and processed by a text encoder. The image is passed through an image encoder with augmented views. We apply self-supervised SimSiam-losses (SSV), sentence-level local contrast between sentences and an associated image embedding (L), as well as report-level global contrast between report embeddings and image (G). Furthermore, we apply the losses building on local and global contrast on the augmented views used in SSV.

Given this, it might seem natural to utilize MILNCE [180] as the MIL-based objective for integrating sentences in training. This formulation minimizes the distances between matching sets of two feature modalities, leading to a natural symmetry. As we face a matching between a single (the image) and a set of representations (sentences) for our local setting, we redesign it by splitting its symmetry:

$$\mathscr{L}_L(I_i,R_i) = -\log \frac{\sum_{k=1}^{n} \exp(\sigma(pr_i^L, pr_{ik}^S)/\tau_L)}{\sum_{j=1}^{N}\sum_{m=1}^{n} \exp(\sigma(pr_i^L, pr_{jm}^S)/\tau_L)} - \sum_{k=1}^{n} log \frac{\exp(\sigma(pr_i^L, pr_{ik}^S)/\tau_L)}{\sum_{j=1}^{N} \exp(\sigma(pr_j^L, pr_{ik}^S)/\tau_L)}, \tag{10}$$

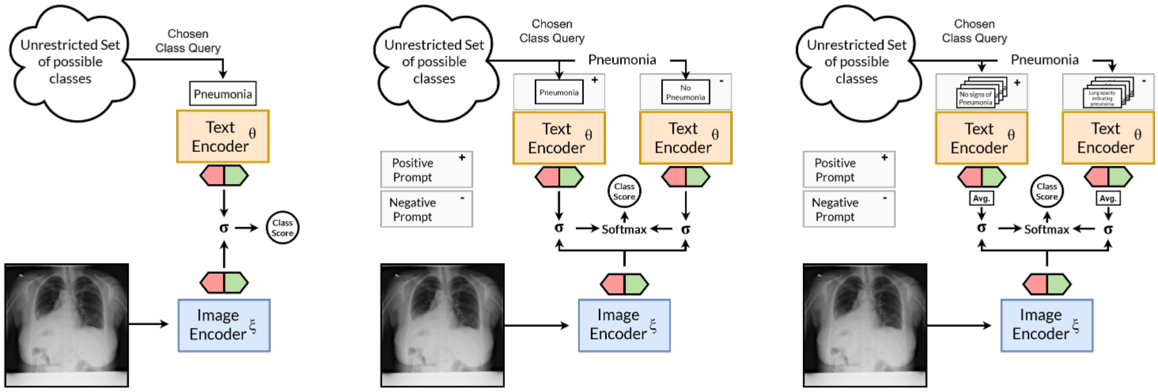
with τ_L being a learned parameter and $\sigma(\cdot,\cdot)$ denoting the cosine similarity.

Global Contrast: For each batch in our training, we assume that each image-report pair is unique and use the following objective which leverages attention-fused reports:

$$\mathscr{L}_G(I_i,R_i) = -\log \frac{\exp(\sigma(pr_i^G, pr_i^R)/\tau_G)}{\sum_{j=1}^{N} \exp(\sigma(pr_i^G, pr_j^R)/\tau_G)} - log \frac{\exp(\sigma(pr_i^G, pr_i^R)/\tau_G)}{\sum_{j=1}^{N} \exp(\sigma(pr_j^G, pr_i^R)/\tau_G)} \tag{11}$$

Self-Supervision: CLIP has been shown to be a data-intensive algorithm [197, 205]. Several recent methods have combined self-supervision signals with the CLIP objective [156, 184]. As, in the medical domain, we often have access to significantly smaller datasets compared to the natural image domain, we follow the approach of Li *et al.* [156] and incorporate SimSiam [37] to address this issue. To do so, we generate two augmented versions of the input image $A_1(I)$ and $A_2(I)$ and add a three-layer encoder-head pr^E and a two-layer prediction-head pr^P on top of the visual backbone ψ to enforce similarity between the two views:

$$\mathscr{L}_S(A_1(I), A_2(I)) = -\sigma(pr^P_{A_1(I)}, \text{detach}^1(pr^E_{A_2(I)})) - \sigma(pr^P_{A_2(I)}, \text{detach}(pr^E_{A_1(I)})) \tag{12}$$



(a) *Cosine* (b) *Basic* (c) *Detailed*

Figure 14: Illustration of inference schemes for open set multi-label classification. (a) *Cosine* computes just the similarity between image features and the class prompt. (b) *Basic* uses prompts for both presence and absence of the class and normalizes the similarity to either. (c) *Detailed* uses a set of positive and negative prompts for the class.

The augmentation follow SimSiam [32] and, thus, incorporate random resize and cropping, color jittering and blurring. We omit random flipping as orientation is a major denotion in medical reports (i.e. right lung vs. left lung). In addition, we use the augmented images from the self-supervised objective and apply our local and global contrastive objectives in the same manner.

$$\mathscr{L}_M(I_i, R_i) = \mathscr{L}_G(A_1(I_i), R_i) + \mathscr{L}_L(A_1(I_i), R_i) + \mathscr{L}_G(A_2(I_i), R_i) + \mathscr{L}_L(A_2(I_i), R_i) \quad (13)$$

The final objective for report-guided training is as follows:

$$\mathscr{L}(I_i, R_i) = \lambda_1 * \mathscr{L}_L(I_i, R_i) + \lambda_2 * \mathscr{L}_G(I_i, R_i) + \lambda_3 * \mathscr{L}_S(I_i, R_i) + \lambda_4 * \mathscr{L}_M(I_i, R_i) \quad (14)$$

with $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_4 = 0.25$. We set these values intuitively based on the inverse of the number of present contrastive loss terms in each loss formulation similar to Radford et al.'s CLIP, or Chen and He's SimSiam.

3.2.2 *Model Inference*

In traditional fixed set classification models, identifying the class of a given image is a straightforward process: the image is passed through a network and the class with the highest confidence is chosen as the prediction. However, when the model architecture or training procedure does not allow for a classification layer, vision-language often utilize zero-shot-like inference [106, 205, 294], which involves searching for the nearest neighbor search in a semantic space based on text embeddings [71] in a multi-class setting. As such, one pre-extracts fitting text embeddings representing each class and for a given image finds the closest matching query. This setting has been also been applied for zero-shot pathology recognition [106], where an image had to be assigned to one out of a set of five diseases. However, this method becomes problematic when trying to detect multiple diseases, as pathologies are not mutually exclusive and modeling each possible combination as individual classes is infeasible. Similarly, modeling co-occurrence as individual classes is also infeasible due to the exponentially rising number of possible class combinations. To address this issue, we propose a multi-label classification method that queries an image with

---

[1] "detach" stops gradient computation for a variable in backpropagation.

class-related prompts and utilizes their similarity scores as prediction probabilities for the respective disease class. We distinguish between three potential settings and illustrate these in Fig. 14.

The first setting, *Cosine*, uses the class c directly as query q. We compute sentence-level and report-level embeddings of the text query and compute their cosine similarity to the global and local image embedding. We, then, use the average of these scores as class prediction score $P(c,I)$.

$$P_{Cosine}(c,I) = \frac{\sigma(pr^L, pr^S_q)/\tau}{2} + \frac{\sigma(pr^G, pr^R_q)/\tau}{2} \tag{15}$$

We recognize that the similarity of a single query for the class presence can be ambiguous due to the proximity of opposing semantics (e.g. negations) in the feature space, leading to potential confusion between the presence and absence of a class. To overcome this, we perform inference over two sets of queries for each class c - one indicating the presence of the class (q^p_c) and the other indicating its absence (q^n_c). For example, the query "opacities consistent with pneumonia" would indicate the presence of pneumonia, while "the lungs are clear" would indicate its absence.

In our *Basic* setting, we consider prompts such as q^p_c = "{class}" and q^n_c = "no {class}". We calculate the cosine similarity between the image and both queries, and the final prediction for class c is defined as the ratio of the exponential of the similarity score for the presence query to the sum of the exponentials of the similarity scores for both the presence and absence queries, with the learned scaling factor τ depending on the used projection.

$$\begin{aligned} P_{Basic}(c,I) = & \frac{\exp(\sigma(pr^L, pr^S_{q^p})/\tau)}{\exp(\sigma(pr^L, pr^S_{q^n})/\tau) + \exp(\sigma(pr^L, pr^S_{q^p})/\tau)} \\ & + \frac{\exp(\sigma(pr^G, pr^R_{q^p})/\tau)}{\exp(\sigma(pr^G, pr^R_{q^n})/\tau) + \exp(\sigma(pr^G, pr^R_{q^p})/\tau)} \end{aligned} \tag{16}$$

To improve the effectiveness of language-vision models, it is important to align the downstream task to the training [205]. In zero-shot classification, the features of the class name are typically extracted through a word embedding model [291], which can lack important context about the class. Similarly, while we narrow down our observed space in our *Basic* setting, we found that a more detailed prompt design can improve performance. To address this issue, we model a set of positive and negative prompts specific to pathologies to enhance the matching process between visual and textual projections.

We consider a set of prompts following the templates Q^p = '{adverb}{indication_verb} {effect}* {location}* {class_synonym}' and Q^n = '{adverb} {indication_verb} {absence} {class_synonym}'. We utilize all combinations of these categories to create a variety of prompts. During inference, the features of all queries in the same set are averaged. We term this setting *Detailed*:

$$\begin{aligned} P_{Detailed}(c,I) = & \frac{\exp(\sigma(pr^L, mean(pr^S_{Q^p}))/\tau)}{\exp(\sigma(pr^L, mean(pr^S_{Q^n}))/\tau) + \exp(\sigma(pr^L, mean(pr^S_{Q^p}))/\tau)} + \\ & \frac{\exp(\sigma(pr^G, mean(pr^R_{Q^p}))/\tau)}{\exp(\sigma(pr^G, mean(pr^R_{Q^n}))/\tau) + \exp(\sigma(pr^G, mean(pr^R_{Q^p}))/\tau)}, \end{aligned} \tag{17}$$

with $pr_Q = \bigcup_{q \in Q} pr_q$ denoting the set of projections on every query within the templates Q.

3.2.3 *Localization with Vision-Language Models*

As we do not have an explicit classification layer, we cannot directly transfer Class Activation Mappings to our method. Instead, during inference, we compute the final local/global representation by applying a linear layer to the output of the QKV-attention pooling and compute the class score based on class query embeddings. To now compute a saliency map, we follow Zhou *et al.* [335] and split up the attention pooling by discarding the Query and Key linear layers. We apply the weights of the Value layer and subsequent projection individually to all elements of our feature map to result in a map in the local/global feature space. Thus, we compute the class score by applying the different inference settings described in Sec. 3.2.2 for each tile of our feature map to acquire our final saliency map. We illustrate this process in Fig. 15.

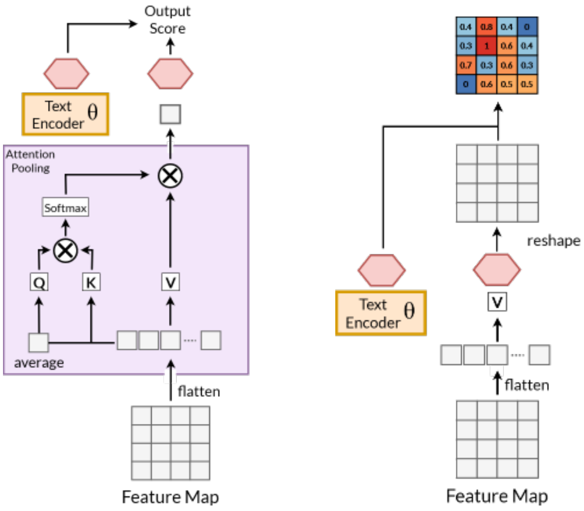


Figure 15: Illustration of the QKV-attention pooling (left) and the restructuring for localization (right) [335].

3.2.4 *Prompt-based Dataset Extension*

While medical reports are more commonly used in practice, most large-scale datasets are only publicly available with fixed sets of labels. By reversing our prompt engineering process, we can investigate the effect of additional data in the training of our method. To study the impact of additional data on the training of our method, we generate synthetic reports for the PadChest and ChestX-Ray14 datasets based on their class labels. This allows us to sample sentences indicating the presence or absence of a class and create over 200k additional image-report pairs.

3.2.5 *Experimental Setup*

3.2.5.1 *Datasets*

MIMIC-CXR: It contains 377,110 chest X-rays taken from 65179 patients with 14 disease labels and 227,835 reports. We use the splits provided by [121]. Unless further specified all models were trained on this dataset.
CheXpert: It contains 224,316 chest X-rays taken from 65,240 patients with 14 disease labels. The labels are shared with MIMIC-CXR. We only consider the validation split provided by [110].
ChestX-ray14: It contains 112,120 frontal-view chest X-rays taken from 30,805 patients with 14 disease labels. We use the splits provided by [245].
PadChest: It consists 160k chest X-rays of 67k patients with 174 findings. We use this dataset solely for training purposes in the pseudo-label scheme.

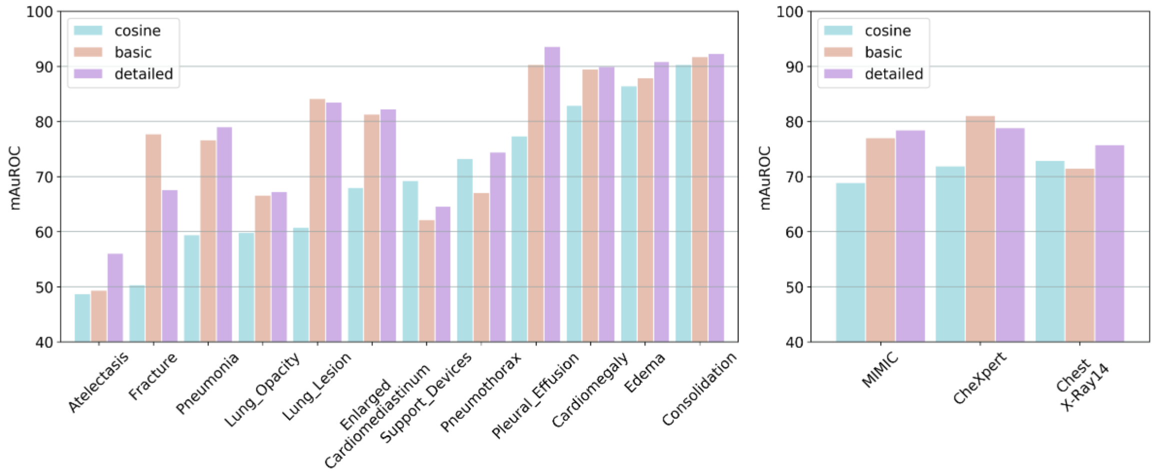


Figure 16: Performance changes based on differences in prompt generation. Class wise performance on the left. Mean performance to the right. Models trained on MIMIC.

3.2.5.2 *Evaluation Setup*

We evaluate the multi-label classification ability of all networks via the AUROC and show the performance over MIMIC-CXR, CheXpert and ChestX-Ray14. For all experiments except Table **??** we consider validation performance. Labels with value -2 and -1 are ignored for the calculation of the metric as their state is not certain. For all ablations, we use the "basic"-prompting scheme, while for further experiments the "detailed"-scheme is used.

3.2.5.3 *Implementation Details:*

For all experiments we use the same ResNet50 and Transformer as Redford et al. [205] as backbones. We optimize with AdamW [170], a learning rate of 0.0001 with a cosine schedule and 10000 warm up steps. We trained classification models with a learning rate of 0.0005 as this has shown slightly better performance. During training, we resize the images to the inference size of 320 × 320 and randomly crop by 288x288. For sentence-based models, we use a context length of 72 tokens, while for others we use 144 tokens. Each sentence model uses up to 10 sentences.

3.2.6 *Results*

Ablation - Effect of Heads: We examine the effect of using both prediction heads during inference. To begin, we present the performance of each head individually, then compare different fusion approaches in Table 4. For feature fusion, we combine the local and global features of the same modality. For score fusion, we calculate scores as described and aggregate the class predictions

Inf.	MIMIC	CheXpert	CXR14	Avg.
Local	77.81	78.09	71.72	75.87
Global	76.24	80.42	71.00	75.88
Max	76.85	71.29	78.22	75.45
Cat	77.29	80.30	71.72	76.43
Mean	77.06	81.08	71.50	76.54

Table 4: We evaluate the impact of the scores from the model heads and their combinations.

Parts	MIMIC	CheXpert	CXR14	Avg.
$\mathcal{L}_G$	75.47	77.24	69.22	73.97
$\mathcal{L}_G+\mathcal{L}_L$	76.20	82.24	69.26	75.90
$\mathcal{L}_G+\mathcal{L}_L+\mathcal{L}_S$	76.10	76.08	74.24	75.47
$\mathcal{L}_G+\mathcal{L}_L+\mathcal{L}_M$	77.03	77.36	71.72	75.37
Ours	77.06	81.08	71.50	76.54

Table 5: Ablation on the impact of the different loss parts described in Section 3.2.1.2.

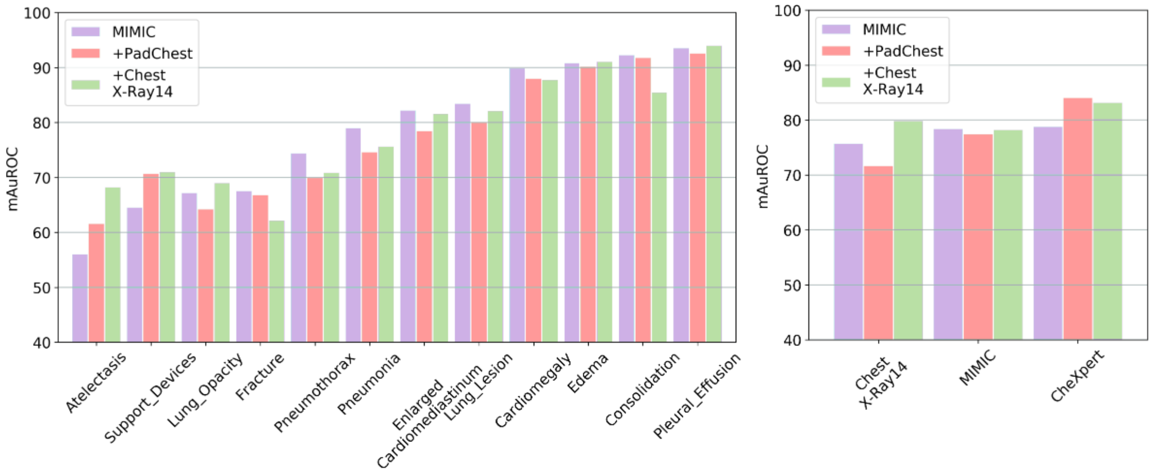


Figure 17: Contributions of data scaling for chest radiograph dataset. Performance change of adding additional chest X-ray datasets with prompt-based captions.

based on either the maximum or average score. We find that the global and local head show almost the same performance for our method. While max-score fusion results in a 0.4% drop in across-dataset performance, mean-score fusion shows a 0.6% improvement.

Ablation - Effect of Losses: Table 5 illustrates the influence of various objective functions on the model. As we can see, incorporating local contrast enhances the model's performance by 2%. However, using both the self-supervised and mirrored objectives results in a 0.45% decrease in performance. Interestingly, the self-supervised loss alone significantly improves performance on ChestX-Ray14 by over 2%. Combining both objectives yields a 0.6% increase in performance across all datasets.

Multi-label Inference and Prompt Engineering: We show the impact of our proposed inference schemes in Figure 16. We see that performance improves considerably when switching from a simple similarity score to our *Basic* scheme. Especially classes such as fractures were unable to be categorized just using cosine-similarity, whereas when positive and negative prompts these can be better identified. When the detailed prompt the mean performance further improves, however, we can see that the performance worsens for some classes, indicating that the correct choice of prompts is also of importance.

Data size Impact: As depicted in Figure 17, the use of prompt-based reports during training significantly enhances the validation performance of ChestX-Ray14 through the incorporation of

Method	MIMIC-CXR (in-domain) val	MIMIC-CXR (in-domain) test	CheXpert (out-of-domain) val	ChestX-Ray14 (out-of-domain) val	ChestX-Ray14 (out-of-domain) test
Label-Supervised	77.26	77.42	78.90	79.70	76.47
CLIP	73.23	70.25	75.85	68.03	63.34
SLIP	72.45	72.44	78.49	71.45	67.55
$MILNCE_{local}$	69.30	69.18	74.98	67.56	63.06
LoCo	77.03	78.15	81.71	71.92	68.14
GloCo	75.47	76.58	77.24	69.22	65.86
Ours	**78.46**	79.40	78.86	75.77	71.23
Ours*	78.30	**80.40**	**83.24**	**79.90**	**78.33**

Table 6: Classification performance on MIMIC, CheXpert and Chest-XRay14. * indicates that the model was trained with additional PadChest and ChestX-Ray14 data.

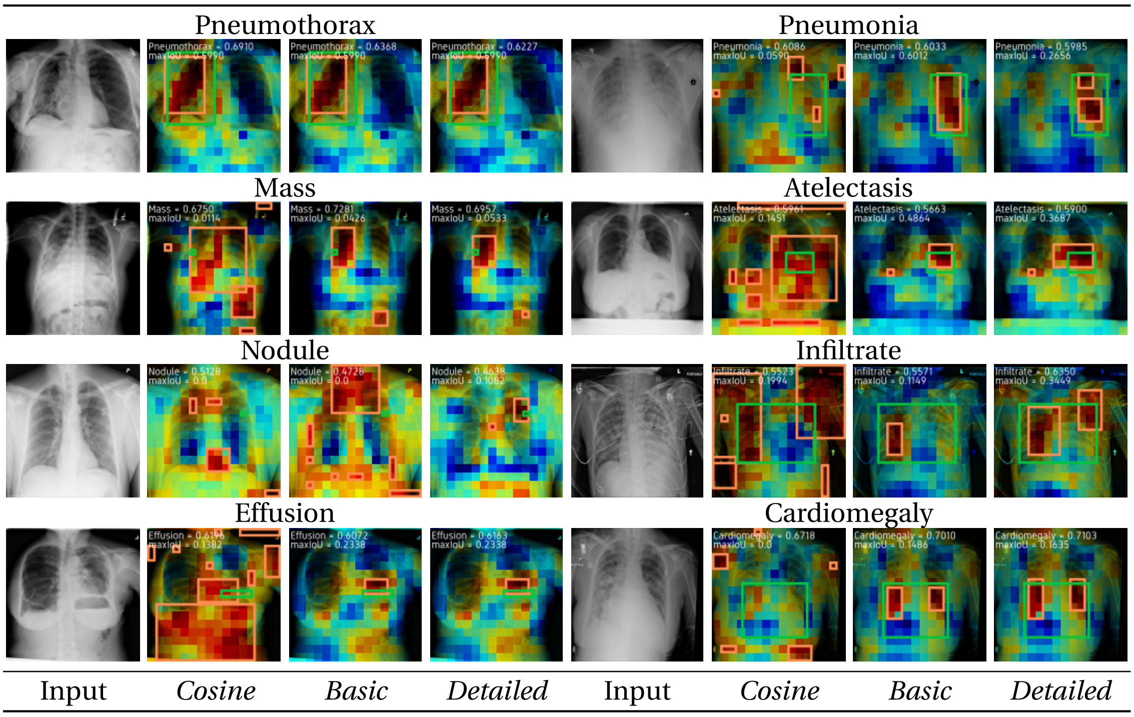


Figure 18: We compare the patch-wise predictions of the considered inference schemes for different diseases. The value ranges from 0 (blue) to 1 (red). We show prediction boxes (orange) around the connected component of the prediction over a 0.8 threshold and Ground-Truth bounding boxes (green).

artificial training data. While some classes may experience a decline in performance, the overall improvement is apparent when artificial additional data is added.

Comparison with Other Approaches: We pit our method approach against a vision network trained with label supervision on its corresponding dataset, as well as other methods trained using the MIMIC-CXR dataset. CLIP uses contrastive losses between the entire report and the image [205]. SLIP expands on CLIP and incorporates self-supervision through a SIMCLR-like objective [184]. $MILNCE_{local}$, which is trained with the MILNCE objective alone [180]. We also tested our method using both the local and global objectives, referred to as LoCo and GloCo. Using the "detailed" prompt scheme, we evaluated the results, which are shown in Table 6.

Our local contrastive loss consistently outperformed the MILNCE version across all datasets. Our complete proposed method outperformed the CLIP and SLIP baselines, achieving similar performance to the supervised ResNet for domains similar to MIMIC. However, it underperformed on the ChestX-Ray14 dataset. When we added the additional artificial report datasets of PadChest and ChestX-Ray14, our method managed to beat label-supervised performance across all datasets.

3.2.6.1 *Saliency Mappings*

We show the patch-wise activations of our network for different findings in Fig. 15. We highlight the differences between the our inference schemes for cases of ChestX-Ray14. We can see that for nearly all classes, the *Cosine* inferences are more spread out and do not allow the network to focus on essential patterns. In contrast, the *Basic* inference allow the network to focus on more specific regions, which get expanded on by *Detailed*.

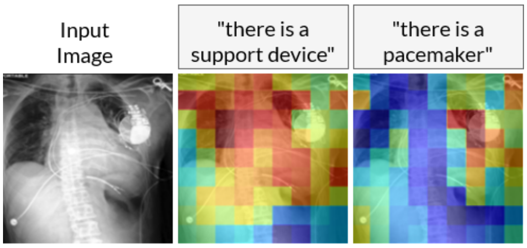


Figure 19: Illustration of the QKV-attention pooling (left) and the restructuring for localization (right) [335].

Effect of the prompt: The choice of prompt has an immense effect on not only the classification performance but also the localization. In Fig. 19, we show the saliency maps for two prompts corresponding to the class 'support devices'. For the prompt *There is a support device*, the network is unable to pick up any specific corresponding pattern in the image. For a specific support device, however, this is turns out differently. By prompting for a pacemaker instead, the network activates on the main body of the pacemaker, albeit not on the leads. This can be ascribed to the fact that support devices are typically referred to by their actual names instead of their supercategory. As such, the supercategory appears less in the medical reports, which hinders the training process to build proper text-image associations.

3.3 CHAPTER CONCLUSION

To generate holistic radiological reports, we require to not only identify particular diseases, but also recognize where they occur. In this chapter, we propose a novel loss formulation in which one gathers auxiliary supervision from the network's predictions to provide instance-level supervision. This enables the model to find a trade-off between classification and localization performance more efficiently without any additional supervision such as bounding boxes or pixel-wise annotations. In comparison to existing MIL-based loss functions, we do not rely on initialization and still provide pixel-wise supervision driving the network. Due to the design of this loss, it can support any MIL-setting such as patch-based pathology diagnosis. We demonstrate our method on two MIL-based datasets as well as the challenging NIH ChestX-Ray14 dataset. We display promising classification and localization performance qualitatively and quantitatively.

However, the need for prior label extraction restricts the capabilities of a network, as we cannot identify findings that were not specifically selected. Expanding a dataset to fit new needs is an expensive process. To combat this issue, we subsequently proposed an approach to make networks less reliant to label supervision through contrastive language-image pre-training on report level. By directly utilizing the medical report during training, we can teach the network to identify any finding that occurs in clinical practice. As the information of these models is not directly accessible, we introduced a novel way of constructing inference. By querying for the presence and absence of a class, we can not only precisely identify but also localize any finding in a manner that is competitive with label-supervised models across different datasets.

While we believe that our work opens more possible uses for report-supervised models, we also have to note potential pitfalls. The quality of the resulting feature space is heavily dependent on the training data. In the natural image domain, many methods train on millions to billions of image-text or image-label pairs, whereas we are vastly limited in dataset size. As such, the dataset cannot model the entirety of the radiological information and results in a heavy long tail distribution with certain findings only being present a handful of times in contrast to a vast majority of healthy patients. In addition, similar to other CLIP-based methods the classification performance of the models heavily depends on the similarity of the chosen inference prompts to the natural text occurrence of the respective classes. We noticed that the proposed detailed prompting schemes can perform worse than the basic scheme for certain classes if the text is not aligned to the target domain. Similarly, the way “support devices” are occurring in reports does not align with the prompting scheme as it describes a super class of a variety of appliances and will not appear directly as term in medical reports.

In summary, we present promising methods for the identification and localization of radiological findings with no need for manual supervision as we can make efficient use of available medical reports. We outline our contributions as:

Contribution 1: We provide a novel loss function that applies prediction maps for self-guidance to achieve better classification and localization performance without the necessity to expand a given fully convolutional network architecture.

Contribution 2: We develop a training procedure for vision-language models on medical reports that considers individual sentences and the entire report to better encode findings. This way, we can make better use of the longer text structure of medical reports compared to other vision-language models.

Contribution 3: We show, how these models, for the first time, can be used for open set classification and localization through the use of contrasting inference setting, making use of presence and absence prompts. We, hereby, achieve performance that is comparable with label-supervised methods.

Our methods can note the occurrence of various findings and localize them in the image, however, this does not capture the all relevant information in a human-readable format. To properly convey present abnormalities, they have to be set into the anatomical context of the patient. Anatomical structures cannot easily be learned implicitly as they occur for every patient and thus cannot be inferred from absence or presence. In the following chapter, we take a look at the intricacies of anatomy for segmentation models and show how these models can be learned with few annotations.

4

LEARNING FROM LITTLE AMOUNTS OF ANNOTATIONS

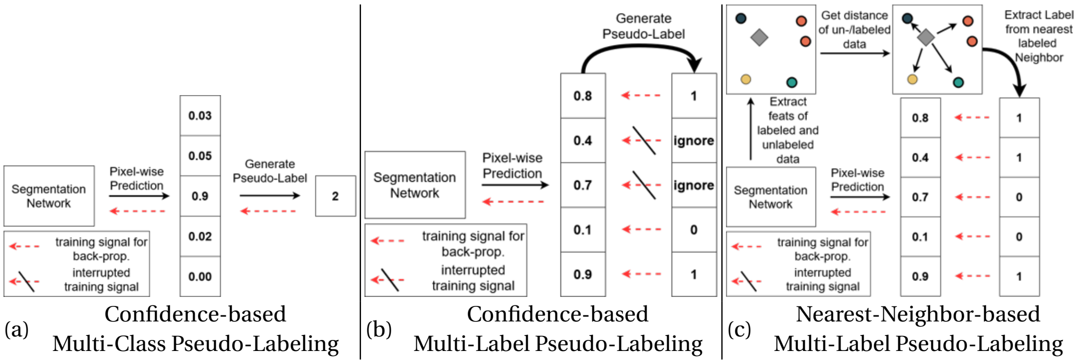
Accurate and comprehensive reporting of medical findings relies on precise recognition of anatomical landmarks. However, collecting these annotations can be challenging, especially when certain positions have multiple interleaving semantics. This leaves us with a limited number of labeled images and a vast amount of unlabeled data. In this chapter, we investigate the intricacies of semi-supervised learning with focus on chest x-rays to make use of both labeled and unlabeled data. Here, the overlap of visible structures turns the task from a multi-class to a multi-label segmentation problem. To overcome the issues of the medical domain, we introduce reference-guided pseudo-labeling, where we perform nearest-neighbor comparisons between unlabeled data and a set of labeled reference images to assign pseudo-labels, which are weighted based on their class entropy. We show through our experiments that this approach excels in the multi-label and also is applicable in the standard multi-class segmentation setting.

The following sections are built on our *AAAI 2022* publication [242].

4.1 UNDERSTANDING CXRS: MULTI-LABEL SEGMENTATION

Through the presence and absence of image features, we, sooner or later, can often learn which features correlate with different findings by resorting to large amounts of weakly annotated data such as image-text pairs. However, to properly convey these findings, we also have to identify associated anatomical landmarks. As these landmarks are visible throughout all samples as every human shares the same underlying anatomy, learning through a simple occurrence signal becomes difficult. Thus, in order to properly teach machines anatomical understanding, we require explicit dense annotations.

Acquiring detailed annotations for tasks such as semantic segmentation is a challenging and time-consuming process [46], particularly in the medical field due to the specialized knowledge needed [179]. Additionally, the demands of clinical practice often make it difficult for doctors to provide a large amount of detailed annotations. Therefore, it is desirable to achieve accurate semantic segmentation with minimal annotated data. In order to improve the performance of semantic segmentation with limited annotated data, one solution is semi-supervised learning (SSL). SSL extrapolates from small amounts of labeled data through different objectives involving also a larger amount of unannotated data during training. Some approaches include student-teacher frameworks [33, 79, 198, 307], consistency regularization [191, 212, 260], and pseudo-labels [111, 147, 223]. Pseudo-label methods typically use network predictions for unlabeled data either online as targets for training during the same iteration, often in conjunction with generating predictions from perturbed versions of an input image [14, 15, 260], or they are saved for later re-training of the model [308]. These techniques have been investigated particularly in the medical field to segment certain organ structures or pathological patterns for modalities such as CT due to the annotation difficulty [297].


Segmentation Network
Pixel-wise Prediction
0.03
0.05
0.9
0.02
0.00
Generate Pseudo-Label
2
training signal for back-prop.
interrupted training signal
Generate Pseudo-Label
0.8
0.4
0.7
0.1
0.9
1
ignore
ignore
0
1
Get distance of un-/labeled data
Extract Label from nearest labeled Neighbor
Extract feats of labeled and unlabeled data
0.8
0.4
0.7
0.1
0.9
1
1
0
0
1

(a) Confidence-based Multi-Class Pseudo-Labeling (b) Confidence-based Multi-Label Pseudo-Labeling (c) Nearest-Neighbor-based Multi-Label Pseudo-Labeling

Figure 20: Comparison of pseudo-labeling by examples. On the left, we display confidence-based pseudo-labeling for multi-class settings. In the center, we show confidence-based pseudo-labeling for multi-label settings. On the right, we show our proposed nearest-neighbor-based pseudo-labeling for multi-label segmentation. We note that an imbalance in confidence can lead to several ignored classes.

In contrast to related work, we, when considering chest X-Rays, cannot assume that one pixel belongs to a singular class as it would be the case in a CT volume. As the X-Ray penetrates the body it delivers a projection of all structures it passes through based on their attenuation [177]. When now trying to assign a semantic to a certain pixel, we have to consider multiple classes simultaneously. This leads us to a multi-label problem as we now would have to predict several classes at the same time instead of pitting them against each other in a multi-class setting.

When we consider standard confidence-based multi-class pseudo-labeling, we apply a threshold to the softmax predictions of a network for an unlabeled sample. If the prediction passes the threshold, we use its argmax as a target. Otherwise, we ignore the sample. To transfer to the multi-label setting, we get an independent sigmoid prediction for each class. By applying thresholding to the network predictions, we now potentially get pseudo-labels as well as ignore labels within the same sample. Thus, we backpropagate information for confident classes much more often than for potentially associated less confident ones. This enforces the predictions' pre-existing biases while neglecting valid co-occurring samples, which might get a low prediction due to being rarer overall, which can lead to the issue of confirmation bias [223]. We display this problem in the center of Fig. 20.

In order to address the issue of confirmation bias in multi-label pseudo-labeling, we propose a new approach in which we compare embeddings between predictions of unlabeled and labeled reference images to instill supervision. Instead of directly forming the supervision from the class predictions, we extract the label vector from the nearest-neighbors of an unlabeled pixel among the pixels in a labeled small reference set. This method allows us to regulate the bias towards large classes and bypass the problems of using direct class predictions as supervision.

We show the effectiveness of our method different scenarios with small amounts of labeled data. Our method shines when segmenting overlapping labels with scarce data as we reach fully supervised performance from just six labeled images on the JSRT [257] dataset. In comparison to other pseudo-labeling strategies, our method also enables the segmentation of high-class scenarios as we display for a small set of fine-grained annotations for CXR anatomy. We, further, show its applicability for multi-class segmentation on the RETOUCH [20] dataset, achieving competitive results especially for minimal amounts of samples.

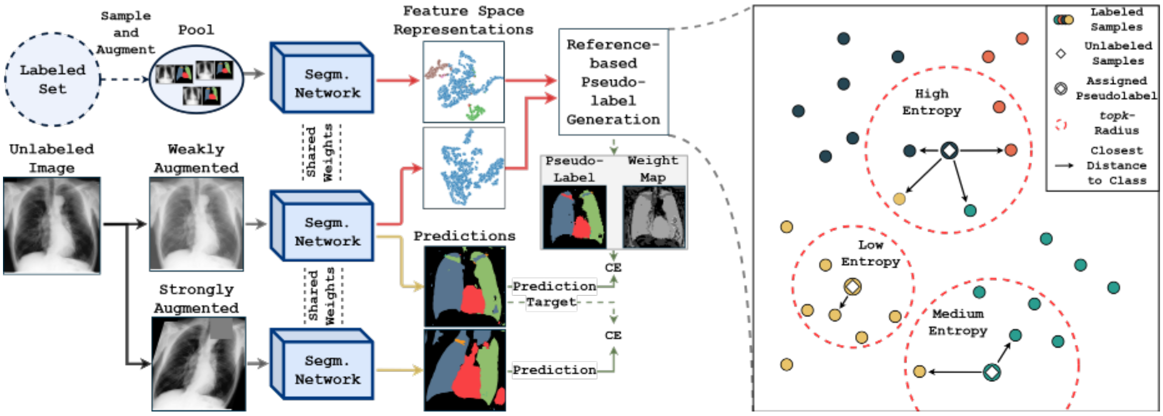


Figure 21: Overview of the proposed training step for unlabeled images. For each unlabeled image, we extract its features in addition to a pool of sampled annotated images to generate pseudo-labels. In parallel, we use the predictions of a weakly augmented sample to act as supervision for a strongly augmented version of the image. On the right, we illustrate our reference- based pseudo-label generation for k = 5.

4.2 REFERENCE-GUIDED PSEUDO-LABEL GENERATION

In this section, we present a new approach to generating online pseudo-labels by using label-wise feature similarities from a set of annotated reference samples. We start by introducing some preliminary definitions, then describe our Reference-guided Pseudo-Label Generation (RPG) method, and finally enhance RPG with augmentation-based consistency regularization.

4.2.1 *Preliminary Definitions*

In the context of semi-supervised semantic segmentation, we are given a small set of labeled images $\mathscr{S}_{\mathscr{L}} = \{(x_i, y_i)\}_{i=1}^{N_l}$, where each image is represented as $x_i \in \mathbb{R}^{ch \times h \times w}$ with ch channels, height h, and width w. The labels are represented as $y_i \in \{0, \ldots, c-1\}^{h \times w}$ for c-class segmentation or $y_i \in \{0,1\}^{c \times h \times w}$ for multi-label segmentation. Additionally, we are given a large number of unlabeled images $\mathscr{S}_{\mathscr{U}} = \{x_i\}_{i=1}^{N_u}$. The goal is to use $\mathscr{S}_{\mathscr{L}}$ and $\mathscr{S}_{\mathscr{U}}$ to train a model that can accurately predict labels on previously unseen images.

To this end, we define the segmentation model as a two-part process: a dense feature extractor $f_{\text{feat}} : \mathbb{R}^{ch \times h \times w} \to \mathbb{R}^{d \times h \times w}$ and a subsequent pixel-wise classifier $f_{cls} : \mathbb{R}^{d \times h \times w} \to [0,1]^{c \times h \times w}$. The feature extractor is a neural network, while the classifier is implemented using a 1×1 convolution and normalization function (either sigmoid or softmax, depending on the format of the labels y_i).

4.2.2 *Nearest-Neighbors from References for Pseudo-Label Generation*

In our approach to generate pseudo-labels for semantic segmentation, we utilize image-reference-pairs and transfer the semantics of the labeled image to the unlabeled image. Instead of making confidence-based predictions, we find the best fit in feature space by discovering most similar data point from a pool of references. Using these pseudo-labels, we can then train our model using the cross-entropy loss. Our method is illustrated in Fig. 21.

4.2.2.1 *Building a Reference Pool*

As we build pseudo-labels from labeled references by making feature space comparisons, we project both labeled and unlabeled pixels into the same feature space using f_{feat}. However, processing all $h \times w$ d-dimensional pixel-wise representations for each image in $\mathscr{S}_{\mathscr{L}}$ is not feasible due to limited memory. Integrating solutions like a memory-bank [306] becomes difficult due to the large number of pixel-wise representations and coinciding semantics. To address these issues, we randomly sample a pool $\mathscr{P}$ of labeled images from $\mathscr{S}_{\mathscr{L}}$ in each mini-batch iteration:

$$\mathscr{P} = \{(x, y) \sim \mathscr{S}_{\mathscr{L}}\}^p \tag{18}$$

Since we will later generate pseudo-labels from $\mathscr{P}$, it is important that all classes are present in the pool, otherwise it will not be possible to recover missing class labels. To ensure this, we assign each labeled image a global class label indicating whether a single pixel of a particular class is present, and sample p images such that each class occurs at least once in $\mathscr{P}$.

We create the reference set $\mathscr{R}_{\mathscr{P}}$ by extracting the pixel-wise features of each image in $\mathscr{P}$ to get pairs of pixel-representations and -labels:

$$\mathscr{R}_{\mathscr{P}} = \{(f_{\text{feat}}(x), y) : (x, y) \in \mathscr{P}\} \tag{19}$$

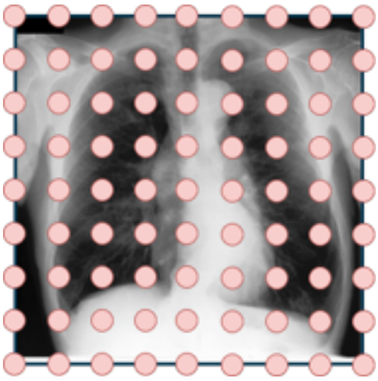


Figure 22: Example of equidistant pixel-sampling via Nearest-Neighbor interpolation for the generation of our reference pool.

To reduce memory requirements, we sub-sample equidistant pixel-level representations and labels to a smaller size $s \times s$ using nearest-neighbor interpolation as shown in Fig 21. Doing so also lets us retain the correct semantics of a feature vector as information does not get blurred which would be the case for linear interpolation.

We discard the spatial relationships between pixels and treat $\mathscr{R}_{\mathscr{P}}$ as a set of d dimensional feature vector-label pairs with $|\mathscr{R}_{\mathscr{P}}| = p \cdot h \cdot w$. By sampling continuously from $\mathscr{P}$ in each iteration during training, the labeled images can be subject to various data augmentation techniques (i.e. cropping, rotation, ...), resulting in a diverse set of pixel-level representations in $\mathscr{R}_{\mathscr{P}}$.

4.2.2.2 *Nearest-Neighbor Label Assignment*

We create pseudo-labels for each unlabeled pixel by finding the closest labeled pixels in feature space from a reference pool $\mathscr{R}_{\mathscr{P}}$. We extract pixel-wise features $\hat{u} = f_{\text{feat}}(u)$ for each unlabeled image $u \in \mathscr{S}_{\mathscr{U}}$ in the mini-batch, where $\hat{u}$ is a feature map of size $h \times w$ and dimension d. We then assign a target to each unlabeled vector $\hat{u}_{\chi,v}$ with spatial coordinates $\chi, v \in \mathbb{N}^{h \times w}$ based on the contextually closest vector in $\mathscr{R}_{\mathscr{P}}$. The proximity measure between labeled and unlabeled pixel-representations is the clipped cosine distance $\mathscr{D}$, resulting in:

$$\mathscr{D}(r, \hat{u}_{\chi,v}) = 1 - \max\left(\frac{\sum_{i=1}^{d} r_i \cdot \hat{u}_{\chi,v,i}}{\sqrt{\sum_{i=1}^{d} r_i^2} \cdot \sqrt{\sum_{i=1}^{d} \hat{u}_{\chi,v,i}^2} + \epsilon}, 0\right), \tag{20}$$

where $r \in \mathscr{R}_{\mathscr{P}}$, y is the label, $\epsilon = 1e^{-8}$, and the subscript i indexes the i-th dimension of a vector. The distance between two feature vectors is zero if they are identical and one if they are orthogonal or contrary to each other.

For each unlabeled pixel $u_{\chi,v}$, we assign a pseudo-label $l(u_{\chi,v})$ based on the label of its closest sample in the reference pool:

$$l(u_{\chi,v}) = y : \operatorname*{argmin}_{(r,y)\in\mathscr{R}_{\mathscr{P}}} \mathscr{D}(r, \hat{u}_{\chi,v}) \tag{21}$$

The labels of the entire image are given by $l(u) = \{l(u_{\chi,v}) : u_{\chi,v} \in u\}$. Note that this way, y can be either a one-hot vector or a complex multi-label vector. Our approach differs from traditional pseudo-labeling, where a network has to surpass manually designed thresholds for each class, as we directly utilize the label assignment for each corresponding point. Our nearest-neighbor target assignment method is similar to previous work [111, 165, 178], but we operate online and access multiple reference images simultaneously.

4.2.2.3 *Density-based Class Entropy*

With an fitting pool size p, our nearest-neighbor label assignment has been shown to be effective for semantic segmentation. However, we have observed that this method can be problematic when the features of an unlabeled pixel are similar to multiple classes, as this can mislead the network during training. To address this issue, we weigh the impact of an unlabeled pixel based on the ambiguity of its surroundings in the feature space.

Specifically, for an unlabeled pixel with feature $\hat{u}_{\chi,v}$, we compute the minimum distance $\delta^j_{\hat{u}_{\chi,v}}$ to each class j among the k nearest neighbors $\mathscr{R}^k_{\mathscr{P}}$ in the feature space:

$$\delta^j_{\hat{u}_{\chi,v}} = \min_{(r,y)\in\mathscr{R}^k_{\mathscr{P}} \wedge y=j} \mathscr{D}(r, \hat{u}_{\chi,v}) \tag{22}$$

If class j is not present in the reference pool $\mathscr{R}^k_{\mathscr{P}}$, we set the distance $\delta^j\, \hat{u}_{\chi,v}$ to one. We then use these class distances to compute the j class probabilities $P^j_{u_{\chi,v}}$ for the unlabeled pixel via the class-wise normalization:

$$P^j_{u_{\chi,v}} = \frac{1-\delta^j_{u_{\chi,v}}+\epsilon}{\sum_{j'=1}^{c} 1-\delta^{j'}_{u_{\chi,v}}+\epsilon}, \tag{23}$$

where ϵ is a small constant added to avoid division by zero. We then calculate the weight $W_{u_{\chi,v}}$ for the unlabeled pixel using the normalized entropy of the class probabilities:

$$W_{u_{\chi,v}} = 1 + \sum_{j=1}^{c} P^j_{u_{\chi,v}} \frac{\log P^j_{u_{\chi,v}}}{\log c} \tag{24}$$

This factor $W_{u_{\chi,v}}$ lowers the weight of pseudo-labeled pixels in highly ambiguous regions of the feature space. It encourages the inclusion of more classes in the pseudo-labels, rather than just choosing the most common one, because the distances to all classes are taken into account. Additionally, this weighting handles the case where $\delta^j_{u_{\chi,v}} = 1$ for all classes, in which the entropy is maximal and the unlabeled pixel is ignored ($W_{u_{\chi,v}} = 0$). We illustrate further cases in Figure 21. Finally, our method is formalized as the loss function $\mathscr{L}_{RPG}$:

$$\begin{aligned}\mathscr{L}_{RPG} =\,& \mathbb{E}_{(x,y)\in\mathscr{S}_{\mathscr{L}}}[\mathrm{CE}(f^c_{\mathrm{cls}}(f_{feat}(x)), y)] \\ &+ \mathbb{E}_{x\in\mathscr{S}_{\mathscr{U}}}[\mathrm{CE}(f^c_{\mathrm{cls}}(f_{feat}(x)), l(x)) \cdot W_x],\end{aligned} \tag{25}$$

with CE denoting binary or multi-class cross-entropy depending on the type of segmentation task.

4.2.3 *Consistency Regularization*

To demonstrate that our approach is compatible with consistency regularization techniques in semantic segmentation, we adapt it with FixMatch [260]. we generate pseudo-labels from network predictions on weakly augmented images and use them as labels for strongly augmented versions of the same image, thus enforcing consistency between them. The weak augmentations used are typical perturbations such as flipping, while for strong augmentations, we follow the RandAugment method [48]. Our setup is similar to that of Sohn *et al.* [260], but since we are dealing with segmentation instead of classification, we create pixel-level pseudo-labels and set the designated label for areas affected by the CutOut augmentation [51] to 'background'. For one-hot targets y, we use the standard pseudo-label formulation [260] shown in Equation 26.

$$l'(u_{\chi,v}) = \begin{cases} \text{argmax}_c\, f^c_{cls}(\hat{u}_{\chi,v}) & \text{, if } f^c_{cls}(\hat{u}_{\text{x,y}}) > \tau \\ \text{ignore} & \text{, else} \end{cases} \tag{26}$$

We also extend the FixMatch formulation to enable multi-label segmentation as shown in Equation 27, where τ is a scalar threshold value that separates labeled and ignored pixels.

$$l'(u_{\chi,v}) = \begin{cases} \lfloor f^c_{cls}(\hat{u}_{\chi,v}) \rceil & \text{, if } |f^c_{cls}(\hat{u}_{\chi,v}) - 0.5| > |0.5 - \tau| \\ \text{ignore} & \text{, else} \end{cases} \tag{27}$$

The whole image is labeled by choosing based on the task the respective $l'(\cdot)$.

$$l'(u) = \{l'(u_{\chi,v}) : u_{\chi,v} \in u\} \tag{28}$$

We denote the final consistency regularized loss term $\mathscr{L}_{RPG^+}$ as:

$$\mathscr{L}_{RPG^+} = \mathscr{L}_{RPG} + \mathbb{E}_{x \in \mathscr{S}_{\mathscr{U}}}[\text{CE}(f^c_{cls}(f_{feat}(a_s(x))), a_s(l'(x)))] \tag{29}$$

Our RPG enables the network to learn co-occurring targets, which leads to more confident predictions in weakly augmented images. This compliments the FixMatch-based approach as we now also gather pseudo-labels for more complex class combinations for consistency regularization.

4.3 EXPERIMENTAL SETUP

In this experimental section, we evaluate the performance of several methods for semantic segmentation in the context of limited annotated data and the applicability of our RPG for fine-grained anatomy segmentation. To do this, we will compare several approaches, beginning with a standard UNet Baseline model that is trained only on the available annotated data. However, due to the small number of annotated images, we run these models for a fixed number of iterations rather than a fixed number of epochs. We will consider the confidence-based Pseudo-Label method, which was originally proposed by Lee et al. [147] This method uses pseudo-labels for predictions above a certain threshold τ. We stack these against a naive Nearest Neighbor label assignment approach, which does not employ a weighting function. Furthermore, we also compare to two recent methods: MLDS (Multi-Label Deep Supervision) [217], which combines deep supervision with a Mean-Teacher [273] setup, and "FixMatch" [260], which leverages strongly and weakly augmented prediction comparisons. We use mean Intersection over Union (mIoU) as the performance metric to compare the performance of these methods.

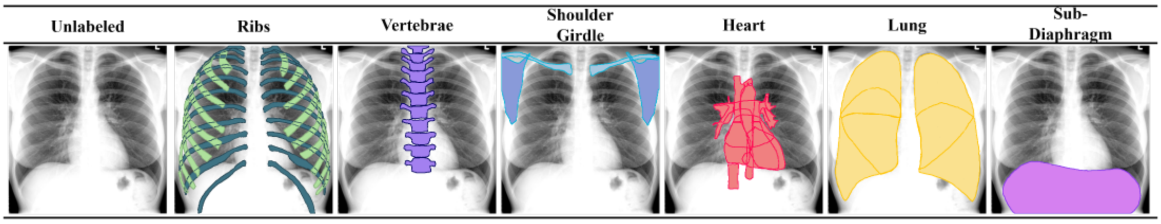


Figure 23: Display of super-classes of our annotation scheme of anatomy in chest radiographs.

4.3.1 *Datasets*

Our method is evaluated on two medical tasks: multi-label segmentation of CXR anatomy and multi-class segmentation of retinal fluid. For the CXR anatomy task, we use the public JSRT dataset, which consists of five classes that may overlap at the pixel level: right and left clavicle, right and left lung, and heart. The dataset includes two sets of images (123 and 124). For each set, we generate five random splits, each using a different number of labeled images ($N_l \in \{3, 6, 12, 24\}$). For each split, we use five images from the first set for validation and the second set for testing. To further demonstrate the effectiveness of our method for overlapping labels in small data settings, we expand upon the JSRT dataset by using more detailed annotations, resulting in a total of 72 labels belonging to the following supercategories: *Ribs*, *Vertebrae*, *Shoulder Girdle*, *Heart*, *Lung* and *Sub-Diaphragm*. These labels include finer-grained classes within and between different super-categories, such as the different lobes of the lung and the anterior and posterior parts of each rib. Our label scheme is outlined with the (super-) classes illustrated in Fig. 23. To create this expanded dataset, a medical expert annotated two CXRs, taking up to 3 hours per image. We evaluate the performance on this task by fusing our fine-grained classes into the corresponding JSRT labels and using five JSRT-labeled images from the first set for validation and the second set for testing.

For the retinal fluid segmentation task, we use the Spectralis vendor of the RETOUCH dataset, which includes 14 optical coherence tomography volumes with 49 b-scans each. We follow the setup of a previous study [217] and perform 10-fold cross-validation, with training sets using a different number of labeled images ($N_l \in 3, 6, 12, 24$) in each split and validation and test sets of roughly equal size. We ensure that in each split, all diseases are represented at least once in the mask labels.

4.3.2 *Implementation Details*

For our image segmentation model, we utilize the UNet architecture [227] with batch normalization [109] and bilinear upscaling blocks. Each convolutional layer, except for the final layer, has a kernel size of 3 and a bias. The feature extraction function, f_{feat}, describes the network up to the second to last layer. The classification function, f_{cls}, consists of a 1×1 convolution followed by a sigmoid function for JSRT data and a Softmax function for RETOUCH data. We initialize the network using Xavier initialization [77] with a gain of 0.01 and use Adam optimization [133] with default parameters, a learning rate of 0.0005, and a weight decay of 0.0005. The model is trained for 100 epochs on JSRT data, 200 epochs on extended JSRT data, and 50 epochs on RETOUCH data.For data augmentation, we randomly rescale images to between 90% and 110% of their original size and crop them to 80% of their original size. We also apply additive noise sampled from a normal distribution with mean 0 and standard deviation 0.1 with a probability of

35%, jitter images with an intensity of 0.15 and a probability of 80%, and rotate images by up to 12% with a probability of 70%. We use a batch size of 5 and image size of 512 for JSRT and a batch size of 8 for RETOUCH data, following the preprocessing described in [217]. In all experiments, the data batches consist of a combination of $\mathscr{P}$ with $p = 3$ and randomly sampled images from the full dataset. We set the number of nearest neighbors considered to be $k = 7000$ and the representation map size to be $s = 64$. All experiments were run on a single 11GB NVIDIA GeForce RTX 2080 using the PyTorch framework [195]. We evaluate the model every 10 epochs, apply the best-performing model on the validation set to the test set, and report the mean and standard deviation of the results.

4.3.3 *Ablation Studies*

4.3.3.1 *Pool Size for Pseudo-Label Assignment*

In order to assess the effectiveness of using nearest-neighbor-based pseudo-label generation for semantic segmentation, we conducted experiments to analyze the segmentation performance with different combinations of pool size and the number of annotated images. The results, presented in Table 7, indicate that when only a single image is used, the network's segmentation performance is relatively poor. However, as we increase the number of images in the pool, we see a consistent improvement in performance for all amounts of annotated images. We found that while increasing the pool size overall has a positive impact on performance, the improvement beyond $p = 3$ is significantly smaller and the required memory increases consistently. Based on these findings, we conclude that $p = 3$ is a suitable trade-off between performance and memory consumption, and we maintain this value for all subsequent experiments.

p	$N_l = 3$	$N_l = 6$	$N_l = 12$	$N_l = 24$
1	0.52 ± 0.02	0.66 ± 0.04	0.74 ± 0.02	0.78 ± 0.01
2	0.58 ± 0.04	0.72 ± 0.03	0.77 ± 0.03	0.82 ± 0.01
3	0.64 ± 0.05	0.76 ± 0.02	0.81 ± 0.02	0.84 ± 0.01
4	0.65 ± 0.04	0.76 ± 0.02	0.79 ± 0.03	0.84 ± 0.01
5	0.66 ± 0.03	0.77 ± 0.02	0.82 ± 0.02	0.84 ± 0.01

Table 7: Comparison of Nearest-Neighbor performance for different image pool sizes.

4.3.3.2 *Amount of Observed Neighbors*

To combat potential ambiguity in feature space, we integrate our class entropy weighting scheme into the nearest neighbor assignment. Thus, we explored the effect of the effective radius in feature space on our density-based class entropy weighting for different pool sizes p and relative amounts of considered neighbors k for $N_l = 3$. The results of our analysis are presented in Table 8 with each column displaying the relative amount of all considered features in the reference pool $\mathscr{R}_{\mathscr{P}}$.

As can be seen in the table, the performance of our method increases as k is increased, regardless of the pool size. This improvement is observed until 50% of all features in the reference pool, at which point the performance peaks and then falls off significantly for larger values of k. This suggests that an increased search radius is generally beneficial, but finding the optimal k can be challenging. We were unable to test larger values of k for larger pool sizes due to memory constraints.

p	$k = 25\%$	$k = 50\%$	$k = 75\%$	$k = 100\%$
1	0.55 ± 0.03	0.57 ± 0.02	0.56 ± 0.01	0.52 ± 0.04
2	0.60 ± 0.04	0.62 ± 0.03	0.58 ± 0.03	0.51 ± 0.04
3	0.66 ± 0.03	0.70 ± 0.02	0.67 ± 0.02	0.52 ± 0.06
4	0.68 ± 0.03	0.68 ± 0.03	OOM	OOM
5	0.68 ± 0.02	OOM	OOM	OOM

Table 8: Comparison of RPG for different k and image pool sizes. 'OOM' denotes '**O**ut **O**f **M**emory'.

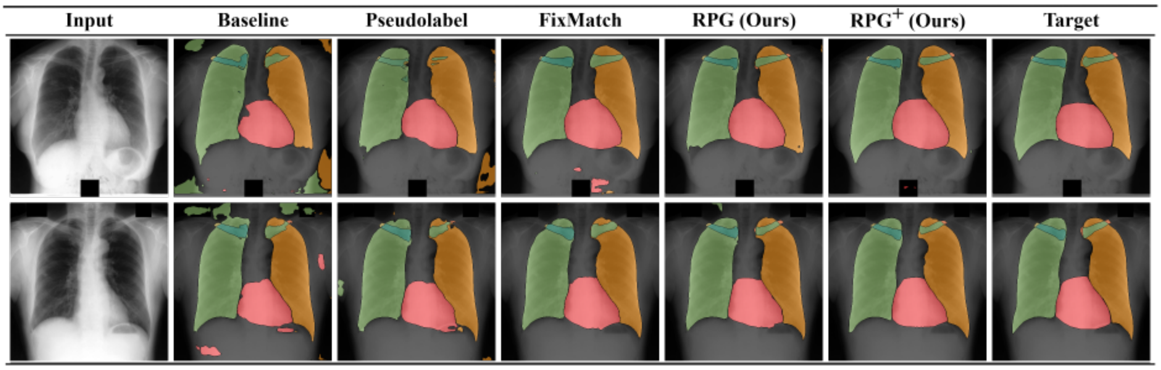


Figure 24: Qualitative Segmentation Results on the JSRT [257] dataset for $N_l = 6$.

It is also worth noting that the results in Table 8 show a clear improvement over the nearest-neighbor assignments alone, as demonstrated in the first column of Table 7. Overall, our weighting scheme appears to be effective in enhancing the performance of RPG when applied to the nearest neighbor assignment.

4.3.4 *Quantitative Results*

4.3.4.1 *Results on JSRT*

In Table 9, we present the mIoU scores of the compared methods for multi-label anatomy segmentation. Standard pseudo-labeling demonstrates varying performance depending on the chosen threshold, with better results for more annotated images but worse performance than the baseline in the low data case for any chosen threshold. Nearest Neighbor pseudo-labels improve upon the baseline by 3-5% for few samples, but show slightly worse results for 24 annotated examples. FixMatch achieves a 12% improvement above the baseline for $N_l = 3$, but its performance deteriorates for more annotations despite taking longer to converge. Both FixMatch and standard Pseudo-labeling exhibit varying performance for different values of τ. Our proposed method,

Methods	$N_l = 3$	$N_l = 6$	$N_l = 12$	$N_l = 24$
Baseline	0.59 ± 0.04	0.73 ± 0.02	0.81 ± 0.01	0.85 ± 0.01
Pseudolabel$_{\tau=0.8}$ [147]	0.56 ± 0.04	0.73 ± 0.04	0.81 ± 0.03	0.87 ± 0.01
Pseudolabel$_{\tau=0.95}$ [147]	0.57 ± 0.03	0.74 ± 0.03	0.82 ± 0.02	0.87 ± 0.01
Nearest Neighbor	0.64 ± 0.05	0.76 ± 0.02	0.81 ± 0.02	0.84 ± 0.01
FixMatch$_{\tau=0.8}$ [260]	0.71 ± 0.05	0.79 ± 0.02	0.80 ± 0.01	$0.85 \pm 0.00^*$
FixMatch$_{\tau=0.95}$ [260]	0.67 ± 0.05	0.77 ± 0.02	0.81 ± 0.02	$0.85 \pm 0.01^*$
RPG (Ours)	0.71 ± 0.02	0.79 ± 0.02	0.83 ± 0.02	0.87 ± 0.01
RPG$^+$ (Ours)	$\mathbf{0.77 \pm 0.05}^*$	$\mathbf{0.85 \pm 0.01}^*$	$\mathbf{0.87 \pm 0.00}^*$	$\mathbf{0.88 \pm 0.01}^*$
Full Access ($N_l = 123$)		0.85		

Table 9: Performance comparison of our work to related work on the datasets JSRT. * denotes that due to lacking convergence the model was trained twice the iterations. Bold and underlines denote best and second best performance.

RPG, performs equally well as FixMatch for smaller N_l and outperforms it for larger N_ls, while not using strong augmentations. We also find that integrating consistency regularization through strongly augmented images in RPG^+ improves the performance of RPG for all settings, resulting in a 18% improvement over the baseline for $N_l = 3$ and matching fully supervised performance with six labeled samples.

On the top of Table 10, we provide class-wise performance for three annotated samples. We see that both the baseline and pseudo-labeling struggle with less common classes such as the clavicles. FixMatch exhibits considerable improvements for classes with more annotated pixels, while the performance for the clavicles only slightly improves. RPG also improves upon the baseline for heart and lungs, but shows significant improvements for the difficult clavicles. Additionally, RPG^+ combines the aspects of RPG and augmentation-based consistency regularization, significantly improving all categories except for the *right clavicle*, with gains of up to 26% over the baseline.

In Fig. 24, we display segmentation predictions, highlighting the class-wise shortcomings of the different methods. We see the baseline and pseudo-labels, although to lesser extent, predict lungs on the outskirts of the image due to visible dark patterns. Fixmatch shows good segmentations for lungs and heart but oversegments the clavicle areas. RPG misses lung segmentations at the costophrenic angle but has slimmer predictions of the clavicles. RPG^+ noticeably has less oversegmentations than confidence-based predictions and improves on RPG.

4.3.4.2 *Results on Extended JSRT*

The results of our fine-grained annotation of the JSRT dataset are displayed in the bottom half of Table 10. When training solely on the annotated images, the baseline model achieved a mIoU of 10.43%. Both prediction-based pseudo-labeling methods performed worse than the baseline in this more complex low data environment. On the other hand, our RPG method was able to correctly predict the classes of the super-categories of the left and right lungs and the heart, resulting in an mIoU of 31.53%, an improvement of 21.10% over the baseline. The RPG^+ method noticeably improved upon all categories and also is able to produce predictions for the clavicle classes. This leads to RPG^+ achieving an mIoU of 35.31%.

In Fig. 25, we display just the extended anatomy segmentations obtained using RPG and RPG^+ as the predictions of the other methods resulted in unstructured visual noise. RPG pro-

Methods	Data	Right Lung	Left Lung	Heart	Right Clavicle	Left Clavicle	Mean
Baseline	JSRT $N_l = 3$	0.83 ± 0.02	0.81 ± 0.01	0.59 ± 0.02	0.47 ± 0.05	0.42 ± 0.07	0.59 ± 0.04
Pseudolabel$_{\tau=0.8}$		0.86 ± 0.02	0.87 ± 0.02	0.65 ± 0.10	0.35 ± 0.06	0.28 ± 0.06	0.56 ± 0.04
FixMatch$_{\tau=0.8}$		0.94 ± 0.00	0.93 ± 0.00	0.82 ± 0.03	0.50 ± 0.09	0.44 ± 0.14	0.71 ± 0.05
RPG (Ours)		0.91 ± 0.01	0.90 ± 0.01	0.71 ± 0.02	0.55 ± 0.04	**0.55 ± 0.03**	0.71 ± 0.02
RPG^{+*} (Ours)		**0.95 ± 0.00**	**0.95 ± 0.00**	**0.85 ± 0.02**	**0.60 ± 0.09**	0.50 ± 0.15	**0.77 ± 0.05**
Baseline	Custom $N_l = 2$	0.1105	0.0994	0.2335	0.0526	**0.0256**	0.1043
Pseudolabel$_{\tau=0.8}$		0.2335	0.0847	0.0920	0.0000	0.0000	0.0820
FixMatch$_{\tau=0.8}$		0.0504	0.0463	0.0041	0.0000	0.0000	0.0246
RPG (Ours)		0.6065	0.4592	0.5108	0.0000	0.0000	0.3153
RPG$^+$ (Ours)		**0.6326**	**0.4852**	**0.5636**	**0.0671**	0.0168	**0.3531**

Table 10: Performance comparison on JSRT and our extended annotations (Custom). * denotes a training of twice the iterations. Bold and underlined denote best and second best performance respectively.

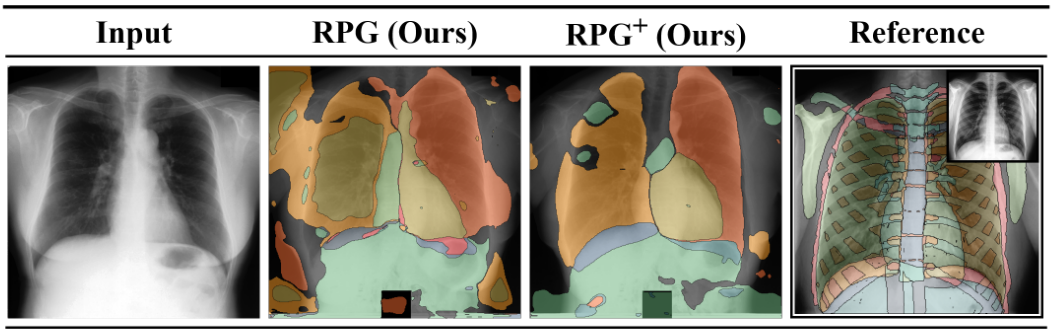


Figure 25: Qualitative Segmentation Results on our extended anatomical X-Ray annotations. We display segmentations for RPG and RPG^{+} in the center, while we show one of the reference images used for training on the right.

duces coarse oversegmentations but we can delineate the outlines of the different anatomical classes. We see that the RPG^{+}, albeit still producing errors, was able to roughly reconstruct the lung-subcategories and the ventricles of the heart, as well as the sub-diaphragm, but struggled with explicit predictions of bone structures.

4.3.4.3 *Results on Spectralis*

Our results for the Spectralis dataset are presented in Table 11. As shown in the table, the RPG method consistently outperforms the baseline for all values of N_l. This demonstrates it is applicable in the multi-class segmentation setting. Additionally, we see that RPG^{+} performs significantly better than the other methods for low data schemes with $N_l = 3$ and $N_l = 6$, achieving an improvement of up to 15%. However, it exhibits similar performance to MLDS for $N_l = 24$. Overall, despite our method noticably improving over the baseline, we do not see as immense boosts compared to other methods when compared to the multi-label setting of JSRT or our extended anatomy annotations.

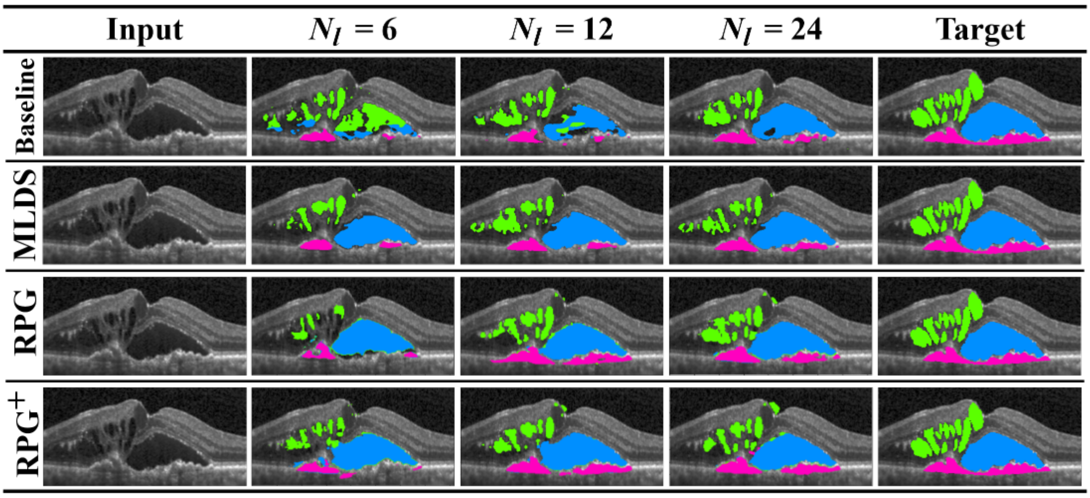


Figure 26: Qualitative Results for $MLDS$,RPG, and RPG^{+} on RETOUCH for different amounts of training examples.

Methods	$N_l = 3$	$N_l = 6$	$N_l = 12$	$N_l = 24$
Baseline	0.15 ± 0.07	0.27 ± 0.08	0.35 ± 0.06	0.49 ± 0.05
IIC [115]	0.22 ± 0.09	0.32 ± 0.07	0.41 ± 0.07	0.53 ± 0.06
Perone and Cohen-Adad (2018)	0.21 ± 0.09	0.31 ± 0.10	0.39 ± 0.07	0.50 ± 0.08
MLDS [217]	0.16 ± 0.15	0.35 ± 0.11	0.54 ± 0.09	**0.59 ± 0.07**
RPG (Ours)	0.21 ± 0.10	0.30 ± 0.08	0.45 ± 0.08	0.54 ± 0.08
RPG$^+$ (Ours)	**0.31 ± 0.11**	**0.45 ± 0.10**	**0.55 ± 0.08**	**0.59 ± 0.08**
Full Access ($N_l = 415$)	0.62 ± 0.05			

Table 11: Performance comparison on Retouch. Bold and underlined denote best and second best performance.

To provide additional insight, we present qualitative comparisons in Fig. 26. For $N_l = 6$, we see the baseline confusing several fluid regions. *RPG* seems to undersegment, whereas this is less the case for MLDS and RPG^+. When increasing the number of annotated training samples to 12 all models seem to become more accurate, although there are till noticeable errors in the baseline and *MLDS* slightly oversegments fluids while undersegmenting epithelial detachments. For 24 samples the methods show visually similar results.

4.4 CHAPTER CONCLUSION

In this chapter, to address the lack of anatomical annotations and the difficulty to implicitly learn such information, we investigate how we can utilize few dense annotations to train deep learning models. We introduced a new method for generating artificial supervision for semantic segmentation. Our approach utilizes labeled images as reference material, allowing us to match pixels in an unlabeled image to their corresponding labels in a set of reference images. To avoid ambiguous targets to negatively impact the training, we employ a class-entropy weighting term, through which impacts the loss based on the class diversity of the neighborhood an unlabaled sample. This method avoids the pitfalls of confidence-based pseudo-labeling, such as confirmation bias, and due to extracting completely annotated target vectors perfectly fits the problem of multi-label segmentation. As we require no additional parameters, our *RPG* can be easily integrated into any existing framework without the need for any changes to the architecture.

Our approach is suitable for medical image analysis as it makes use of the structural similarity provided by underlying human anatomy. We show the effectiveness of our method through thorough experiments on chest X-ray anatomy segmentation and retinal fluid segmentation, achieving fully supervised performance with a small number of samples and significantly reducing the annotation cost. We also show results on manual fine-grained anatomical annotations for chest X-Rays, and while existing methods struggle bring forward coherent results, our *RPG* enables coarse segmentations with few samples.

In summary, our proposed method represents a promising solution for generating supervision in a cost-effective and efficient manner, particularly in the context of medical image analysis. We summarize our contributions as:

Contribution 1: We illustrate a different view on online pseudo-labels. By enforcing consistency between predictions and the feature space, we cover cases not handled by standard pseudo-labeling approaches, considerably improving performance in multi-label segmentation.

Contribution 2: We provide detailed sensitivity studies investigating the importance of reference pool size for our pseudo-labels and the impact of the neighborhood sizes for the computation of class entropy.

Contribution 3: We show the effectiveness of our method on different datasets and various low data settings. Thereby, we demonstrate its use for handling the challenging segmentation of overlapping labels from scarce data as we reach fully supervised performance from six labeled images.

However, while our approach to semi-supervised segmentation works well in the segmentation of anatomical structures, we still require some form of manual annotation to get the process started. Detailed annotations as done in our experiments for chest X-rays take an immense amount of time (in our example three hours per image), which leads to even small scale datasets being difficult to collect. In the following chapter, we investigate how we can scale the dataset collection process in a manner, that allows the training of deep neural networks and does not require manual annotation effort by making use of partially annotated domains.

5

LEARNING FROM PARTIALLY ANNOTATED DOMAINS

A medical report has to convey a vast range of information to cover the any state of visible organs. However, most medical datasets focus strictly on a specific set findings or organs, but do not consider their surroundings. In this chapter, we aim to unify the available information to provide necessary context to different imaging domains. We start by paving the way for CT to X-Ray translation and show how we can make use of a variety of partial labels from the CT domain to segment anatomy of X-Rays in Section 5.1. Afterwards, we expand on this prove of concept and display ways to scale this data collection process through self-training in Section 5.2. Finally, in Section 5.3, we make the connection between vision and language as we ground reported findings in X-Rays via their associated anatomical descriptions.

While the incorporation of unlabeled data can heavily benefit the training of deep learning models, one still has to acquire enough data to kickstart the process. For our goal of achieving a holistic understanding of a given image, this can pose as quite troublesome since we have to gather annotations for a vast variety of labels. For segmentations in X-rays, this issue is exacerbated due to the body absorbing radiation to a highly varying degree leading to anatomical structures in two-dimensional images being visibly overlayed with each other. As the complete anatomic annotation for a single image already took up to three hours due to the ambiguity of overlapping structures [242], an entire dataset appears as a daydream.

Is there any way we can gather the required labels without hiring an army of medical staff for annotation purposes? We take a look at other radiology subdomains and notice most datasets focus on a specific subset relevant to their specialty, i.e. segmentations of Thoracic Organs at Risk [143] or the spine [247] in CT volumes. These datasets partially comprising of less than a 100 patients have shown to enable the development of robust networks transferrable to new data [112]. We assume that by aggregating all these different annotations, we can achieve a rather complete outlook on the human body.

Assuming this is achievable, we now face one final obstacle in order to utilize these annotations for our CXR report generation models: As most considered datasets are in the CT domain,

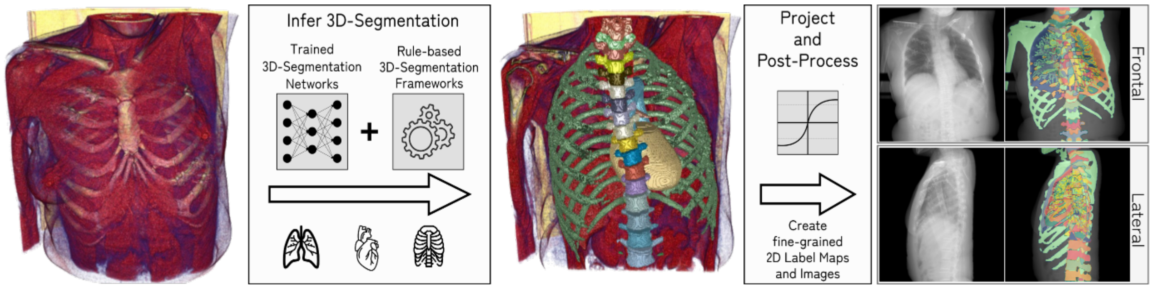

Infer 3D-Segmentation
Trained 3D-Segmentation Networks
+
Rule-based 3D-Segmentation Frameworks
Project and Post-Process
Create fine-grained 2D Label Maps and Images
Frontal
Lateral

Figure 27: Dataset creation protocol [243]. We apply 3D segmentation methods to generate comprehensive annotations of a CT dataset. Afterwards, the CTs and their labels are projected to 2D, emulating X-rays.

we have to find way to translate this information to CXRs. While this task of domain adaptation is typically difficult, we make use of the fact that the CT albeit being a volumetric imaging modality directly build upon the two-dimensional X-Ray. Therefore, as these modalities share several visual features, we can formulate a back-projection to map not only the CT but also the associated annotations to a two-dimensional plane, simulating an X-Ray.

The following sections build upon the idea of utilizing the CT domain to aggregate various annotations and, by using the corresponding projections, enable the accurate segmentation of chest radiographs.

5.1 ANNOTATIONS OF CHEST X-RAYS VIA COMPUTER TOMOGRAPHY PROJECTION

The following section is based on our publication in *BMVC 2022* [243].

As we have shown, most datasets in the CXR domain rely on either automatically parsed pathology labels or complete medical reports. While this has rapidly advanced the field of automated X-Ray processing, dense masks, which are essential in many computer vision approaches, have taken a backseat due to the immense difficulty of gathering precise annotations for such overlapping structures. As CTs allow for easier recognition of specific structures due to their volumetric characteristic, we propose a new method for generating annotations for X-Rays using them to address this challenge. Our approach, which we illustrate in Fig. 27, involves collecting annotations from existing CT models, projecting them onto 2D images that mimic the X-Ray domain, and creating a finely labeled dataset for anatomy segmentation in both frontal and lateral views. Using the shared imaging basis, we leverage the similarities in anatomy between different imaging modalities.

5.1.1 *Automated Generation of Projected CXR Datasets*

In previous chapters, we built deep learning models to recognize anatomies or findings on the X-Ray domain in an end-to-end manner without considering the mechanics underlying the imaging process. This narrows our understanding of potential connections to other imaging modalities. Therefore, we will first examine the similarities and differences between CT and X-Ray imaging and then demonstrate how we can use this information to create accurately labeled datasets.

5.1.1.1 *CT and X-Ray: Two birds of a feather*

X-Ray imaging is a basis for several non-invasive imaging techniques that use ionizing radiation in the form of X-Rays to produce images of the body's internal structures. When X-Rays pass through the body, dense tissues, such as bone, absorb more radiation than softer tissues, such as muscle or fat. This difference in absorption allows X-Rays to depict the body's internal structures.

In conventional chest X-Ray exams, the erect patient is positioned between an X-Ray tube and a detector. The collected data by the detector is then used to create a comprehensive image. The main distinctions in positioning (such as anteroposterior (AP) and posteroanterior (PA), both of which we refer to a *Frontal*, or lateral (L)), are related to which view enables the most insight.

In a similar manner, CTs are an imaging technique building on ionizing radiation to produce detailed images of the body's internal structures. However, instead of using a fixated X-Ray tube and detector as with X-Rays, CT scanners use multiple X-Ray sources and detectors that rotate around the body, producing a series of cross-sectional images to compute a slice. This process is done until all slices of a desired body part are taken, resulting in a detailed three-dimensional

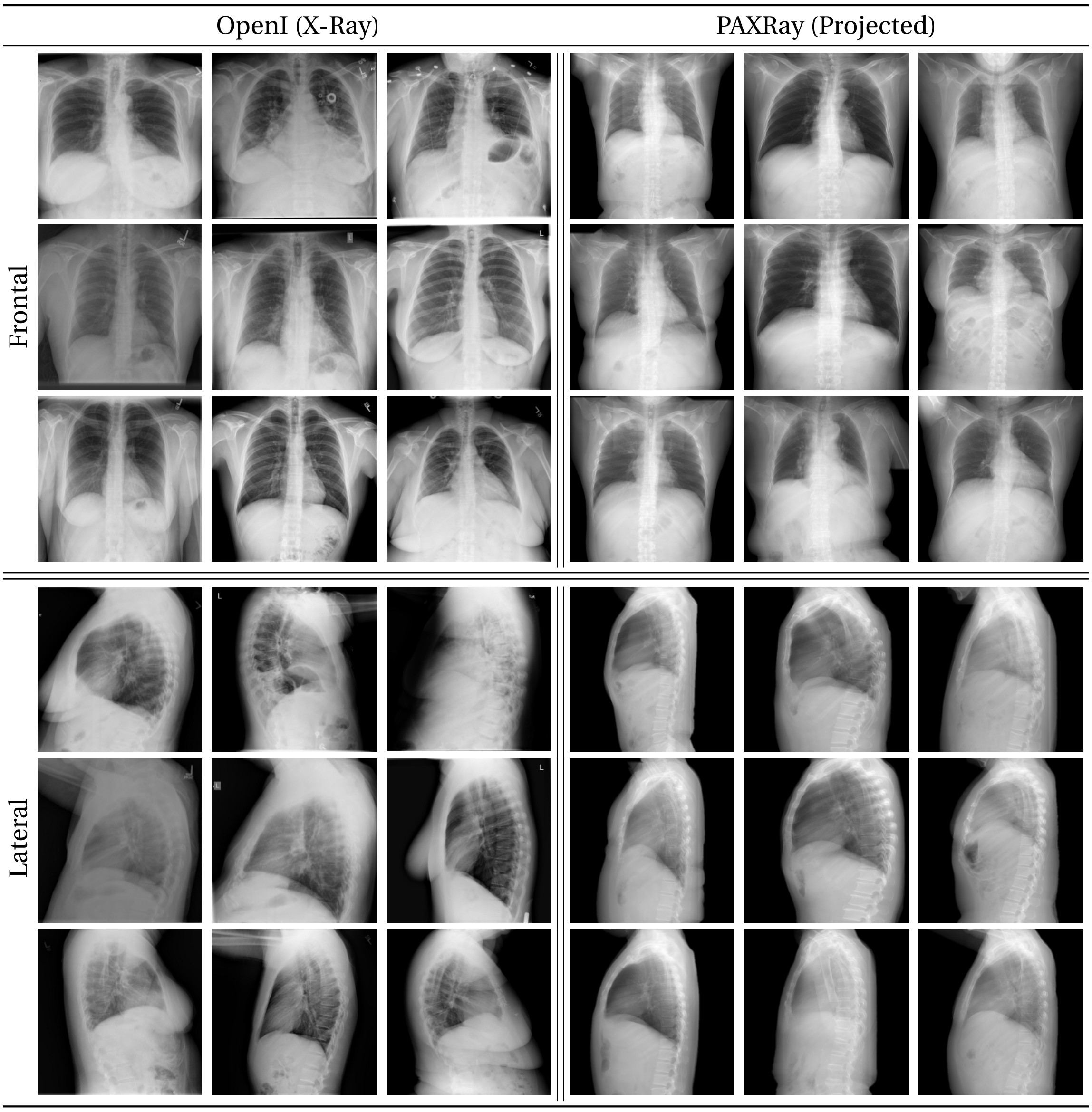


Figure 28: Comparison of different real and projected samples for the frontal and lateral view

volume. This eliminates superimposition common in standard X-Rays, enabling a radiologist to more easily interpret a patient. While the CT scan enables better diagnostic possibilities than the X-Ray, it typically involves higher radiation doses.

Often, an X-Ray is one of the first steps in the radiologic diagnostic process. If not enough information to diagnose a particular condition can be extracted, a CT scan may be ordered. Hence, it is crucial to enable systems to extract the maximum information to avoid unnecessary procedures.

It is vital to minimize radiation exposure to reduce the risk of harmful side effects. We argue that this can be achieved by directly utilizing CT's information depth in the training process of CXR segmentation models by projecting CTs and their semantic context back to a two-dimensional plane.

5.1.1.2 *Projecting CT back to X-Ray*

The back-projection is related to the concepts of Mean Intensity Projection and Maximum Intensity Projection [66], which allow one to gain quick insights on a patient without requiring a thorough look through the CT. Unlike in the Mean- or Maximum-Intensity-projection, we want to generate images that look visually similar to real X-Rays.

We start the back-projection process by first collecting the outline of the patient to separate it from the table he is lying on. Let V be the volume of a CT scan, which we clip to the standard 12-bit range. We get the body mask via first thresholding $M_{Body} = V \geq -100$, filling each contour and extracting the connected component of the highest volume. Afterwards, we collect the $\mathscr{M}_{bone}$ through slicewise generalized histogram thresholding [13]. We standardize the volume along the axis at which it is to be reduced, map it to the range of $[0,1]$ via a sigmoid function σ and sharpen:

$$V'_{Body} = M_{Body} \cdot \sigma\left(\frac{V - mean(V)}{std(V)}\right). \tag{30}$$

We repeat this for the bone region to get V'_{Bone}. Afterward, the V's are summed, min-max-feature scaled and rescaled to 8-bit. We then average the volume along the coronal plane to get a frontal projected X-Ray along the sagittal plane for a lateral one. This projection is similar to Matsubara *et al.* [172]. We split the projection into full body and bone projection. However, we found that utilizing the additional body masks assists with excluding unwanted structures, such as the table.

We compare our projected X-Rays with samples from the OpenI dataset for frontal and lateral views in Figure 28. The differences in the frontal view are due to the different positioning of the shoulder girdle. In the X-Rays, the arms are usually placed alongside the body, while in the projected images, the arms are raised due to the nature of the CT scan. In the lateral view, the X-Rays show a more comprehensive range of orientation and pose. However, the projected images, typically taken while the patient is lying down, result in similar poses between the different images. This leads to visual differences between images of female patients in both frontal and lateral views, such as the third column and second row of real X-Ray images and the first column and second row of projected images.

5.1.1.3 *Automated Label Generation*

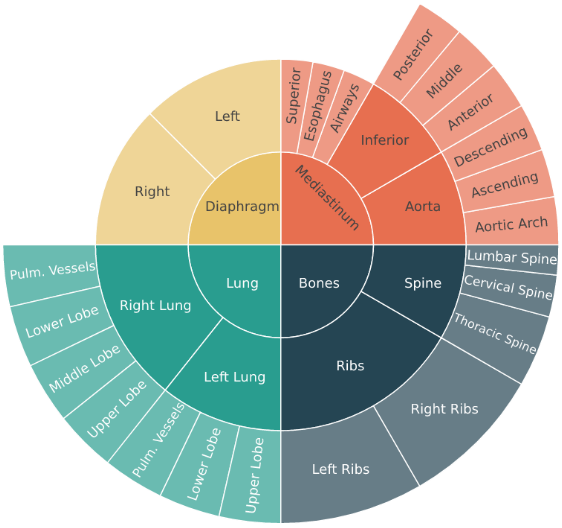


Figure 29: Compressed diagram of our initial proposed label structure for PAX-Ray.

The emergence of the UNet has significantly improved the accuracy of 3D medical image segmentation [40, 112, 228]. In our annotation process, we use both proven traditional methods [144, 145] and the state-of-the-art nnUNet [112], which has shown great potential as a black-box segmentation model.

Our annotation scheme is based on a label hierarchy with each label mask represented by $\mathscr{M}_l$, where $l \in L$ and L is the set of considered labels. We begin by generating a body mask $\mathscr{M}_{body}$ to separate it from the CT-detector backplate by selecting the largest connected component after thresholding as described previously. We then consider the

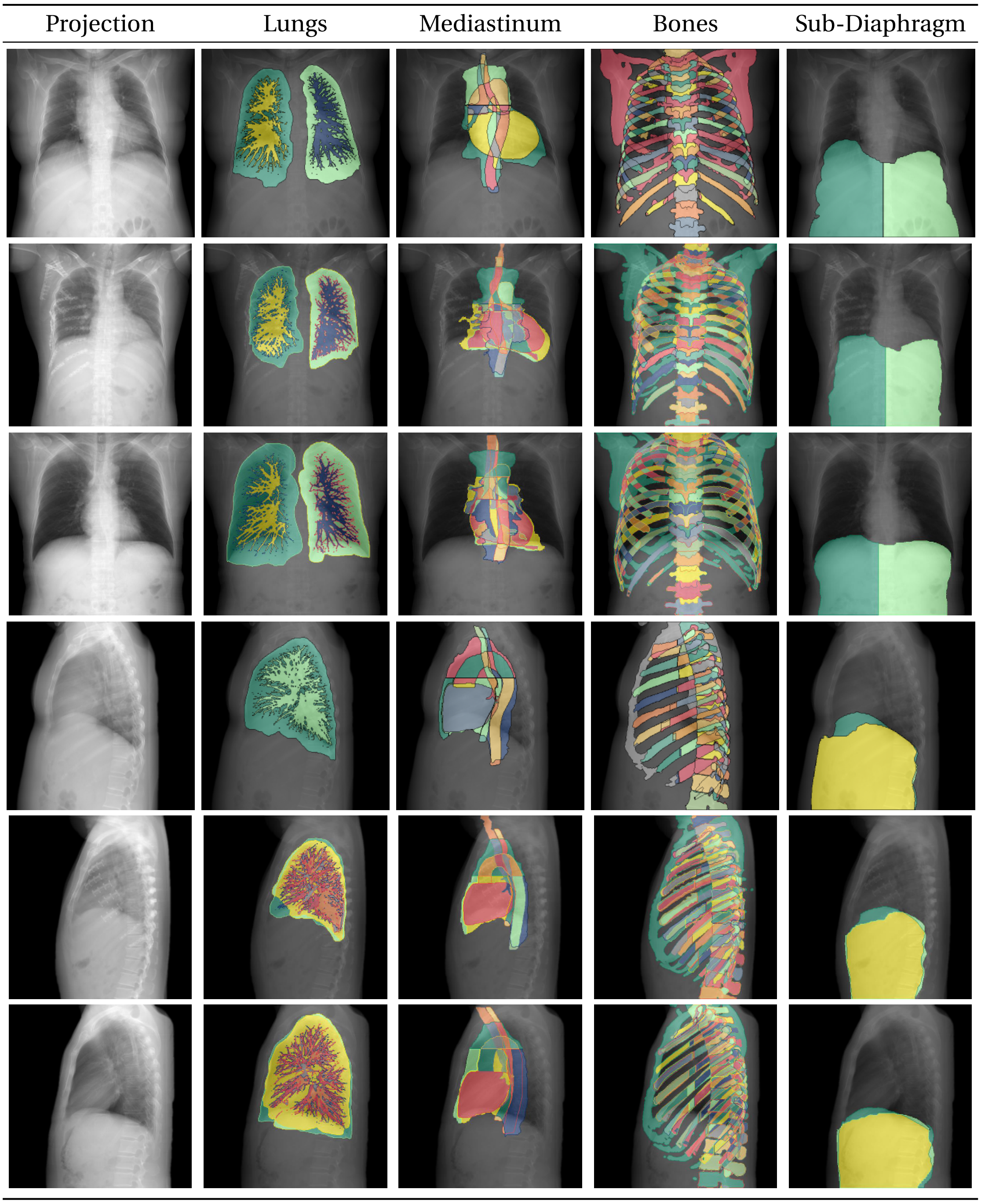


Figure 30: Examples of overlapping annotations in both lateral and frontal view of PAX-ray

super-categories of *bones*, *lung*, *mediastinum* and *sub-diaphragm*. In general, we assume that each fine-grained class is a subset of its parent class, such as the spine within the bone structure ($\mathcal{M}_{spine} \subset M_{bone}$). We obtain $\mathcal{M}_{bone}$ through slicewise generalized histogram thresholding [13]. Within the bone structure, we segment the individual vertebrae using a nnUNet [112] trained on the Verse dataset [162, 169, 247] and on ribs from RibFrac [313]. For Verse, the final validation performance ranges between 0.82 and 0.84 across all splits. We extend the rib annotation of Yang *et al.* [313] by distinguishing individual ribs as well as posterior and anterior parts based on their center and horizontal inflection.

For the lungs, we use the lung lobe segmentation model by Hofmanninger *et al.* [101]. The reported dice score for their lung segmentation model of 0.99, 0.94, and 0.98 on the LTRC [1], LCTSC [315], and VESS12 [232] datasets respectively. The merger of the individual lobes results in the lung halves. We further obtain the pulmonary vessels and total lungs through calculated thresholding and post-processing strategies [144].

For the mediastinum, we consider the area between the lung halves. To segment this area, we utilize Koitka *et al.* 's Body Composition Analysis (BCA) [135] and split it into superior and inferior along the 4th T-spine following medical definitions[45]. Koitka *et al.* [135] show a performance for their BCA of 0.96 dice on their internal dataset. We extract annotations for the *heart*, *aorta*, *airways*, and *esophagus* using the SegThor dataset [143]. The final validation performance for the nnUNet trained on SegThor ranges between 0.91 and 0.93 across all splits.

As for the sub-diaphragm, we consider the area below the diaphragm. This area can be extracted from the soft tissue region segmented using the BCA, which we split centrally into the left and right hemidiaphragm as there is no anatomical indicator.

To generate the label set L, we apply the combination of mentioned networks and rulesets on the publicly available RibFrac [118] dataset, which is suitable for this projection process due to its focus on the thoracic area, the high axial resolution, and the scans being recorded without contrast agents similar to X-Rays. We ignore volumes with conflicting segmentations and manually remove volumes with noticeable errors.

We project these labels to 2D using the max operation along the desired dimension and apply post-processing steps based on the observed anatomy (i.e., restricting specific classes to their largest connected component). Doing so, lets us collect the **P**rojected **A**natomy for **X-Ray** dataset (PAX-Ray) consisting of 852 images with associated masks.

5.1.1.4 *Further Visual Examples of Generated Labels*

We show examples of the annotations of our dataset in Fig. 30. We can see that lung halves can overlap in the second frontal row. While this might seem contradictory at first, when considering the lung as a 3D volume, the classes can overlap along one dimension, i.e., the lung halves in front and behind the heart, as seen in Fig. 31. As we intend to capture the entire anatomy of the upper body, we consider this labeling assumption a more holistic approach.

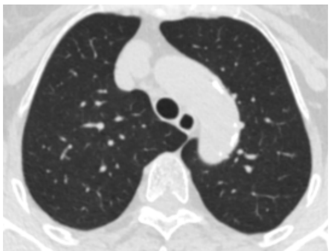


Figure 31: Axial slice of two lung halves overlapping sagittally.

5.1.2 *Experimental Setup*

Evaluation Setup: To assess the accuracy of our segmentation method, we utilize the mean Intersubsection over Union (IoU) measure on the PAXRay dataset. We follow the train/val/test split established in the RibFrac dataset [118], resulting in a total of 598, 74, and 180 images in the train, validation, and test sets, respectively. During the training process, we evaluate the performance of our model on the validation set every 10 epochs and ultimately choose the model that performs best on the validation set for our final test.

Implementation Details: To demonstrate the effectiveness of our fine-grained multi-label dataset, we trained segmentation models using different backbone networks, including the UNet (commonly used in the medical field) and the SFPN with a ResNet-50 backbone. We also compare the use of pre-training on the VinBigDataset [188] with using the networks from scratch. Since the labels may overlap, we used binary cross-entropy and a binary dice loss during training. We also employed random resize-and-cropping (with a range of $[0.8, 1.2]$) as an augmentation technique. We used AdamW optimization with a learning rate of 0.001 for 110 epochs, decaying by a factor of 10 at epochs 60, 90, and 100. The base image size for this process was 512.

5.1.3 *Results*

We see quantitative results in Table 12. We show the performance of selected super-classes and the mean over all 166 classes due to the immense number of classes. We display the complete performance over all classes in Fig. 34.

In this setting, we find that using pre-trained networks leads to an improvement of up to ~ 14% in mIoU. Additionally, we see that the UNet outperforms the SFPN by approximately 9%. However, some classes, such as the heart, lobes, and aorta, see similar performance between the two architectures, while others, such as mediastinal regions, airways, and ribs, show notable differences.

Overall, we can achieve up to ~ 90% mIoU on certain classes like the spine or heart, but segmentation of lung vessels proves particularly difficult with an IoU of 52%. The most significant discrepancy in segmentation quality between the two networks is observed in rib cage segmentation, where the UNet has a gain of 6.1%. Figure 32 shows qualitative examples of the segmentation results. The vessel tree and tracheal ends towards the bronchi are challenging to segment accurately, while lobe-, intermediastinal-, and bone-related classes appear as expected.

We demonstrate the usefulness of our proposed dataset for segmenting anatomy in real chest X-Rays by presenting qualitative results on the OpenI dataset in Fig. 33. The segmentation results for various structures, such as the mediastinal classes, sub-diaphragm, lung regions, spine, and ribs, are accurate in both views. However, we observe occasional errors in the prediction of individual ribs, particularly in the overlapping areas with the heart, which the overall brighter region may cause in these areas. Additionally, the predictions of lung vessels appear coarse compared to the annotations, which is expected due to the difficulty in visualizing vessels in chest X-Rays. Although errors similar to those in projected X-Rays can be seen (e.g., in the lateral rib or diaphragm segmentations), the results are encouraging even without using a domain adaptation method [275].

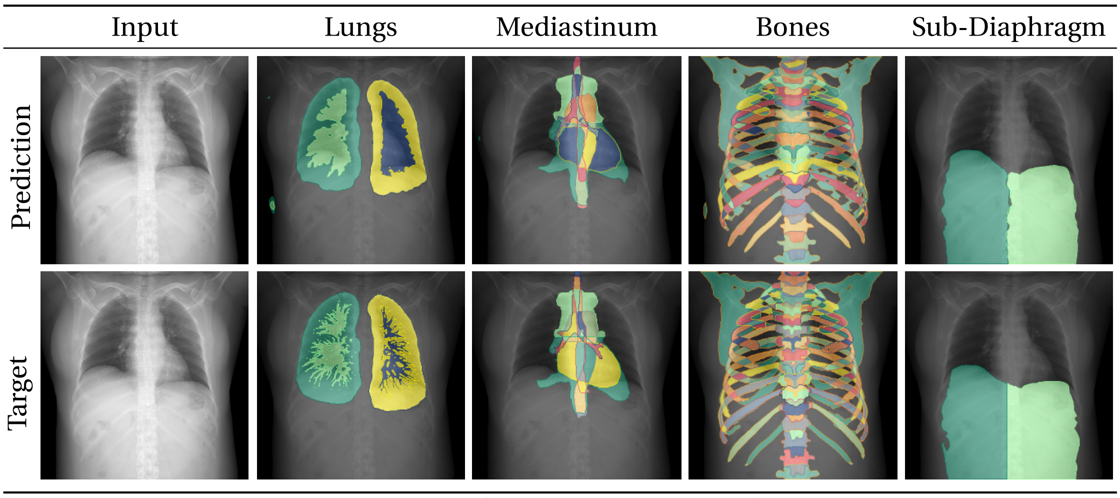


Figure 32: Qualitative results of a UNet on the test set of the PAXRay dataset

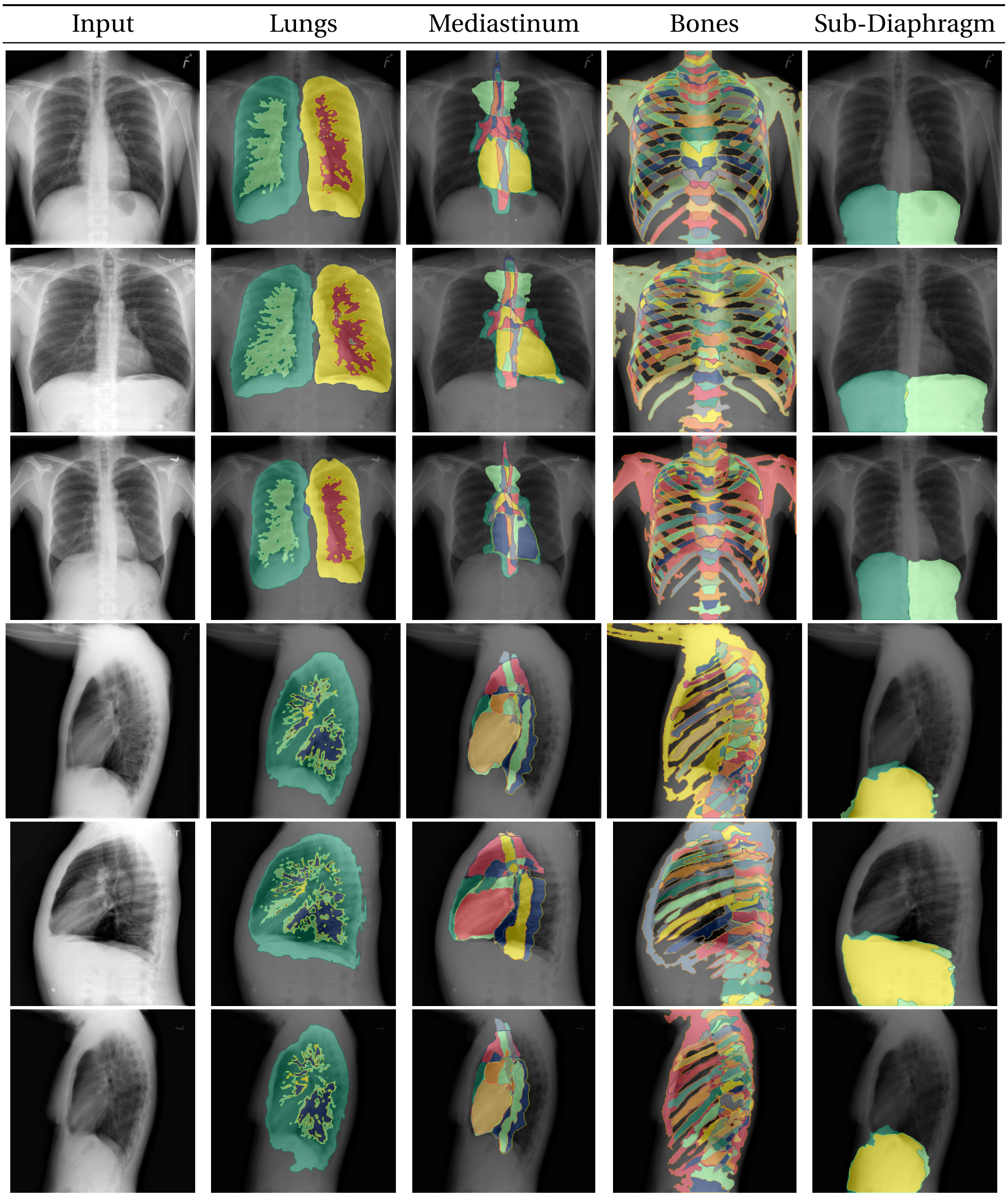


Figure 33: Qualitative results of a UNet trained on our PAX-ray dataset for a patient in OpenI

5.2 EXTENDING THE PAXRAY DATASET

As we can see, the introduced PAX-Ray dataset enabled glances into the potential of anatomical segmentations in CXR. However, due to the limited size of the dataset, the segmentation of specific sub-classes, such as the individual vertebrae or ribs, becomes difficult to train. Similarly, the human thoracic region consists of more relevant organs, such as the upper abdominal region, than are currently gathered.

Thus, we propose to extend the PAX-Ray dataset through multiple steps. First, we adapt the translation step between the two modalities to fit the physical underpinnings of Lambert-Beer's law. Secondly, we integrate additional classes by training on other datasets. Previously, we in-

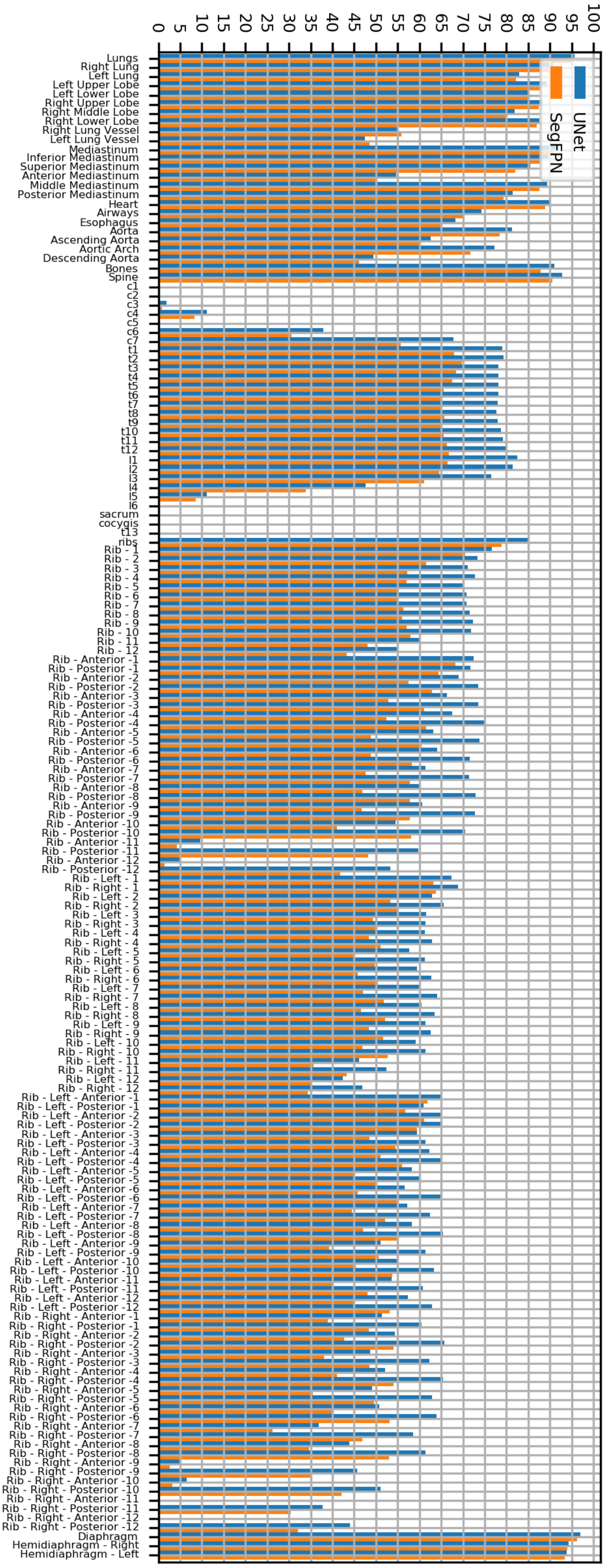


Figure 34: Segmentation performance in IoU on the test split for all classes of our proposed PAXRay dataset

ferred a set of individual models on the entire dataset. This, however, is a massive time sink as depending on the volume spacing single inference with the nnUNet might take more than 30 minutes. To reduce inference time, we aggregate all classes first on an intermediate full-body CT dataset. Retraining on this merged dataset can translate the knowledge more efficiently to a large set of unannotated thorax CT datasets. We can cut the number of required inferences by a factor of twelve. Lastly, we translate these pseudo-labeled CTs to projected CXR to generate our PAX-Ray++ consisting of more than 14,000 images and 2,000,000 annotated instances. We illustrate the entire process in Fig. 35.

5.2.1 *Adjusted Label Accumulation*

For the generation of PAX-Ray, we utilized a combination of a set of learned and non-learned approaches for a diverse set of classes jointly predicted for the RibFrac dataset [118]. When expanding the dataset, we look at two directions. First, increasing the number of observed labels. While the PAX-Ray dataset has many observed classes, it lacks in abdominal organs, heart and lung regions, and several bones. We, thus, gather ten different sources to train networks for subsequent prediction of desired additional classes. As labels from these sources can overlap, we first merge the predictions on a single dataset to sort out inconsistencies between annotation schemes and merge similar class labels. Doing so allows us to train a single model on the resulting dataset, significantly reducing inference and postprocessing time when applied to large-scale datasets. Secondly, we extend the dataset through the amount of considered source CT volumes. Initially, we used a single source dataset of 4̃40 volumes after filtering, leading to a dataset with less than a thousand images. This amount of images is small compared to several computer vision datasets. Consequently, we collect several large-scale thorax CT datasets to increase the number of projected X-Rays. Subsequently to these steps, we map these three-dimensional masks to a plane as done with PAX-Ray and filter cases with implausible mask predictions, such as anatomical structures with sizes massively deviating from the mean volume or having multiple connected components despite it being single structure. The masks are then post-processed with standard morphological operations and rescaled to the desired size.

5.2.1.1 *Considered Label Sources*

We consider datasets focusing on different body systems and available segmentation tools to provide a more leveled label distribution. We use these to train five folds of full resolution nnUNets [112] and keep a separate set for validation purposes. For inference, we use an ensemble of all folds with test time augmentation, such as horizontal flipping. We list the source datasets, their volume information, and the segmentation performance of their resulting networks in Table 13.

	Init.	Lung		Mediastinum				Bones		Sub-Dia.	Mean
		Lobes	Vessels	Regions	Heart	Aorta	Airw.	Spine	Ribs		
SFPN	(Random)	82.3	49.5	68.6	81.8	67.8	55.6	84.8	69.4	93.9	37.8
	(VBData)	86.3	**52.1**	74.6	88.9	79.0	70.0	90.5	78.8	96.2	51.9
UNet	(Random)	85.0	49.8	74.8	87.7	77.9	68.8	90.0	81.5	95.6	54.5
	(VBData)	**86.9**	50.8	**77.3**	**89.9**	**80.8**	**73.2**	**92.5**	**84.9**	**96.7**	**60.6**

Table 12: Category-wise segmentation performance in IoU on the test split of our proposed PAXRay dataset

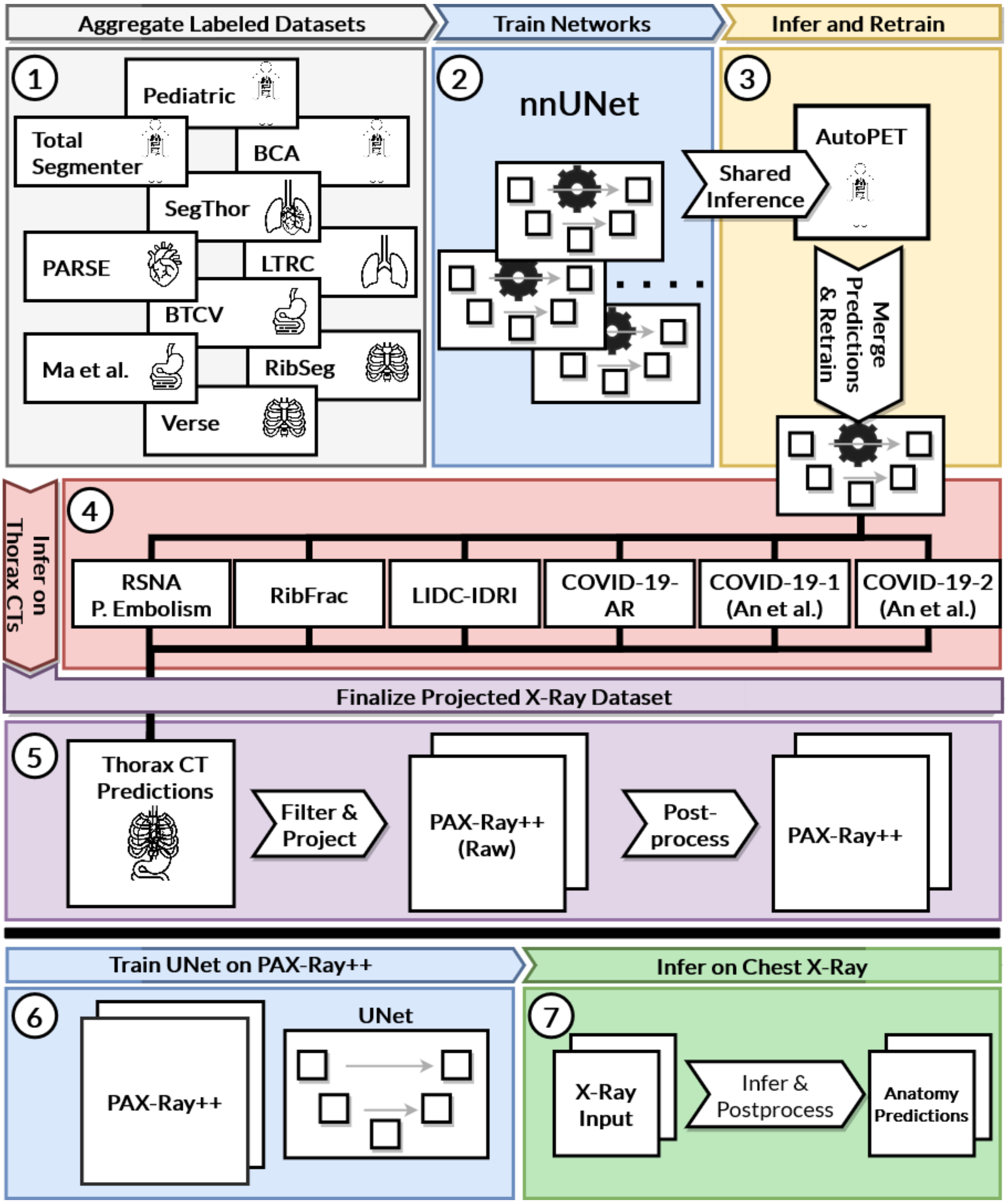


Figure 35: The workflow for scaling up PAX-Ray. We collect labeled datasets describing different categories, such as bones or abdominal regions. These datasets are used to train nnUNet ensembles, which are then applied to AutoPET to generate a shared label space. This shared label space is then used to re-train the model and generate pseudo-labels for several thorax CT datasets. The CT volumes and pseudo-labels are projected to a two-dimensional plane and post-processed using anatomical priors and morphological operations, resulting in PAX-Ray++. Finally, 2D segmentation models such as UNet are trained on the projected data to predict anatomical structures in CXR accurately.

Source	Series	Labels	Label Domain	Spacing	Slices	IoU
Pediatric [125]	287/72	29	Full Body	$[0.52 \pm 0.11, 0.52 \pm 0.11, 1.76 \pm 0.52]$	306.6 ± 245.4	0.833
Total Segmentator [299]	581/146	104	Full Body	$[1.5 \pm 0.0, 1.5 \pm 0.0, 1.5 \pm 0.0]$	259.0 ± 130.3	0.9251
SegTHOR [143]	40/20	5	Thoracic	$[1.00 \pm 0.09, 1.00 \pm 0.09, 2.39 \pm 0.23]$	184.7 ± 30.35	0.9232
PARSE [289]	100/30	1	P. Artery	$[0.67 \pm 0.07, 0.67 \pm 0.07, 0.99 \pm 0.01]$	301.4 ± 31.22	0.751
BTCV [185]	30/20	14	Abdominal	$[0.78 \pm 0.09, 0.78 \pm 0.09, 3.87 \pm 1.00]$	123.3 ± 32.19	0.8511
Ma et al [173]	40/10	13	Abdominal	$[0.81 \pm 0.07, 0.81 \pm 0.07, 2.63 \pm 0.52]$	95.88 ± 9.000	0.9093
RibSeg [313]	420/160	24	Ribs	$[0.74 \pm 0.07, 0.74 \pm 0.07, 1.13 \pm 0.14]$	359.5 ± 59.06	0.8996
Verse [247]	112/100	23	Spine	$[0.79 \pm 0.23, 0.79 \pm 0.23, 1.29 \pm 0.65]$	444.8 ± 348.6	0.8387
RSNA PE [44]	7302	-	-	$[0.67 \pm 0.08, 0.67 \pm 0.08, 1.25 \pm 0.22]$	237.6 ± 54.00	-
RibFrac [118]	660	-	-	$[0.74 \pm 0.07, 0.74 \pm 0.07, 1.13 \pm 0.14]$	359.5 ± 59.06	-
LIDC-IDRI [8]	1036	-	-	$[0.68 \pm 0.08, 0.68 \pm 0.08, 1.74 \pm 0.83]$	237.9 ± 132.7	-
COVID-19-AR [54]	176	-	-	$[1.57 \pm 0.79, 1.34 \pm 0.77, 1.42 \pm 1.05]$	438.8 ± 178.0	-
COVID-19-1 [91]	215	-	-	$[0.77 \pm 0.10, 0.77 \pm 0.10, 4.76 \pm 0.88]$	68.74 ± 36.24	-
COVID-19-2 [130]	632	-	-	$[0.77 \pm 0.10, 0.77 \pm 0.10, 4.72 \pm 0.94]$	72.21 ± 48.63	-

Table 13: Comparison of considered CT datasets for PAX-Ray++ regarding size, number of labels, label domain, volume spacing, number of slices and segmentation performance of a nnUNet [112] in 5-fold cross validation. Top half displays datasets used for label aggregation, while the bottom part was used as imaging source for PAX-Ray++.

The Pediatric dataset [125] contains 287/72 (training/validation) volumes of pediatric patients, with 29 labels spanning the entire body. The segmentation model achieved an IoU of 0.833. The Total Segmentator dataset [299] includes 581/146 volumes, with 104 labels for whole body segmentation, including classes such as humerus and the atria of the heart. We achieve an IoU of 0.9251. The SegTHOR dataset [143] includes 40/20 volumes, with 5 labels for thoracic organs at risk. We achieve an IoU of 0.9232. The PARSE dataset [289] includes 100/30 volumes for pulmonary artery segmentation. On the challenge benchmark, we achieve an IoU of 0.751. The BTCV dataset [185] includes 30/20 images, with 14 labels for abdominal segmentation, on which we achieve an IoU of 0.8511. The dataset by Ma et al. [173] includes 40/10 testing images, with 13 labels for abdominal segmentation. With an IoU of 0.9093. The RibSeg dataset [313] includes 420/160 volumes. While the original annotations segment the ribs as a single class along the outskirts of the ribs, we fill the ribs via thresholding and morphological operations and subsequently partition these via connected components into the 12 left and right ribs, with 24 labels for rib segmentation with an IoU of 0.8996. The Verse dataset [247] includes 112/100 volumes, with 23 labels for vertebrae segmentation. We achieve an IoU of 83.87%. We further utilize pre-trained classification models such as the Body Composition Analysis by Koitka *et al.* [135] and the lung lobe segmentation by Hoffmaninger *et al.* [1].

5.2.1.2 *Finalizing the Label Structure*

After training the nnUNet ensembles on the individual dataset, we infer the individual models on 560 whole-body volumes of AutoPET [75], a large CT dataset for tumor segmentation covering various body regions. We merge classes of overlapping semantics and discard unnecessary classes or ones of insufficient quality by manual assessment, such as the humerus classes. We further apply postprocessing techniques such as morphological operations and abiding anatomical biases such as maintaining the largest connected component for organs that are supposed to contain a single one.

We then retrain the nnUNet on the resulting dataset. This allows us to generate predictions of the complete label set with just a single model, which is significant as inference and postprocessing for a single ensemble can span up to an hour per volume, making the generation of a large-scale dataset a tall order. We integrate anatomical regions by modifying several anatomical classes:

1. The mediastinum stems from the BCA predictions and is split into upper and lower mediastinum along the T4 vertebra.
2. We differentiate the anterior and posterior mediastinum along the heart segmentation boundary.
3. We create the upper and middle lung regions and bases by splitting along the vertical thirds. The apical lung region is defined by the start of the individual clavicles.
4. We define the tracheal bifurcation as starting from 10 slices above the split of the trachea.
5. We maintain the process of the aorta splits and sub-hemidiaphragm from PAX-Ray.

Furthermore, when projecting the pseudo-labels to a frontal or lateral view, we merge classes to accommodate the characteristics of the view. For example, distinguishing between the left and right breast is impossible in the lateral view, whereas this is trivial in the frontal view. On the other hand, the lower mediastinum is better distinguishable in the lateral view, while not in the frontal one. We display the final classes in Fig. 37.

5.2.1.3 *Adjusted CT to X-Ray Projection*

Our original way of CT back-projection for the generation of the PAX-Ray dataset was focused on visual similarity. However, it was not based on the properties of the attenuation of light. We still split the volume into a body and bone volume for separate processing and subsequent merging, but instead of Eq. 30, we follow Matsubara *et al.* [172]. This allows us to come closer to a definition of Lambert-Beer's Law of the attenuation of light relating to the material properties it travels through. However, as CXR generally uses a harder type of radiation, we found that, for the frontal view CXR, further applying histogram equalization [201] leads to more visually pleasing results. We display the differences in translation in Fig. 36. We can see that, especially in the basilar lung regions, PAX-Ray++ is darker, while the abdomen is brighter than in our initial work.

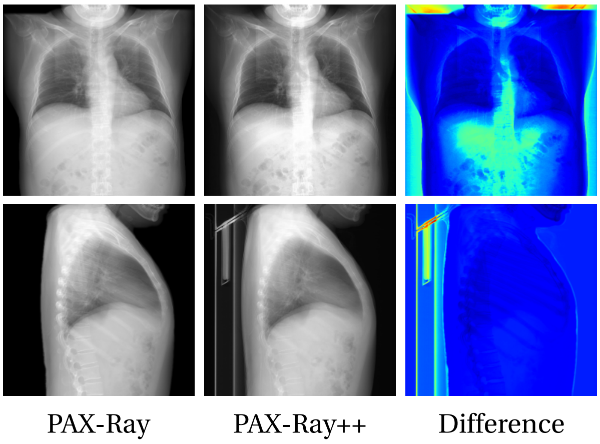


Figure 36: Illustration of projections of PAX-Ray 5.1.1.2 (left) and PAX-Ray++5.2.1.3 (center). We display the difference image on the right.

5.2.1.4 *Sources of PAX-Ray++*

To generate the PAX-Ray++ dataset, we require CT volumes as imaging data. We choose large unlabeled CT datasets focusing on the thoracic region, which is displayed in the bottom part of Table 13.

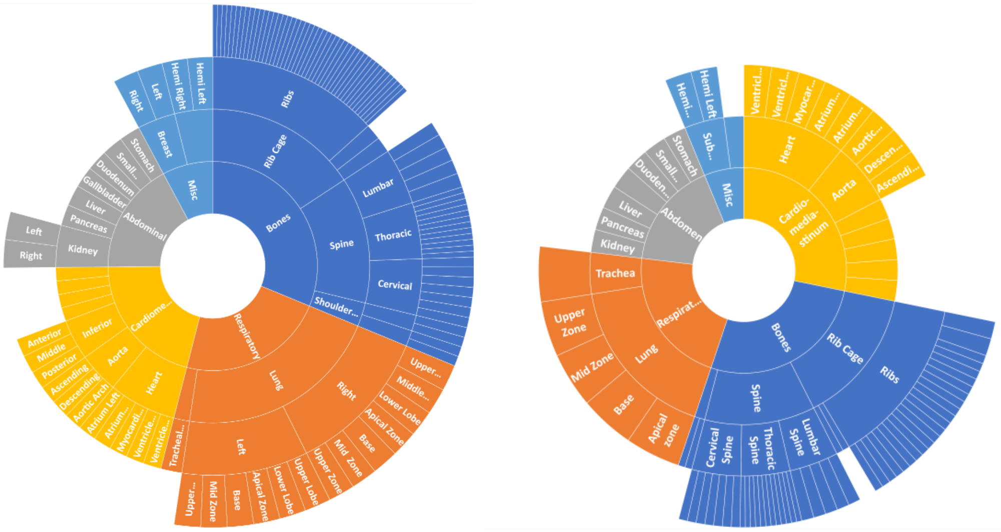


Figure 37: Compressed diagram of our proposed label structure for frontal and lateral images in PAX-Ray.

The datasets include RSNA PE, RibFrac, LIDC-IDRI, COVID-19-AR, COVID-19-1, and COVID-19-2. The RNSA Pulmonary Embolism dataset, a contrast CT dataset for identifying pulmonary embolisms, is the largest, with more than 7000 volumes with an average axial spacing of 1.25 mm and 237 slices. The RibFrac dataset for detecting rib fractures, which we used for the initial PAX-Ray, consists of 660 volumes with the smallest spacing of 1.13 mm and 359 slices. The LIDC-IDRI dataset was collected to detect pulmonary nodules and consists of 1036 volumes with a spacing of 1.74 and, on average, 237 slices. The COVID-19-AR dataset has, on average, the most slices of 438 and 176 volumes. However, it has a relatively large coronal and sagittal spacing. The COVID-19-1 and -2 datasets contain 215 and 632 volumes but have the smallest amount of slices and the largest axial spacing.

We infer the nnUNet ensemble trained on the aggregated pseudo-labels on Autopet on these datasets (see Fig. 35). We apply the same postprocessing methods as done with the initial predictions on AutoPET and filter volumes with predictions with anatomical deviations, i.e., too few predicted ribs. We then project the image and label files to a frontal and lateral view and resize to a uniform size of 512 × 512 using nearest interpolation for masks and Lanczos [277] for images.

5.2.2 *Dataset Analysis*

We next investigate the characteristics of PAX-Ray++. As each image is generated from a CT volume, the spacing and the slice amount have a noticeable effect on the label and projection quality of the resulting sample. This can be seen in the frontal and lateral qualitative samples in Fig. 38 and Fig. 39. We see the original projections and labels corresponding to the categories *Airways*, *Vascular system*, *Bones* and *abdomen* for the different datasets. The different datasets offer a variety of shades when emulating an X-Ray. The RibFrac and the RSNA PE dataset contain larger abdominal areas, while the remaining datasets focus more on the thoracic region. Due to the lower resolution and slice amount of the COVID-19-1 and -2 datasets, we can note ridges for the segmentations.

We show in Fig. 40 the number of instances per class in the PAX-Ray++ dataset and compare those with the equivalent classes in PAX-Ray. Apart from PAX-Ray++ containing more explicit classes, we can see that overall the instance amount is larger by a magnitude. Furthermore, while

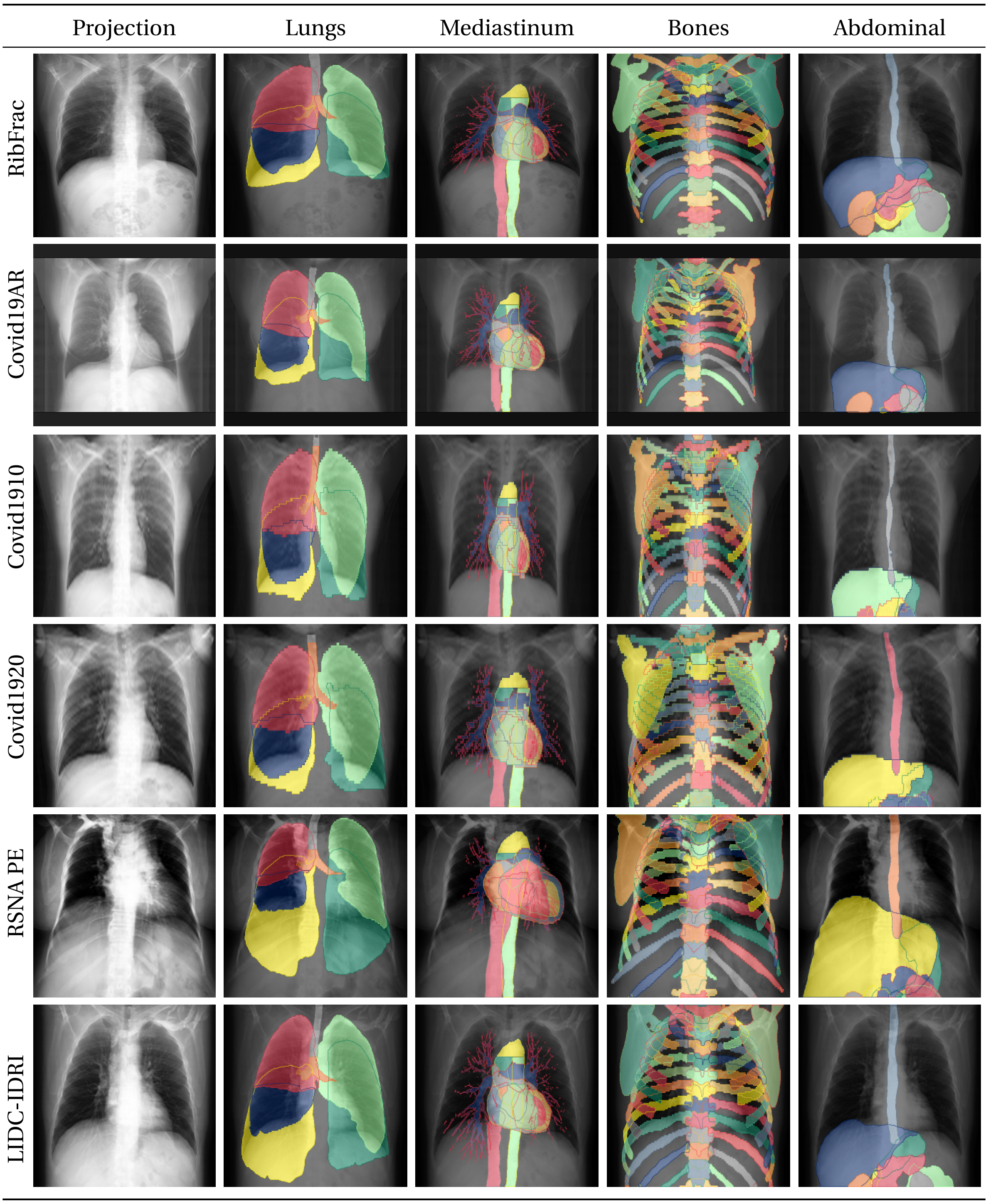


Figure 38: Sample of frontal projected x-rays from the RibFrac, COVID-19-AR, COVID-19-1, COVID-19-2, RSNA PE, and LIDC-IDRI dataset. We show labels belonging to the categories *Airways*, *Vascular System*, *Bones*, and *Abdomen*.

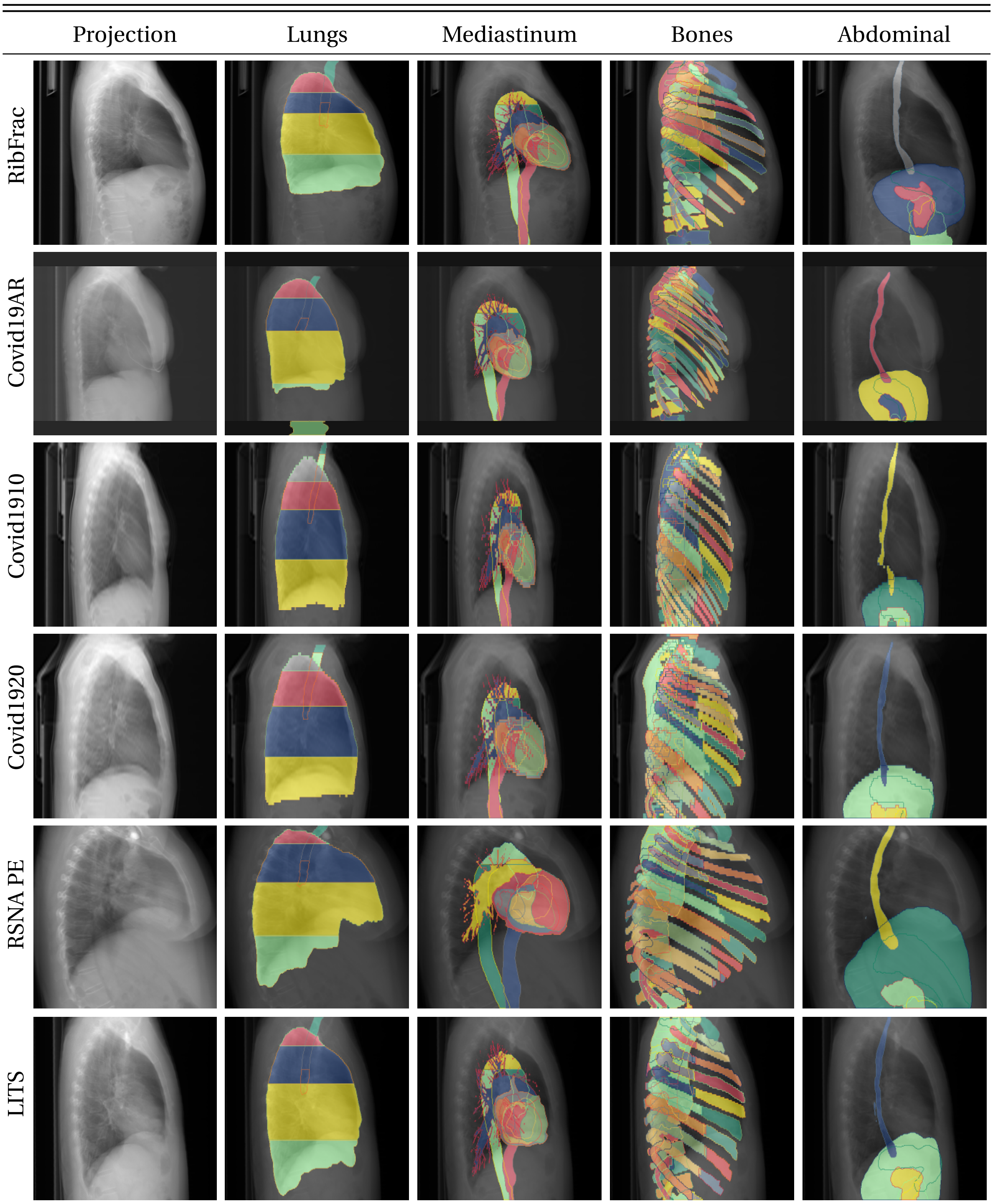
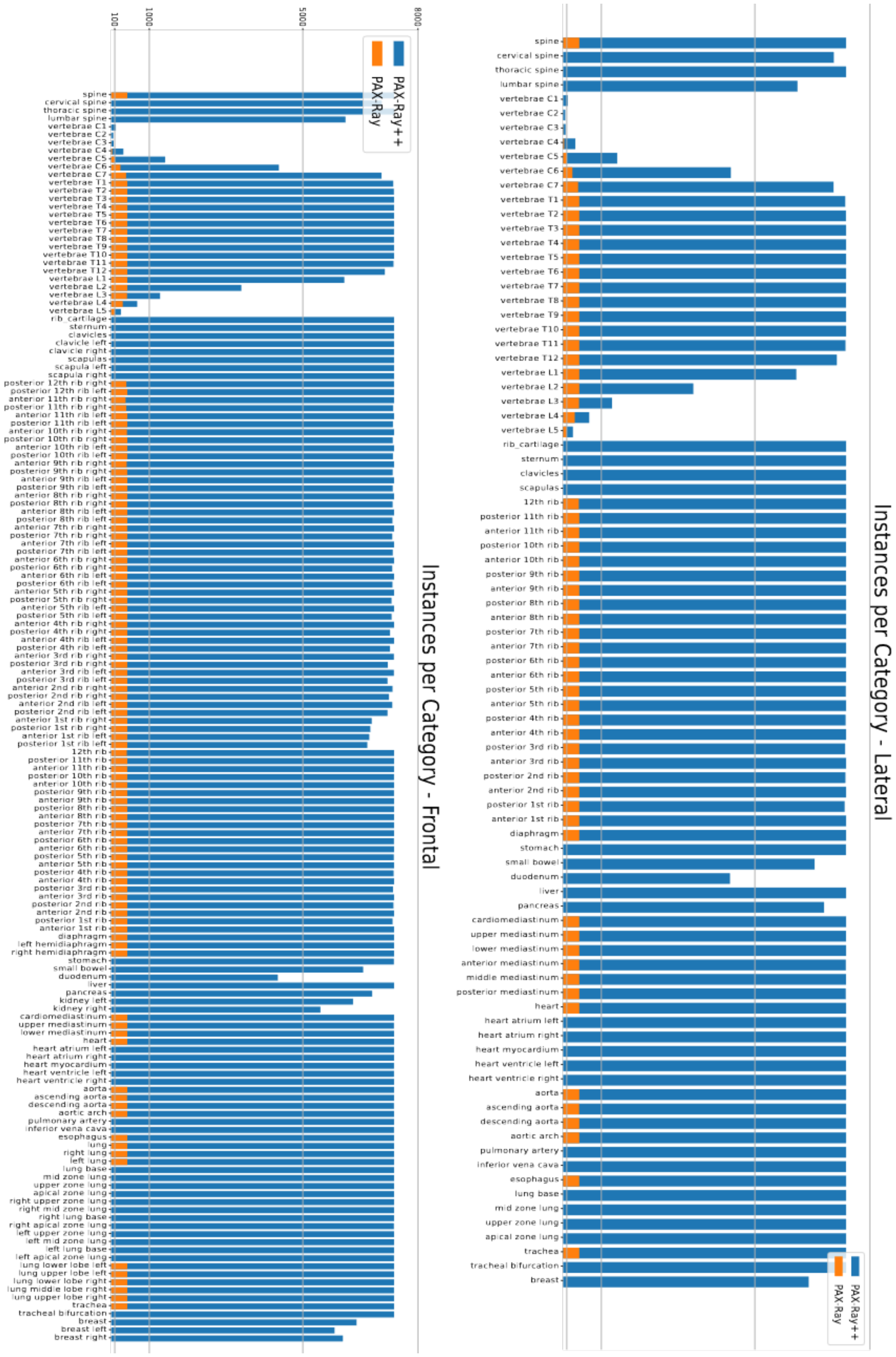
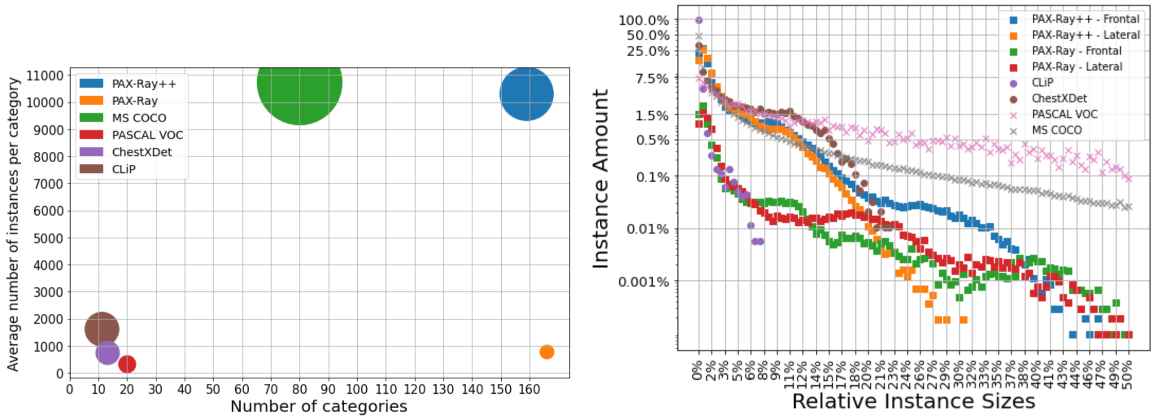


Figure 39: Sample of lateral projected x-rays from the RibFrac, COVID-19-AR, COVID-19-1, COVID-19-2, RSNA PE, and LIDC-IDRI dataset. We show labels belonging to the categories *Airways, Vascular System, Bones,* and *Abdomen.*



(a) (b)

Figure 40: (a) #instances per category for frontal images. (b) #instances per category for lateral images.



(a) (b)

Figure 41: (a) Number of instances per category vs. number of categories. Position indicates mean, size indicates standard deviation. (b) Plot indicating distribution of instance sizes. We compare X-Ray and real world datasets. PAX-Ray++ shows a large amount of labeled instances which tend to be smaller than labeled instances in natural image datasets like COCO or PASCAL VOC.

all instances occur precisely once per image, the lower ribs and the lower and upper vertebrae do not occur in every image. Similarly, abdominal classes such as the duodenum or stomach are not always visible. Finally, it has to be noted that the breast segmentations are independent of the sex of the patient, as also male participants can show sufficient tissue constellations.

We compare our PAX-Ray++ dataset against other datasets within and without the CXR domain. These datasets include ChestXDet [166], a dataset with mask annotations for pulmonary diseases, CLiP [268], a dataset with annotations for catheters and lines, our PAX-Ray, as well as the popular natural image datasets PASCAL VOC [61] and MSCOCO [163].

In Fig. 41a) we compare the number of instances per category against the number of categories. We see that within the domain, PAX-Ray++ provides vastly more different annotations and, in general, provides more annotations than other datasets within the domain, even comparable to real-world datasets like MS COCO.

In Fig. 41b) we display the size of each instance annotation relative to the image size. The natural image datasets show more instances of taking up more prominent parts of the image, with at least 75% and 50% of annotations taking up more than 1% of the image for PASCAL VOC and MS COCO. On the other hand, CXR datasets mainly consist of smaller annotations. We see that nearly all annotations of CLiP are at most 1% of the image size and at most 0.01% of annotations taking up 7%. ChestXDet is more balanced. However, less than 0.1% of annotations take up more than 20% of the image. Related to the domain, PAX-Ray++ also consists majorly of more minor instances. However, it also shows instances of up to 70% of the image.

5.2.3 *Experimental Setup*

5.2.3.1 *CT to X-Ray Projection*

Interpreting the quality of the CT to X-Ray backprojection is non-trivial. We seek a correct translation that maintains all physiological features of the CT. Conversely, we want the resulting images to be as close to the X-Ray domain as possible to ease training and improve downstream segmentation tasks. Unfortunately, these properties can not always be maintained due to the differences in the imaging procedure, such as the patient's position since a CT is taken with the patient lying down while in the typical CXR the patient is standing. To study the quality of the backprojection and the effects of different parts, we utilize the OpenI dataset [52] for the computation of the Fréchet-Inception Distance [99]. We calculate the multivariate normal distributions of real ($\mathcal{N}(\mu, \Sigma)$) and projected images ($\mathcal{N}(\mu_w, \Sigma_w)$) estimated from Inception v3 [265] features.

$$\text{FID} = |\mu - \mu_w|_2^2 + tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{\frac{1}{2}}) \tag{31}$$

We evaluate the FID scores on the projection of the RibFrac dataset. We compute the metrics separately for frontal and lateral views. For the extraction of features, all images are resized to 299×299.

5.2.3.2 *Anatomy Segmentation*

We use 30 frontal and lateral images of PAX-Ray++ for the segmentation of anatomical structures as an internal validation set. These images were manually inspected to be visually correct.

To test the real-world applicability of PAX-Ray++ for developing anatomy segmentation models, we prepare a test set of 30 frontal and lateral X-Rays of the PadChest dataset. First, we use our best segmentation model based on the validation performance to generate preliminary anatomical segmentations. Then, we tasked two medical experts to independently examine and correct these densely annotated X-Rays using the annotation tool CVAT [2]. On average, it was reported that even in this simplified annotation process, a single image could take up to 1.5 hours.

5.2.3.3 *Implementation Details*

For the projection, we utilize Python 3.10 and packages such as NumPy [92] and SciPy [284]. All image projections were originally resized to 512×512 using Lanczos interpolation.

For our segmentation models, we chose the UNet [228] with a ResNet-50 [96] and ViT-B [58] backbone. We used pre-trainings on ImageNet [53] as we did not see a difference in performance to other pre-training methods, i.e., MAE [94] on CXRs. As before, we train with binary cross-entropy and employ an additional binary dice loss. For our base augmentation, we used random resize-and-cropping of range [0.8, 1.2] as augmentation with an image size of 512 and optimized using AdamW [170] with a learning rate of 0.001 for 110 epochs decaying by a factor of 10 at {60, 90, 100} epochs. We utilize RandAugment [48] with $N = 9$ and randomized magnitude for heavy augmentation.

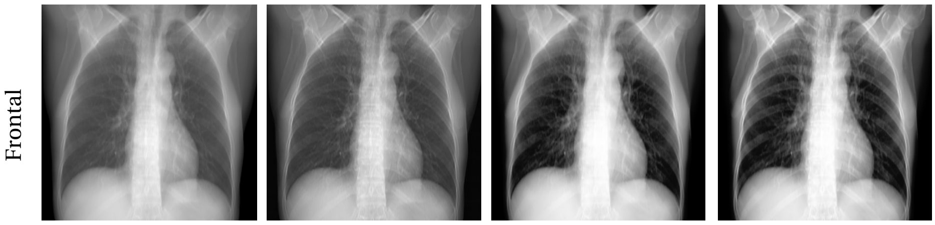
5.2.4 *Results*

5.2.4.1 *CT to X-Ray Projection*

Effect of Translation Method: We first investigate the effect of different translation methods in Fig. 42. Then, we compare the mean intensity projection with Matsubara *et al.* [172] methods and the other effect of histogram equalization (HE). For the frontal view, we see that the MIP and Matsubara *et al.* 's method have nearly the same FID to real X-Rays, but when added with HE, Matsubara *et al.* shows the best FID with strong visible contrast compared to the originally more faded look. In the lateral view, HE is not a positive addition, and the original translation by Matsubara *et al.* leads to the best results. As such, for further projections, we use Matsubara *et al.* 's method with HE for the frontal case and without for the lateral one.

Effect of Volume Spacing: Next, we want to investigate the effect that the spacing of a volume can have on the resulting image quality. This is relevant as the dataset we translate can have a variety of axial spacing, as shown in Table 13.

	Mean-IP	Matsubara *et al.*	Mean-IP + HE	Matsubara *et al.* + HE
FID	145.62	149.42	134.3	126.62



FID	161.41	152.74	164.52	200.207

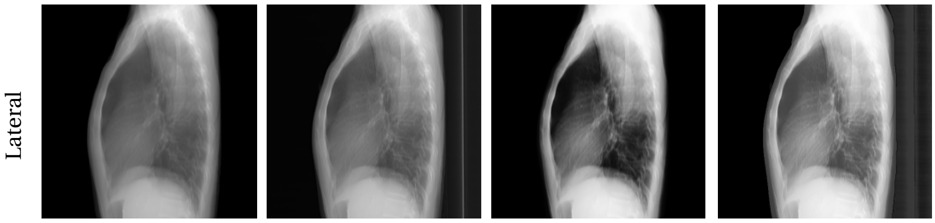
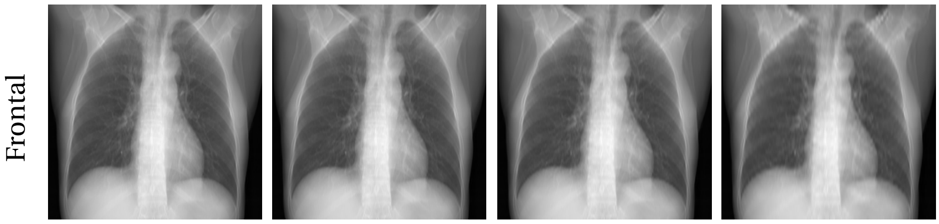


Figure 42: FID scores and qualitative examples of MIP and Matsubara *et al.* 's method with and without histogram equalization (HE) for frontal and lateral views.

We test the spacing Matsubara *et al.* 's method. We see that the FID continually increases with the axial spacing. While the difference between 1 and 2 mm is relatively small, further increasing the spacing leads to noticeable differences metrics-wise, while there it is less noticeable from a qualitative perspective. These results might indicate difficulties for networks processing images from COVID-19-1/2.

	1mm	2mm	3mm	5mm
FID	149.42	160.55	191.11	237.64



FID	152.74	157.54	173.32	214.11

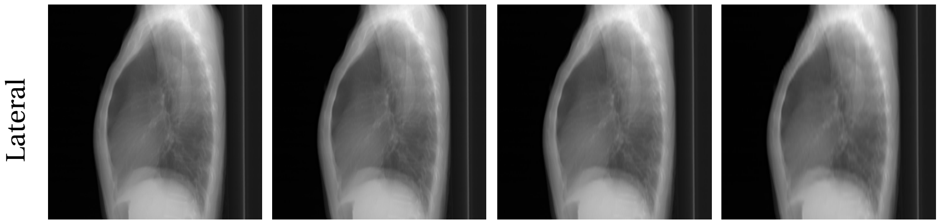


Figure 43: FID scores and qualitative examples of Matsubara *et al.* 's method applied on CT's with different axial spacings for frontal and lateral views. We can see that for CTs with higher spacing the FID noticably increases.

5.2.4.2 *Projected Anatomy Segmentation*

We use our manually examined validation set to develop the best segmentation model to apply to real-world data and show the results in Fig. 44. We first compare the ResNet-50 vs. the ViT-B backbone without using data augmentation. We see that the Resnet-50 outperforms the ViT-B in all classes. Furthermore, we can see minor improvements when adding simple augmentation, which is enhanced when switching to a heavy augmentation setup.

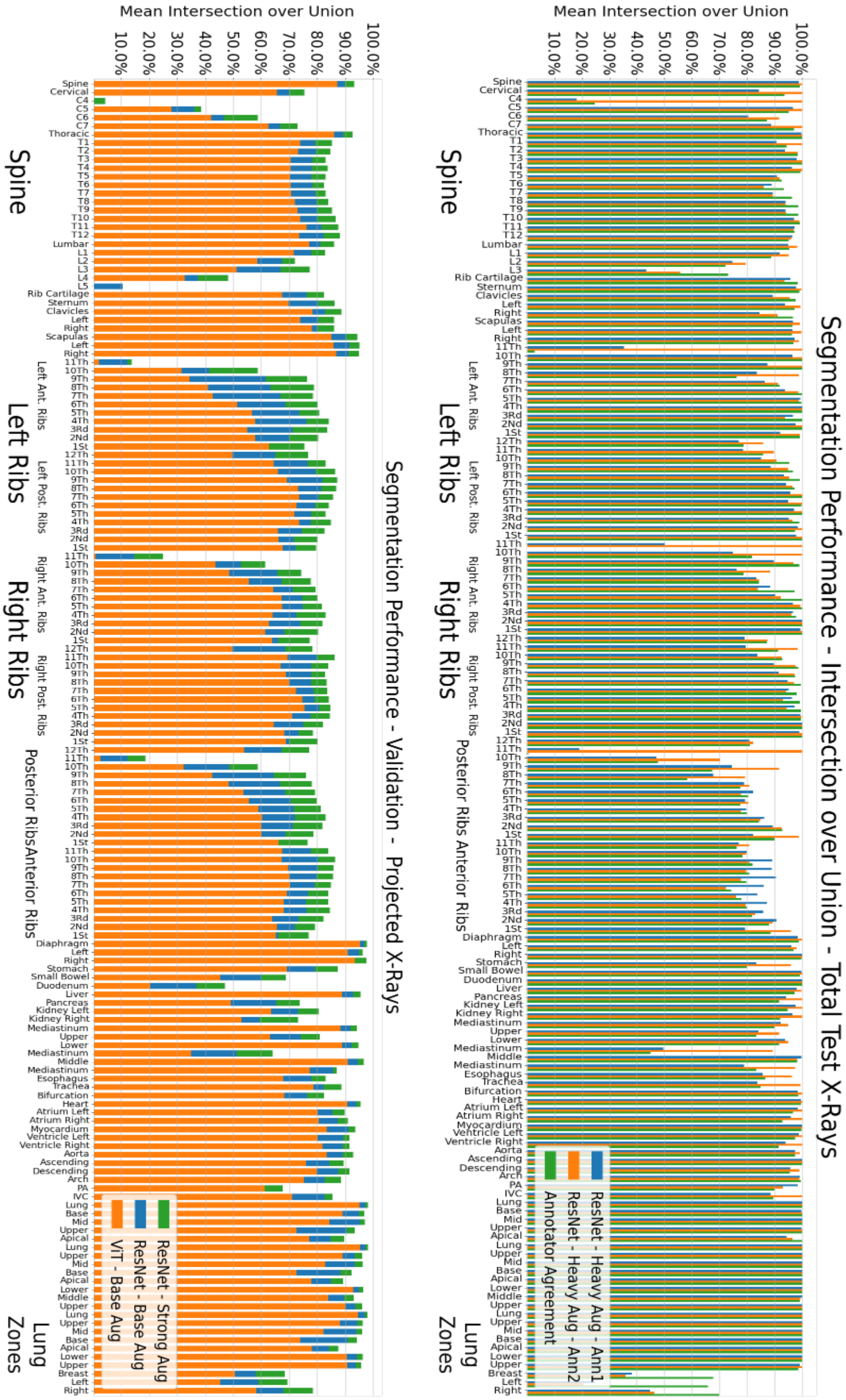
Generally, we see good performances for standard spinal classes, such as the thoracic vertebrae, while rare vertebrae in the visible thoracic region are more complex. It can be noted that classes in the lateral view tend to have better scores than their frontal counterparts, which is also attributed to the fact that frontal classes tend to be finer granular (i.e., left/right split). Abdominal classes can vary in segmentation quality. For example, while the liver or stomach are typically well-segmented, the duodenum and kidneys are more complex. This, however, might stem from the abdominal region in CXR being relatively bright without strongly noticeable delineations, making it difficult to accurately segment manually. Heart-related classes show to have near-perfect segmentations with scores above 90% IoU. Individual lung borders are more challenging to identify, leading to scores between 50-60% IoU.

We show qualitative results in Fig. 45 and Fig. 46. Here, we can see classes belonging to the supercategories lungs, vascular systems, bones and abdomen/digestive system. We see that apart from minor deviations at the arterial structure, the segmentations show near to no deviation to the ground truth for both frontal and lateral.

5.2.4.3 *X-Ray Anatomy Segmentation*

We show the segmentation performance against medical expert annotations for all classes in Fig. 44 b). As the task for the human annotators was not to annotate from scratch but to correct wrong pixel-wise predictions, we can see a high agreement for most classes. The most significant disagreements exist for rare bone structures such as L3 and C4, the mediastinal distribution or the breasts. We show the view-wise performance in Fig. 47. For the frontal classes, high agreement with both annotators can be reached for most classes apart from rare bone structures and the breasts. In the lateral case, this distinction needs to be clarified. For abdominal and mediastinal classes, annotator 2 aligns more with the model's predictions, while annotator 1 notices some flaws. At the same time, the opposite can be seen for the ribs.

We display qualitative results in Fig. 48 and Fig. 49. The annotators tend to be content with most annotations. Edits come with extensions of the esophagus, trachea and aorta and corrections of the lower ribs. There is little consensus for some classes, such as the stomach, as can be seen in Fig. 44 b). In the lateral view, rib segmentations can become quite hard to interpret. While both annotators disagree with the rib segmentations, they do not agree on how they should look. Similarly to the frontal view, tube-like structures like the esophagus are extended as they can appear fractured at times.



(a)

(b)

Figure 44: (a) Segmentation performance in IoU on the validation split for all classes of our proposed PAXRay++ dataset for the settings using ViT-B and ResNet-50 backbones, and the impact data augmentations. (b) Segmentation performance in IoU on real CXR from the PadChest dataset for all classes compared to medical expert annotations.

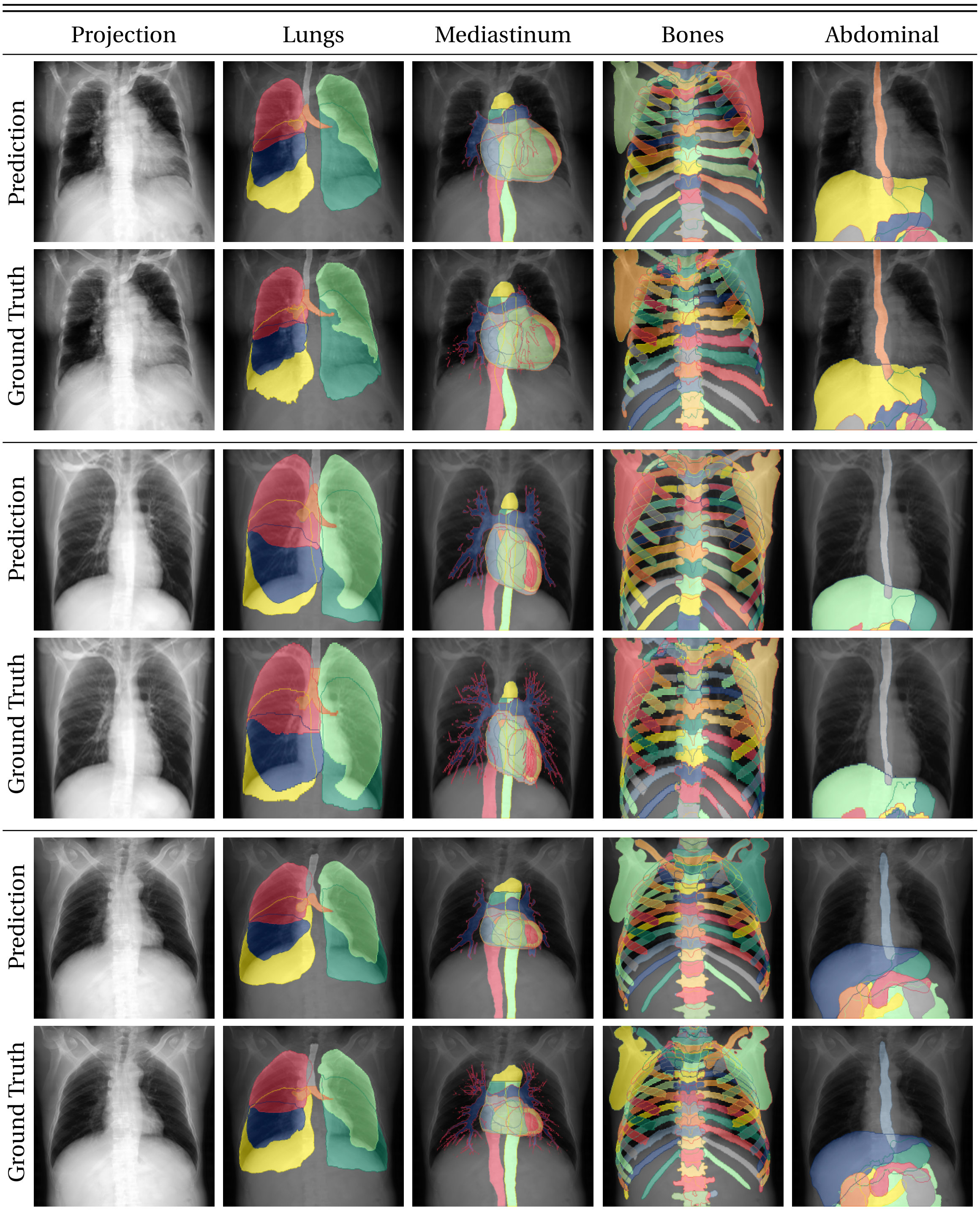


Figure 45: Qualitative segmentation results for lateral projections for the class categories of lungs, vascular systems, bones and abdomen.

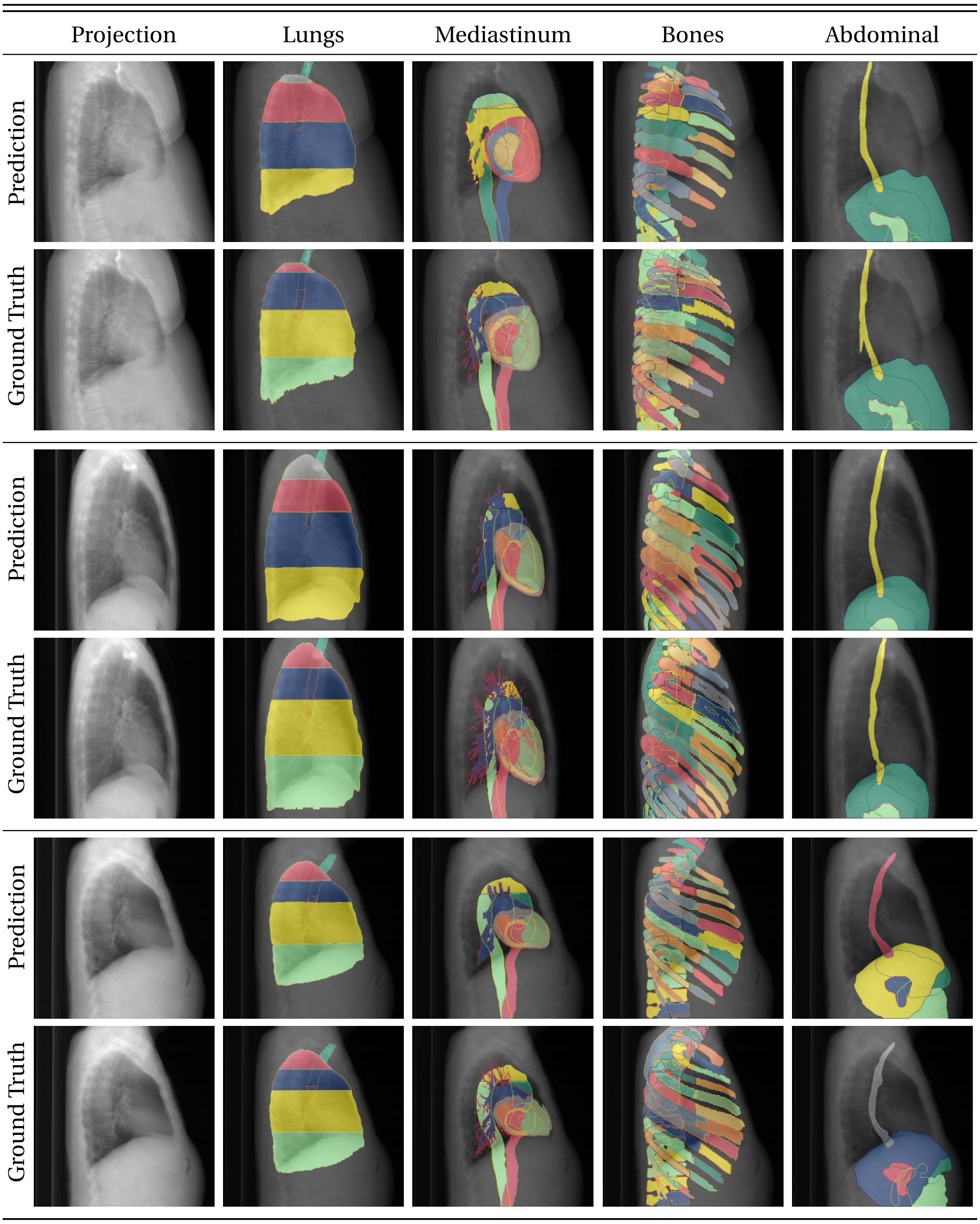
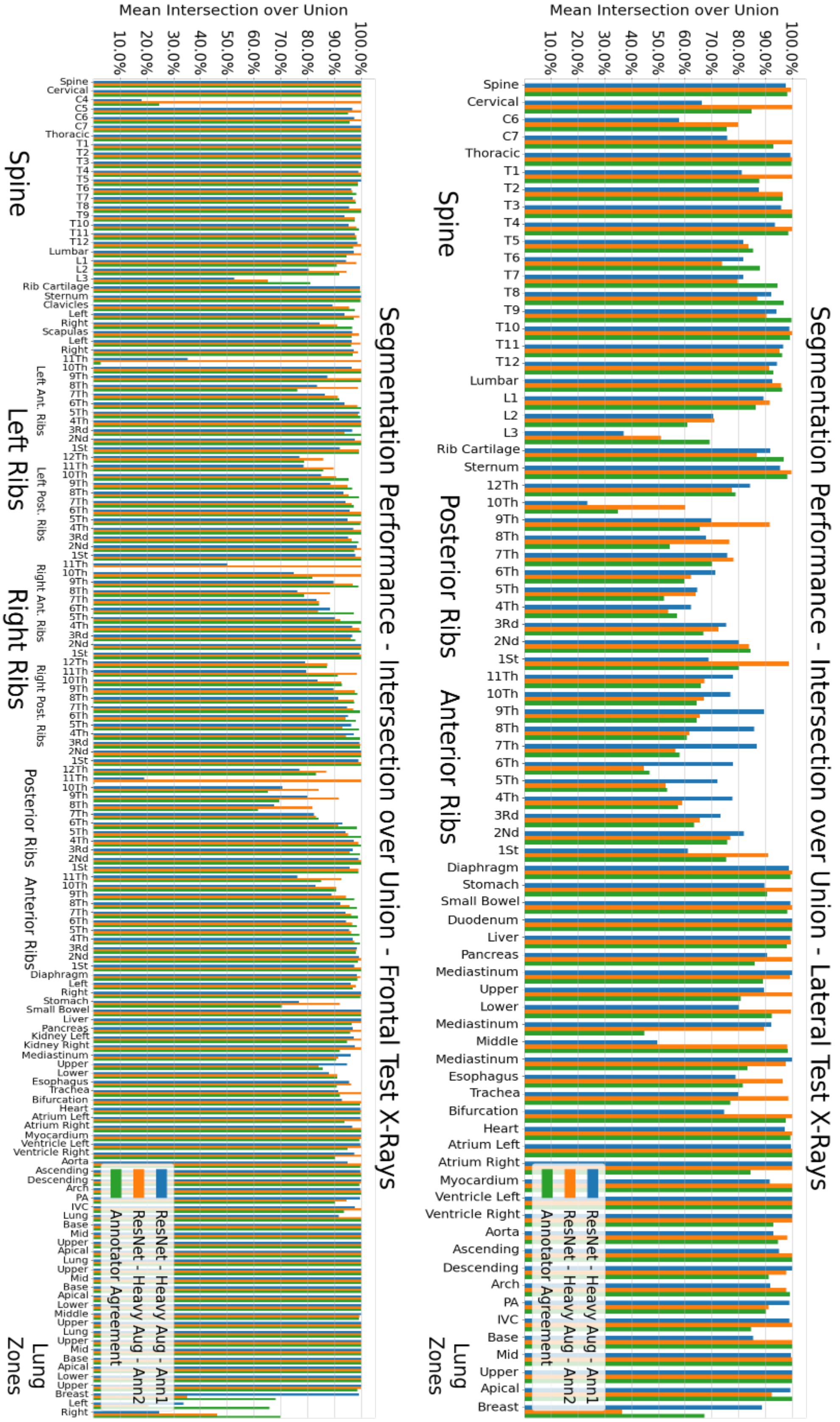


Figure 46: Qualitative segmentation results for lateral projections for the class categories of lungs, vascular systems, bones and abdomen.



(a)

(b)

Figure 47: (a) Segmentation performance in IoU on real CXR from the PadChest dataset for frontal classes. (b) Segmentation performance in IoU on real CXR from the PadChest dataset for lateral classes.

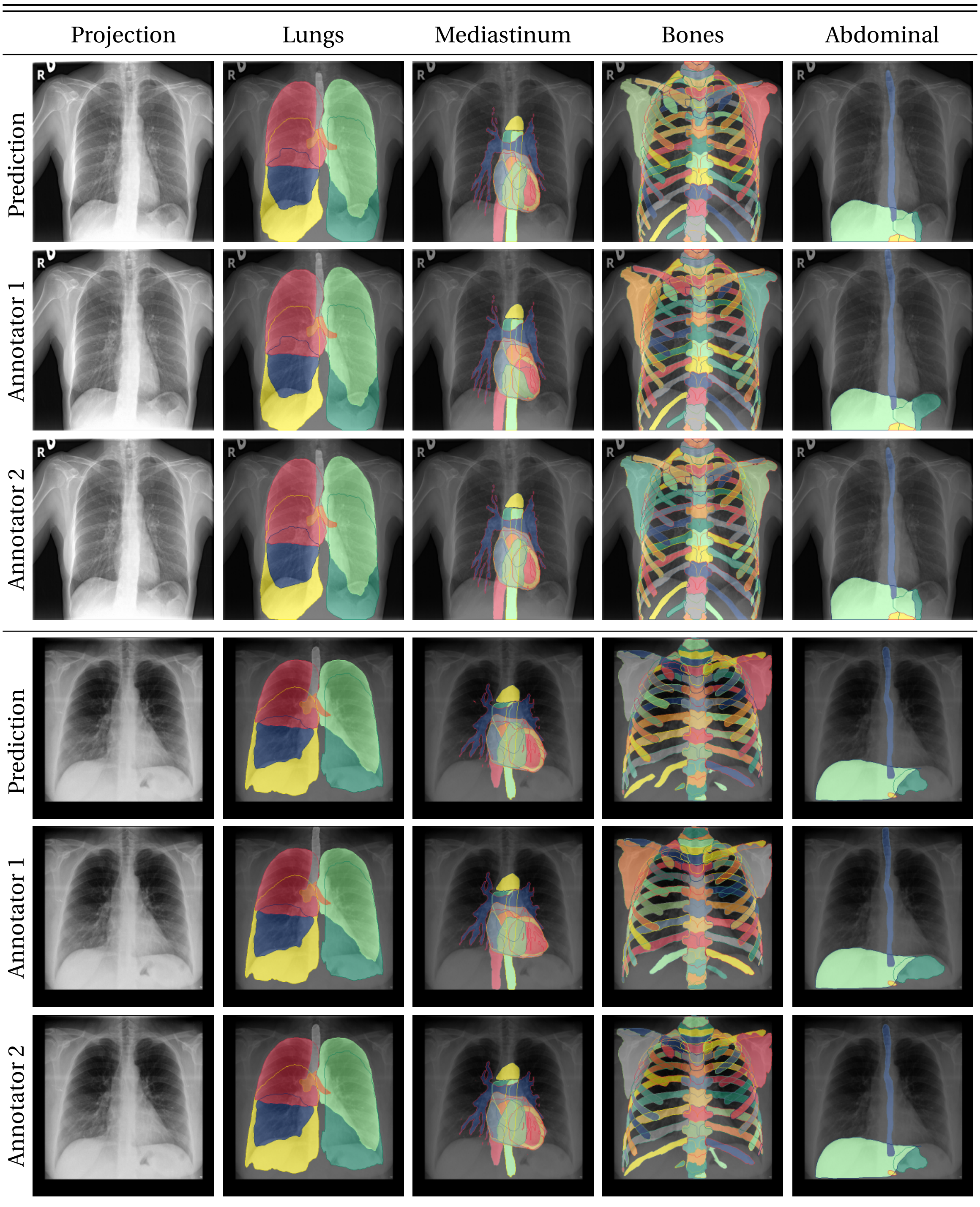


Figure 48: Qualitative segmentation results for frontal CXR for the class categories of lungs, vascular systems, bones and abdomen. We compare against the corrections of Annotator 1 and 2.

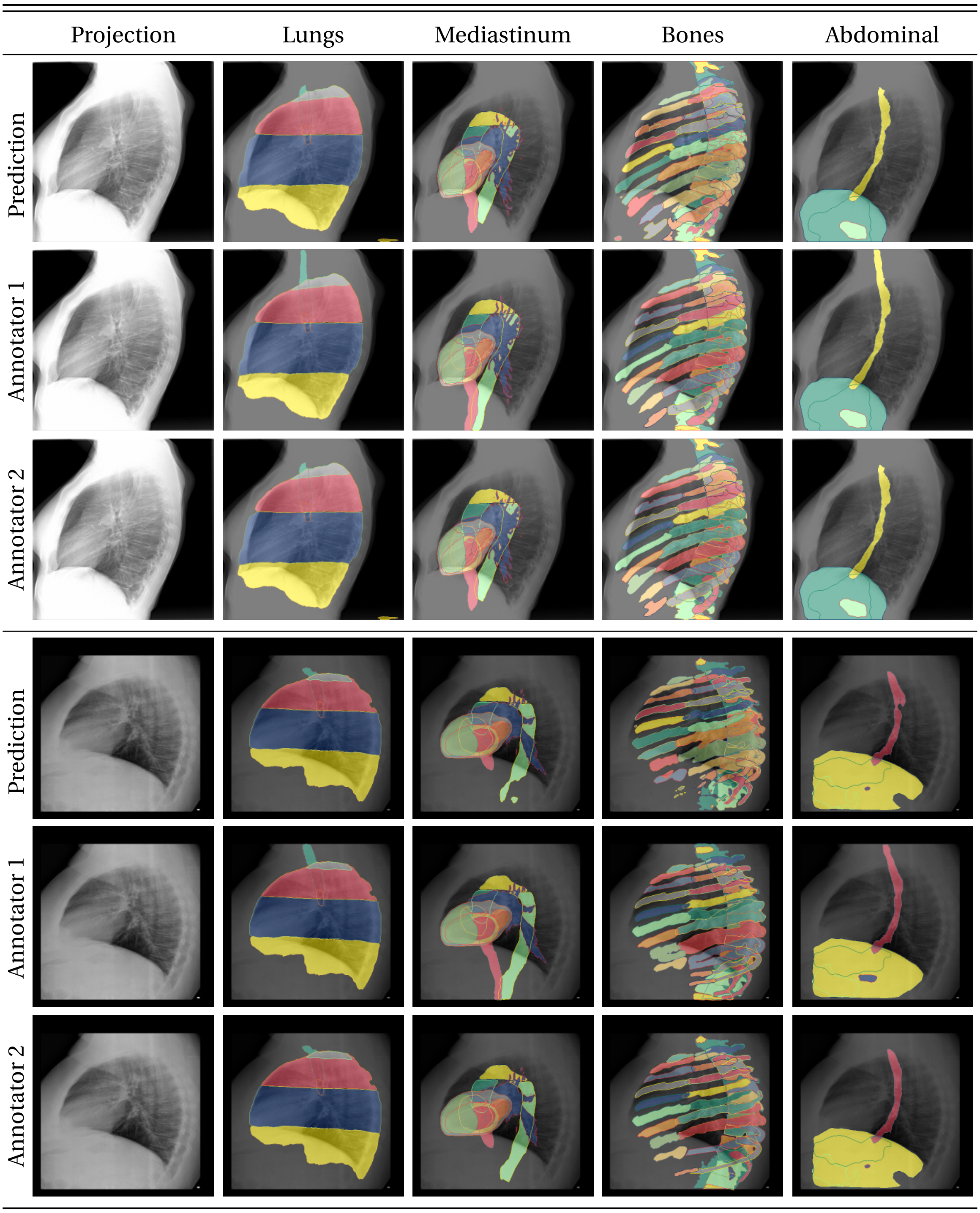


Figure 49: Qualitative segmentation results for frontal CXR for the class categories of lungs, vascular systems, bones and abdomen. We compare against the corrections of Annotator 1 and 2.

5.3 GROUNDING MEDICAL REPORTS VIA ANATOMY DESCRIPTORS

In medical reports, frequently, medical observations are paired with anatomical regions to refer to their respective positions. For example, the finding PULMONARY NODULE OVERLYING THE POSTERIOR SIXTH RIB can be localized using previous automatic anatomical segmentation of the posterior sixth rib. We highlight the potential use for anatomical information through the co-occurrence of anatomy and observations in the case of the Chest X-Ray Dataset OpenI in Fig. 50.

Beginning with the assumption that certain diseases tend to occur in specific anatomical regions, as shown in Fig. 51, we create a simple baseline to argue for the usefulness of incorporating anatomical information when grounding medical findings.

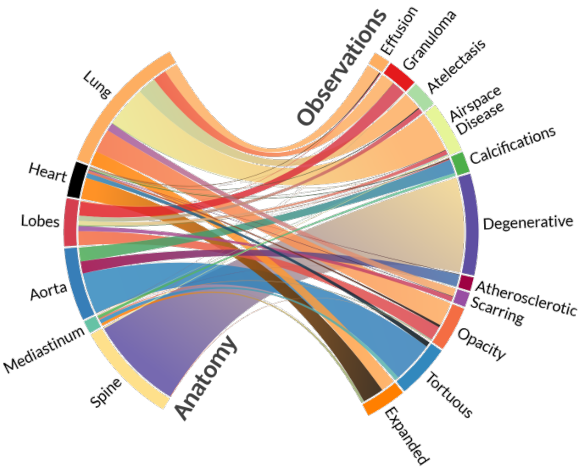


Figure 50: Chord diagramm of the co-occurences of *Anatomies* and *Findings* in medical reports for patients in the OpenI dataset

For each image-report pair $(\mathcal{I}_i, \mathcal{R}_i) \in \{(\mathcal{I}_1, \mathcal{R}_1), (\mathcal{I}_2, \mathcal{R}_2), \ldots, (\mathcal{I}_N, \mathcal{R}_N)\}$ in a dataset consisting of N pairs, we consider the finding-subsection of the report which contains descriptions of visual observations in the corresponding image. As depicted in the top branch of Fig. 51, we use a named-entity-recognition (NER) model to process the report $\mathcal{R}_i$ sentence by sentence, identifying medically relevant phrases $\mathcal{P}_{ij} \in \mathcal{P}_i, 0 \leq j \leq |\mathcal{P}_i|$ that are classified as problem or treatment and discard the rest [122, 279, 332]. We then filter the words $w \in \mathcal{P}_{ij}$ using the NER-model $\mathcal{C}$ to classify w into *Anatomy A* (e. g.heart, . . .), *Anatomy-modifier AM* (e. g.posterior,. . .), *Observation O* (e. g.pneumothorax, . . .) or *Observation-modifier OM* (e. g.above, . . .) [122, 332]. These categories are grouped, and any words that do not belong to any of these categories are omitted, resulting in a filtered phrase $\mathcal{P}^*_{ij}$ containing groups of words W^x_{ij} for each category $x \in \{A, AM, O, OM\}$:

$$\begin{aligned} \mathcal{P}^*_{ij} &= \{W^A_{ij}, W^{AM}_{ij}, W^O_{ij}, W^{OM}_{ij}\} \\ W^x_{ij} &= \{w \mid \mathcal{C}(w) = x, w \in \mathcal{P}_{ij}\}, \end{aligned} \tag{32}$$

In order to obtain embeddings for the filtered phrase $\mathcal{P}^*_{ij}$, we use a pre-trained word-embedding model $\mathcal{E}$ to extract d-dimensional embeddings for all words in the phrase that are classified as anatomy and anatomy modifiers:

$$\begin{aligned} F^A_{ij} &= \{\mathcal{E}(w) \mid w \in W^A_{ij}\} \\ F^{AM}_{ij} &= \{\mathcal{E}(w) \mid w \in W^{AM}_{ij}\} \end{aligned} \tag{33}$$

For phrases that do not contain either anatomy or anatomy modifiers, we set $F^x_{ij} = 0^d$ with $x \in \{A, AM\}$. In cases where multiple words in a phrase belong to the same category (either anatomy or anatomy modifier), we consider the category representation to be the mean of all the word embeddings belonging to that category.

The final phrase embedding is then calculated by adding the category embeddings together, as shown in the following equation:

$$F_{ij} = \text{mean}(F^A_{ij}) + \text{mean}(F^{AM}_{ij}) \tag{34}$$

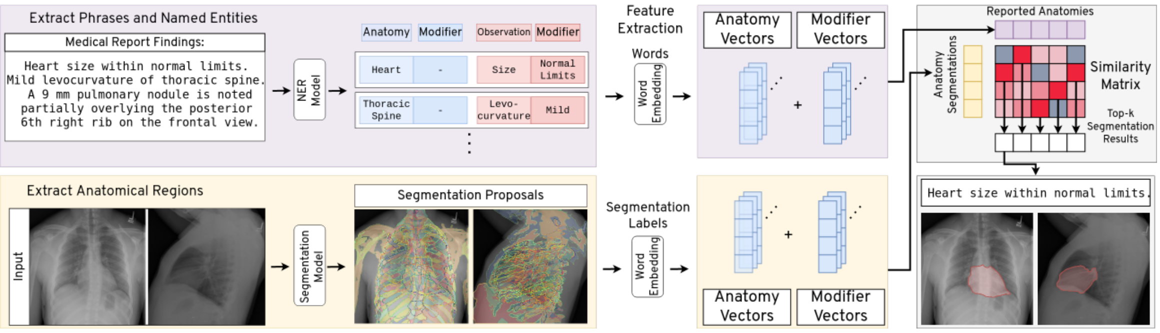


Figure 51: Anatomy Grounding Baseline. Using NER divide phrases into *Anatomy*, *Anatomy-Modifier*, *Observation*, *Observation-Modifier*. For the images, we generate proposals using our anatomy segmentation model. We extract word-wise embeddings for the phrase/anatomy labels, aggregate them and retrieve the most similar region for each phrase

In the bottom branch of Abb. 51, we use our segmentation network trained on PAX-Ray to extract predictions for each class with its class label denoted as $l \in L$ for each view. These predictions are thresholded to obtain mask regions, then refined using anatomical constraints and post-processing steps similar to those described in subsection 5.1.1.3. Finally, for all of the resulting segmentation masks, we separate the label text l into its anatomy and anatomy modifier components. For example, the label 'left lung' will be split into 'lung' as the anatomy and 'left' as its modifier. We extract features T_l in a similar manner as previously mentioned:

$$\begin{aligned} T_l^A &= \{\mathscr{E}(w) \mid \mathscr{C}(w) = A, w \in l\} \\ T_l^{AM} &= \{\mathscr{E}(w) \mid \mathscr{C}(w) = AM, w \in l\} \\ T_l &= mean(T_l^A) + mean(T_l^{AM}) \end{aligned} \tag{35}$$

To find the similarity between image regions and phrases in the lateral and frontal views, we calculate the cosine-similarity matrix $S^i \in [-1,1]^{|\mathscr{P}_i| \times 166}$ using the feature vectors. Each entry in the matrix is defined by $S^i jl = cos(F_{ij}, T_l)$. When a phrase query is given, we use the top-k similarities to return the corresponding k-th segmentation proposal. We return the entire image if the phrase does not contain any anatomy.

5.3.0.1 *Medical Phrase Grounding*

Grounding Dataset: To evaluate medical phrase grounding, we utilized the OpenI dataset [52], which contains medical reports and frontal and lateral chest X-ray images. In order to create a diverse validation dataset for CXR phrase grounding, two radiologists identify medical phrases in the findings-section of 100 medical reports from the OpenI dataset and mark the corresponding part within the image with a bounding box. This resulted in a total of 178 frontal and 146 lateral bounding box annotations. These annotations were then used as ground truth for the task. After applying NER techniques using Stanza, these 100 reports resulted in annotated phrases, including 90 different observations, 70 observation modifiers, 88 different anatomies, and 42 anatomy modifiers.

Evaluation Setup: we evaluate the usefulness of anatomy segmentations for medical phrase grounding by calculating the average hit rate at various IoU thresholds [317]. A hit is a candidate region that overlaps with the ground truth annotation with an IoU above a set threshold. We compare our anatomy segmentations with common region proposal methods used in phrase grounding algorithms [50, 72, 182] in natural images such as EdgeBoxes [339], Selective Search [278], and

	Method(N=200)	HR$_{25}$	HR$_{50}$	HR$_{75}$
Frontal	Whole Image	16.5	5.8	0.4
	Selec. Search [278]	72.8	16.5	7.7
	EdgeBoxes [339]	18.9	4.8	0.9
	RPN [218]	53.8	18.9	1.4
	Anatomy Segm.	**93.2**	**66.9**	**20.8**

	Method(N=200)	HR$_{25}$	HR$_{50}$	HR$_{75}$
Lateral	Whole Image	23.1	8.4	1.0
	Selec. Search [278]	80.7	47.7	19.2
	EdgeBoxes [339]	35.7	11.9	1.8
	RPN [218]	68.8	24.7	0.9
	Anatomy Segm.	**88.0**	**62.3**	**20.1**

Table 14: Hit rates of region proposals for different IoU thresholds (denoted by subscript).

Region Proposal Networks [218]. For each labeled image, we extract the top 200 scoring boxes, following the approach of most phrase grounding methods [50, 72, 317].

Next, we assess the performance of our proposed baseline in terms of top-1/5/10 region retrieval at IoU thresholds of 25/50/75% and compare it to using the entire phrase and just the anatomy itself for comparison with our label. We compare the usage of features of both anatomy and anatomy modifiers (*ModAnat*) from word features for the comparisons of report parts and segmentation labels with a set of baselines. As lower bound, we always return the (*Whole Image*) regardless of input. We include the (*Oracle*) performance using selective search, which would always return a correct box if available, to provide context for the value of our anatomy-based segmentations. This represents the upper bound of weakly supervised methods. In other words, if the proposal method cannot provide good initial hints, the grounding method will not be able to match phrases with their corresponding image region. Next, we take a look at using features extracted from the entire phrase using phrase feature extractors (*PhraseDist*) and just using anatomy text features using word features (*AnatDist*).

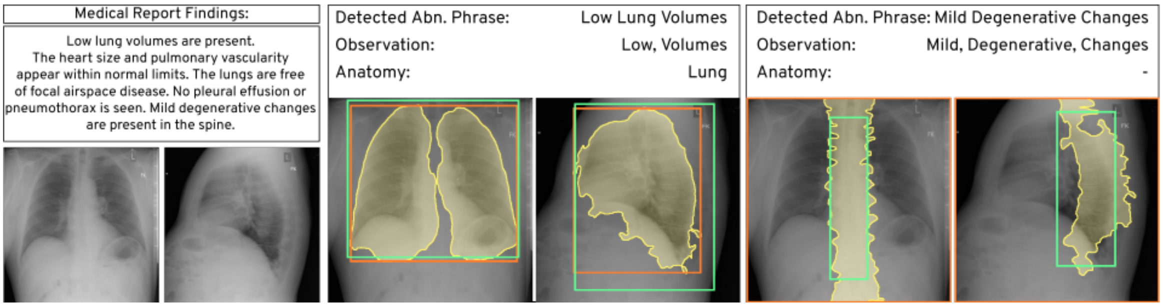
Implementation Details: We use Stanza [331] to process our reports and infer observations and treatments using the i2b2-2010 corpus [279] and anatomies and observations from the Radiology corpus [122, 331]. Additionally, we utilize ChexBert [259] to extract an "is-anomaly" token for each phrase. We use BioWordVec [328] and BioSentVec [31] to extract the word and phrase features. In order to evaluate grounding through bounding box comparison, we extract corresponding bounding boxes for each segmentation result.

5.3.1 *Results for Medical Phrase Grounding*

Hit Rate Analysis: The hit rate (HR) results in Table 14 demonstrate that the selective search algorithm provided the best proposals for traditional approaches in both the frontal and lateral view.

	Method	Box Proposals	Text Features	Top-1$_{25}$	Top-1$_{50}$	Top-1$_{75}$	Top-5$_{50}$	Top-10$_{50}$
Frontal	Whole Image	None	None	18.5	7.1	0.5	7.1	7.1
	Oracle	Sel. Search	None	72.8	16.5	7.7	16.5	16.5
	PhraseDist	Anat. Seg.	BioSent	36.5	17.9	2.9	23.3	27.5
	Anat.Dist	Anat. Seg.	BioWord	34.7	13.1	0.5	26.3	**28.1**
	ModAnat.	Anat. Seg.	BioWord	**38.9**	**21.5**	**4.7**	**27.5**	**28.1**
Lateral	Whole Image	None	None	23.1	8.4	1.0	8.4	8.4
	Oracle	Sel. Search	None	80.7	47.7	19.2	47.7	47.7
	PhraseDist	Anat.Seg.	BioSent	47.3	22.1	4.2	26.3	30.5
	Anat.Dist.	Anat.Seg.	BioWord	45.2	17.8	2.1	30.5	31.5
	ModAnat.	Anat. Seg.	BioWord	**49.4**	**26.3**	**8.4**	**32.6**	**32.6**

Table 15: Medical phrase grounding performance on OpenI showing Top-k region retrieval performance at different IoU thresholds (denoted by the subscript).



Medical Report Findings:

Low lung volumes are present. The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. Mild degenerative changes are present in the spine.

Detected Abn. Phrase: Low Lung Volumes
Observation: Low, Volumes
Anatomy: Lung

Detected Abn. Phrase: Mild Degenerative Changes
Observation: Mild, Degenerative, Changes
Anatomy: -

Figure 52: We show ground truth (green), retrievals (orange) and expected retrieval (yellow). If anatomy phrases are identified a result is provided. Otherwise, the segmentation, albeit accurate, is not retrieved.

However, we notice a significant decline in quality when increasing the IoU threshold. For example, in the frontal view, the hit rate decreases by almost 56%. This is in contrast to the Flicker30K dataset, where the hit rate of selective search at a 50% IoU threshold for 200 boxes was reported as 85.68% [317]. On the other hand, our approach, which is not trained in the segmentation of observations but rather in anatomies, leads to improvements in HR across all categories. For instance, we achieve a 50% improvement in HR for the frontal view at an IoU of 50%. This suggests that anatomy guidance can be a more effective starting point for the localization of observations, especially when considering the HR in relation to the oracle's performance.

Grounding Results: In Table 15, we present our quantitative results for medical phrase grounding. The results demonstrate that the direct sentence comparison and our proposed method outperform the oracle's performance based on selective search proposals for the frontal view at a commonly used IoU threshold of 50%. In addition, utilizing anatomy and its modifiers makes a noticeable improvement over using complete sentence embeddings. Similarly, using features of the anatomy modifier improves over just using the anatomy features with an increase of more than 4% for both views in $Top-1_{25}$-accuracy and 1-2% for $Top-1_{50}$-accuracy. In Figure 52, we provide qualitative results. We highlight anatomy and medical phrases, and we can see that even without direct reference to disease, anatomical regions can be used to retrieve medical findings. We show qualitative results in Figure 52. We highlight anatomy and medical phrases. Despite not directly referring to the disease, anatomical regions can be utilized to retrieve medical findings.

5.4 CHAPTER CONCLUSION

In this chapter, we follow the idea that we can leverage anatomical annotations from various sources in a pixel-wise distinct CT-domain to develop segmentation models in the difficult to annotate X-Ray domain. We build on the shared underlying source of radiation, which allows to construct a projection from a CT volume that resembles a common X-Ray. This allows us to aggregate a vast variety of anatomical labels in CT domain and then project these with the volume to a two dimensional plane to simulate X-ray training data.

We first show in a small scale experiment that this setup can to produce anatomy segmentation models. These can, in turn, be used to ground anatomical information from medical reports. We have shown, that such anatomical segmentations can serve as better region proposals for the grounding of medical phrases than common proposal methods.

After establishing the validity of this approach, we scaled up the PAX-Ray dataset by integrating more relevant labels as well as using more thorax CT's as a basis for translation. We merge labels from ten sources on a seperate dataset via a set of rules and label post-processing. After training on this joint label set, we run inference on six sufficiently large thorax ct datasets to in the end gather more than 7000 pseudolabeled CTs. Training on the projections of these volumes presented us with segmentation models that have a high agreement with medical personal on real CXR data.

We summarize our contributions as:

Contribution 1: We show how we can utilize existing CT annotations for the analysis of CXR. By merging annotations of multiple partially annotated datasets and projecting these onto a two dimensional plane we can effectivly train deep learning based models. This allows us to build the first robust anatomy segmentation model for CXR.

Contribution 2: We build the PAX-Ray++ dataset, the first dataset with dense anatomical annotations in the CXR domain spanning label categories from abdominal and mediastinal regions to specific bones and organs. This datasets contains more than 14.000 images and is with more than 2 million instances comparable to real world datasets such as MS COCO.

Contribution 3: We show the use of anatomical segmentations by applying them to the task of medical phrase grounding. While common region proposal mechanisms fail on this medical domain, anatomical segmentations provide a better anchor for vision-language understanding.

While our approach has enabled the first holistic segmentation of the human anatomy in CXR without the need of *any* manual annotation in that domain, we have to note potential shortcomings. CTs are typically a much scarcer resource than the common CXR making it difficult to scale this approach to the levels of natural image datasets with millions of images due the overall lack of volumetric data. CXR also tend to use a slightly harder type of radiation compared to standard CTs. This fact paired with the common use of contrast agents in CTs might lead to a divergence of visual quality between the projected X-rays and real ones. We, however, believe that this can be salvaged through a more sophisticated use of backprojection either via physical simulations or generative models.

Part III

INSIGHTS

6

IMPACT ON THE FIELD

We require reliable information to ensure proper patient care in the medical field. This thesis has advanced the research in automated medical reporting in several ways. While a large portion of contemporary work focuses on the generation of report texts in an auto-regressive, we aimed to take the field closer to a way of holistic, automated medical report generation by enabling the localization of anatomical and pathological structures. We highlighted new ways to handle unsolved problems for chest X-Ray analysis without the requirement of consulting additional manual annotations and compared our methods when applicable against existing related work. This chapter reflects on our main contributions from the perspectives of opened research directions and datasets.

6.1 NEW RESEARCH DIRECTIONS

OPEN SET PATHOLOGY RECOGNITION: Pathology recognition in chest X-rays is typically bound to automated label extraction from medical reports. This automated labeling process introduces inherent noise into the training data and fixes the model to recognize a predefined set of classes. We show that these issues can be circumvented by training directly on medical reports in a contrastive manner. By formulating an inference scheme based on vision-language similarities, we can directly infer the occurrence of any pathology in question. In Section 3.2, we show the effects of text-based inference in chest radiographs.

MULTI-LABEL PSEUDO-LABELING: Many weakly-semi- and self-supervised learning scenarios employ pseudo-labels to train models on data without strong annotations. These pseudo-labels are typically derived from simple thresholds applied to class predictions. This assumption might be insufficient for multi-label scenarios as we ignore states of other co-occurring classes. We propose to investigate pseudo-label specific for multi-label segmentation and break away from the commonly investigated multi-class setup. We explore this direction in Chapter 4 and show that the other ways of formulating pseudo-labels can be highly effective in domains with overlapping label structures.

CT BACKPROJECTION FOR AUTOMATED X-RAY ANALYSIS: Gathering dense annotations for visually ambiguous data such as X-Rays is a complex and time-consuming task where medical experts need help to reach a unanimous conclusion. This is exaggerated for the generation of large scale-datasets for developing deep learning models for the pixel-wise analysis of chest X-rays. To circumvent this issue, we found that annotated computer tomographs can be utilized as training data for X-Ray analysis due to their visual similarity when projected to a two-dimensional plane. We show in Chapter 5, how we can utilize various annotations in the CT domain to segment chest X-rays, which we published at BMVC 2022 [243].

6.2 NEW DATASETS

At the beginning of this thesis, we established the necessity of anatomical information for the correct assessment of the physiological state of a patient. However, detailed delineation of such is extremely difficult and time-consuming in X-Rays when done manually due to the nature of imaging modality. In this thesis, we pave the way for automated analysis of such structures through the generation of the PAX-Ray and PAX-Ray++ datasets in Chapter 5. Furthermore, we trace 160 anatomical structures in the thoracic region in more than 7000 frontal and lateral images by utilizing annotations in the CT domain. This dataset enables the first-ever detailed anatomical segmentation of chest radiographs. With more than two million annotated instances, PAX-Ray++ is the most detailed dataset in the X-Ray domain, comparable to natural image datasets such as MS COCO. Different natural image datasets, however, that many structures overlap and large instances occur less often.

We have evaluated the potential of our datasets via downstream tasks such as medical phrase grounding and manual evaluation of model predictions on external X-Ray datasets.

6.3 NEW METHODS

SELF-GUIDANCE FOR MULTI-LABEL RECOGNITION AND LOCALIZATION: Knowing which part of the image leads to a model's decision is necessary for the use of pathology recognition models. Unfortunately, many deep learning models inaccurately activate image regions when applied to pathology recognition. In Section 3.1 based on our ACCV 2020 paper [245], we propose our self-guiding loss to aid the network during training by providing artificial supervision on a patch level. Doing so led us to achieve improved classification and localization results.

CONTRASTIVE PROMPTING FOR OPEN SET MULTI-LABEL RECOGNITION: Dual encoder Vision-Language models trained in a contrastive way allow for strong backbones for further fine-tuning in vision or language tasks. However, it is non-trivial to extract the ingrained knowledge from these models directly without architectural changes. Many existing works have proposed to extract text prompts for all classes in question, pit them against each other and choose the prompt most similar to a given image. However, this setting is not applicable in the medical setting as multiple pathologies can co-occur. To handle this, we proposed a contrastive prompting setup in Section 3.2 based on our MICCAI 2022 publiction [246], where we pit each class against its absence to adapt to the multi-label setting. We further show how to extend this for open-set pathology localization.

REFERENCE-GUIDED PSEUDO-LABELING FOR MULTI-LABEL SEGMENTATION: As mentioned, multi-label segmentation can have multiple classes co-occurring at the same pixel compared to multi-class segmentation. While one can still apply standard pseudo-label formulations based on confidence thresholds, other classes are not considered in the prediction. This might lead the network to ignore less common classes. In Section 4, which is based on our AAAI 2022 publication [242], we propose a pseudo-labeling strategy based on nearest neighbors in feature space, which allows us to utilize the full range of supervision for a single pixel. Furthermore, by integrating a weighting scheme based on class entropy, we can effectively let noisy unlabeled samples have less impact during training.

7

FUTURE WORK

In this thesis, we have discussed several of the problems one faces when developing automated reporting systems for chest radiographs. We have developed several approaches that make use of the available data to ease the development process and aid models to become more interpretable through the use of dense predictions. In this chapter, we discuss how our work can build a basis for development of future work that uses our methods for report generation and brings forth additional usable data the training of more sophisticated models.

7.1 AUTOMATED REPORTING

Due to the lack of medical personal, automated reporting systems will play a crucial role in modern healthcare systems, as they will allow for efficient and accurate communication between clinicians and other parties while improving overall time management. Fine-grained dense predictions, which enable the interpretation of medical image data, are a key tool in this process. By leveraging these grounded predictions, deep learning models will better identify subtle patterns and relationships within patient data, enabling more precise and personalized medical reports, while their findings will become well traceable compared to implicitely trained models. In the following, we will take a short look at potential future research directions in this field.

7.1.1 *Dense Predictions for Report Generation*

As described in Section 2.1, a majority of work on automated report generation builds upon implicit training methods due to the lack of dense annotations. This is in contrast to the natural image domain, where the use of dense annotations and resulting data structures provides the foundation of reliable captioning models [123, 189]. We believe that the process of CT backprojection has the potential to alleviate this issue as cases can be confidently labeled in CT and the resulting projections will provide sufficient training data for the analysis of CXR. Our resulting PAX-Ray++ dataset provides a vast amount of available annotations for the development of downstream tasks by identifying subtle patterns and relationships within patient data, enabling more precise and personalized medical reports. This in return, may help reduce the errors and inconsistencies, while provide better interpretability of the used models.

7.1.2 *Automated Extraction of Visual Biomarkers*

Due to the autoregressive training nature most the current reporting systems fail to pick up on quantifiable information or perform reasoning similar to state of the art large language models like GPT [67]. However, doctors can extract visual biomarkers such as the size, brightness or similar features for regions of interest. These can provide valuable information for diagnosis, disease progression, and evaluation of treatment efficacy. This process, while too time-consuming to perform manually, might lead to an improved analysis of patient cohorts and potentially the investigation of diseases. Our work builds a basis for the development of fine-grained densely annotated datasets for CXR. Resulting segmentation networks may provide the required mask information in an accurate and timely manner, thus, improving the reporting process.

7.2 MODELLING OF STRUCTURES FOR COMPUTER TOMOGRAPHY BACK-PROJECTION

In this thesis, we have shown the potential of using CT volumes as a basis for training data for CXR segmentation. This approach worked well, achieving anatomy segmentation performances close to human annotator agreement. Pathological 3D annotations can extend this dataset for CXR instance segmentation. As volume annotations of pathologies can be rare, data augmentation techniques such as Copy-Paste [76]. By transferring annotations with their associated image region to other volumes one can generate a vast amount of CTs viable for backprojection.

Part IV

APPENDIX

A

AUTHORED PUBLICATIONS

This doctoral research resulted in the following thesis-related publications:

1. Seibold, Constantin, Jens Kleesiek, Heinz-Peter Schlemmer, and Rainer Stiefelhagen. "Self-Guided Multiple Instance Learning for Weakly Supervised Thoracic DiseaseClassification and Localizationin Chest Radiographs." In Proceedings of the Asian Conference on Computer Vision. 2020.
2. Seibold, Constantin, Matthias A. Fink, Charlotte Goos, Hans-Ulrich Kauczor, Heinz-Peter Schlemmer, Rainer Stiefelhagen, and Jens Kleesiek. "Prediction of low-keV monochromatic images from polyenergetic CT scans for improved automatic detection of pulmonary embolism." In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1017-1020. IEEE, 2021.
3. Fink, Matthias A., Victoria L. Mayer, Thomas Schneider, Constantin Seibold, Rainer Stiefelhagen, Jens Kleesiek, Tim F. Weber, and Hans-Ulrich Kauczor. "CT angiography clot burden score from data mining of structured reports for pulmonary embolism." Radiology 302, no. 1 (2022): 175-184.
4. Seibold, Constantin Marc, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. "Reference-guided pseudo-label generation for medical semantic segmentation." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, pp. 2171-2179. 2022.
5. Fink, Matthias A., Constantin Seibold, Hans-Ulrich Kauczor, Rainer Stiefelhagen, and Jens Kleesiek. "Jointly Optimized Deep Neural Networks to Synthesize Monoenergetic Images from Single-Energy CT Angiography for Improving Classification of Pulmonary Embolism." Diagnostics 12, no. 5 (2022): 1224.
6. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J. (2022). Breaking with Fixed Set Pathology Recognition Through Report-Guided Contrastive Training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13435. Springer, Cham.

The following scientific publications were co-authored by Constantin Seibold during his PhD research but these are not directly related to the thesis.

1. Reiß, Simon, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. "Graph-Constrained Contrastive Regularization for Semi-weakly Volumetric Segmentation." In European Conference on Computer Vision, pp. 401-419. Springer, Cham, 2022.
2. Reiß, Simon, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. "Every annotation counts: Multi-label deep supervision for medical image segmentation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9532-9542. 2021.

3. Salg, Gabriel Alexander, Maria-Katharina Ganten, Andreas Michael Bucher, Hannes Goetz Kenngott, Matthias Alexander Fink, Constantin Seibold, Ricarda Elisabeth Fischbach et al. "A reporting and analysis framework for structured evaluation of COVID-19 clinical and imaging data." NPJ digital medicine 4, no. 1 (2021): 1-9.

4. Marinov, Zdravko, Stanka Vasileva, Qing Wang, Constantin Seibold, Jiaming Zhang, and Rainer Stiefelhagen. "Pose2Drone: A Skeleton-Pose-based Framework for Human-Drone Interaction." In 2021 29th European Signal Processing Conference (EUSIPCO), pp. 776-780. IEEE, 2021.

5. Roitberg, Alina, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. "Let's play for action: Recognizing activities of daily living by learning from life simulation video games." In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8563-8569. IEEE, 2021.

6. Tan, Haobin, Chang Chen, Xinyu Luo, Jiaming Zhang, Constantin Seibold, Kailun Yang, and Rainer Stiefelhagen. "Flying guide dog: Walkable path discovery for the visually impaired utilizing drones and transformer-based semantic segmentation." In 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1123-1128. IEEE, 2021.

7. Sarfraz, Saquib, Marios Koulakis, Constantin Seibold, and Rainer Stiefelhagen. "Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 336-345. 2022.

8. Sarfraz, M. Saquib, Constantin Seibold, Haroon Khalid, and Rainer Stiefelhagen. "Content and colour distillation for learning image translations with the spatial profile loss." arXiv preprint arXiv:1908.00274 (2019).

9. Roitberg, Alina, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, and Rainer Stiefelhagen. "A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding." arXiv preprint arXiv:2204.04734 (2022).

10. Schölch, Lukas, Jonas Steinhäuser, Maximilian Beichter, Constantin Seibold, Kailun Yang, Merlin Knaeble, Thorsten Schwarz, Alexander Maedche, and Rainer Stiefelhagen. "Towards Automatic Parsing of Structured Visual Content through the Use of Synthetic Data." In 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1607-1613. IEEE, 2022.

BIBLIOGRAPHY

[1] https://ltrcpublic.com/ (cit. on pp. 62, 68).

[2] https://www.cvat.ai/ (cit. on p. 75).

[3] American Board of Radiology. `https://www.theabr.org/diagnostic-radiology`. Dec. 2022 (cit. on p. 3).

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. “Bottom-up and top-down attention for image captioning and visual question answering.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 6077–6086 (cit. on p. 12).

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. “Vqa: Visual question answering.” In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 2425–2433 (cit. on p. 12).

[6] Jonas Appelberg, Gunnar Nordahl, and Christer Janson. “Lung volume and its correlation to nocturnal apnoea and desaturation.” In: Respiratory medicine 94.3 (2000), pp. 233–239 (cit. on p. 11).

[7] Arbeitsbedingungen im Krankenhaus: Burn-out schon beim Nachwuchs. `https://www.aerzteblatt.de/archiv/211051/Arbeitsbedingungen-im-Krankenhaus-Burn-out-schon-beim-Nachwuchs`. Dec. 2022 (cit. on p. 4).

[8] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. “The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans.” In: Medical physics 38.2 (2011), pp. 915–931 (cit. on p. 68).

[9] Ärztemangel in Deutschland. `https://beyondhealth.de/corporate-blog/aerztemangel-in-deutschland`. Dec. 2022 (cit. on p. 4).

[10] Ärztemangel in Deutschland. `https://www.praktischarzt.de/magazin/aerztemangel-deutschland/`. Dec. 2022 (cit. on p. 4).

[11] Zaheer Babar, Twan van Laarhoven, and Elena Marchiori. “Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines.” In: Plos one 16.11 (2021), e0259639 (cit. on pp. 11, 15).

[12] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. “Comparison of deep learning approaches for multi-label chest X-ray classification.” In: Scientific reports 9.1 (2019), pp. 1–10 (cit. on pp. 10, 21, 22).

[13] Jonathan T Barron. “A generalization of otsu’s method and minimum error thresholding.” In: European Conference on Computer Vision. Springer. 2020, pp. 455–470 (cit. on pp. 60, 61).

[14] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. “Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring.” In: arXiv preprint arXiv:1911.09785 (2019) (cit. on pp. 14, 43).

[15] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. “Mixmatch: A holistic approach to semi-supervised learning.” In: arXiv preprint arXiv:1905.02249 (2019) (cit. on pp. 14, 43).

[16] Rossella Bettini, Stefano Partelli, Letizia Boninsegna, Paola Capelli, Stefano Crippa, Paolo Pederzoli, Aldo Scarpa, and Massimo Falconi. “Tumor size correlates with malignancy in nonfunctioning pancreatic endocrine tumor.” In: Surgery 150.1 (2011), pp. 75–82 (cit. on p. 11).

[17] Riddhish Bhalodia, Ali Hatamizadeh, Leo Tam, Ziyue Xu, Xiaosong Wang, Evrim Turkbey, and Daguang Xu. “Improving Pneumonia Localization via Cross-Attention on Medical Images and Reports.” In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 571–581. ISBN: 978-3-030-87196-3 (cit. on p. 10).

[18] Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. “REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays.” In: arXiv e-prints (2021), arXiv–2109 (cit. on pp. 15, 16).

[19] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. “Baselines for chest x-ray report generation.” In: Machine Learning for Health Workshop. PMLR. 2020, pp. 126–140 (cit. on p. 11).

[20] Hrvoje Bogunović et al. “RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge.” In: IEEE Transactions on Medical Imaging 38.8 (2019), pp. 1858–1874. DOI: `10.1109/TMI.2019.2901398` (cit. on p. 44).

[21] Adrian P Brady. “Error and discrepancy in radiology: inevitable or avoidable?” In: Insights into imaging 8.1 (2017), pp. 171–182 (cit. on p. 4).

[22] Burn-out Krise in der Radiologie. `https://www.aerztezeitung.de/Wirtschaft/Burn-out-Krise-in-der-Radiologie-404853.html`. Dec. 2022 (cit. on p. 4).

[23] 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.” In: arXiv preprint arXiv:1901.07441 (2019) (cit. on pp. 10, 15, 21).

[24] 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.” In: Medical image analysis 66 (2020), p. 101797 (cit. on p. 10).

[25] Jinzheng Cai, Le Lu, Adam P Harrison, Xiaoshuang Shi, Pingjun Chen, and Lin Yang. “Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2018, pp. 589–598 (cit. on pp. 10, 21, 22).

[26] Peng Cao, Fulong Ren, Chao Wan, Jinzhu Yang, and Osmar Zaiane. "Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis." In: Computerized Medical Imaging and Graphics 69 (2018), pp. 112–124 (cit. on p. 16).

[27] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. "Machine learning and the physical sciences." In: Reviews of Modern Physics 91.4 (2019), p. 045002 (cit. on p. 4).

[28] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. "Curriculum labeling: Self-paced pseudo-labeling for semi-supervised learning." In: arXiv e-prints (2020), arXiv–2001 (cit. on p. 14).

[29] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE. 2018, pp. 839–847 (cit. on p. 13).

[30] Haomin Chen, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. "Deep Hierarchical Multi-label Classification of Chest X-ray Images." In: Proceedings of Machine Learning Research 1 (2019), p. 13 (cit. on p. 10).

[31] Qingyu Chen, Yifan Peng, and Zhiyong Lu. "BioSentVec: creating sentence embeddings for biomedical texts." In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE. 2019, pp. 1–5 (cit. on p. 86).

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In: International conference on machine learning. PMLR. 2020, pp. 1597–1607 (cit. on p. 34).

[33] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. "Big self-supervised models are strong semi-supervised learners." In: arXiv preprint arXiv:2006.10029 (2020) (cit. on pp. 14, 43).

[34] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. "Pali: A jointly-scaled multilingual language-image model." In: arXiv preprint arXiv:2209.06794 (2022) (cit. on p. 12).

[35] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. "Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 2613–2622 (cit. on p. 14).

[36] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco captions: Data collection and evaluation server." In: arXiv preprint arXiv:1504.00325 (2015) (cit. on p. 12).

[37] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 15750–15758 (cit. on p. 33).

[38] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "Uniter: Universal image-text representation learning." In: European conference on computer vision. Springer. 2020, pp. 104–120 (cit. on p. 12).

[39] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. “BING: Binarized normed gradients for objectness estimation at 300fps.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 3286–3293 (cit. on p. 13).

[40] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. “3D U-Net: learning dense volumetric segmentation from sparse annotation.” In: International conference on medical image computing and computer-assisted intervention. Springer. 2016, pp. 424–432 (cit. on p. 60).

[41] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. “Weakly supervised object localization with multi-fold multiple instance learning.” In: IEEE transactions on pattern analysis and machine intelligence 39.1 (2016), pp. 189–203 (cit. on p. 16).

[42] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. “Multi-column deep neural networks for image classification.” In: 2012 IEEE conference on computer vision and pattern recognition. IEEE. 2012, pp. 3642–3649 (cit. on p. 4).

[43] Clinical radiology UK workforce census 2016 report. `https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-2016-report`. Dec. 2022 (cit. on p. 4).

[44] Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. “The RSNA pulmonary embolism CT dataset.” In: Radiology: Artificial Intelligence 3.2 (2021), e200254 (cit. on p. 68).

[45] Mark S Cook and Anthony J Weinhaus. “Anatomy of the Thoracic Wall, Pulmonary Cavities, and Mediastinum.” In: Handbook of Cardiac Anatomy, Physiology, and Devices (2015), pp. 35–60 (cit. on p. 62).

[46] Marius Cordts et al. “The cityscapes dataset for semantic urban scene understanding.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 3213–3223 (cit. on pp. 13, 43).

[47] Heather D Couture, James Stephen Marron, Charles M Perou, Melissa A Troester, and Marc Niethammer. “Multiple instance learning for heterogeneous images: Training a cnn for histopathology.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2018, pp. 254–262 (cit. on p. 16).

[48] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. “Randaugment: Practical automated data augmentation with a reduced search space.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, pp. 702–703 (cit. on pp. 14, 48, 75).

[49] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. “The epic-kitchens dataset: Collection, challenges and baselines.” In: IEEE Transactions on Pattern Analysis and Machine Intelligence 43.11 (2020), pp. 4125–4141 (cit. on p. 13).

[50] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. "Align2ground: Weakly supervised phrase grounding guided by image-caption alignment." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 2601–2610 (cit. on pp. 85, 86).

[51] Terrance DeVries and Graham W Taylor. "Improved regularization of convolutional neural networks with cutout." In: arXiv preprint arXiv:1708.04552 (2017) (cit. on p. 48).

[52] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. "Preparing a collection of radiology examinations for distribution and retrieval." In: Journal of the American Medical Informatics Association 23.2 (2016), pp. 304–310 (cit. on pp. 10, 15, 21, 75, 85).

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: 2009 IEEE conference on computer vision and pattern recognition. Ieee. 2009, pp. 248–255 (cit. on p. 75).

[54] Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Piroon Jenjaroenpun, Thomas Powell, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri Blake, et al. "Chest imaging representing a COVID-19 positive rural US population." In: Scientific data 7.1 (2020), pp. 1–6 (cit. on p. 68).

[55] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. "Weakly supervised cascaded convolutional networks." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 914–922 (cit. on p. 13).

[56] Die Sorge um Ärztemangel wächst. `https://www.aerztezeitung.de/Politik/Die-Sorge-um-Aerztemangel-waechst-253732.html`. Dec. 2022 (cit. on p. 4).

[57] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." In: Artificial intelligence 89.1-2 (1997), pp. 31–71 (cit. on p. 16).

[58] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." In: arXiv preprint arXiv:2010.11929 (2020) (cit. on p. 75).

[59] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. "Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation." In: IEEE Access 7 (2019), pp. 99065–99076 (cit. on p. 14).

[60] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. "Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model." In: Machine Learning for Health. PMLR. 2021, pp. 209–219 (cit. on p. 11).

[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html (cit. on p. 74).

[62] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. "Discovering faster matrix multiplication algorithms with reinforcement learning." In: Nature 610.7930 (2022), pp. 47–53 (cit. on p. 4).

[63] Luis Felipe Zeni and Claudio R Jung. "Distilling Knowledge From Refinement in Multiple Instance Detection Networks." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, pp. 768–769 (cit. on p. 16).

[64] Ross W Filice, Anouk Stein, Carol C Wu, Veronica A Arteaga, Stephen Borstelmann, Ramya Gaddikeri, Maya Galperin-Aizenberg, Ritu R Gill, Myrna C Godoy, Stephen B Hobbs, et al. "Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset." In: Journal of digital imaging 33.2 (2020), pp. 490–496 (cit. on pp. 15, 16).

[65] Matthias A Fink, Victoria L Mayer, Thomas Schneider, Constantin Seibold, Rainer Stiefelhagen, Jens Kleesiek, Tim F Weber, and Hans-Ulrich Kauczor. "CT angiography clot burden score from data mining of structured reports for pulmonary embolism." In: Radiology 302.1 (2022), pp. 175–184 (cit. on pp. 3, 4, 11).

[66] Elliot K Fishman, Derek R Ney, David G Heath, Frank M Corl, Karen M Horton, and Pamela T Johnson. "Volume rendering versus maximum intensity projection in CT angiography: what works best, when, and why." In: Radiographics 26.3 (2006), pp. 905–922 (cit. on p. 60).

[67] Luciano Floridi and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." In: Minds and Machines 30 (2020), pp. 681–694 (cit. on p. 93).

[68] Daniel Forsberg, Beverly Rosipko, and Jeffrey L Sunshine. "Radiologists' variation of time to read across different procedure types." In: Journal of digital imaging 30.1 (2017), pp. 86–94 (cit. on p. 4).

[69] Daniel G French, Michael Dilena, Simon LaPlante, Farid Shamji, Sudhir Sundaresan, James Villeneuve, Andrew Seely, Donna Maziak, and Sebastien Gilbert. "Optimizing postoperative care protocols in thoracic surgery: best evidence and new technology." In: Journal of Thoracic Disease 8.Suppl 1 (2016), S3 (cit. on p. 3).

[70] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. "Consistency regularization and cutmix for semi-supervised semantic segmentation." In: arXiv preprint arXiv:1906.01916 2.4 (2019), p. 5 (cit. on p. 14).

[71] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. "Devise: A deep visual-semantic embedding model." In: Advances in neural information processing systems 26 (2013) (cit. on p. 34).

[72] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. "Multimodal compact bilinear pooling for visual question answering and visual grounding." In: arXiv preprint arXiv:1606.01847 (2016) (cit. on pp. 85, 86).

[73] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. "Vision-language pre-training: Basics, recent advances, and future trends." In: arXiv preprint arXiv:2210.09263 (2022) (cit. on p. 12).

[74] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. "C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 9834–9843 (cit. on p. 13).

[75] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. "A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions." In: Scientific Data 9.1 (2022), p. 601 (cit. on p. 68).

[76] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. "Simple copy-paste is a strong data augmentation method for instance segmentation." In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, pp. 2918–2928 (cit. on p. 94).

[77] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256 (cit. on p. 49).

[78] Cristina González, Nicolás Ayobi, Isabela Hernandez, José Hernández, Jordi Pont-Tuset, and Pablo Arbelaez. "Panoptic Narrative Grounding." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 1364–1373 (cit. on p. 12).

[79] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. "Knowledge distillation: A survey." In: International Journal of Computer Vision 129.6 (2021), pp. 1789–1819 (cit. on p. 43).

[80] Ruffin J. Graham, Moulay A. Meziane, Thomas W. Rice, Thirugnanam Agasthian, Neil Christie, Kathleen Gaebelein, and Nancy A. Obuchowski. "Postoperative portable chest radiographs: Optimum use in thoracic surgery." In: The Journal of Thoracic and Cardiovascular Surgery 115.1 (1998), pp. 45–52. ISSN: 0022-5223. DOI: https://doi.org/10.1016/S0022-5223(98)70441-6. URL: https://www.sciencedirect.com/science/article/pii/S0022522398704416 (cit. on p. 3).

[81] Yves Grandvalet and Yoshua Bengio. "Semi-supervised learning by entropy minimization." In: Advances in neural information processing systems 17 (2004) (cit. on p. 14).

[82] Yves Grandvalet, Yoshua Bengio, et al. "Semi-supervised learning by entropy minimization." In: CAP. 2005, pp. 281–296 (cit. on p. 14).

[83] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. "A survey of deep learning techniques for autonomous driving." In: Journal of Field Robotics 37.3 (2020), pp. 362–386 (cit. on p. 13).

[84] Michaela Grosser. "Burnout im Krankenhaus: Ursachen, Folgen und Prävention." In: Die auszehrende Organisation. Springer, 2014, pp. 209–237 (cit. on p. 4).

[85] Qingji Guan and Yaping Huang. "Multi-label Chest X-ray Image Classification via Category-wise Residual Attention Learning." In: Pattern Recognition Letters (2018). DOI: https://doi.org/10.1016/j.patrec.2018.10.027 (cit. on pp. 10, 21).

[86] Sebastian Guendel, Florin C Ghesu, Sasa Grbic, Eli Gibson, Bogdan Georgescu, Andreas Maier, and Dorin Comaniciu. “Multi-task Learning for Chest X-ray Abnormality Classification on Noisy Labels.” In: arXiv preprint arXiv:1905.06362 (2019) (cit. on p. 10).

[87] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. “On calibration of modern neural networks.” In: International Conference on Machine Learning. PMLR. 2017, pp. 1321–1330 (cit. on p. 14).

[88] Agrim Gupta, Piotr Dollar, and Ross Girshick. “Lvis: A dataset for large vocabulary instance segmentation.” In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, pp. 5356–5364 (cit. on p. 15).

[89] Ralph Gutmann, Michael Leunig, Jens Feyh, Alwin E Goetz, Konrad Messmer, Ernst Kastenbauer, and Rakesh K Jain. “Interstitial hypertension in head and neck tumors in patients: correlation with tumor size.” In: Cancer research 52.7 (1992), pp. 1993–1995 (cit. on p. 11).

[90] Tarek N Hanna, Haris Shekhani, Kiran Maddu, Chao Zhang, Zhengjia Chen, and Jamlik-Omari Johnson. “Structured report compliance: effect on audio dictation time, report length, and total radiologist study time.” In: Emergency radiology 23.5 (2016), pp. 449–453 (cit. on pp. 4, 5, 9).

[91] Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. “Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets.” In: Nature communications 11.1 (2020), pp. 1–7 (cit. on p. 68).

[92] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. “Array programming with NumPy.” In: Nature 585.7825 (2020), pp. 357–362 (cit. on p. 75).

[93] Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. “How to create a great radiology report.” In: RadioGraphics 40.6 (2020), pp. 1658–1670 (cit. on pp. 4, 5).

[94] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. “Masked autoencoders are scalable vision learners.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 16000–16009 (cit. on p. 75).

[95] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn.” In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 2961–2969 (cit. on p. 21).

[96] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770–778 (cit. on pp. 12, 27, 75).

[97] JF Healthcare. “Object-CXR - Automatic detection of foreign objects on chest X-rays.” In: (). URL: `https://web.archive.org/web/20201127235812/https://jfhealthcare.github.io/object-CXR/` (cit. on p. 15).

[98] Carl P Herbort Jr, Ilknur Tugal-Tutkun, Piergiorgio Neri, Carlos Pavésio, Sumru Onal, and Phuc LeHoang. “Failure to integrate quantitative measurement methods of ocular inflammation hampers clinical practice and trials on new therapies for posterior uveitis.” In: Journal of Ocular Pharmacology and Therapeutics 33.4 (2017), pp. 263–277 (cit. on p. 11).

[99] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. “Gans trained by a two time-scale update rule converge to a local nash equilibrium.” In: Advances in neural information processing systems 30 (2017) (cit. on p. 75).

[100] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory.” In: Neural computation 9.8 (1997), pp. 1735–1780 (cit. on pp. 11, 12).

[101] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem.” In: European Radiology Experimental 4.1 (2020), pp. 1–13 (cit. on p. 62).

[102] Benjamin Hou, Georgios Kaissis, Ronald M. Summers, and Bernhard Kainz. “RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting.” In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 293–303. ISBN: 978-3-030-87234-2 (cit. on p. 10).

[103] Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. “Ratchet: Medical transformer for chest x-ray diagnosis and reporting.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2021, pp. 293–303 (cit. on p. 11).

[104] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. “Self-erasing network for integral object attention.” In: Advances in Neural Information Processing Systems. 2018, pp. 549–559 (cit. on p. 24).

[105] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. “Densely connected convolutional networks.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 4700–4708 (cit. on p. 11).

[106] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. “Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition.” In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 3942–3951 (cit. on p. 34).

[107] Maximilian Ilse, Jakub M Tomczak, and Max Welling. “Attention-based deep multiple instance learning.” In: arXiv preprint arXiv:1802.04712 (2018) (cit. on pp. 16, 22, 26, 27).

[108] Maximilian Ilse, Jakub M Tomczak, and Max Welling. “Deep multiple instance learning for digital histopathology.” In: Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier, 2020, pp. 521–546 (cit. on p. 16).

[109] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: International conference on machine learning. PMLR. 2015, pp. 448–456 (cit. on p. 49).

[110] 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.” In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019, pp. 590–597 (cit. on pp. 10, 15, 21, 22, 36).

[111] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. “Label propagation for deep semi-supervised learning.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 5070–5079 (cit. on pp. 14, 43, 47).

[112] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.” In: Nature methods 18.2 (2021), pp. 203–211 (cit. on pp. 57, 60, 61, 66, 68).

[113] Allan Jabri, Armand Joulin, and Laurens van der Maaten. “Revisiting visual question answering baselines.” In: European conference on computer vision. Springer. 2016, pp. 727–739 (cit. on p. 12).

[114] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. “Automatic tuberculosis screening using chest radiographs.” In: IEEE transactions on medical imaging 33.2 (2013), pp. 233–245 (cit. on p. 15).

[115] Xu Ji, João F Henriques, and Andrea Vedaldi. “Invariant information clustering for unsupervised image classification and segmentation.” In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 9865–9874 (cit. on pp. 14, 54).

[116] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. “Scaling up visual and vision-language representation learning with noisy text supervision.” In: International Conference on Machine Learning. PMLR. 2021, pp. 4904–4916 (cit. on pp. 12, 31).

[117] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. “Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation.” In: International conference on medical image computing and computer-assisted intervention. Springer. 2018, pp. 777–785 (cit. on p. 14).

[118] Liang Jin et al. “Deep-Learning-Assisted Detection and Segmentation of Rib Fractures from CT Scans: Development and Validation of FracNet.” In: EBioMedicine (2020) (cit. on pp. 15, 62, 66, 68).

[119] Baoyu Jing, Pengtao Xie, and Eric Xing. “On the automatic generation of medical imaging reports.” In: arXiv preprint arXiv:1711.08195 (2017) (cit. on p. 11).

[120] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. `https://doi.org/10.13026/8360-t248`. 2019 (cit. on p. 15).

[121] Alistair EW 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 de-identified publicly available database of chest radiographs with free-text reports." In: Scientific data 6.1 (2019), pp. 1–8 (cit. on pp. 3, 15, 16, 21, 36).

[122] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. "MIMIC-III, a freely accessible critical care database." In: Scientific data 3.1 (2016), pp. 1–9 (cit. on pp. 84, 86).

[123] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 4565–4574 (cit. on p. 93).

[124] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. "Image retrieval using scene graphs." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 3668–3678 (cit. on p. 13).

[125] Petr Jordan, Philip M Adamson, Vrunda Bhattbhatt, Surabhi Beriwal, Sangyu Shen, Oskar Radermecker, Supratik Bose, Linda S Strain, Michael Offe, David Fraley, et al. "Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours." In: Medical Physics (2022) (cit. on p. 68).

[126] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold." In: Nature 596.7873 (2021), pp. 583–589 (cit. on p. 4).

[127] Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. "Toward best practices in radiology reporting." In: Radiology 252.3 (2009), pp. 852–856 (cit. on pp. 4, 5, 9).

[128] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. "Universal semi-supervised semantic segmentation." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 5259–5270 (cit. on p. 14).

[129] Neena Kapoor, Ronilda Lacson, and Ramin Khorasani. "Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools." In: Journal of the American College of Radiology 17.11 (2020). Focus on Practical AI, pp. 1363–1370. ISSN: 1546-1440. DOI: `https://doi.org/10.1016/j.jacr.2020.08.016`. URL: `https://www.sciencedirect.com/science/article/pii/S1546144020308760` (cit. on p. 4).

[130] Michael T Kassin, Nicole Varble, Maxime Blain, Sheng Xu, Evrim B Turkbey, Stephanie Harmon, Dong Yang, Ziyue Xu, Holger Roth, Daguang Xu, et al. "Generalized chest CT and lab curves throughout the course of COVID-19." In: Scientific reports 11.1 (2021), pp. 1–13 (cit. on p. 68).

[131] Brendan S Kelly, Louise A Rainford, Sarah P Darcy, Eoin C Kavanagh, and Rachel J Toomey. "The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification." In: Radiology 280.1 (2016), pp. 252–260 (cit. on pp. 3, 4).

[132] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." In: Advances in neural information processing systems 31 (2018) (cit. on p. 12).

[133] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." In: arXiv preprint arXiv:1412.6980 (2014) (cit. on pp. 27, 49).

[134] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. "Panoptic segmentation." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 9404–9413 (cit. on p. 12).

[135] Sven Koitka, Lennard Kroll, Eugen Malamutmann, Arzu Oezcelik, and Felix Nensa. "Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks." In: European radiology 31.4 (2021), pp. 1795–1804 (cit. on pp. 62, 68).

[136] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D Plumbley. "Sound event detection and time–frequency segmentation from weakly labelled data." In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.4 (2019), pp. 777–787 (cit. on p. 16).

[137] Soheil Kooraki, Melina Hosseiny, Lee Myers, and Ali Gholamrezanezhad. "Coronavirus (COVID-19) outbreak: what the department of radiology should know." In: Journal of the American college of radiology 17.4 (2020), pp. 447–451 (cit. on p. 4).

[138] Philipp Krähenbühl. "Free supervision from video games." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 2955–2964 (cit. on p. 14).

[139] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: International journal of computer vision 123.1 (2017), pp. 32–73 (cit. on p. 12).

[140] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images." In: (2009) (cit. on p. 27).

[141] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: Communications of the ACM 60.6 (2017), pp. 84–90 (cit. on pp. 4, 12, 21).

[142] Harold L Kundel and Calvin F Nodine. "Interpreting chest radiographs without visual search." In: Radiology 116.3 (1975), pp. 527–532 (cit. on p. 4).

[143] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. "SegTHOR: Segmentation of thoracic organs at risk in CT images." In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE. 2020, pp. 1–6 (cit. on pp. 57, 62, 68).

[144] Bianca Lassen, Jan-Martin Kuhnigk, Michael Schmidt, Stefan Krass, and Heinz-Otto Peitgen. “Lung and lung lobe segmentation methods at Fraunhofer MEVIS.” In: Proceedings of the Fourth International Workshop on Pulmonary Image Analysis. 2011, pp. 185–199 (cit. on pp. 4, 60, 62).

[145] Bianca Lassen, Eva M van Rikxoort, Michael Schmidt, Sjoerd Kerkstra, Bram van Ginneken, and Jan-Martin Kuhnigk. “Automatic segmentation of the pulmonary lobes from chest CT scans based on fissures, vessels, and bronchi.” In: IEEE transactions on medical imaging 32.2 (2012), pp. 210–222 (cit. on p. 60).

[146] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition.” In: Proceedings of the IEEE 86.11 (1998), pp. 2278–2324 (cit. on pp. 21, 26, 27).

[147] Dong-Hyun Lee et al. “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.” In: Workshop on challenges in representation learning, ICML. Vol. 3. 2013 (cit. on pp. 14, 43, 48, 51).

[148] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. “Weakly supervised multiple instance learning histopathological tumor segmentation.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2020, pp. 470–479 (cit. on p. 16).

[149] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. “Knowledge-driven encode, retrieve, paraphrase for medical image report generation.” In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019, pp. 6666–6673 (cit. on pp. 10, 21).

[150] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. “Rethinking pseudo labels for semi-supervised object detection.” In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 2. 2022, pp. 1314–1322 (cit. on p. 14).

[151] Jun Li, Shibo Li, Ying Hu, and Huiren Tao. “A Self-guided Framework for Radiology Report Generation.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2022, pp. 588–598 (cit. on p. 11).

[152] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. “Align before fuse: Vision and language representation learning with momentum distillation.” In: Advances in neural information processing systems 34 (2021), pp. 9694–9705 (cit. on p. 12).

[153] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. “Visualbert: A simple and performant baseline for vision and language.” In: arXiv preprint arXiv:1908.03557 (2019) (cit. on p. 12).

[154] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. “Weakly supervised object detection with segmentation collaboration.” In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 9735–9744 (cit. on p. 13).

[155] Xin Li, Rui Cao, and Dongxiao Zhu. “Vispi: Automatic Visual Perception and Interpretation of Chest X-rays.” In: arXiv preprint arXiv:1906.05190 (2019) (cit. on pp. 10, 21).

[156] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. “Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm.” In: arXiv preprint arXiv:2110.05208 (2021) (cit. on p. 33).

[157] Yazhao Li, Yanwei Pang, Jianbing Shen, Jiale Cao, and Ling Shao. “NETNet: Neighbor Erasing and Transferring Network for Better Single Shot Object Detection.” In: arXiv preprint arXiv:2001.06690 (2020) (cit. on pp. 10, 21).

[158] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. “Hybrid retrieval-generation reinforced agent for medical image report generation.” In: Advances in neural information processing systems. 2018, pp. 1530–1540 (cit. on p. 11).

[159] 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.” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 8290–8299 (cit. on pp. 10, 16, 21, 22, 30).

[160] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. “A Structure-Aware Relation Network for Thoracic Diseases Detection and Segmentation.” In: IEEE Transactions on Medical Imaging 40.8 (2021), pp. 2042–2052 (cit. on pp. 15, 16).

[161] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. “Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network.” In: IEEE transactions on neural networks and learning systems 30.11 (2019), pp. 3484–3495 (cit. on pp. 16, 22).

[162] Hans Liebl, David Schinz, Anjany Sekuboyina, Luca Malagutti, Maximilian T Löffler, Amirhossein Bayat, Malek El Husseini, Giles Tetteh, Katharina Grau, Eva Niederreiter, et al. “A Computed Tomography Vertebral Segmentation Dataset with Anatomical Variations and Multi-Vendor Scanner Data.” In: arXiv preprint arXiv:2103.06360 (2021) (cit. on p. 61).

[163] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context.” In: European conference on computer vision. Springer. 2014, pp. 740–755 (cit. on pp. 12, 15, 74).

[164] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. “Gps-net: Graph property sensing network for scene graph generation.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 3746–3753 (cit. on pp. 12, 13).

[165] Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. “Deep metric transfer for label propagation with limited annotated data.” In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019, pp. 0–0 (cit. on pp. 14, 47).

[166] Jingyu Liu, Jie Lian, and Yizhou Yu. ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities. 2020. arXiv: `2006.10550v3 [eess.IV]` (cit. on p. 74).

[167] 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: Proceedings of the IEEE International Conference on Computer Vision. 2019, pp. 10632–10641 (cit. on pp. 10, 21, 22, 30).

[168] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "Ssd: Single shot multibox detector." In: European conference on computer vision. Springer. 2016, pp. 21–37 (cit. on p. 21).

[169] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. "A vertebral segmentation dataset with fracture grading." In: Radiology: Artificial Intelligence 2.4 (2020), e190138 (cit. on p. 61).

[170] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization." In: arXiv preprint arXiv:1711.05101 (2017) (cit. on pp. 37, 75).

[171] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. "Weakly-supervised action localization with expectation-maximization multi-instance learning." In: European conference on computer vision. Springer. 2020, pp. 729–745 (cit. on p. 16).

[172] Naoki MATSUBARA, Atsushi TERAMOTO, Kuniaki SAITO, and Hiroshi FUJITA. "Generation of pseudo chest X-ray images from computed tomographic images by nonlinear transformation and bone enhancement." In: Medical Imaging and Information Sciences 36.3 (2019), pp. 141–146 (cit. on pp. 60, 69, 75).

[173] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. "Abdomenct-1k: Is abdominal organ segmentation a solved problem." In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) (cit. on p. 68).

[174] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images." In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 1–9 (cit. on p. 12).

[175] MD Mamlouk, PC Chang, and RR Saket. "Contextual radiology reporting: a new approach to neuroradiology structured templates." In: American Journal of Neuroradiology 39.8 (2018), pp. 1406–1414 (cit. on p. 4).

[176] Brian McFee, Justin Salamon, and Juan Pablo Bello. "Adaptive pooling operators for weakly labeled sound event detection." In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.11 (2018), pp. 2180–2193 (cit. on pp. 16, 22, 26).

[177] Des McMorrow and Jens Als-Nielsen. Elements of modern X-ray physics. John Wiley & Sons, 2011 (cit. on p. 44).

[178] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. "The contextual loss for image transformation with non-aligned data." In: Proceedings of the European conference on computer vision (ECCV). 2018, pp. 768–783 (cit. on p. 47).

[179] Bjoern H Menze et al. “The multimodal brain tumor image segmentation benchmark (BRATS).” In: IEEE transactions on medical imaging 34.10 (2014), pp. 1993–2024 (cit. on p. 43).

[180] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. “End-to-end learning of visual representations from uncurated instructional videos.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 9879–9889 (cit. on pp. 33, 39).

[181] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” In: arXiv preprint arXiv:1301.3781 (2013) (cit. on p. 12).

[182] Mehdi Moradi, Ali Madani, Yaniv Gur, Yufan Guo, and Tanveer Syeda-Mahmood. “Bimodal network architectures for automatic generation of image annotation from text.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2018, pp. 449–456 (cit. on p. 85).

[183] Veronica Morfi and Dan Stowell. “Data-efficient weakly supervised learning for low-resource audio event detection using deep learning.” In: arXiv preprint arXiv:1807.06972 (2018) (cit. on pp. 16, 23).

[184] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. “SLIP: Self-supervision meets Language-Image Pre-training.” In: arXiv preprint arXiv:2112.12750 (2021) (cit. on pp. 33, 39).

[185] Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge. `https://doi.org/10.7303/syn3193805`. Dec. 2022 (cit. on pp. 4, 68).

[186] NHS England: Diagnostic imaging dataset statistical release. `https://www.england.nhs.uk/`. Apr. 2020 (cit. on p. 21).

[187] Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. “Variational Topic Inference for Chest X-Ray Report Generation.” In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 625–635. ISBN: 978-3-030-87199-4 (cit. on p. 10).

[188] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. “VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations.” In: arXiv preprint arXiv:2012.15029 (2020) (cit. on pp. 15, 16, 62).

[189] Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. “In defense of scene graphs for image captioning.” In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 1407–1416 (cit. on p. 93).

[190] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. “Im2text: Describing images using 1 million captioned photographs.” In: Advances in neural information processing systems 24 (2011) (cit. on p. 12).

[191] Yassine Ouali, Céline Hudelot, and Myriam Tami. "Semi-supervised semantic segmentation with cross-consistency training." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 12674–12684 (cit. on pp. 14, 43).

[192] Sascha Pahernik, Stefanie Ziegler, Frederik Roos, Sebastian W Melchior, and Joachim W Thüroff. "Small renal tumors: correlation of clinical and pathological features with tumor size." In: The Journal of urology 178.2 (2007), pp. 414–417 (cit. on p. 11).

[193] Beomhee Park, Yongwon Cho, Gaeun Lee, Sang Min Lee, Young-Hoon Cho, Eun Sol Lee, Kyung Hee Lee, Joon Beom Seo, and Namkug Kim. "A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities." In: Scientific reports 9.1 (2019), pp. 1–9 (cit. on pp. 10, 21, 22).

[194] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in PyTorch." In: NIPS-W. 2017 (cit. on p. 27).

[195] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library." In: Advances in neural information processing systems 32 (2019) (cit. on p. 50).

[196] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation." In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014, pp. 1532–1543 (cit. on p. 12).

[197] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. "Combined Scaling for Zero-shot Transfer Learning." In: arXiv preprint arXiv:2111.10050 (2021) (cit. on pp. 12, 31, 33).

[198] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. "Meta pseudo labels." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 11557–11568 (cit. on pp. 14, 43).

[199] Viet-Quoc Pham, Nao Mishima, and Toshiaki Nakasu. "Improving Visual Question Answering by Semantic Segmentation." In: Artificial Neural Networks and Machine Learning – ICANN 2021. Ed. by Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter. Cham: Springer International Publishing, 2021, pp. 459–470. ISBN: 978-3-030-86365-4 (cit. on p. 10).

[200] Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. "Clinically correct report generation from chest x-rays using templates." In: International Workshop on Machine Learning in Medical Imaging. Springer. 2021, pp. 654–663 (cit. on p. 9).

[201] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. "Adaptive histogram equalization and its variations." In: Computer vision, graphics, and image processing 39.3 (1987), pp. 355–368 (cit. on p. 69).

[202] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 2641–2649 (cit. on p. 12).

[203] Dan Popa, Florin Pop, Cristina Serbanescu, and Aniello Castiglione. “Deep learning model for home automation and energy reduction in a smart home environment platform.” In: Neural Computing and Applications 31.5 (2019), pp. 1317–1337 (cit. on p. 13).

[204] Chinmay Prabhakar, Anjany Sekuboyina, Johannes C Paetzold, Hongwei Bran Li, Tamaz Amiranashvili, Jens Kleesiek, and Bjoern Menze. “Improving Generalized Zero-shot Learning Using Knowledge Graphs for Multi-label Chest X-ray Classification.” In: (2022) (cit. on p. 11).

[205] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. “Learning transferable visual models from natural language supervision.” In: International Conference on Machine Learning. PMLR. 2021, pp. 8748–8763 (cit. on pp. 12, 31–35, 37, 39).

[206] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. “Language models are unsupervised multitask learners.” In: OpenAI blog 1.8 (2019), p. 9 (cit. on p. 11).

[207] Radiology review. `https://www.cqc.org.uk/sites/default/files/20180718-radiology-reporting-review-report-final-for-web.pdf`. Dec. 2022 (cit. on p. 4).

[208] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists.” In: PLoS medicine 15.11 (2018), e1002686 (cit. on pp. 10, 21).

[209] 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.” In: arXiv preprint arXiv:1711.05225 (2017) (cit. on pp. 10, 21, 22).

[210] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text-conditional image generation with clip latents.” In: arXiv preprint arXiv:2204.06125 (2022) (cit. on p. 4).

[211] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. “Zero-shot text-to-image generation.” In: International Conference on Machine Learning. PMLR. 2021, pp. 8821–8831 (cit. on p. 4).

[212] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. “Semi-supervised learning with scarce annotations.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, pp. 762–763 (cit. on p. 43).

[213] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement.” In: arXiv preprint arXiv:1804.02767 (2018) (cit. on p. 21).

[214] Jeremie Reeb, Pierre-Emmanuel Falcoz, Anne Olland, and Gilbert Massard. "Are daily routine chest radiographs necessary after pulmonary surgery in adult patients?" In: Interactive cardiovascular and thoracic surgery 17.6 (2013), pp. 995–998 (cit. on p. 3).

[215] Bruce I. Reiner, Nancy Knight, and Eliot L. Siegel. "Radiology Reporting, Past, Present, and Future: The Radiologist's Perspective." In: Journal of the American College of Radiology 4.5 (2007), pp. 313–319. ISSN: 1546-1440. DOI: `https://doi.org/10.1016/j.jacr.2007.01.015`. URL: `https://www.sciencedirect.com/science/article/pii/S1546144007000270` (cit. on p. 3).

[216] Simon Reiss, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. "Every Annotation Counts: Multi-Label Deep Supervision for Medical Image Segmentation." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021, pp. 9532–9542 (cit. on p. 13).

[217] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. "Every annotation counts: Multi-label deep supervision for medical image segmentation." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 9532–9542 (cit. on pp. 48–50, 54).

[218] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In: Advances in neural information processing systems 28 (2015), pp. 91–99 (cit. on pp. 12, 21, 86).

[219] Mohammad Hosein Rezazade Mehrizi, Peter van Ooijen, and Milou Homan. "Applications of artificial intelligence (AI) in diagnostic radiology: a technography study." In: European radiology 31.4 (2021), pp. 1805–1811 (cit. on p. 4).

[220] Michael von Rhein, Andreas Buchmann, Cornelia Hagmann, Reto Huber, Peter Klaver, Walter Knirsch, and Beatrice Latal. "Brain volumes predict neurodevelopment in adolescents after surgery for congenital heart disease." In: Brain 137.1 (2014), pp. 268–276 (cit. on p. 11).

[221] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList." In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 4902–4912 (cit. on p. 4).

[222] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. "Playing for data: Ground truth from computer games." In: European conference on computer vision. Springer. 2016, pp. 102–118 (cit. on p. 14).

[223] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning." In: arXiv preprint arXiv:2101.06329 (2021) (cit. on pp. 14, 43, 44).

[224] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. “A Primer in BERTology: What We Know About How BERT Works.” In: Transactions of the Association for Computational Linguistics 8 (2020), pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54 (cit. on p. 4).

[225] Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. “Let’s play for action: Recognizing activities of daily living by learning from life simulation video games.” In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2021, pp. 8563–8569 (cit. on p. 14).

[226] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-resolution image synthesis with latent diffusion models.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 10684–10695 (cit. on p. 4).

[227] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: International Conference on Medical image computing and computer-assisted intervention. Springer. 2015, pp. 234–241 (cit. on p. 49).

[228] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: International Conference on Medical image computing and computer-assisted intervention. Springer. 2015, pp. 234–241 (cit. on pp. 60, 75).

[229] Röntgendiagnostik: Häufigkeit und Strahlenexposition für die deutsche Bevölkerung. https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html. Dec. 2022 (cit. on p. 4).

[230] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 3234–3243 (cit. on p. 14).

[231] Eyal Rozenberg, Daniel Freedman, and Alex Bronstein. “Localization with Limited Annotation for Chest X-rays.” In: (2019) (cit. on pp. 10, 21, 22).

[232] Rina D Rudyanto, Sjoerd Kerkstra, Eva M Van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, Ilkay Öksüz, et al. “Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study.” In: Medical image analysis 18.7 (2014), pp. 1217–1232 (cit. on p. 62).

[233] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. “Imagenet large scale visual recognition challenge.” In: International journal of computer vision 115.3 (2015), pp. 211–252 (cit. on p. 27).

[234] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinski. "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning." In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021, pp. 1721–1730 (cit. on p. 16).

[235] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks." In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2018, pp. 990–994 (cit. on p. 11).

[236] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. "Learning from synthetic data: Addressing domain shift for semantic segmentation." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 3752–3761 (cit. on p. 14).

[237] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. "Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization." In: arXiv preprint arXiv:2106.16093 (2021) (cit. on p. 14).

[238] Teri Sippel Schmidt. Template: Rad Chest 2 Views. `https://radreport.org/home/50271/2017-11-14%2017:22:09` (cit. on p. 9).

[239] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." In: arXiv preprint arXiv:2210.08402 (2022) (cit. on pp. 12, 15).

[240] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." In: arXiv preprint arXiv:2111.02114 (2021) (cit. on pp. 12, 15).

[241] Lawrence H Schwartz, David M Panicek, Alexandra R Berk, Yuelin Li, and Hedvig Hricak. "Improving communication of diagnostic radiology findings through structured reporting." In: Radiology 260.1 (2011), p. 174 (cit. on p. 4).

[242] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. "Reference-guided pseudo-label generation for medical semantic segmentation." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 2. 2022, pp. 2171–2179 (cit. on pp. 7, 43, 57, 92).

[243] Constantin Marc Seibold, Simon Reiß, M. Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. "Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding." In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press, 2022. URL: `https://bmvc2022.mpi-inf.mpg.de/0058.pdf` (cit. on pp. 7, 15, 57, 58, 91).

[244] Constantin Seibold, Jens Kleesiek, Heinz-Peter Schlemmer, and Rainer Stiefelhagen. "Self-Guided Multiple Instance Learning for Weakly Supervised Thoracic DiseaseClassification and Localizationin Chest Radiographs." In: Proceedings of the Asian Conference on Computer Vision (ACCV). 2020 (cit. on p. 22).

[245] Constantin Seibold, Jens Kleesiek, Heinz-Peter Schlemmer, and Rainer Stiefelhagen. "Self-guided Multiple Instance Learning for Weakly Supervised Disease Classification and Localization in Chest Radiographs." In: Asian Conference on Computer Vision. Springer. 2020, pp. 617–634 (cit. on pp. 7, 10, 36, 92).

[246] Constantin Seibold, Simon Reiß, M. Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. "Breaking with Fixed Set Pathology Recognition Through Report-Guided Contrastive Training." In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, 2022, pp. 690–700. ISBN: 978-3-031-16443-9 (cit. on pp. 7, 13, 31, 92).

[247] Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. "VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images." In: Medical image analysis 73 (2021), p. 102166 (cit. on pp. 57, 61, 68).

[248] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 618–626 (cit. on pp. 13, 14, 22).

[249] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, and Jie Yang. "AMIL: Adversarial Multi-instance Learning for Human Pose Estimation." In: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16.1s (2020), pp. 1–23 (cit. on pp. 16, 17).

[250] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, pp. 2556–2565 (cit. on p. 12).

[251] Yan Shen and Mingchen Gao. "Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization." In: International Workshop on Machine Learning in Medical Imaging. Springer. 2018, pp. 389–397 (cit. on pp. 10, 21, 22).

[252] Yunhang Shen, Liujuan Cao, Zhiwei Chen, Feihong Lian, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, and Rongrong Ji. "Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 16694–16705 (cit. on p. 16).

[253] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. "Explainable and explicit visual reasoning over scene graphs." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 8376–8384 (cit. on p. 16).

[254] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. "Transductive semi-supervised deep learning using min-max features." In: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 299–315 (cit. on p. 14).

[255] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia." In: Radiology. Artificial intelligence 1.1 (2019) (cit. on p. 15).

[256] Wataru Shimoda and Keiji Yanai. "Self-supervised difference detection for weakly-supervised semantic segmentation." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 5208–5217 (cit. on p. 13).

[257] Junji Shiraishi et al. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules." In: American Journal of Roentgenology 174.1 (2000), pp. 71–74 (cit. on pp. 15, 44, 51).

[258] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: arXiv preprint arXiv:1409.1556 (2014) (cit. on p. 12).

[259] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. "CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT." In: arXiv preprint arXiv:2004.09167 (2020) (cit. on pp. 10, 86).

[260] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence." In: arXiv preprint arXiv:2001.07685 (2020) (cit. on pp. 14, 43, 48, 51).

[261] Evgeniya Sokolovskaya, Tejas Shinde, Richard B. Ruchman, Andrew J. Kwak, Stanley Lu, Yasmeen K. Shariff, Ernest F. Wiggins, and Leizle Talangbayan. "The Effect of Faster Reporting Speed for Imaging Studies on the Number of Misses and Interpretation Errors: A Pilot Study." In: Journal of the American College of Radiology 12.7 (2015), pp. 683–688. ISSN: 1546-1440. DOI: `https://doi.org/10.1016/j.jacr.2015.03.040`. URL: `https://www.sciencedirect.com/science/article/pii/S1546144015002033` (cit. on p. 4).

[262] Assaf B Spanier, D Cohen, and Leo Joskowicz. "A new method for the automatic retrieval of medical cases based on the RadLex ontology." In: International journal of computer assisted radiology and surgery 12.3 (2017), pp. 471–484 (cit. on p. 11).

[263] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for simplicity: The all convolutional net." In: arXiv preprint arXiv:1412.6806 (2014) (cit. on p. 13).

[264] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. "Multiple instance learning for histopathological breast cancer image classification." In: Expert Systems with Applications 117 (2019), pp. 103–111 (cit. on p. 16).

[265] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. “Rethinking the inception architecture for computer vision.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 2818–2826 (cit. on pp. 24, 75).

[266] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. “Self-Supervised Wasserstein Pseudo-Labeling for Semi-Supervised Image Classification.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 12267–12277 (cit. on p. 14).

[267] Hao Tan and Mohit Bansal. “Lxmert: Learning cross-modality encoder representations from transformers.” In: arXiv preprint arXiv:1908.07490 (2019) (cit. on p. 12).

[268] Jennifer SN Tang, Jarrel CY Seah, Adil Zia, Jay Gajera, Richard N Schlegel, Aaron JN Wong, Dayu Gai, Shu Su, Tony Bose, Marcus L Kok, et al. “CLiP, catheter and line position dataset.” In: Scientific Data 8.1 (2021), pp. 1–7 (cit. on pp. 15, 74).

[269] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. “Multiple instance detection network with online instance classifier refinement.” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 2843–2851 (cit. on p. 16).

[270] Wen Tang, Han Kang, Ying Cao, Pengxin Yu, Hu Han, Rongguo Zhang, and Kuan Chen. “M-SEAM-NAM: Multi-instance Self-supervised Equivalent Attention Mechanism with Neighborhood Affinity Module for Double Weakly Supervised Segmentation of COVID-19.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2021, pp. 262–272 (cit. on p. 16).

[271] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. “Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs.” In: International Workshop on Machine Learning in Medical Imaging. Springer. 2018, pp. 249–258 (cit. on pp. 10, 21, 22).

[272] Ajay K Tanwani, Joelle Barral, and Daniel Freedman. “RepsNet: Combining Vision with Language for Automated Medical Reports.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2022, pp. 714–724 (cit. on p. 11).

[273] Antti Tarvainen and Harri Valpola. “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.” In: arXiv preprint arXiv:1703.01780 (2017) (cit. on pp. 14, 48).

[274] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning.” In: Nature Biomedical Engineering (2022), pp. 1–8 (cit. on p. 13).

[275] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. “Unsupervised domain adaptation in semantic segmentation: a review.” In: Technologies 8.2 (2020), p. 35 (cit. on p. 63).

[276] Ilknur Tugal-Tutkun, Carl P Herbort Jr, Alessandro Mantovani, Piergiorgio Neri, and Moncef Khairallah. "Advances and potential new developments in imaging techniques for posterior uveitis. Part 1: noninvasive imaging methods." In: Eye 35.1 (2021), pp. 33–51 (cit. on p. 11).

[277] Ken Turkowski. "Filters for common resampling tasks." In: Graphics gems (1990), pp. 147–165 (cit. on p. 70).

[278] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. "Selective search for object recognition." In: International journal of computer vision 104.2 (2013), pp. 154–171 (cit. on pp. 13, 85, 86).

[279] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." In: Journal of the American Medical Informatics Association 18.5 (2011), pp. 552–556 (cit. on pp. 84, 86).

[280] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. "Attention is All you Need." In: Advances in Neural Information Processing Systems. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf` (cit. on p. 32).

[281] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: Advances in neural information processing systems 30 (2017) (cit. on p. 21).

[282] Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, et al. "Your classifier can secretly suffice multi-source domain adaptation." In: Advances in Neural Information Processing Systems 33 (2020), pp. 4647–4659 (cit. on p. 14).

[283] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 3156–3164 (cit. on p. 12).

[284] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." In: Nature Methods 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2` (cit. on p. 75).

[285] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. "C-mil: Continuation multiple instance learning for weakly supervised object detection." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 2199–2208 (cit. on p. 16).

[286] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems." In: Advances in neural information processing systems 32 (2019) (cit. on p. 4).

[287] Hongyu Wang and Yong Xia. "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography." In: arXiv preprint arXiv:1807.03058 (2018) (cit. on pp. 10, 21, 22).

[288] Jun Wang, Abhir Bhalerao, and Yulan He. “Cross-modal prototype driven network for radiology report generation.” In: European Conference on Computer Vision. Springer. 2022, pp. 563–579 (cit. on p. 11).

[289] Kuanquan Wang et al. Pulmonary Artery Segmentation Challenge 2022. `https://parse2022.grand-challenge.org/` (cit. on p. 68).

[290] Qingfeng Wang, Jie-Zhi Cheng, Ying Zhou, Hang Zhuang, Changlong Li, Bo Chen, Zhiqin Liu, Jun Huang, Chao Wang, and Xuehai Zhou. “Low-shot multi-label incremental learning for thoracic diseases diagnosis.” In: International Conference on Neural Information Processing. Springer. 2018, pp. 420–432 (cit. on pp. 10, 21).

[291] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. “A survey of zero-shot learning: Settings, methods, and applications.” In: ACM Transactions on Intelligent Systems and Technology (TIST) 10.2 (2019), pp. 1–37 (cit. on p. 35).

[292] 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: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 2097–2106 (cit. on pp. 10, 15, 16, 21, 22, 27, 30).

[293] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. “Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 9049–9058 (cit. on pp. 10, 11, 21).

[294] Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. “Self-supervised Image-text Pre-training With Mixed Data In Chest X-rays.” In: arXiv preprint arXiv:2103.16022 (2021) (cit. on pp. 31, 34).

[295] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. “Relaxed multiple-instance SVM with application to object discovery.” In: Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 1224–1232 (cit. on pp. 16, 17).

[296] Yun Wang, Juncheng Li, and Florian Metze. “A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling.” In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. 2019, pp. 31–35 (cit. on pp. 16, 22, 26).

[297] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. “A Medical Semantic-Assisted Transformer for Radiographic Report Generation.” In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2022, pp. 655–664 (cit. on pp. 11, 43).

[298] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. “Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning.” In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 5. 2022, pp. 5914–5922 (cit. on p. 13).

[299] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. “TotalSegmentator: robust segmentation of 104 anatomical structures in CT images.” In: arXiv preprint arXiv:2208.05868 (2022) (cit. on p. 68).

[300] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 1568–1576 (cit. on p. 13).

[301] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. “Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection.” In: Proceedings of the European conference on computer vision (ECCV). 2018, pp. 434–450 (cit. on p. 13).

[302] Alistair JK Williams, Sally L Thrower, Iara M Sequeiros, Alexandra Ward, Alex S Bickerton, Jessica M Triay, Mark P Callaway, and Colin M Dayan. “Pancreatic volume is reduced in adult patients with recently diagnosed type 1 diabetes.” In: The Journal of Clinical Endocrinology & Metabolism 97.11 (2012), E2109–E2113 (cit. on p. 11).

[303] Ronald J Williams and David Zipser. “A learning algorithm for continually running fully recurrent neural networks.” In: Neural computation 1.2 (1989), pp. 270–280 (cit. on p. 11).

[304] Di Wu, Tiantian Wu, Qun Liu, and Zhicong Yang. “The SARS-CoV-2 outbreak: what we know.” In: International journal of infectious diseases 94 (2020), pp. 44–48 (cit. on p. 4).

[305] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. “Domain adaptation with asymmetrically-relaxed distribution alignment.” In: International conference on machine learning. PMLR. 2019, pp. 6872–6881 (cit. on p. 14).

[306] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. “Unsupervised feature learning via non-parametric instance discrimination.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 3733–3742 (cit. on p. 46).

[307] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. “Self-Training With Noisy Student Improves ImageNet Classification.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 (cit. on p. 43).

[308] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. “Self-training with noisy student improves imagenet classification.” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 10687–10698 (cit. on pp. 14, 43).

[309] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. “Show, attend and tell: Neural image caption generation with visual attention.” In: International conference on machine learning. PMLR. 2015, pp. 2048–2057 (cit. on p. 12).

[310] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. “Dash: Semi-supervised learning with dynamic thresholding.” In: International Conference on Machine Learning. PMLR. 2021, pp. 11525–11536 (cit. on p. 14).

[311] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays." In: ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018, pp. 103–110 (cit. on pp. 10, 21, 22).

[312] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P Harrison, Mohammadhadi Bagheri, and Ronald M Summers. "Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 9261–9270 (cit. on p. 22).

[313] Jiancheng Yang, Shixuan Gu, Donglai Wei, Hanspeter Pfister, and Bingbing Ni. "RibSeg Dataset and Strong Point Cloud Baselines for Rib Segmentation from CT Scans." In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2021, pp. 611–621 (cit. on pp. 61, 68).

[314] Jing Yang, Ya Zheng, Xi Gou, Ke Pu, Zhaofeng Chen, Qinghong Guo, Rui Ji, Haojia Wang, Yuping Wang, and Yongning Zhou. "Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis." In: Int J Infect Dis 94.1 (2020), pp. 91–95 (cit. on p. 4).

[315] Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter van Elmpt, Andre Dekker, Tim Lustberg, and Mark Gooding. "Data from lung CT segmentation challenge." In: The cancer imaging archive 20 (2017) (cit. on p. 62).

[316] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. "Auto-encoding scene graphs for image captioning." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 10685–10694 (cit. on pp. 12, 13).

[317] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. "A fast and accurate one-stage approach to visual grounding." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 4683–4693 (cit. on pp. 12, 85–87).

[318] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. "Stacked attention networks for image question answering." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 21–29 (cit. on p. 12).

[319] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. "Learning to diagnose from scratch by exploiting dependencies among labels." In: arXiv preprint arXiv:1710.10501 (2017) (cit. on p. 22).

[320] Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. "Weakly supervised medical diagnosis and localization from multiple resolutions." In: arXiv preprint arXiv:1803.07703 (2018) (cit. on pp. 10, 16).

[321] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." In: Advances in neural information processing systems 31 (2018) (cit. on p. 16).

[322] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. "AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation." In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Ed. by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert. Cham: Springer International Publishing, 2021, pp. 72–82. ISBN: 978-3-030-87199-4 (cit. on p. 10).

[323] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. 4. 2021, pp. 3208–3216 (cit. on p. 12).

[324] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." In: arXiv preprint arXiv:1710.09412 (2017) (cit. on p. 14).

[325] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. "Top-down neural attention by excitation backprop." In: International Journal of Computer Vision 126.10 (2018), pp. 1084–1102 (cit. on p. 22).

[326] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. "Vinvl: Revisiting visual representations in vision-language models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 5579–5588 (cit. on p. 12).

[327] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. "Self-produced guidance for weakly-supervised object localization." In: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 597–613 (cit. on p. 24).

[328] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. "BioWordVec, improving biomedical word embeddings with subword information and MeSH." In: Scientific data 6.1 (2019), pp. 1–9 (cit. on p. 86).

[329] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. "When radiology report generation meets knowledge graph." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 07. 2020, pp. 12910–12917 (cit. on p. 11).

[330] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. "Contrastive learning of medical visual representations from paired images and text." In: arXiv preprint arXiv:2010.00747 (2020) (cit. on pp. 12, 31).

[331] Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. "Biomedical and clinical English model packages for the Stanza Python NLP library." In: Journal of the American Medical Informatics Association (June 2021). ISSN: 1527-974X (cit. on p. 86).

[332] Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. "Biomedical and clinical English model packages for the Stanza Python NLP library." In: Journal of the American Medical Informatics Association 28.9 (June 2021), pp. 1892–1899. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab090. eprint: https:

//academic.oup.com/jamia/article-pdf/28/9/1892/39731803/ocab090.pdf. URL: https://doi.org/10.1093/jamia/ocab090 (cit. on p. 84).

[333] Lin Zhao, Minglei Li, Jinqiao Kou, Jian Zhang, and Yang Zhang. "A framework for event-oriented text retrieval based on temporal aspects: a recent review." In: Proceedings of the 2020 12th International Conference on Machine Learning and Computing. 2020, pp. 39–46 (cit. on p. 4).

[334] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 2921–2929 (cit. on pp. 13, 22).

[335] Chong Zhou, Chen Change Loy, and Bo Dai. "Extract free dense labels from clip." In: European Conference on Computer Vision. Springer. 2022, pp. 696–712 (cit. on pp. 36, 40).

[336] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. "Conditional prompt learning for vision-language models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 16816–16825 (cit. on p. 12).

[337] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. "Learning to prompt for vision-language models." In: International Journal of Computer Vision 130.9 (2022), pp. 2337–2348 (cit. on p. 12).

[338] Yizhou Zhou, Xiaoyan Sun, Dong Liu, Zhengjun Zha, and Wenjun Zeng. "Adaptive pooling in multi-instance learning for web video annotation." In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017, pp. 318–327 (cit. on pp. 16, 17, 22, 23, 27).

[339] C Lawrence Zitnick and Piotr Dollár. "Edge boxes: Locating object proposals from edges." In: European conference on computer vision. Springer. 2014, pp. 391–405 (cit. on pp. 13, 85, 86).

[340] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. "PseudoSeg: Designing Pseudo Labels for Semantic Segmentation." In: International Conference on Learning Representations. 2021 (cit. on p. 14).

[341] . Template Library Advisory Panel (TLAP). https://www2.rsna.org/timssnet/About/committee.cfm?c=00673210 (cit. on p. 9).