Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding

Constantin Seibold[1]
constantin.seibold@kit.edu
Simon Reiß[1]
simon.reiss@kit.edu
Saquib Sarfraz[1,2]
muhammad.sarfraz@kit.edu
Matthias A. Fink[3]
matthias.fink@uni-heidelberg.de
Victoria Mayer[3]
victoria.mayer@med.uni-heidelberg.de
Jan Sellner[4]
j.sellner@dkfz-heidelberg.de
Moon Sung Kim[5]
moon-sung.kim@uk-essen.de
Klaus H. Maier-Hein[4]
k.maier-hein@dkfz-heidelberg.de
Jens Kleesiek[4]
jens.kleesiek@uk-essen.de
Rainer Stiefelhagen[1]
rainer.stiefelhagen@kit.edu

[1] Institute of Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2] Autonomous Systems
Daimler TSS
Karlsruhe, Germany

[3] University Hospital Heidelberg
Heidelberg, Germany

[4] Medical Image Computing
German Cancer Research Center
Heidelberg, Germany

[5] Institute for Artificial Intelligence in Medicine
University Clinic Essen
Essen, Germany

Abstract

In clinical radiology reports, doctors capture important information about the patient's health status. They convey their observations from raw medical imaging data about the inner structures of a patient. As such, formulating reports requires medical experts to possess wide-ranging knowledge about anatomical regions with their normal, healthy appearance as well as the ability to recognize abnormalities. This explicit grasp on both the patient's anatomy and their appearance is missing in current medical image-processing systems as annotations are especially difficult to gather. This renders the models to be narrow experts *e.g.* for identifying specific diseases. In this work, we recover this missing link by adding human anatomy into the mix and enable the association of content in medical reports to their occurrence in associated imagery (medical phrase grounding). To exploit anatomical structures in this scenario, we present a sophisticated automatic pipeline to gather and integrate human bodily structures from computed tomography datasets, which we incorporate in our PAXRAY: A Projected dataset for the segmentation of Anatomical structures in X-RAY data. Our evaluation shows that methods that

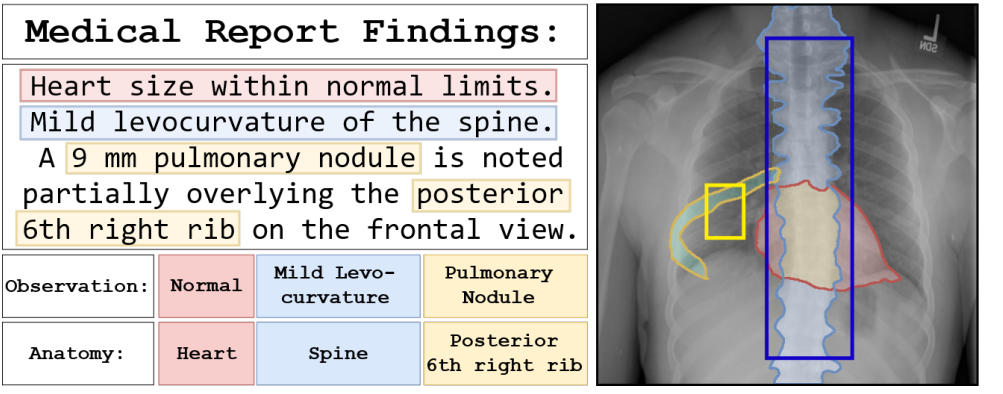

Figure 1: Overlap between the segmentation of anatomies and expert annotations on a sample of OpenI [16] indicating the necessity of anatomical understanding. Boxes are radiologists' annotation of findings. Masks show predictions for '6th right rib', 'spine' and 'heart'

take advantage of anatomical information benefit heavily in visually grounding radiologists' findings, as our anatomical segmentations allow for up to absolute 50% better grounding results on the OpenI dataset as compared to commonly used region proposals.

1 Introduction

With millions of images being produced every year, chest radiographs (CXR) are an essential part of daily clinical practice for initial diagnosis of pathologies such as rib fractures [2], pneumothoraces [40] or pulmonary infections [8]. For their interpretation, medical experts undergo extensive training to understand the present body structure and its consequent deviations for a radiologic image of a patient [2]. Subsequently, the radiologist summarizes the relevant visual information as a medical report for the further clinical workflow.

In Fig. , we display an example of a medical report. We display in the CXR on the right, that the radiolist's report follows anatomical structures to localize and describe anomalies similar to the prominent ABCDE-scheme [58]. We argue the utilization between these correlations between anomalous findings and anatomical regions can be beneficial in the understanding of medical reports. For example, the finding PULMONARY NODULE OVERLYING THE POSTERIOR SIXTH RIB can be localized using automatic anatomical segmentation.

However, the arising challenge now becomes how to get hold of these segmentations? Dense annotations for natural [14] and medical images [22] are challenging to collect. 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. This leads to ambiguous, inextricable, visually blended patterns in X-rays that even with expert knowledge annotating fine-grained anatomy structures are unfeasible. This is also stated by Seibold *et al*. [52] as the fine-grained mask annotation of a single CXR takes up to three hours. Due to this immense cost, most datasets stick to either a minimal mask labels [29, 56], or strictly rely on image-level labels [9, 16, 27, 31, 53].

To bypass these issues, we find inspiration in three related facts: Firstly, computed tomography (CT) being aggregated multi-view 2D X-Rays [23]. Secondly, the immense advantages in identifying anatomy in CTs [10], i.e. the esophagus can easily be tracked in a CT

whereas it is harder in CXRs. Lastly, the consistent body structure of a patient throughout modalities. Building upon these observations, we contribute threefold:

We propose the use pipeline which makes use of proven segmentation methods in CTs to generate accurate anatomical annotation and subsequently transfers the 3D labels with the respective CT scan to 2D leading to simplified gathering of accurate CXR annotations.

Using this pipeline, we present the *first* fine-grained anatomy dataset: *PAXRay*. Based on high quality predictions in the CT space, we display 92 individual labels of anatomical structures, which, when including super-classes, lead to a total of 166 labels in both lateral and frontal view. We make the dataset available for the community here.

Finally, we show that the usage of fine-grained anatomical structures can noticeably assist in matching medical observations and image regions. We, hereby, outperform commonly used region proposal methods by up to 50% Hitrate for grounding methods.

2 Related Work

Medical Image Understanding. The amount of CXR datasets [9, 16, 27, 32, 63] allowed for a massive development of deep learning approaches [4, 25, 45, 46, 47, 50, 52, 69]. These datasets are typically automatically annotated by a text classifier trained on a fixed set of diseases [27, 57, 63]. Many works exist for the identification of diseases [5, 47, 50, 63], automated generation of reports [45, 64] or visual question answering [55]. While there have been methods which move away from fixed set training through multi-modal contrastive training to become more flexible [26, 51, 59, 72], deep learning algorithms in this area is widely regarded as a black box [8]. Several of these methods integrated interpretability through the use of class activation mappings [52, 53, 75] or attention [45] which, however, diverges from a doctor's anatomy-based approach [58]. While some approaches emerged that utilize anatomical information [1, 20], the level of detail is restricted to bounding boxes [65] or the heart and lung area as found in i.e. the JSRT dataset [56], thus narrowing down the potential field of application. Through the generation of our fine-grained PAX-Ray, the largest anatomy segmentation dataset at the time, we propose the usage of anatomical information in CXR to enable further interpretability of medical image analysis, and the diagnoses of physicians.

Visual Phrase Grounding. Visual grounding seeks to encode informative content in natural language with visual features to localize visual content referenced in the text [15, 17, 18, 19, 66, 68]. Most of such methods are two-stage methods [15]. In the first stage, a region proposal method such as EdgeBoxes [76], Selective Search [61] or trained detectors like Faster-RCNN [48] generates potential regions of interest. In the second stage, one tries to match queries to a fitting region based on their affinity [15, 19, 39]. As the two-stage model performance directly relies on the usability of the proposal methods, they can be seen as their upper bound [68]. In this work, we notice that proposal methods are suffering in the X-ray domain and propose to offset the shortcomings through the use of anatomical segmentations.

3 Automated Generation of Projected CXR Datasets

Due to the immense difficulty of gathering precise pixel-wise annotations in the CXR domain, most datasets rely on either automatically parsed pathology labels or complete medical reports. In contrast, as we extract information from much easier to annotate CTs, we propose a novel pipeline for generating annotations assisting the CXR domain and provide a densely

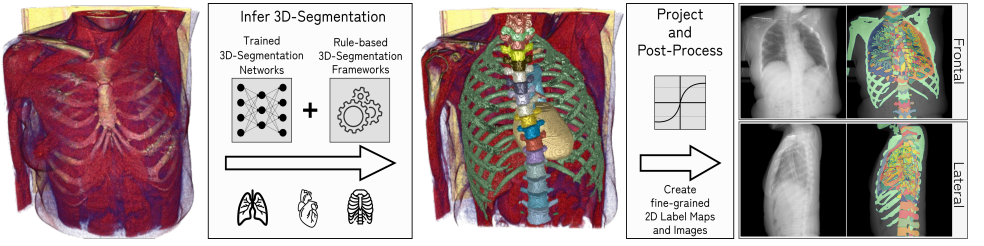

Figure 2: Dataset creation protocol. We apply established 3D segmentation methods to generate comprehensive annotations of a CT dataset. Afterwards, the CTs and their labels are projected to 2D using post-processing techniques to emulate X-ray characteristics.

labeled fine-grained dataset for anatomy segmentation containing both frontal and lateral views. Here, we leverage the consistency of anatomy between imaging domains to collect annotations from established models in the CT domain and then project these automatically generated annotations and images to 2D imitating the X-ray domain as shown in Fig. 2.

3.1 Automated Label Generation

With the emergence of the UNet [12, 28, 49] the quality of 3D segmentation models for medical imaging has shown to be surprisingly reliable. For our annotation process, we utilize conventional segmentation methods [37, 38] as well as the recent nnUNet [28].

We build the annotation scheme based on a label hierarchy with each label mask being denoted by $\mathcal{M}_l, l \in L$ with L being the set of considered labels. We start with the generation of a body mask $\mathcal{M}_{body}$ to separate it from the CT-detector backplate by selecting the largest connected component after thresholding. We then consider the four super-categories of *bones*, *lung*, *mediastinum* and *sub-diaphragm*. Generally, we assume each fine-grained class as a subset of its parent class, i.e. the spine within the bone structure ($\mathcal{M}_{spine} \subset M_{bone}$). We gather $\mathcal{M}_{bone}$ through a slicewise generalized histogram thresholding [3]. Within the bone structure, we segment the individual vertebrae [40, 41, 53] and overall ribs [67]. We expand the rib annotation of Yang *et al.* [67] by discerning individual ribs as well as posterior and anterior parts based on their center and horizontal inflection.

For the lungs, we utilize the lung lobe segmentation model by Hofmanninger *et al.* [24]. The merger of the individual lobes leads to the lung halves. We further gather the pulmonary vessels and total lungs through calculated thresholding and post-processing strategies [37].

For the mediastinum, we considered the area between the lung halves. To segment this area we utilized Koitka *et al.*'s Body Composition Analysis (BCA) [35] and split it into superior and inferior along the 4th T-spine following medical definitions [13]. We extract annotations for the *heart*, *aorta*, *airways*, and *esophagus* using the SegThor dataset [36].

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/right hemidiaphragm as no anatomical indicator exists.

To generate the label set L, we apply the combination of mentioned networks and rule-sets on the publically available RibFrac [30] dataset which is fitting for such a 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 contradicting segmentations and manually remove volumes with noticeable errors.

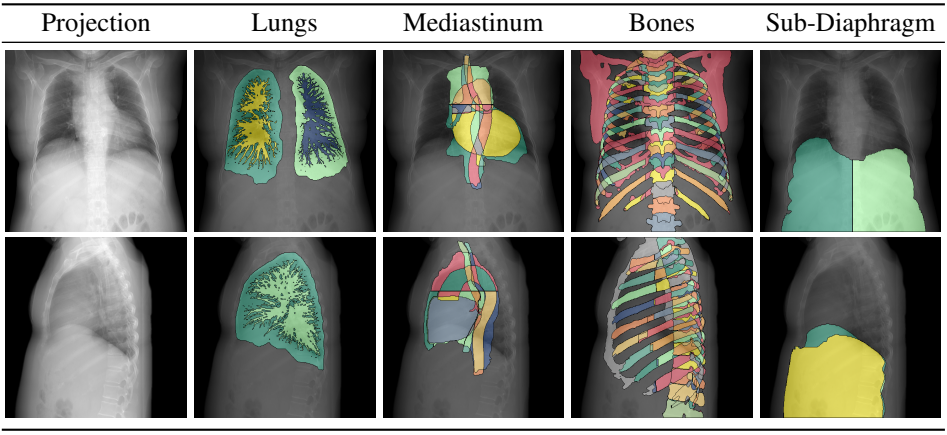

Figure 3: Examples of overlapping annotations in both lateral and frontal view of PAX-ray

We project these labels to 2D using the max-operation along the desired dimension and apply morphological post-processing steps based on the observed anatomy. We provide the full list of labels and the segmentation performance of the approaches in the supplementary.

3.2 CT to X-ray projection

We project the CTs in similar to Kausch *et al.* [33] and Matsubara *et al.* [43]. Let V be the volume of a CT scan and M_{Body}, M_{Bone} the body and bone masks we gathered prior. We clip the V to the common 12-bit range. 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:

$$V'_{Body} = M_{Body} \cdot \sigma\left(\frac{V - mean(V)}{std(V)}\right). \tag{1}$$

We repeat this for the bone region to get V'_{Bone}. Afterwards, the V's are summed, min-max-feature scaled, and rescaled to the desired range. We average along the desired dimension to get the image. We show exemplary image-label-pairs in Fig. 3 and the supplemental.

4 Anatomy-guided Phrase Grounding of Medical Reports

In medical reports, oftentimes medical observations are paired with anatomical regions to refer to their respective positions. Starting from the assumption of co-occurrence between diseases and anatomical regions within the text, we build a straightforward baseline to indicate the usability of anatomy guidance for the grounding of observations as seen in Fig. 4.

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*-section of the report containing the description of visual observations in the image. As shown in the top branch of Fig. 4, we process the report $\mathcal{R}_i$ sentence-wise to split it into medically relevant phrases $\mathcal{P}_{ij} \in \mathcal{P}_i, 0 \leq j \leq |\mathcal{P}_i|$ that are classified as problem or treatment by the named-entity-recognition (NER) model and discard the rest [51, 52, 73]. Subsequently, we 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, ...) [51, 73],

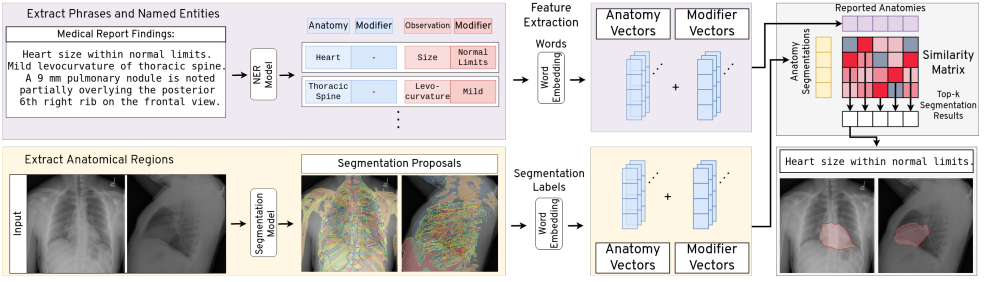

Figure 4: Anatomy Grounding Baseline. Using NER divide phrases into *Anatomy*, *Anatomy-Modifier*, *Observation*, *Observation-Modifier*. For the images, we generate proposals using our anatomy segmentation model. We extract word-wise embeddings for the phrase/anatomy labels, aggregate them and retrieve the most similar region for each phrase

group them and omit w if it doesn't belong to any of these categories. Thus, we get a filtered phrase $\mathcal{P}^*_{ij}$ which contains groups of the words W^x_{ij} of each category $x \in \{A, AM, O, OM\}$:

$$\begin{aligned} \mathcal{P}^*_{ij} &= \{W^A_{ij}, W^{AM}_{ij}, W^O_{ij}, W^{OM}_{ij}\} \\ W^x_{ij} &= \{w \mid \mathcal{C}(w) = x, w \in \mathcal{P}_{ij}\}, \end{aligned} \tag{2}$$

We utilize a pre-trained word-embedding model $\mathcal{E}$ to extract d-dimensional embeddings for all words in the filtered phrase $\mathcal{P}^*_{ij}$ occurring as anatomy and anatomy modifier:

$$\begin{aligned} F^A_{ij} &= \{\mathcal{E}(w) \mid w \in W^A_{ij}\} \\ F^{AM}_{ij} &= \{\mathcal{E}(w) \mid w \in W^{AM}_{ij}\} \end{aligned} \tag{3}$$

For phrases occurring without an anatomy or anatomy modifier, we set $F^x_{ij} = 0^d$ with $x \in \{A, AM\}$. As multiple words for a phrase can occur as anatomy or anatomy modifier we consider the category representation as mean of all word embeddings belonging to that specific category.

The final phrase embedding is then the sum of both category embeddings.

$$F_{ij} = \text{mean}(F^A_{ij}) + \text{mean}(F^{AM}_{ij}) \tag{4}$$

In the bottom branch of Fig. 4, we extract anatomical regions using our segmentation network. Doing so we get 166 binary predictions with their associated class label text $l \in L$ for each view, which we threshold to get mask regions. We refine these segmentation masks through similar anatomical constraints of their parent classes and post-processing steps as in Section 3.1. For all segmentation masks we split their label text l into anatomy and its modifier and we extract features T_l in a similar manner as above:

$$\begin{aligned} T^A_l &= \{\mathcal{E}(w) \mid \mathcal{C}(w) = A, w \in l\} \\ T^{AM}_l &= \{\mathcal{E}(w) \mid \mathcal{C}(w) = AM, w \in l\} \\ T_l &= mean(T^A_l) + mean(T^{AM}_l) \end{aligned} \tag{5}$$

Utilizing these feature vectors, we compute the cosine-similarity matrix $S^i \in [-1,1]^{|\mathcal{P}_i| \times 166}$ between both image regions and phrases individually for the lateral- and frontal view with

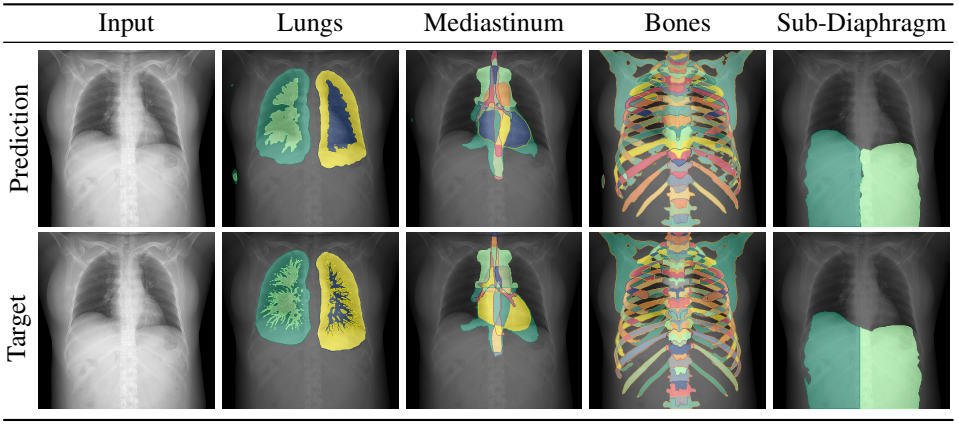

Figure 5: Qualitative results of a UNet on the test set of the *PAXRay* dataset

each entry being defined by $S^i_{jl} = cos(F_{ij}, T_l)$. Then, for a given phrase query we return the segmentation proposal based on the top-k similarities. For phrases without anatomy we simply return the whole image.

4.1 Implementation Details

Anatomy Segmentation: To show the usability of our fine-grained multi-label dataset, we train segmentation models with differently trained backbone networks. We chose the in the medical domain commonly utilized UNet[49] and the SFPN[34] with a ResNet-50[21] backbone. As the labels can overlap we train with binary cross-entropy and employ an additional binary dice loss. We used random resize-and-cropping of range $[0.8, 1.2]$ as augmentation with an image size of 512 and optimize using AdamW[42] with a learning rate of 0.001 for 110 epochs decaying by a factor of 10 at $\{60, 90, 100\}$ epochs.

Phrase Grounding: We process our reports using Stanza [74] to infer observations/treatments using the i2b2-2010 corpus [62] as well as anatomies and observations through the Radiology corpus [51, 74]. We utilize ChexBert [57] to extract an additional *is-anomaly* token for each phrase. To extract word and phrase features we utilize BioWordVec [71] and BioSentVec [1]. As we evaluate grounding in this task via bounding box comparison, for each segmentation result we extract a corresponding bounding box.

5 Experiments

5.1 Anatomy Segmentation

Experimental Setting: We evaluate the segmentation quality quantitatively on the PAXRay dataset using the typically used mean Intersection over Union (IoU). We maintain the train/val/test splits of the RibFrac dataset [30] as such we are left with 598/74/180 images. We validate our models every 10th epoch and test the model which performed best on validation.

Results: We see quantitative results in Table 1. We show the performance on selected superclasses and the mean over all 166 classes due to the immense number of classes. We show the complete performance over all classes in the supplementary material.

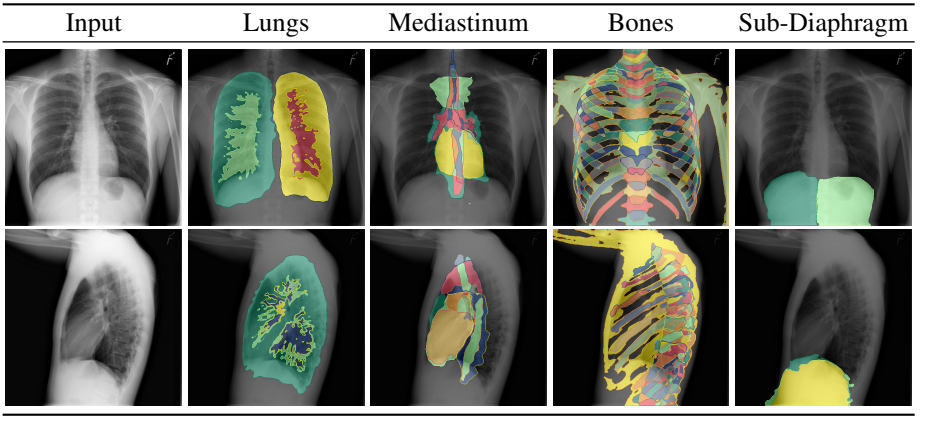

Figure 6: Qualitative results of a UNet trained on our PAX-ray dataset for a patient in OpenI

We see that in the observed setting we profit from pre-trained networks with a gain from up to $\sim 14\,\%$ mIoU. We see a difference in performance based on the architecture as the UNet outperforms the SFPN by roughly $9\,\%$. While for classes like the heart, lobes, or aorta these architectures perform similarly there are noticeable differences for individual mediastinal regions, airways, and ribs.

While we are able to segment several classes with up to $\sim 90\,\%$ mIoU on classes like the spine or heart, the correct segmentation of lung vessels is especially difficult with an IoU of $52\,\%$. Furthermore, the largest difference in segmentation quality between the two networks lies in the rib cage segmentation where the UNet has a gain of 6.1%. We observe qualitative examples in Fig. 5. The vessel tree and tracheal ends towards the bronchi pose as difficult, whereas lobe-, intermediastinal-, and bone-related classes appear as expected.

To show the applicability of our proposed dataset for the anatomy segmentation in real CXR, we display qualitative results on the OpenI dataset in Fig. 6. While similar errors to the projected x-rays can be noticed i.e. for the lateral rib or diaphragm segmentations, the results show to be quite promising albeit no domain adaptation method [60] was applied.

5.2 Medical Phrase Grounding

Experimental setting: For our evaluation of medical phrase grounding, we use the OpenI dataset [16] which consists of medical reports paired with frontal and lateral chest X-rays. We tasked two radiologists to highlight phrases within 100 medical reports in the OpenI

	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 1: Segmentation performance in IoU on the test split of our proposed *PAXRay* dataset

	Method(N=200)	HR_{25}	HR_{50}	HR_{75}
Frontal	Whole Image	16.5	5.8	0.4
	Selec. Search [61]	72.8	16.5	7.7
	EdgeBoxes [76]	18.9	4.8	0.9
	RPN [48]	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 [61]	80.7	47.7	19.2
	EdgeBoxes [76]	35.7	11.9	1.8
	RPN [48]	68.8	24.7	0.9
	Anatomy Segm.	**88.0**	**62.3**	**20.1**

Table 2: Hit rates of region proposals for different IoU thresholds (denoted by subscript).

dataset resulting in 178 frontal and 146 lateral bounding box annotations.

We evaluate the usability of anatomy segmentations for medical phrase grounding in two parts. First, we investigate the upper bound achievable by computing the average hit rate (HR) at different IoU thresholds [68]. A hit is considered as a candidate region overlapping with the ground truth annotation with an IoU above the set threshold. We compare our anatomy segmentations with common region proposal methods utilized by phrase grounding algorithms [15, 18, 44] in natural images such as EdgeBoxes [76], Selective Search [61] and Region Proposal networks [48]. We extract the top 200 scoring boxes for each labeled image following most phrase grounding methods [15, 18, 68].

Afterwards, we show 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 for the comparison with our label as well as just the anatomy in itself. We also display the *oracle*'s performance utilizing selective search to put the value of the proposed anatomy-based segmentations into perspective as it poses as the upper bound of weakly supervised methods, i.e. if the proposal method is unable to provide good initial hints the grounding method itself cannot match phrases with their image region.

Hit Rate Analysis: We show hit rate (HR) results in Table 2. We see that for the traditional approaches in both the frontal and the lateral view the selective search algorithm provided the best proposals, however, we observe an extreme loss in quality when increasing the IoU threshold, i.e. in the frontal view the hit rate drops by nearly 56 %. These 16.5 % stand in comparison to the Flicker30K dataset where the hit rate of selective search at a 50 % IoU

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

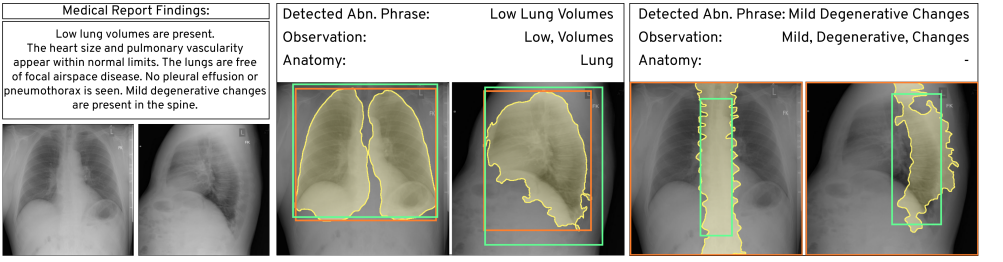

Figure 7: We show ground truth, retrievals and expected retrieval. If anatomy phrases are identified a result is provided. Otherwise, the segmentation, albeit accurate, is not retrieved.

threshold for 200 boxes was reported as 85.68 % [68]. In contrast, without being trained in the segmentation of observations but rather anatomies, we achieve improvements across all categories with *i.e.* a 50 % improvement in HR for the frontal view at an IoU of 50 %. This indicates that anatomy guidance can be a better starting point for the localization of observations as the HR relates to the oracle's performance.

Grounding Results: We show our quantitative results for medical phrase grounding in Table 3. We see that both the direct sentence comparison as well as our proposed method surpass the oracle's performance based on proposals by selective search for the frontal view on the commonly used IoU threshold of 50 %. Utilizing both anatomy and their modifiers improves noticeably over using complete sentence embeddings. We show qualitative results in Figure 7. We highlight anatomy and medical phrases. We see that despite not directly referring to disease, anatomical regions can be utilized to retrieve medical findings.

6 Discussion and Conclusion

In this paper, we propose a method for the automatic generation of anatomy labels for chest X-rays through the projection of CT data and their respective annotations via established segmentation methods to enable more complex downstream tasks such as medical phrase grounding. As the required time for fine-grained mask annotations in medical images is massive, our scheme can be considered to be an immense time save for the generation of CXR annotations and could be extended to any annotation type, be it anatomy or pathology.

We introduce the PAXRay dataset which consists of projected CXR paired with a large amount of fine-grained anatomical structures. The information richness of dense pixel-wise annotations and the shared anatomical context between X-Rays allows us to train fine-grained anatomy segmentation models. Furthermore, with our proposed method the PAXRay dataset can be extended arbitrarily by utilizing additional CT datasets. We show in our experiments that the resulting models can segment anatomical regions on not only projected but also real CXR images, thus, enabling precise anatomy localization to build reliable region proposals for CXR analysis. This allows us to outperform prior oracle-like performance with a simple baseline method. We anticipate that our work allows the community to develop improved methods for generating more interpretable computer-assisted diagnosis tools.

7 Acknowledgements

The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health".

References

[1] Nkechinyere N. Agu, Joy T. Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. Anaxnet: Anatomy aware multi-label finding classification in chest x-ray. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 804–813, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3.

[2] Muhammad Awais, Basit Salam, Naila Nadeem, Abdul Rehman, and Noor U Baloch. Diagnostic accuracy of computed tomography scout film and chest x-ray for detection of rib fractures in patients with chest trauma: a cross-sectional study. *Cureus*, 11(1), 2019.

[3] Jonathan T Barron. A generalization of otsu's method and minimum error thresholding. In *European Conference on Computer Vision*, pages 455–470. Springer, 2020.

[4] 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 Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 571–581, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

[5] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 571–581. Springer, 2021.

[6] Andrea Borghesi and Roberto Maroldi. Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. *La radiologia medica*, 125(5):509–513, 2020.

[7] William E Brant and Clyde A Helms. Fundamentals of diagnostic radiology. 2012.

[8] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

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

[10] Brandon C Chapman, Douglas M Overbey, Feven Tesfalidet, Kristofer Schramm, Robert T Stovall, Andrew French, Jeffrey L Johnson, Clay C Burlew, Carlton Barnett, Ernest E Moore, et al. Clinical utility of chest computed tomography in patients with rib fractures ct chest and rib fractures. *Archives of trauma research*, 5(4), 2016.

[11] Qingyu Chen, Yifan Peng, and Zhiyong Lu. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE, 2019.

[12] 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*, pages 424–432. Springer, 2016.

[13] Mark S Cook and Anthony J Weinhaus. Anatomy of the thoracic wall, pulmonary cavities, and mediastinum. *Handbook of Cardiac Anatomy, Physiology, and Devices*, pages 35–60, 2015.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[15] 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*, pages 2601–2610, 2019.

[16] 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. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[17] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018.

[18] 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. *arXiv preprint arXiv:1606.01847*, 2016.

[19] 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*, pages 1364–1373, 2021.

[20] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In *International Conference on Computer Science, Engineering and Education Applications*, pages 638–647. Springer, 2018.

[21] 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*, pages 770–778, 2016.

[22] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[23] Gabor T Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media, 2009.

[24] 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. *European Radiology Experimental*, 4(1):1–13, 2020.

[25] Benjamin Hou, Georgios Kaissis, Ronald M. Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 293–303, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2.

[26] 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*, pages 3942–3951, 2021.

[27] 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*, volume 33, pages 590–597, 2019.

[28] 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. *Nature methods*, 18(2):203–211, 2021.

[29] 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. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.

[30] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, Jiajun Chen, and Ming Li. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. *EBioMedicine*, 2020.

[31] 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. *Scientific data*, 3(1):1–9, 2016.

[32] 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. *Scientific data*, 6(1):1–8, 2019.

[33] Lisa Kausch, Sarina Thomas, Holger Kunze, Tobias Norajitra, André Klein, Jan Siad El Barbari, Maxim Privalov, Sven Vetter, Andreas Mahnken, Lena Maier-Hein, and Klaus H. Maier-Hein. C-arm positioning for spinal standard projections in different intra-operative settings. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 352–362, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87202-1.

[34] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.

[35] 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. *European radiology*, 31(4):1795–1804, 2021.

[36] 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)*, pages 1–6. IEEE, 2020.

[37] 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*, pages 185–199, 2011.

[38] 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. *IEEE transactions on medical imaging*, 32(2):210–222, 2012.

[39] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[40] 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. *arXiv preprint arXiv:2103.06360*, 2021.

[41] 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. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[43] Naoki MATSUBARA, Atsushi TERAMOTO, Kuniaki SAITO, and Hiroshi FUJITA. Generation of pseudo chest x-ray images from computed tomographic images by non-linear transformation and bone enhancement. *Medical Imaging and Information Sciences*, 36(3):141–146, 2019.

[44] 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*, pages 449–456. Springer, 2018.

[45] Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational topic inference for chest x-ray report generation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 625–635, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

[46] Viet-Quoc Pham, Nao Mishima, and Toshiaki Nakasu. Improving visual question answering by semantic segmentation. In Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter, editors, *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 459–470, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86365-4.

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

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[49] 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*, pages 234–241. Springer, 2015.

[50] 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*, 2020.

[51] Constantin Seibold, Simon Reiß, M. Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. Breaking with fixed set pathology recognition through report-guided contrastive training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 690–700, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16443-9.

[52] 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*, volume 36, pages 2171–2179, 2022.

[53] 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. *Medical image analysis*, 73:102166, 2021.

[54] 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*, pages 618–626, 2017.

[55] Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):1–18, 2021.

[56] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. 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. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[57] 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. *arXiv preprint arXiv:2004.09167*, 2020.

[58] Troels Thim, Niels Henrik Vinther Krarup, Erik Lerkevang Grove, Claus Valter Rohde, and Bo Løfgren. Initial assessment and treatment with the airway, breathing, circulation, disability, exposure (abcde) approach. *International journal of general medicine*, 5:117, 2012.

[59] 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. *Nature Biomedical Engineering*, pages 1–8, 2022.

[60] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies*, 8(2):35, 2020.

[61] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

[62] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[63] 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*, pages 2097–2106, 2017.

[64] 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*, pages 9049–9058, 2018.

[65] Joy Wu, Yaniv Gur, Alexandros Karargyris, Ali Bin Syed, Orest Boyko, Mehdi Moradi, and Tanveer Syeda-Mahmood. Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 799–803. IEEE, 2020.

[66] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017.

[67] 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*, pages 611–621. Springer, 2021.

[68] 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*, pages 4683–4693, 2019.

[69] 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 Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 72–82, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

[70] Paul Zarogoulidis, Ioannis Kioumis, Georgia Pitsiou, Konstantinos Porpodis, Sofia Lampaki, Antonis Papaiwannou, Nikolaos Katsikogiannis, Bojan Zaric, Perin Branislav, Nevena Secen, et al. Pneumothorax: from definition to diagnosis and treatment. *Journal of thoracic disease*, 6(Suppl 4):S372, 2014.

[71] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019.

[72] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[73] 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. *Journal of the American Medical Informatics Association*, 28(9):1892–1899, 06 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab090. URL https://doi.org/10.1093/jamia/ocab090.

[74] 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. *Journal of the American Medical Informatics Association*, 06 2021. ISSN 1527-974X.

[75] 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*, pages 2921–2929, 2016.

[76] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.