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

Constantin Seibold\(^1\)
constantin.seibold@kit.edu

Simon Reiβ\(^1\)
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 [70] or pulmonary infections [6]. 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 [7]. 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 radiologist’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, 63].

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, 64, 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 [48, 53, 54, 75], 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 [54, 63, 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, 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 \(M_l, l \in L\) with \(L\) being the set of considered labels. We start with the generation of a body mask \(M_{body}\) 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 (\(M_{spine} \subset M_{bone}\)). We gather \(M_{bone}\) 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.
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 $M_{\text{Body}}, M_{\text{Bone}}$ 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 $\sigma$ and sharpen:

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

We repeat this for the bone region to get $V'_{\text{Bone}}$. 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 $(I_i, R_i) \in \{(I_1, R_1), (I_2, R_2), \ldots, (I_N, R_N)\}$ 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 $R_i$ sentence-wise to split it into medically relevant phrases $P_{ij} \in P_i$, $0 \leq j \leq |P_i|$ that are classified as problem or treatment by the named-entity-recognition (NER) model and discard the rest [31, 62, 73]. Subsequently, we filter the words $w \in P_{ij}$ using the NER-model $C$ 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, …) [31, 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.

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:

$$\begin{align*}
F_{ij}^A &= \{ \mathcal{E}(w) \mid w \in W_{ij}^A \} \\
F_{ij}^{AM} &= \{ \mathcal{E}(w) \mid w \in W_{ij}^{AM} \}
\end{align*}$$

For phrases occurring without an anatomy or anatomy modifier, we set $F_{ij}^x = 0^d$ 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.

$$F_{ij} = \text{mean}(F_{ij}^A) + \text{mean}(F_{ij}^{AM})$$

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:

$$\begin{align*}
T_l^A &= \{ \mathcal{E}(w) \mid C(w) = A, w \in l \} \\
T_l^{AM} &= \{ \mathcal{E}(w) \mid C(w) = AM, w \in l \}
\end{align*}$$

Utilizing these feature vectors, we compute the cosine-similarity matrix $S^l \in [-1, 1]^{P_l \times 166}$ between both image regions and phrases individually for the lateral- and frontal view with
each entry being defined by $S'_{ji} = \cos(F_{ij}, T_l)$. 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 [31, 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 [11]. 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 super-classes 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-, intermadiastinal-, 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

<table>
<thead>
<tr>
<th>Init.</th>
<th>Lung</th>
<th>Mediastinum</th>
<th>Bones</th>
<th>Sub-Dia. Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lobes</td>
<td>Regions</td>
<td>Heart</td>
<td>Aorta</td>
</tr>
<tr>
<td>SFPN</td>
<td>(Random)</td>
<td>82.3</td>
<td>68.6</td>
<td>74.6</td>
</tr>
<tr>
<td></td>
<td>(VBData)</td>
<td>86.3</td>
<td>75.6</td>
<td>81.9</td>
</tr>
<tr>
<td>UNet</td>
<td>(Random)</td>
<td>85.0</td>
<td>74.8</td>
<td>74.8</td>
</tr>
<tr>
<td></td>
<td>(VBData)</td>
<td>86.9</td>
<td>77.3</td>
<td>89.9</td>
</tr>
</tbody>
</table>

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


