MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training

Chaoyi Wu[1,2], Xiaoman Zhang[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2], Weidi Xie[1,2]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University [2]Shanghai AI Laboratory
{wtzxxxwcy02, xm99sjtu, ya_zhang, wangyanfeng, weidi}@sjtu.edu.cn
https://chaoyi-wu.github.io/MedKLIP/

Abstract

In this paper, we consider the problem of enhancing self-supervised visual-language pre-training (VLP) with medical-specific knowledge, by exploiting the paired image-text reports from the radiological daily practice. In particular, we make the following contributions: **First**, *unlike existing works that directly process the raw reports, we adopt a novel report filter to extract the medical entities, avoiding unnecessary complexity from language grammar and enhancing the supervision signals;* **Second**, *we propose a novel entity embedding module by querying an external knowledge description base, to exploit the rich context of additional information that the medical domain affords, and implicitly build relationships between entities in the language embedding space;* **Third**, *we propose a novel Transformer-based fusion model for spatially aligning the entity description with visual signals at the image patch level only with self-supervised learning, thus enabling the ability for spatial grounding;* **Fourth**, *we conduct thorough experiments to validate the effectiveness of our proposed architecture, and benchmark on numerous public benchmarks* e.g., *ChestX-ray14, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2, COVID Rural, and EdemaSeverity. In both zero-shot and fine-tuning settings, our model has demonstrated strong performance compared with the former methods on disease classification and grounding.*

1. Introduction

With the rapid development of deep learning, numerous works have been proposed to facilitate computer-aided diagnosis in the medical field [16, 17, 39, 48]. Despite the tremendous progress, these models are normally trained to recognize or segment the structures that fall into a certain closed set of anatomical or disease categories, whenever a new disease comes to be of interest, a costly procedure for data annotation, model re-training, and ethics proof will be required, fundamentally limiting its practical values. As an alternative, recent research considers to train the model by exploiting a large number of multi-modal medical data, that is generated from daily clinical routine, for example, the most common example is the dataset of X-ray images with paired radiological reports [15, 25, 28].

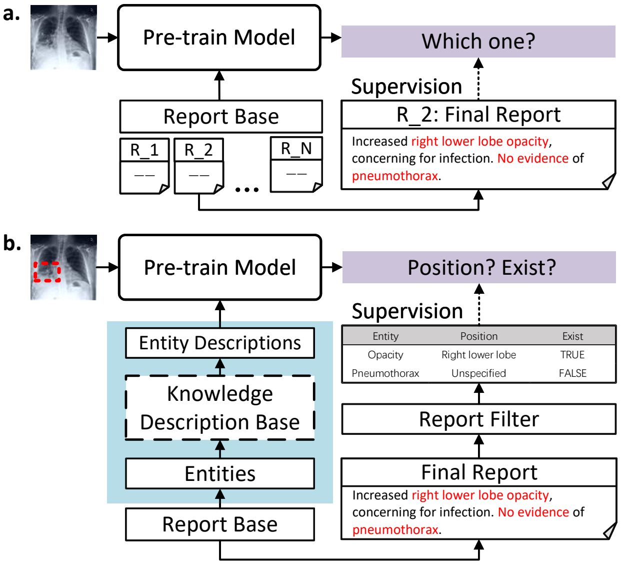

a.
Pre-train Model
Which one?
Supervision
Report Base
R_1
R_2
R_N
R_2: Final Report
Increased right lower lobe opacity, concerning for infection. No evidence of pneumothorax.
b.
Pre-train Model
Position? Exist?
Supervision
Entity Descriptions
Knowledge Description Base
Entities
Report Base
Entity Position Exist
Opacity Right lower lobe TRUE
Pneumothorax Unspecified FALSE
Report Filter
Final Report
Increased right lower lobe opacity, concerning for infection. No evidence of pneumothorax.

Figure 1. Our method mainly considers combining medical knowledge with VLP. **a.** shows the standard VLP flowchart which uses text-image retrieval as a proxy task. **b.** is our MedKLIP flowchart. We adopt a report filter module to decompose raw reports at entity level and further use knowledge descriptions to explain entities. Our model can realize zero-shot classification and grounding.

This paper focuses on self-supervised vision-language representation learning in the medical domain, with the goal of zero-shot disease diagnosis (classification) and grounding. Undoubtedly, such tasks have also been widely investigated in the computer vision community, with significant progress made in the past years, for example, CLIP [43], ALBEF [30], BLIP [29], etc. However, to achieve such a goal in the medical domain, different challenges must be resolved, which requires research efforts from the community: *First*, data availability, training Foundation Models in computer vision normally require over millions of image-text pairs at ease, while in the medical domain, only a few hundred thousand pairs are available [28]. The limited data amount challenges language models to understand the reports in free form [7]. *Second*, the problem considered in medical diagnosis is naturally fine-grained, that requires distinguishing between fine appearance details to understand the disease, as a consequence, domain knowledge is essential; *Third*, robustness is crucial, it is, therefore, preferable to have explainability, where diagnosis results come along with the visual grounding, to help radiologists to understand the system, and build trust between humans and machines.

Unlike existing work in medical VLP (Vision-Language Pre-training) [7, 22, 40, 56] that naïvely matches raw reports with image scans, we propose a novel knowledge-enhanced visual-language model that takes medical prior into consideration and enables us to address the aforementioned challenges explicitly, as shown in Fig. 1: *First*, we propose a report filter to extract useful medical entities, and simplify each report into sets of triplets, denoted as $\{\text{entity}, \text{position}, \text{exist}\}$. Consequently, decomposing reports into entities leads to an effective representation of the reports with minimal information loss, enriching supervision signals at the detailed entity level; *Second*, we map these entities into fine-grained descriptions by querying a well-defined medical knowledge base, and compute the text embedding for these descriptions, to implicitly build relationships between entities; *Third*, we adopt a transformer-based architecture for aligning the image patches with entity descriptions, that simultaneously infer the likelihood of certain diseases and the visual evidence in the form of a spatial heatmap, *i.e.*, providing grounding for explainability purpose.

We train the model on the most widely-used medical image-report dataset MIMIC-CXR [28] and rigorously evaluate on numerous public benchmarks, *e.g.*, ChestX-ray14 [50], RSNA Pneumonia [44], SIIM-ACR Pneumothorax [1], COVIDx CXR-2 [41], COVID Rural [12, 47], and EdemaSeverity [8]. We get a state-of-the-art zero-shot classification and grounding performance on different diseases, spanning different image distributions, with further fine-tuning, our model still exceeds previous models significantly.

2. Related Work

Vision-Language Pre-training Models. Vision-Language Pre-training (VLP) models have achieved great success in natural scenarios. Generally, there are two typical structures for VLP models. One is two-stream methods [5,27,30]. The other is single-stream methods [10, 32]. These impressive results promote VLP methods in medical. Different from natural data, medical VLP suffers from a serious Lack of Data (224k [28] vs 400M [5]). Most medical VLP methods follow the two-stream methods [8, 22, 40, 56]. ConVIRT [56] first proposed to use contrastive learning as a proxy task in medical. LoVT and GLoRIA then focus on improving the local alignment performance [8, 22]. BioViL notices the language pattern in reports is different from other natural texts and re-designs the language model used for medical VLP [7]. These works have greatly contributed to the development of this aspect. However, they still view medical texts and images as common natural data and use a classical pipeline to handle them, instead of leveraging the rich prior knowledge in medical.

Medical Named-Entity-Recognition Models. Various natural language processing (NLP) approaches have been proposed to extract useful information from radiology reports [25, 37, 42, 45]. These early methods considered only the disease, causing they lose a lot of information. Some Analysis tools [4, 6] have also been developed to extract key clinical concepts and their attributes from biomedical text. Further state-of-the-art works [26, 51] are proposed to extract relationship between different entities without demand of pre-defined close disease set, retaining most of useful information with high accuracy. This progress inspires us a lot and provide a new perspective for VLP. However, how to take advantage of this Named-Entity-Recognition (NER) models have not been discussed sufficiently in VLP field.

Medical Knowledge Enhanced Models. Leveraging external medical knowledge to enhance deep learning models is not a new topic [52]. Depending the approaches of using medical knowledge, They can be classified into model-based and input-based two types. In model-based types, the authors may imitate the radiological practice to design the model [20, 23, 31, 49] or change the model structure based on diagnostic patterns [11, 18, 38]. In input-based types, the knowledge is viewed as an extra input to calculate features [46, 53, 54] or to guide the final loss [9, 24]. Multi-task learning and attention mechanism [33, 34, 36] are often adopted to realize medical knowledge combination. However, most of these works are task-specific [52] because medical knowledge lacks an effective shared representation space across various diseases while we demonstrate that leveraging medical entity descriptions with text encoding has the potential to provide such a space.

3. Method

In this section, we start by describing our considered problem scenario in Sec. 3.1, followed by the details of our proposed Transformer-based architecture in Sec. 3.2, including, the visual encoder, knowledge-enhanced text encoder, and the fusion module for aligning the visual and language signals. In Sec. 3.3, we describe the training procedure for our proposed model with the paired image-reports

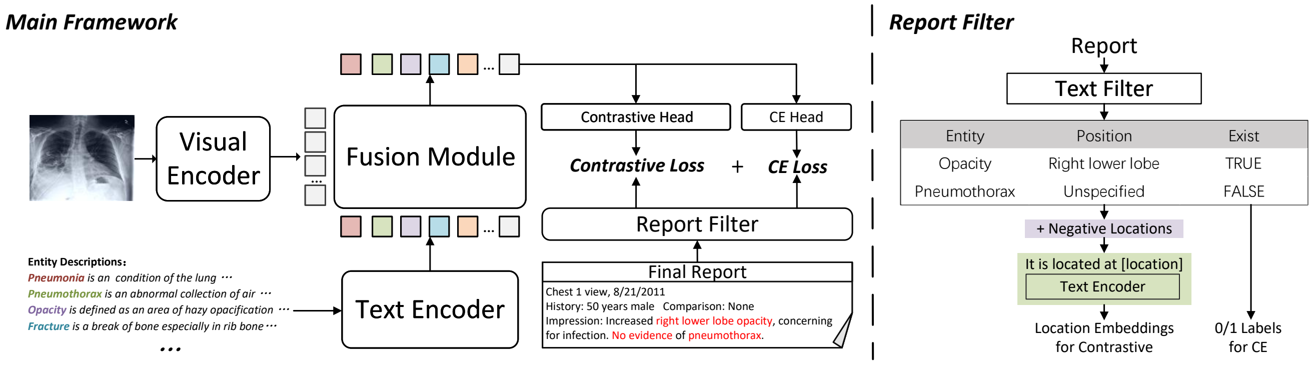

Main Framework
Visual Encoder
Fusion Module
Contrastive Head
CE Head
Contrastive Loss
+
CE Loss
Report Filter
Final Report
Chest 1 view, 8/21/2011
History: 50 years male Comparison: None
Impression: Increased right lower lobe opacity, concerning for infection. No evidence of pneumothorax.
Text Encoder
Entity Descriptions:
Pneumonia is an condition of the lung ···
Pneumothorax is an abnormal collection of air ···
Opacity is defined as an area of hazy opacification ···
Fracture is a break of bone especially in rib bone ···
···
Report Filter
Report
Text Filter
Entity
Position
Exist
Opacity
Right lower lobe
TRUE
Pneumothorax
Unspecified
FALSE
+ Negative Locations
It is located at [location]
Text Encoder
Location Embeddings for Contrastive
0/1 Labels for CE

Figure 2. The whole framework of our method. The figure mainly contains the fourth module: *Visual Encoder*, *Knowledge-enhanced Language Encoding*, *Fusion Module*. *Knowledge-enhanced Language Encoding* contains *Text Encoder* and *Report Filter*. *Report Filter* extracts entities from the raw reports and *Text Encoder* further embeds them. *Visual Encoder* is used to encoder the input of visual modalities and *Fusion Module* is used for cross-modality interaction. The details of *Report Filter* can be found in the right sub-figure. A report is first filtered by a pre-trained filter and viewed as a set of triplets. The *"Position"* part is mixed with some negative positions for contrastive loss and the *"Exist"* part is used for CE loss.

sourced from the daily routine X-ray scans.

3.1. Problem Scenario

Assuming we are given a training set with N samples, *i.e.*, $\mathcal{D}_{\text{train}} = \{(\mathcal{X}_1, \mathcal{T}_1), \dots, (\mathcal{X}_N, \mathcal{T}_N)\}$, where $\mathcal{X}_i, \mathcal{T}_i$ refer to the X-ray image and its corresponding medical report generated in the daily routine scans, respectively, our goal is to train a visual-language model that enables us to diagnose the existence of certain diseases and localize the visual evidence spatially. Specifically, at inference time, we can freely ask the system to identify the likelihood of the patient getting a certain disease (may or may not be seen during training), with its visual description for the disease of interest:

$$\hat{s}_i, \hat{m}_i = \Phi_{\text{fusion}}(\Phi_{\text{visual}}(\mathcal{X}_i), \Phi_{\text{textual}}([\text{description}])), \quad (1)$$

where $\mathcal{X}_i \in \mathbb{R}^{H \times W \times 3}$ refers to an image sample from the **test set**, with H, W denoting height and width respectively. $\hat{s}_i \in [0, 1]$ refers to the inferred likelihood of the patient having a certain disease indicated by the input description, and $\hat{m}_i \in \mathbb{R}^{H \times W \times 1}$ denotes a predicted spatial heatmap, with high activation on pixels that potentially provide the visual indication for such disease. In the following section, we will detail the individual components of our architecture, namely, the visual encoder, text encoder, and fusion module, and training them with the available training set ($\mathcal{D}_{\text{train}}$).

3.2. Architecture

In this section, we detail our proposed framework, consisting of three main components, namely, visual encoding, knowledge-enhanced language encoding, and fusion module, as shown in Fig. 2. Note that, we hereon only consider single sampled image-reports pair $(\mathcal{X}_i, \mathcal{T}_i)$, and ignore the subscript in notations for simplicity.

3.2.1. Visual Encoding

Given an X-ray image scan $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$, we can compute the features with a visual backbone:

$$\mathcal{V} = \Phi_{\text{visual}}(\mathcal{X}) \in \mathbb{R}^{h \times w \times d}, \quad (2)$$

h, w, d refer to the height, width, and feature dimension of the output feature map, in our case, we adopt a standard ResNet-50 as the visual backbone, and take the output from the 4th residual block. Note that, we make the such an architectural choice for a fair comparison with existing work [7,22,40,56], while other visual backbones, *e.g.*, ViT [14], can equally be applied.

3.2.2. Knowledge-enhanced Language Encoding

The goal of this module is to extract useful information from the text report, by incorporating medical domain knowledge. In particular, we design two stages, namely report filtering, and entity encoding.

Report Filtering. To start with, we propose to condense the report and transform it into a set of entity triplets, *i.e.*, removing the unnecessary complexity from language grammar, as shown in Figure 2 (right). In detail, we use a pre-trained text filter [26, 55] to extract valuable information from the report, for example, the medical entities and their corresponding positions on the image.

Specifically, given a report $\mathcal{T}$ with a set of sentences, $\mathcal{T} = \{s_1, s_2, ..., s_M\}$, the filter independently operates on each of the sentences, and construct a number of triplets from the report, with the extracted entity (most are diseases), spatial position, and a label indicating the existence of the disease:

$$\Phi_{\text{filter}}(s_j) = \{\text{entity}_n, \text{position}_n, \text{exist}_n\}, n \in [0, t_j], \quad (3)$$

where t_j represents the total number of entities contained in one sentence, with $n = 0$ indicating the special case that there is no valid entity. **Note that**, the position refers to the spatial position of the entity lying in an image, it is not to be confused with the positional embedding in Transformer.

Entity Encoding. Here, we replace the entities by querying detailed visual descriptions from a medical-purpose knowledge base, for example, "Pneumonia" → "It is a condition of the lung primarily affecting the small air sacs known as alveolar. It may present with opacities and pleural effusion and it can increase the diagnostic accuracy of lung consolidation". Note that, such descriptions for the medical terminologies can easily be sourced from either existing educational textbooks or online resources[1][2]. Despite its simplicity, converting the entities into descriptions is crucial for more reliable and open-vocabulary diagnosis, as it further decomposes the medical entities into visual attributes that are shared by different diseases, encouraging the model to capture a deep understanding of the visual evidence.

To encode the entity, we use the ClinicalBERT [3] as a pre-trained text encoder, to first compute the embedding for the entity description and position, and then adopt a linear MLP to flexibly project the embedding to desired dims:

$$e = \Phi_{\text{textual}}([\text{description}]) \in \mathbb{R}^{d}, \tag{4}$$

$$p = \Phi_{\text{textual}}(\text{“it is located at [position]”}) \in \mathbb{R}^{d'}. \tag{5}$$

Each triplet has now been converted into $\{e, p, l\}, l \in \{0, 1\}$ denotes the existence of the entity.

Discussion We have made two major differences compared to the existing visual-language models in computer vision, *First*, the information in medical reports is often more condensed, normally describing the existence of abnormality and their positions in the image, thus, adopting the **filter** operation can avoid unnecessary complexity from grammar, while still retaining most of the useful information in reports. *Second*, entities tend to be medical terminologies that are only understandable to audiences with a medical background, enrich the encoding by visual descriptions can significantly help the model to capture a deep understanding of the visual evidence for diseases, specifically, for seen diseases, such shared visual attributes enable to build the implicit relationship, while for unseen diseases, their visual evidence may have already been well understood by processing the descriptions of the seen ones, as they tend to be shared among diseases.

3.2.3. Fusion Module

After extracting all the entities and their corresponding positions from the reports, we select the top $|Q|$ most commonly appearing entities in reports, and compute the textual embeddings for their corresponding descriptions, denoted as an entity set $Q = \{e_1, e_2, ..., e_{|Q|}\}$, and top $|P|$ position embeddings as a position set $P = \{p_1, p_2, ..., p_{|P|}\}$. For a certain image, its computed visual representation and the entity set will be passed into a fusion module for alignment, consisting of multiple Transformer Decoder layers. We treat the entity set Q as Query, and the image features $\mathcal{V}$ as Key and Value into the Transformer decoders, the outputs from the fusion module are further fed into two linear MLP layers, one is used for inferring the existence of the entity, and the other generates an embedding to indicate the entity's spatial position:

$$\{\hat{s},\ \hat{p},\ \hat{m}\} = \Phi_{\text{fusion}}(\mathcal{V},\ Q), \tag{6}$$

where $\hat{s} \in \mathbb{R}^{|Q|}$ represents the existence prediction for each entity query, and $\hat{p} \in \mathbb{R}^{|Q| \times d'}$ represents the prediction embedding of spatial position for the entities. $\hat{m}$ denotes the average of the cross-attention maps sourced from Transformer Decoder layers. The training procedure will be detailed in Sec. 3.3.

Discussion. In contrast to the existing approaches [56] that aligns the reports with the entire image, our adopted Transformer decoder enables to compute correspondences between text and image at the patch level. Consequently, the image features $\mathcal{V}$ are more suitable for downstream segmentation tasks and the average of the cross-attention map in each layers can be used directly for **zero-shot** grounding.

3.3. Training

To train the proposed model, we leverage the corresponding triplets, for entities that are not mentioned in the considered report, we simply ignore them while computing loss. For simplicity, the following formulations are based on the assumption that the considered query do have corresponding triplets. In specific, for the existence prediction $\hat{s}$, we use binary cross-entropy with the existence label, denoted as $\mathcal{L}_{\text{cls}}$; to supervise the position prediction for each entity query, we adopt contrastive learning, randomly sample M position embeddings from the position set:

$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|} \sum_{k=1}^{|Q|} \frac{e^{\langle \hat{p}_k, p_k \rangle}}{e^{\langle \hat{p}_k, p_k \rangle} + \sum_{u=1}^{M} e^{\langle \hat{p}_k, P_{\mathcal{I}(k,u)} \rangle}}, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors and $\mathcal{I}(\cdot, \cdot)$ is a random index sampling function. P is unnormalized in calculation.

The final loss is the sum of the two:

$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{loc}} + \alpha_2 \mathcal{L}_{\text{cls}}, \tag{8}$$

[1] Wikipedia https://en.wikipedia.org/wiki/Wiki

[2] UMLS [6] https://www.nlm.nih.gov/research/umls/index.html

where α_1, α_2 refer to two hyper-parameters controlling the ratio of the two losses, and we set them to be 1.0 by default.

4. Experiment

In this section, we start by introducing the dataset used for experiments, *e.g.*, pre-training, and various downstream datasets. Then we describe the implementation details and the considered baselines.

4.1. Pre-training Dataset

MIMIC-CXR v2 [19, 28] consists of over 227k studies of paired image-report data, they are sourced from 65,379 patients at different scanning. Each study can have one or two images (different scan views), totaling 377,110 images.

4.2. Datasets for Downstream Tasks

ChestX-ray14 [50] contains 112,120 frontal-view X-ray images of 30,805 unique patients, collected from the year of 1992 to 2015 by NIH(National Institutes of Health), with labels of 14 common diseases provided. We split the dataset into $0.8/0.1/0.1$ for train/valid/test.

RSNA Pneumonia [44] contains more than 260k frontal-view chest X-rays with corresponding pneumonia opacity masks collected by RSNA (Radiological Society of North America). Commonly, it is treated as a classification tasks [7, 22]. We split the dataset into $0.6/0.2/0.2$ for train/valid/test.

SIIM-ACR Pneumothorax [1] contains more than 12k frontal-view chest X-rays with pneumothorax masks collected by SIIM-ACR (Society for Imaging Informatics in Medicine and American College of Radiology). Similarly to RSNA Pneumonia dataset, it can be both used as classification and segmentation tasks. We split the dataset into $0.6/0.2/0.2$ for train/valid/test.

COVIDx CXR-2 [41] and COVID Rural [12, 47] aim to evaluate on diagnosing COVID-19. COVIDx CXR-3 contains 29,986 images from 16,648 patients with COVID-19 classification labels. We use it as a classification dataset and split it into $0.7/0.2/0.1$ for train/valid/test. Additionally, we use COVID Rural dataset for COVID-19 segmentation. It contains more than 200 chest X-rays with segmentation masks, and we split it into $0.6/0.2/0.2$ for train/valid/test.

Edema Severity [8] contains 6,524 examples from MIMIC-CXR with pulmonary edema severity labels (0 to 3, increasing severity) extracted from the radiology reports. Of these, 141 radiologists were examined by radiologists, and consensus was reached on severity level. It can be seen as a typical fine-grained classification task. We split the dataset into $0.6/0.2/0.2$ for train/valid/test.

4.3. Implementation Details

This section details the implementation for architecture, pre-training, zero-shot inference and fine-tuning procedure.

Model architecture. As input to the model, images are resized into $224 \times 224 \times 3$. We use the first four layers of ResNet50 [21] as our visual backbone (Φ_{visual}), and adopt a MLP layer to transform the output feature dimension into $d = 256$. As a result, the output feature maps from visual encoder is $\mathcal{V} \in \mathbb{R}^{14\times14\times256}$. On the report side, we extract the entities with a pre-trained text filter, as described in [26], and compute the entity and position embedding with a pre-trained ClinicalBERT [2], its default embedding dim is $d' = 768$. We obtain $|Q| = 75$ entities and $|P| = 51$ positions that most frequently appear in the reports, following [55]. We sample $M = 7$ negative positions for each entity to calculate contrastive loss for training entity position training. In the fusion module, We adopt 4 Transformer Decoder layers with 4 heads in each layer.

Pre-training. At this stage, both the filtering operation and language encoding use pre-trained networks, while the visual encoder and fusion module are trained end-to-end on the image-text pairs. We use AdamW [35] optimizer with $lr = 1 \times 10^{-4}$ and $lr_{\text{warm up}} = 1 \times 10^{-5}$. We train on a GeForce RTX 3090 GPU with batch size 32 for 60 epochs. The first 5 epochs are set for warming up.

Inference. At inference time, given a test image, we aim to infer the existence of certain entity / disease, and ground their visual evidence. For those entities that have appeared at training time, we simply adopt the corresponding elements from the entity query set, while for those unseen ones, we replace the entity with a brief description, and treat that as an added query to the model. The existence output can be directly applied for classification and the average cross-attention of different layers in the transformer-based fusion module between specific query and the visual features are used for grounding.

Fine-tuning. For the downstream tasks, with large amount of training data, we can fine-tune the model end-to-end, with our pre-trained visual backbone as initialization. Specifically, for image classification task, we adopt ResNet50 [21] and initialize its first four layers with our pre-trained visual encoder. For image segmentation task, we use ResUNet [13] as backbone and initialize its encoder with our pre-trained image encoder.

4.4. Baselines

We compare with various existing state-of-the-art medical image-text pre-train methods, namely, ConVIRT [56], GLoRIA [22] and BioViL [7]. Since ConVIRT and GLoRIA are pre-trained on an in-house dataset, we re-train their models on MIMIC-CXR dataset for fair comparison. For BioViL, we use the officially released models by the au-

Dataset	RSNA Pneumonia			SIIM-ACR Pneumothorax			ChestX-ray14		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [56]	0.8042	0.5842	0.7611	0.6431	0.4329	0.5700	0.6101	0.1628	0.7102
GLoRIA [22]	0.7145	0.4901	0.7129	0.5342	0.3823	0.4047	0.6610	0.1732	0.7700
BioViL [7]	0.8280	0.5833	0.7669	0.7079	0.4855	0.6909	0.6912	0.1931	0.7916
Ours	**0.8694**	**0.6342**	**0.8002**	**0.8924**	**0.6833**	**0.8428**	**0.7676**	**0.2525**	**0.8619**

Table 1. Comparison with other state-of-the-art methods on zero-shot classification task. AUC, F1 and ACC scores are reported. For ChestX-ray14, the metrics all refer to the macro average on the 14 diseases.

Prompt Type	Direct Covid-19			Covid-19 Description		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [56]	0.6159	0.7057	0.6113	0.5208	0.6902	0.5266
GLoRIA [22]	0.6319	0.6938	0.5710	0.6659	0.7007	0.6083
BioViL [7]	0.6137	0.6958	0.5461	0.5382	0.6910	0.5375
Ours	0.6561	0.7066	0.5917	**0.7396**	**0.7670**	**0.7006**

Table 2. Comparison with other state-of-the-art methods on zero-shot Covid-19 classification task. AUC, F1 and ACC scores are reported. *"Direct covid-19"* refers to directly use "Covid-19" to construct the prompt sentence while *"Covid-19 Description"* refers to replace the name "Covid-19" with its medical description.

thors. For zero-shot setting, we use the prompt as mentioned by BioViL [7]. For fine-tuning, we all use ResNet50 as the visual encoder as described in Sec. 4.3.

4.5. Metrics

AUC refers to the area under the receiver operating characteristic (ROC) curve, that is commonly used for detection and binary classification tasks.

F1 and ACC are used as supplementary metrics for classification tasks. Specifically, F1 comprehensively measures the recall and precision of the model, and ACC is the short of Accuracy. The final binary prediction threshold is chosen to maximise the F1 score. The ACC score is also calculated under this threshold.

Pointing Game is used for evaluating the grounding performance. In specific, we extract the region with max response in the output heat-map, for one instance, if the region hit the ground-truth mask, it is considered a positive prediction, otherwise negative. Finally, accuracy can be calculated as the pointing game score.

Dice and IOU are commonly used for segmentation tasks. For zero-shot segmentation, we search the segmentation threshold with 0.01 interval for all methods, and report the maximal Dice score for each model.

Precision and Recall refer to the detection Precision and Recall. For medical, it is important that lesions are detected even without fine segmentation. Additionally, in some hard cases, especially for the zero-shot setting, Dice and IOU may be too strict to reflect the performance difference. Precision and recall scores can compensate for these. We choose the IOU threshold as 0.1 to calculate the scores.

5. Results

In this section, we will report the experimental results. In general, we split the results into two parts: zero-shot setting and fine-tuning setting. In the zero-shot case (Sec. 5.1), we carry out the ablation study and compare it with the other SOTA image-text pre-train methods. We mainly consider classification and segmentation tasks; In the fine-tuning case (Sec. 5.2), we evaluate the model's transferability by fine-tuning the model with 1%, 10%, and 100% data portion. Additionally, we also add a disease grading downstream task, which can be seen as a fine-grade classification task, showing that our pre-trained model can be transferred to the downstream tasks at ease.

5.1. Zero-shot

In this section, we compare our method with the other state-of-the-art methods under zero-shot setting, classification, and grounding. Due to the space limitation, we include the entire ablation study in the supplementary material, referring to it for more details and analysis, and all comparisons here are made using our best model with position contrastive loss and entity description encoder.

5.1.1. Classification

Seen Diseases. As shown in Tab. 1, we compare with existing methods on three widely-used datasets, demonstrating consistent performance improvement. Specifically, on pneumonia and pneumothorax datasets, despite the images being collected by different clinics with different diseases, our model improves the AUC score from 0.83 to 0.87 on RSNA pneumonia dataset and from 0.71 to 0.89 on SIIM-ACR pneumothorax dataset, as shown in Tab. 1. This shows that our method can better deal with the multi-center and multi-disease data distribution in medical. While on ChestX-ray14 dataset, we improve the average AUC scores from 0.69 to 0.77, we refer the reader to supplementary material for a more detailed comparison of 14 diseases.

Unseen Diseases. In particular, we use covid-19 to evaluate the systems. **Note that**, our considered setting is different

Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑
GLoRIA [22]	0.7607	0.8330	0.1621	0.2182	0.3468
BioViL [7]	0.8342	0.8521	0.5034	0.3029	0.4386
Ours	**0.8721**	**0.8661**	**0.6420**	**0.3172**	**0.4649**

(a) Zero-shot grounding on Pneumonia

Methods	Pointing Game↑	Recall↑	Precision↑
GLoRIA [22]	0.0651	0.2377	0.0585
BioViL [7]	0.0252	0.1963	0.1429
Ours	**0.1975**	**0.3562**	**0.1940**

(b) Zero-shot grounding on Pneumothorax

Table 3. Comparison with other state-of-the-art methods on zero-shot region grounding tasks. (a) shows the results on RSNA Pneumonia dataset. (b) shows the results on SIIM-ACR Pneumothorax dataset. The pneumothorax region tends to be thin and narrow and much more challenging for grounding, we thus only consider pointing game, recall, and precision. Our method can achieve better performance on different metrics, especially on the pointing game score. ConVIRT as the basic method proposed earliest can not realize this function.

Prompt Type	Direct covid-19					Covid-19 Description				
Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑	Pointing Game↑	AR↑	AP↑	IoU↑	Dice↑
GLoRIA [22]	0.0364	0.2906	0.1073	0.0645	0.1141	0.2727	0.2821	0.1336	0.0596	0.1075
BioViL [7]	0.4000	0.2564	0.2703	0.1198	0.1967	0.1818	0.2393	0.1637	0.0861	0.1427
Ours	0.1818	0.1880	0.1497	0.0747	0.1289	**0.5818**	**0.5214**	**0.4959**	**0.1373**	**0.2278**

Table 4. Comparison with other state-of-the-art methods on zero-shot covid-19 opacity region grounding task. *"Direct covid-19"* refers to directly use "Covid-19" to construct the prompt sentence while *"Covid-19 Description"* refers to replace the name "Covid-19" with its medical description. Our method can achieve better performance on different metrics.

from existing approaches, where all entities have been exposed to the model at training time, and prediction can be made by a retrieval-type approach, *i.e.*, compute the similarity between the image and the entity embedding by encoding the disease name with a language encoder [7], while we are considering a stricter setting for openset classification. Covid-19 is a new disease that only appeared in 2019, MIMIC-CXR reports collected in the year 2015 do not include any entity of covid-19, thus it requires the system to have the ability to diagnose truly unseen diseases.

As shown in Tab. 2, existing approaches that only rely on disease name struggles to make the correct diagnosis. While with our proposed approach by introducing medical knowledge, *i.e.*, using entity descriptions, our methods can understand the complex medical entity descriptions unseen in the training set, and significantly boost the performance 0.66 to 0.74 on AUC and from 0.59 to 0.70 on ACC.

5.1.2. Grounding

In addition to the plain diagnosis, explainability can be equally critical in healthcare, improving the reliability and trustiness of the machine learning systems. Here, we consider providing explainability by grounding the abnormality in the prediction and compare against the existing approaches. Similarly, we split the diseases into seen and unseen ones, depending on whether their names have appeared in the medical reports. Specifically, "Pneumonia" and "Pneumothorax" are viewed as seen, and "Covid-19" is viewed as unseen. Due to the space limitation, we refer the reader to supplementary material for visualization results.

Seen Diseases. We show the results for grounding on RSNA Pneumonia opacity and SIIM-ACR Pneumothorax collapse in Tab. 3. As shown in Tab. 3a, our proposed model surpasses existing approaches on all metrics, for example, we improve the pointing game score from 0.83 to 0.87, the detection Recall from 0.85 to 0.87, the detection precision from 0.50 to 0.64, the IOU from 0.30 to 0.32 and the Dice from 0.44 to 0.46. While on SIIM-ACR dataset (Tab. 3b), the pneumothorax region tends to be thin and narrow, localizing it can often be more challenging than that of opacity grounding [7], we thus only consider pointing game, recall, and precision. Similarly, our method can achieve significantly better performance than prior approaches.

Unseen Diseases. We also conduct the zero-shot grounding experiment on the unseen disease, namely, Covid-19, as shown in Tab. 4. Our model has shown consistent improvements in all metrics, *e.g.*, boosting the pointing game score from 0.40 to 0.58. One observation to be noticed is that, results in Tab. 4 are mostly consistent with those in Tab. 2, *i.e.*, better classification results tend to lead to better grounding. Overall, our model with knowledge-enhanced language encoding has facilitated the visual encoder to learn underlying evidence relating to the diseases, therefore, leading to more interpretable representations than prior approaches.

5.2. Fine-tuning

In this section, we consider the fine-tuning scenario, with the pre-trained model as initialization, and trained end-to-end on the downstream tasks. We test on three different tasks, namely, classification, segmentation, and grading. In classification and segmentation, the test splits and metrics are the same as in the "zero-shot" section. Grading is a new task we introduce in fine-tuning setting, which can be seen as a fine-grained classification task.

Dataset	Pneumonia			Pneumothorax			Covid-19			ChestX-ray14		
Data Portion	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	0.7107	0.8150	0.8626	0.4347	0.6120	0.6571	0.7861	0.9162	0.9554	0.6005	0.7365	0.7924
ConVIRT [56]	0.8398	0.8562	0.8761	0.7134	0.7826	0.9004	0.8675	0.9541	0.9726	0.6615	0.7658	0.8128
GLoRIA [22]	0.8599	0.8666	0.8846	0.7439	0.8538	0.9014	0.9065	0.9381	0.9728	0.6710	0.7642	0.8184
BioViL [7]	0.8233	0.8538	0.8836	0.6948	0.7775	0.8689	0.8989	0.9529	0.9729	0.6952	0.7527	0.8245
Ours	**0.8731**	**0.8799**	**0.8931**	**0.8527**	**0.9071**	**0.9188**	**0.9224**	**0.9657**	**0.9729**	**0.7721**	**0.7894**	**0.8323**

Table 5. Comparison of AUC scores with other state-of-the-art methods on fine-tuning classification task. The macro average of AUC scores on 14 diseases are reported for ChestX-ray14 dataset.

Diseases	Pneumonia			Pneumothorax			Covid-19		
Data Portion	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	0.4347	0.6047	0.7068	0.2133	0.3323	0.7447	0.1481	0.2367	0.3228
ConVIRT [56]	0.5706	0.6491	0.7201	0.5406	0.6121	0.7352	0.1995	0.2724	0.3737
GLoRIA [22]	0.6555	0.6907	0.7328	0.5673	0.5778	0.7694	0.1889	0.2809	0.3869
BioViL [7]	0.6824	0.7038	0.7249	0.6267	0.6998	0.7849	0.2113	0.3239	0.4162
Ours	**0.7064**	**0.7162**	**0.7579**	**0.6659**	**0.7210**	**0.7937**	**0.2445**	**0.3539**	**0.4399**

Table 6. Comparison of Dice scores with other state-of-the-art methods on fine-tuning segmentation tasks. Three diseases are reported, and for each disease, three data portions, 1%, 10%, 100% are adopted to show the performance change under different data amounts.

Methods	0			1			2			3			AVG		
	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
Scratch	0.7631	0.7036	0.6738	0.5383	0.3593	0.3223	0.6692	0.4328	0.7012	0.8420	0.5694	0.8770	0.7031	0.5163	0.6436
ConVIRT [56]	0.8453	**0.7769**	**0.7793**	0.6099	0.3938	0.4629	0.7202	0.4843	0.6445	0.9047	0.6154	0.8809	0.7700	0.5676	0.6919
GLoRIA [22]	0.8304	0.7577	0.7520	0.6208	0.3991	0.4922	0.7339	0.4958	**0.7037**	**0.9246**	**0.6667**	0.9102	0.7774	0.5798	0.7145
BioViL [7]	0.8034	0.7378	0.7148	0.6035	0.3912	0.4570	0.6860	0.4497	0.6777	0.9229	0.6500	0.9160	0.7540	0.5572	0.6914
Ours	**0.8502**	0.7646	0.7539	**0.6641**	**0.4140**	**0.5392**	**0.7605**	**0.5266**	0.7031	0.8845	0.6250	**0.9160**	**0.7898**	**0.5826**	**0.7280**

Table 7. Comparison with other state-of-the-art methods on fine-tuning edema severity grading multi-class classification task. AUC score is reported in the Table. "0,1,2,3" in the table represents the severity level and final macro average scores are reported.

5.2.1. Classification

We experiment on four different datasets, using 1%, 10%, 100% of the data for fine-tuning, that is consistent with the existing work [7, 22, 56]. As shown in Tab. 5, our model has demonstrated substantial improvements in the AUC scores over the existing approaches across all datasets, reflecting that our pre-trained representation is of higher quality than existing models. We refer the readers to the supplementary material for more detailed comparison results.

5.2.2. Segmentation

In Tab. 6, we conduct fine-tuning experiments on three different diseases for segmentation. We pick 1%, 10%, 100% of the data for fine-tuning. For all three different diseases with different image distributions, our methods surpass the existing state-of-the-art methods by a large margin, especially under the low-data regime.

5.2.3. Grading

Besides diagnosis, grading the disease severity level also plays an important role. Here, we adopt our pre-trained features and train them for the multi-class classification task, with 0 to 3 representing different severity levels. As shown in Tab. 7, for each level, the AUC, F1, and ACC scores are calculated as one class against all other ones, for example, 0 vs $\{1, 2, 3\}$. Final macro average scores of four levels are computed. On the majority of severity levels, our method can achieve the best results.

6. Conclusion

In this paper, we introduce a novel medical knowledge enhanced VLP model. First, we propose a report filter to extract useful medical entities with more useful supervision signals, simplifying complex raw reports with minimal information loss. Second, we translate entities into detailed medical descriptions and embed them with a text encoder enabling the network to understand complex medical expert-level knowledge. Finally, a transformer-based structure is proposed to do local region alignment. In experiments, We evaluate our method on different datasets under various settings. Our method shows strong zero-shot classification and grounding abilities, even facing unseen diseases. Besides, in fine-tuning setting, our method still outperforms state-of-the-art methods significantly, showing the superiority of our method.

References

[1] Society for imaging informatics in medicine: Siim-acr pneumothorax segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation. 2019. 2, 5

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. 5

[3] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72, 2019. 4

[4] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010. 2

[5] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*, 2021. 2

[6] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. 2, 4

[7] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21, 2022. Official Implementation: https://github.com/microsoft/hi-ml/tree/main/hi-ml-multimodal. 2, 3, 5, 6, 7, 8, 17

[8] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020. 2, 5

[9] Sihong Chen, Jing Qin, Xing Ji, Baiying Lei, Tianfu Wang, Dong Ni, and Jie-Zhi Cheng. Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in ct images. *IEEE transactions on medical imaging*, 36(3):802–814, 2016. 2

[10] 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*, pages 104–120. Springer, 2020. 2

[11] Hui Cui, Yiyue Xu, Wanlong Li, Linlin Wang, and Henry Duh. Collaborative learning of cross-channel clinical attention for radiotherapy-related esophageal fistula prediction from ct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–220. Springer, 2020. 2

[12] 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. *Scientific data*, 7(1):1–6, 2020. 2, 5

[13] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 5

[14] 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 *International Conference on Learning Representations*, 2020. 3

[15] Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2):100019, 2020. 1

[16] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505, 2017. 1

[17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 1

[18] Leyuan Fang, Chong Wang, Shutao Li, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. *IEEE transactions on medical imaging*, 38(8):1959–1970, 2019. 2

[19] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 5

[20] Ivan Gonzalez-Diaz. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE journal of biomedical and health informatics*, 23(2):547–559, 2018. 2

[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. 5

[22] 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. Official Implementation: `https://github.com/marshuang80/gloria`. 2, 3, 5, 6, 7, 8, 17

[23] Xin Huang, Yu Fang, Mingming Lu, Fengqi Yan, Jun Yang, and Yilu Xu. Dual-ray net: automatic diagnosis of thoracic diseases using frontal and lateral chest x-rays. *Journal of Medical Imaging and Health Informatics*, 10(2):348–355, 2020. 2

[24] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017. 2

[25] 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. 1, 2

[26] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2, 3, 5

[27] 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*, pages 4904–4916. PMLR, 2021. 2

[28] AEWP Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q, 2019. 1, 2, 5

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[30] 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. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2

[31] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: a large-scale database and cnn model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2019. 2

[32] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[33] Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, and Pheng-Ann Heng. Canet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, 39(5):1483–1493, 2019. 2

[34] Qing Liao, Ye Ding, Zoe L Jiang, Xuan Wang, Chunkai Zhang, and Qian Zhang. Multi-task deep convolutional neural network for cancer diagnosis. *Neurocomputing*, 348:66–73, 2019. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[36] Gabriel Maicas, Andrew P Bradley, Jacinto C Nascimento, Ian Reid, and Gustavo Carneiro. Training medical image analysis systems like radiologists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–554. Springer, 2018. 2

[37] Matthew BA McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR, 2020. 2

[38] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 2019. 2

[39] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, 2021. 1

[40] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rückert. Joint learning of localized representations from medical images and reports. *arXiv preprint arXiv:2112.02889*, 2021. 2, 3

[41] Maya Pavlova, Naomi Terhljan, Audrey G Chung, Andy Zhao, Siddharth Surana, Hossein Aboutalebi, Hayden Gunraj, Ali Sabri, Amer Alaref, and Alexander Wong. Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Frontiers in Medicine*, 9, 2022. 2, 5

[42] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018. 2

[43] 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*, pages 8748–8763. PMLR, 2021. 1

[44] 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. *Radiology. Artificial intelligence*, 1(1), 2019. 2, 5

[45] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020. 2

[46] Jiaxing Tan, Yumei Huo, Zhengrong Liang, and Lihong Li. Expert knowledge-infused deep learning for automatic lung nodule detection. *Journal of X-ray Science and Technology*, 27(1):17–35, 2019. 2

[47] Haiming Tang, Nanfei Sun, and Yi Li. Deep learning segmentation model for automated detection of the opacity regions in the chest x-rays of the covid-19 positive patients and the application for disease severity. *medRxiv preprint*, 2020. 2, 5

[48] Joseph J Titano, Marcus Badgeley, Javin Schefflein, Margaret Pain, Andres Su, Michael Cai, Nathaniel Swinburne, John Zech, Jun Kim, Joshua Bederson, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature medicine*, 24(9):1337–1341, 2018. 1

[49] Kun Wang, Xiaohong Zhang, Sheng Huang, Feiyu Chen, Xiangbo Zhang, and Luwen Huangfu. Learning to recognize thoracic disease in chest x-rays with knowledge-guided deep zoom neural networks. *IEEE Access*, 8:159790–159805, 2020. 2

[50] 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. 2, 5

[51] Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[52] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021. 2

[53] Yutong Xie, Yong Xia, Jianpeng Zhang, Yang Song, Dagan Feng, Michael Fulham, and Weidong Cai. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. *IEEE transactions on medical imaging*, 38(4):991–1004, 2018. 2

[54] Wenkai Yang, Juanjuan Zhao, Yan Qiang, Xiaotang Yang, Yunyun Dong, Qianqian Du, Guohua Shi, and Muhammad Bilal Zia. Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification. In *International conference on medical image computing*

and computer-assisted intervention, pages 558–566. Springer, 2019. 2

[55] Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, and Kayhan Batmanghelich. Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer, 2022. 3, 5, 14, 15

[56] 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 *Machine Learning for Healthcare*, 2022. Highest Starred Implementation: https://github.com/edreisMD/ConVIRT-pytorch. 2, 3, 4, 5, 6, 8, 17

Supplementary

A. The Entity Description Base and Position Set

Tab. 8 shows the descriptions we used to translate different entities. We have kept 75 entities in query set Q, following [55]. “Tail_abnorm_obs” entity represents some tailed entities and “exluded_obs” represents some entities useless for diagnosis. The last “covid-19” description row is only referred to for inference since it does not appear in pre-train reports.

Table 8. The Entity description used for translate single entity name. The description can be easily found from the open website.

Entity	Description
normal	It means the absence of diseases and infirmity, indicating the structure is normal.
clear	The lungs are clear and normal. No evidence for other diseases on lung.
sharp	This means that an anatomical structure s boundary or edge is clear and normal, meaning it is free of diseases.
sharply	“Sharply seen means that an anatomical structure is clearly visible.
unremarkable	This represents some anatomical structures are normal, usually modifying cardiac and mediastinal silhouettes.
intact	The bonny structure is complete and normal, meaning no fractures.
stable	The modified anatomical structures are normal and stable. No evidence for diseases.
free	It usually refers to free air and is associate with pneumothorax, atelectasis, pneumoperitoneum and emphysema.
effusion	A pleural effusion is accumulation of excessive fluid in the pleural space, the potential space that surrounds each lung. A pleural effusion infiltrates the space between the visceral pleura and the parietal pleura.
opacity	It is defined as an area of hazy opacification due to air displacement by fluid, airway collapse, fibrosis, or a neoplastic process. It is causes include infections, interstitial lung disease, and pulmonary edema.
pneumothorax	A pneumothorax is an abnormal collection of air in the pleural space between the lung and the chest wall. It may be caused by pneumonia or fibrosis and other diseases.
edema	Pulmonary edema, also known as pulmonary congestion, is excessive liquid accumulation in the tissue and air spaces of the lungs. It will show fluid in the alveolar walls.
atelectasis	It is the collapse or closure of a lung resulting in reduced or absent gas exchange. Findings can include lung opacification and loss of lung volume.
tube	It is a surgical drain that is inserted through the chest wall and into the pleural space or the mediastinum to remove undesired substances such as air.
consolidation	It is a region of normally compressible lung tissue that has filled with liquid instead of air. Consolidation must be present to diagnose pneumonia: the signs of lobar pneumonia are characteristic and clinically referred to as consolidation.
process	Acute process means there is abnormality in the anotomy structure.
abnormality	It means the exist of diseases and infirmity, indicating the structure is abnormal.
enlarge	It usually modifies cardiac silhouette and heart. Cardiomegaly is a medical condition in which the heart is enlarged.
tip	It refers to the top head of the tube.
low	The presence of low lung volumes may be a sign of a restrictive lung condition such as pulmonary fibrosis or sarcoidosis.
pneumonia	Pneumonia is an inflammatory condition of the lung primarily small air sacs known as alveoli. Pneumonia may present with opacities. Complications such as pleural effusion may also be found increasing the diagnostic accuracy of lung consolidation and pleural effusion
line	It refers to venous access line ot PICC lines.
congestion	Pulmonary congestion is defined as accumulation of fluid in the lungs, resulting in impaired gas exchange and arterial hypoxemia.
catheter	catheter is a tube placed in the body to drain and collect urine from the bladder
cardiomegaly	Cardiomegaly (sometimes megacardia or megalocardia) is a medical condition in which the heart is enlarged.
fracture	Fracture is a break in a rib bone.
air	It refers to the free air or gas in pleural space, indicating pneumothorax. Air displacement by fluid may lead to opacity.
tortuous	The Aorta is slightly tortuous. Sometimes it may refer to varicose veins
lead	It refers to the leading head of the tube.
disease	It means the exist of diseases and abnormalty, indicating the structure is abnormal.
calcification	Pulmonary calcification is a common asymptomatic finding. Pulmonary calcifications are caused mainly by two mechanisms: the dystrophic form and the metastatic form
prominence	It means the exist of some observation.
device	It refer to some equipments like picc tub, valve catheter, pacemaker hardware, arthroplastmarker icd defib, device support equipment and mediport
engorgement	Pulmonary vascular engorgement means obstruction of the normal flux of blood within the blood vessel network of the lung resulting in engorgement of pulmonary vessels
picc	A peripherally inserted central catheter (PICC), also called a PICC line, is a long, thin tube that s inserted through a vein in your arm and passed through to the larger veins near your heart.
clip	Surgical clips or vascular clips usually represent the one kind of medical equipments.
elevation	If tissues or anatomical structures are elevated, they are raised up higher than the normal location.
expand	It means the lungs are normally expanded and clear, indicating the absence of pneumothorax.
nodule	A lung nodule or pulmonary nodule is a relatively small focal density in the lung. it may be confused with the projection of a structure of the chest wall or skin, such as a nipple, a healing rib fracture or lung cancer.
wire	Sternotomis wires means the center line of the chest.
fluid	It refers to the water of liquid in the lung and it may indicate edema and other diseases.
degenerative	Degenerative disease is the result of a continuous process based on degenerative cell changes
pacemaker	pacemaker device usually represents the one kind of medical equipments.
thicken	Pleural thickening is an increase in the bulkiness of one or both of the pulmonary pleurae. It may cause by pulmonary Infection, empyema, tuberculosis or lung cancer.

Entity	Description
marking	It represents interstitial markings or bronchovascular markings
scar	A scar (or scar tissue) is an area of fibrous tissue that replaces normal tissues after an injury.
hyperinflate	Hyperinflated lungs are larger-than-normal lungs as a result of trapped air.
blunt	Blunting of the costophrenic angles is usually caused by a pleural effusion, as already discussed. Other causes of costophrenic angle blunting include lung disease in the region of the costophrenic angle, and lung hyperexpansion.
loss	The etiology of lung volume loss can be listed as follow: airway obstruction or compression, obesity, scoliosis, restrictive diseases such as pulmonary fibrosis and interstitial lung disease, tuberculosis.
widen	The mediastinum is not widened or enlarged.
collapse	Collapse lung refers to pneumothorax or atelectasis.
density	The density (more precisely, the volumetric mass density; also known as specific mass), of a substance is its mass per unit volume.
emphysema	Emphysema, or pulmonary emphysema, is a lower respiratory tract disease, characterized by air-filled spaces (pneumatosis) in the lungs, that can vary in size and may be very large.
aerate	Aeration (also called aerification or aeriation) is the process by which air is circulated through, mixed with or dissolved in a liquid or other substances that act as a fluid (such as soil).
mass	A lung mass is an abnormal growth or area in the lungs and it can also view as lung cancer.
crowd	Crowding of the bronchovascular structures is an important direct sign of volume loss. The atelectatic lung enhances densely after contrast administration because of closeness of the pulmonary arteries and arterioles within the collapsed lobe.
infiltrate	A pulmonary infiltrate is a substance denser than air, such as pus, blood, or protein, which lingers within the parenchyma of the lungs. Pulmonary infiltrates are associated with pneumonia, tuberculosis and sarcoidosis.
obscure	Some anatomy structures are not clear and is difficult to understand or see.
deformity	It means some body parts are abnormal or unjuried.
hernia	Lung hernia (Sibson hernia) is a protrusion of lung outside of thoracic wall. the hernia is noted after chest trauma, thoracic surgery or certain pulmonary diseases.
drainage	Tube drainage represents the one kind of medical equipment.
distention	Distension generally refers to an enlargement, dilation, or ballooning effect. It may refer to: Abdominal distension.
shift	The mediastinal shift is the deviation of the mediastinal structures towards one side of the chest cavity, usually seen on chest radiograph. It indicates a severe asymmetry of intrathoracic pressures.
stent	Tracheal stent represents the one kind of medical equipments
pressure	Pulmonary venous pressure is intermediate between mean PAP and LAP over all physiologic pressures
lesion	Lung nodules, pulmonary nodules, white spots, lesions—these terms all describe the same phenomenon: an abnormality in the lungs.
finding	Some observation on body parts, usually indicating abnormalty.
borderline	Borderline size of the cardiac silhouette means the cardiac silhouette is not enlarged and normal.
hardware	It represents the one kind of medical equipments.
dilation	The state of being larger or more open than normal
chf	Heart failure — sometimes known as congestive heart failure — occurs when the heart muscle doesn't pump blood as well as it should. When this happens, blood often backs up and fluid can build up in the lungs, causing shortness of breath.
redistribution	If the pulmonary edema is due to heart failure or fluid overload, you may also see cardiomegaly and distension of the pulmonary veins, particularly in the upper lung fields.
aspiration	Aspiration pneumonia occurs when food or liquid is breathed into the airways or lungs, instead of being swallowed.
tail_abnorm_obs	Some very rare diseases.
excluded_obs	Some meaningless observations.
covid-19	It is a contagious disease caused by a virus. Ground-glass opacities, consolidation, thickening, pleural effusions commonly appear in infection.

Additionally, we keep 51 positive positions, following [55], to form the position set P, as $\{$*trachea, left_hilar, right_hilar, hilar_unspec, left_pleural, right_pleural, pleural_unspec, heart_size, heart_border, left_diaphragm, right_diaphragm, diaphragm_unspec, retrocardiac, lower_left_lobe, upper_left_lobe, lower_right_lobe middle_right_lobe, upper_right_lobe, left_lower_lung, left_mid_lung, left_upper_lung left_apical_lung, left_lung_unspec, right_lower_lung, right_mid_lung, right_upper_lung right_apical_lung, right_lung_unspec, lung_apices, lung_bases, left_costophrenic right_costophrenic, costophrenic_unspec, cardiophrenic_sulcus, mediastinal, spine clavicle, rib, stomach, right_atrium, right_ventricle, aorta, svc, interstitium, parenchymal, cavoatrial_junction, cardiopulmonary, pulmonary, lung_volumes, unspecified, other*$\}$. "Other" is used to reprepresnt some tailed positions.

B. Details of Fusion Module

In the transformer-based fusion module, the queries are first passed through a self-attention layer and then followed by a multi-head attention layer between the modified queries and image features. In each head, the image features are processed by a linear key head and a linear value head as key embeddings and value embeddings independently. The value is weighted-added based on the attention map, which is calculated by the soft-max dot product of the keys and queries. Finally, a feed-forward network gathers the vector of different heads resulting in the output of this layer. The output vector of the former layer is considered as the entity query vector of the next layer. In formulation, if denoting $\Phi^i_{\text{fusion}}(\cdot, \cdot) : R^{N \times D_2} \times R^{P^2 \times D_2} \mapsto R^{N \times D_2}$ as the i-th layer, the procedure is expressed as:

$$\Phi^i_{\text{fusion}}(r_{i-1}, \mathcal{V}) = r_i, r_0 = Q. \tag{9}$$

C. Ablation Study

Our final method mainly contains three key parts, transformer-based fusion module, position location contrastive loss (PosCL), and entity description encoder (DE). We gradually remove the modules to analyze their effectiveness. "w/o (DE)" refers to removing DE module and "w/o (PosCL+DE)" refers to only maintaining the fusion module with basic CE loss. Tab. 9 and Tab. 10 shows the quantitative results.

Transformer Decoder. The lines about "w/o (PosCL + DE)" in tables demonstrate the performance of the basic model modified only by base CE loss. This model can exceed many former methods. This indicates the complex syntax will hurt the network to capture the useful entities significantly and our filtering operation combined with medical NER can greatly relieve the problem.

Position Contrastive Loss. The PosCL can significantly help the network to ground the abnormalities. As shown in the results by adding PosCL the classification results can be further improved, e.g., from 0.75 to 0.76 on AUC in ChestX-ray14 dataset. Besides classification, location contrastive loss can bring more gain in grounding. These results show position is another vital element in reports especially for grounding tasks. Our filtered triplets can conclude and clean the reports with little information loss and make the network learn the report information more straightforward.

Entity Description Encoder. By adding entity descriptions, we want to realize two goals. *First*, in addition to just learning from the image-report data, the network can actively learn the relationship between different entities based on the entity descriptions. As shown in tables, adding descriptions in most scenarios can help the network better understand the entity and bring gain to the final metric scores. *Second*,the description encoder enables our model to **handle openset new diseases**. Since the entity list is a close set during pre-training, our method will be only able to handle the seen diseases without DE, while, with a description encoder, our method can handle unseen diseases and understand complex medical disease knowledge.

Dataset	RSNA Pneumonia			SIIM-ACR Pneumothorax			ChestX-ray14		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
w/o (PosCL + DE)	0.8532	0.6079	0.7669	0.8768	0.6672	0.8187	0.7502	0.2374	0.8541
w/o (DE)	0.8537	0.6241	**0.8146**	**0.9017**	**0.7008**	**0.8584**	0.7621	0.2452	0.8606
Ours	**0.8694**	**0.6342**	0.8002	0.8924	0.6833	0.8428	**0.7676**	**0.2525**	**0.8619**

Table 9. Ablation study on zero-shot classification task. AUC, F1 and ACC scores are reported. For ChestX-ray 14, the metrics all refer to the macro average on the 14 diseases.

Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑
w/o (PosCL + DE)	0.7979	**0.8961**	0.4036	0.2783	0.4230
w/o (DE)	0.8424	0.8226	**0.6520**	0.3118	0.4610
Ours	**0.8721**	0.8661	0.6420	**0.3172**	**0.4649**

(a) Zero-shot grounding on Pneumonia

Methods	Pointing Game↑	Recall↑	Precision↑
w/o (PosCL + DE)	0.1786	0.3151	0.1336
w/o (DE)	**0.2080**	0.3178	0.1711
Ours	0.1975	**0.3562**	**0.1940**

(b) Zero-shot grounding on Pneumothorax

Table 10. Ablation study on zero-shot grounding tasks. (a) shows the results on RSNA Pneumonia dataset. (b) shows the results on SIIM-ACR Pneumothorax dataset.

D. Detailed results on ChestX-ray14

We further show the detailed performance of 14 different diseases on ChestX-ray14 dataset. Tab. 11 shows the results on the zero-shot setting. Our method can exceed the former methods for most diseases. The radar Fig. 3Y shows more visually how our model compares with other solutions under the zero-shot setting. Our method can exceed the former methods for most diseases. Under 100% fine-tuning settings, we achieved similarly excellent results as shown in Tab. 12.

Methods	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
ConVIRT [56]	0.4533	0.4601	0.7262	0.6238	0.6790	0.6322	0.6097	0.6698	0.6855	0.7699	0.4701	0.5293	0.6098	0.6220	0.6101
GLoRIA [22]	0.6680	0.7647	0.7975	0.6159	0.6722	0.5293	0.6755	0.4785	0.7306	0.8212	0.6033	0.5104	**0.6721**	0.7144	0.6610
BioViL [7]	0.5026	0.6328	0.7914	0.5791	0.7029	0.6126	0.6866	0.7516	0.7455	0.8533	0.7136	0.6751	0.6560	0.7692	0.6909
w/o (PosCL + DE)	0.7131	0.8100	0.8635	**0.6361**	0.7776	0.6740	0.6903	0.8124	0.7915	0.8869	0.7480	**0.6780**	0.6429	0.7784	0.7502
w/o (DE)	0.7420	0.8270	**0.8663**	0.6336	0.7867	**0.6974**	**0.7238**	0.8310	0.8037	0.8887	0.7865	0.6715	0.5414	0.8691	0.7621
ours	**0.7506**	**0.8299**	0.8636	0.6280	**0.7885**	0.6947	0.7236	**0.8361**	**0.8079**	**0.8888**	**0.7950**	0.6511	0.5783	**0.9097**	**0.7676**

Table 11. Comparison with other state-of-the-art methods on zero-shot ChestX-ray 14 diseases classification task. For each disease, AUC score is reported and the macro average AUC score is also reported. We use the first three letters to represent one disease but for "pneumonia" and "pneumothorax" we use the first two and the last letters.

Methods	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
Scratch	0.7835	0.8116	0.8563	0.6537	0.7788	0.6912	0.7004	0.8561	0.8090	0.8869	0.8564	0.7534	0.7454	0.9106	0.7924
ConVIRT [56]	0.8012	0.8360	0.8511	0.6613	0.8004	0.7490	0.6998	0.8666	0.8079	0.9023	0.9014	0.7933	0.7468	0.9627	0.8128
GLoRIA [22]	0.8263	0.8326	0.8596	**0.6641**	0.8179	0.7348	0.7104	0.8452	0.8129	0.8977	**0.9310**	0.7886	0.7608	0.9750	0.8184
BioViL [7]	0.8185	0.8543	0.8607	0.6660	0.8302	0.7633	0.7090	0.8595	**0.8287**	0.9031	0.9251	0.7912	0.7638	0.9696	0.8245
ours	**0.8291**	**0.8594**	**0.8719**	0.6565	**0.8382**	**0.7647**	**0.7378**	**0.8807**	0.8275	**0.9083**	0.9224	**0.7977**	**0.7784**	**0.9796**	**0.8323**

Table 12. Comparison with other state-of-the-art methods on fine-tuning ChestX-ray 14 diseases classification task. For each disease, AUC score is reported and the macro average AUC score is also reported. We use the first three letters to represent one disease but for "pneumonia" and "pneumothorax" we use the first two and the last letters.

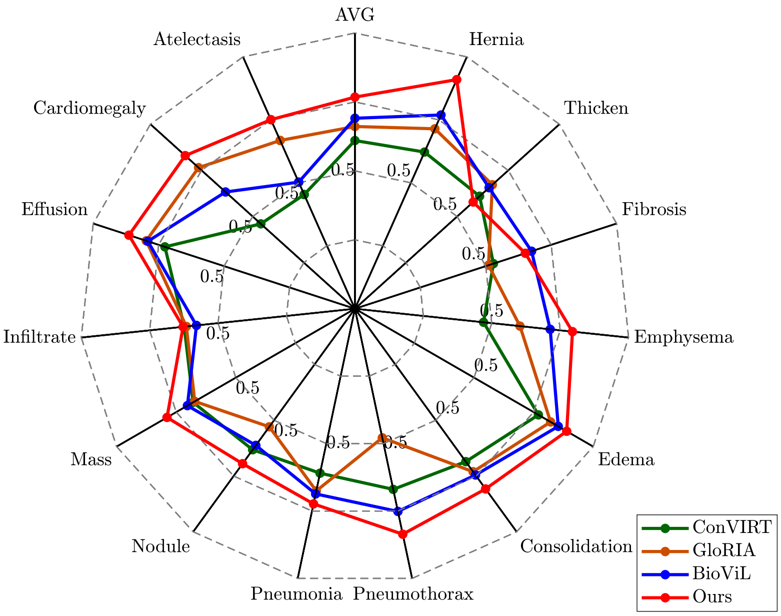
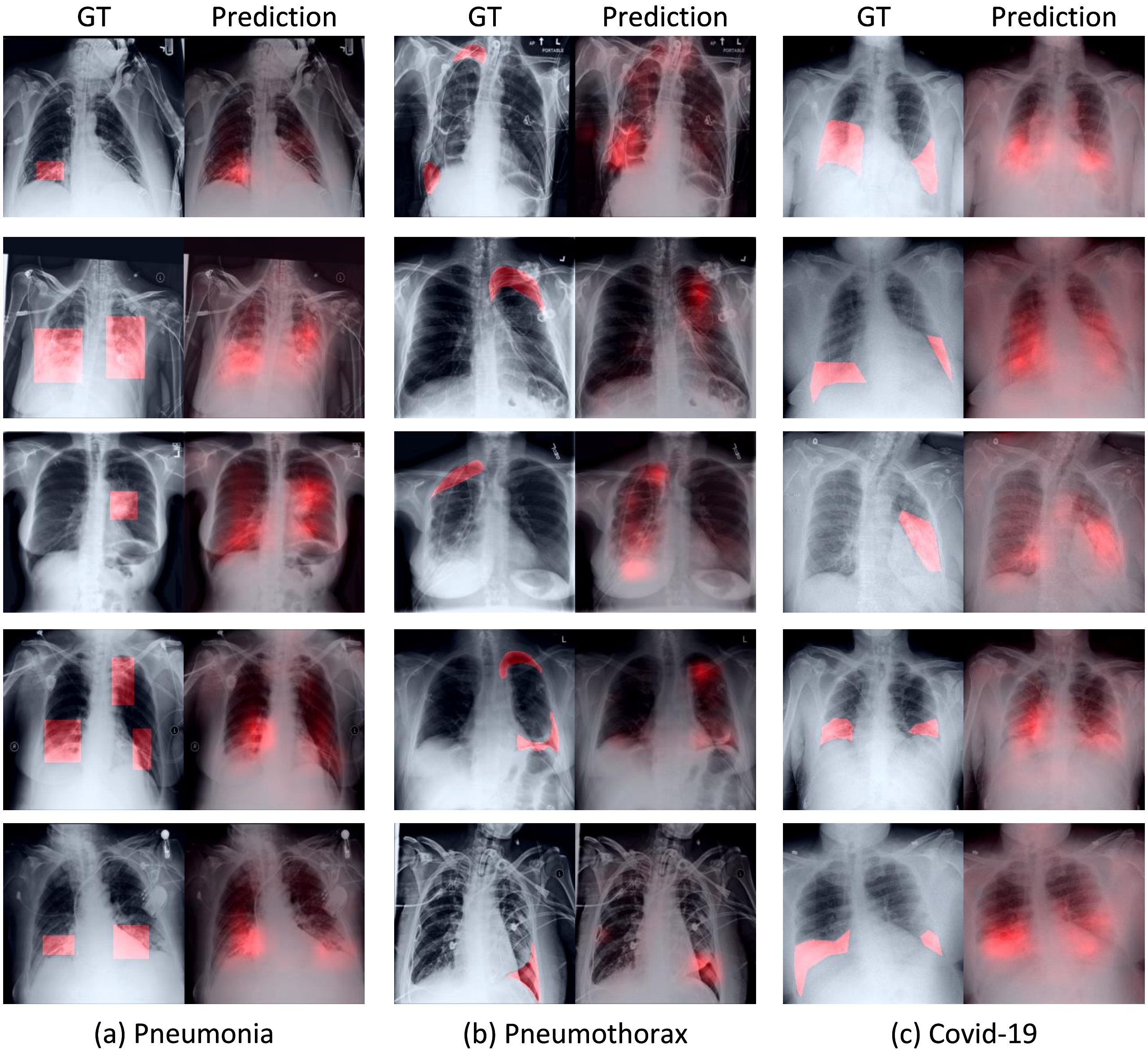


Figure 3. The radar figure of our method and other methods of ChestX-ray14 14 diseases. AUC scores are reported and, as shown, our method exceeds the previous state-of-the-art on most diseases.

E. Visualization Results

Fig. 4 shows visualization results of our model on zero-shot grounding task. As shown in figure, the ground truth of "Pneumonia" is given by bounding box and generally related to a large area region. Thus the metrics on this are higher than other two datasets. Our network captures its regions very well. For "Pneumothorax", its abnormality pattern is different from other diseases, which aim to capturing the collapsed part of the lung, rendering darker areas on the images rather than brighter opacity. Its ground-truth masks are generally thin and narrow while our network can still highlight its location. For "Covid-19", its image textual was similar to "Pneumonia", but since this is a totally new disease, grounding its regions is much more challenging. It requires the model to build relationships between them based on their complex definition and symptoms. The visualization results suggest that our model successfully achieve this, supporting that, for other unseen diseases, our model can also understand their complex descriptions.



(a) Pneumonia (b) Pneumothorax (c) Covid-19

Figure 4. The visualization of zero-shot grounding results of our method. Each column represents the results on one disease and the left in it is the ground-truth and right is the heatmap predication of our model. The brighter the red on the figure, the more likely the model considering this region to be associated with abnormalities.