MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis

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 enhancing medical visual-language pre-training (VLP) with domain-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 triplet extraction module to extract the medical-related information, avoiding unnecessary complexity from language grammar and enhancing the supervision signals;* **Second**, *we propose a novel triplet encoding module with entity translation by querying a knowledge base, to exploit the rich domain knowledge in medical field, and implicitly build relationships between medical entities in the language embedding space;* **Third**, *we propose to use a Transformer-based fusion model for spatially aligning the entity description with visual signals at the image patch level, enabling the ability for medical diagnosis;* **Fourth**, *we conduct thorough experiments to validate the effectiveness of our 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 [46, 20, 55, 19]. 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 are required, fundamentally limiting its practical values. As an alternative, recent research considers to train the model on the corpus, consisting of large amount of multi-modal data, that is generated from daily clinical routine, for instance, the most common example is the dataset of X-ray images with paired radiological reports [18, 28, 31].

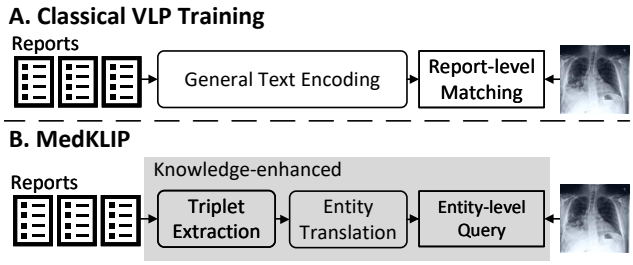

Figure 1: Our method mainly considers combining medical knowledge with VLP. We propose Triplet Extraction and Entity Translation modules, so that the network can be supervised with detailed entity-level signals.

This paper presents our preliminary investigation on vision-language representation learning in the medical domain, with the goal of better zero-shot disease diagnosis (classification) and grounding. Undoubtedly, these tasks have also been widely investigated in the computer vision community, with significant progress made on developing Foundational Models in the past years, for example, CLIP [50], ALBEF [33], BLIP[32], etc. However, to achieve such a goal in the medical domain, different challenges must be resolved, that requires research efforts from the community: *First*, data availability, training Foundation Models in computer vision normally require over millions of image-text pairs, while in the medical domain, only a few hundred thousand pairs are available [31]. The limited data challenges language models to understand the reports in free form [6]. *Second*, the problem considered in computer-aided diagnosis is naturally fine-grained, that requires distinguishing the medical concepts to understand the disease, as a consequence, domain knowledge is essen-

†: Corresponding author.

tial; *Third*, robustness is crucial, it is, therefore, preferable to have explainability, where diagnosis results come along with the visual grounding, to help radiologists understand the system, and build trust between human and machines.

Existing work in medical VLP (Vision-Language Pre-training) [68, 47, 25, 6] follows a straightforward training paradigm by matching raw reports with image scans, as shown in Fig.1A, ignoring the medical prior knowledge, and, thus, we propose a novel knowledge-enhanced visual-language model as shown in Fig. 1B. *First*, we propose a triplet extraction module to extract useful medical entities (keywords) from raw reports, and simplify each report into sets of triplets, denoted as $\{\text{entity}, \text{position}, \text{exist}\}$. Decomposing reports into triplets leads to an effective representation of the reports with minimal information loss due to the structural prior in reports; *Second*, we translate the medical entities into fine-grained descriptions by leveraging a well-defined medical word knowledge base, that tends to explain diseases with common vocabulary. Thus, computing text embeddings for these descriptions enables to implicitly establish relationships between medical entities; *Third*, we view the entities as a query set and adopt a transformer-based architecture for aligning the image patches with entity descriptions, that enables explicit supervision signals at entity level. Consequently, we can simultaneously infer the likelihood of certain diseases with the visual evidence in the form of a spatial heatmap, *i.e.*, providing rough grounding for explainability.

We pre-train the model on one widely-used medical image-report dataset MIMIC-CXR [31], and rigorously evaluate on the task of disease diagnosis across numerous public benchmarks, *e.g.*, ChestX-ray14 [58], RSNA Pneumonia [51], SIIM-ACR Pneumothorax [1], COVIDx CXR-2 [48], COVID Rural [54, 15], and EdemaSeverity [7]. We get state-of-the-art performance on zero-shot classification and grounding on different diseases, spanning different image distributions, with further fine-tuning, our model still exceeds previous models significantly.

2. Related Work

General Vision-Language Pre-training (VLP) Models. In computer vision, such line of research has gained tremendous success in the recent literature, generally speaking, the developed architectures can either be two-stream [4, 33, 30], *i.e.*, dual encoders, or those based on single-stream methods [35, 9], that favors visual-language fusion. In particular, several works [13, 38, 66] consider considers to combine the commonsense knowledge into the vision-language pre-training, however, in this paper, we focus on medical domain, which is clearly more fine-grained and requires significantly more expertise.

Medical Named-Entity-Recognition (NER) Models. Various natural language processing (NLP) approaches have been proposed to extract information from radiology reports [49, 28, 44, 52]. These early methods considered only the disease, thus causing information loss. Further state-of-the-art works [29, 61] are proposed to extract relationship between different entities without demand of pre-defined close disease set, retaining most of useful information with high accuracy. In weakly supervision [67] and report generation fields [14], NER methods have shown great impact, and greatly inspired us for more effective vision-language pre-training with medical domain knowledge injected.

Medical Knowledge Enhanced Models. In general NLP, many works considering combining knowledge, *e.g.*, K-BERT [41] while they condider general knowledge more and focuses on language encoding. In medical community, leveraging external medical knowledge to enhance deep learning models is also a quite important topic [62]. Depending the approaches of using medical knowledge, They can be classified into model-based or input-based. In model-based approaches, the authors aim to imitate the radiological or diagnosis practice to design models [34, 23, 57, 26, 45, 21, 12]. While in input-based approaches, the knowledge is treated as an extra input for computing features [65, 63, 53, 11] or to guide the final training loss [8, 27, 37, 39, 43], commonly used in report generation tasks [59, 64, 40, 11, 36]. However, none of these works are targeting on vision-language pre-training in medical domain with image-report.

Concurrent Works in Medical VLP. Existing medical VLP methods follow the two-stream flow [68, 25, 7, 47, 60, 10], *i.e.*, use contrastive learning and without fusion module, for example, ConVIRT [68] initially proposed to use contrastive loss as a proxy task for aligning the medical scan and corresponding reports, LoVT and GLoRIA then focus on improving the local alignment performance [25, 7]. BioViL notices the language pattern in reports is different from natural texts and re-designs the language model [6]. The recent arxiv preprint, MedCLIP [60], considers leveraging unpaired data to make up data scarcity. The most related to ours is CheXzero [56], which targets at zero-shot diagnosis. Align [10] focuses on leveraging external medical knowledge to improve the performance of VLP model on medical Visual Question Answering (VQA). However, they treat knowledge as an additional loss and we adopt a new training scheme leveraging medical knowledge by aligning entity, instead of raw reports, with images. Despite significant contribution has been made by existing work [68, 25, 7, 47, 60, 56], they still treat medical texts and images as common natural data and do not explicitly leverage the rich prior knowledge from medical domain. In this paper, we consider to incorporate domain knowledge, re-design the pre-training pipeline delicately and target at accurate diagnosis in X-ray scans.

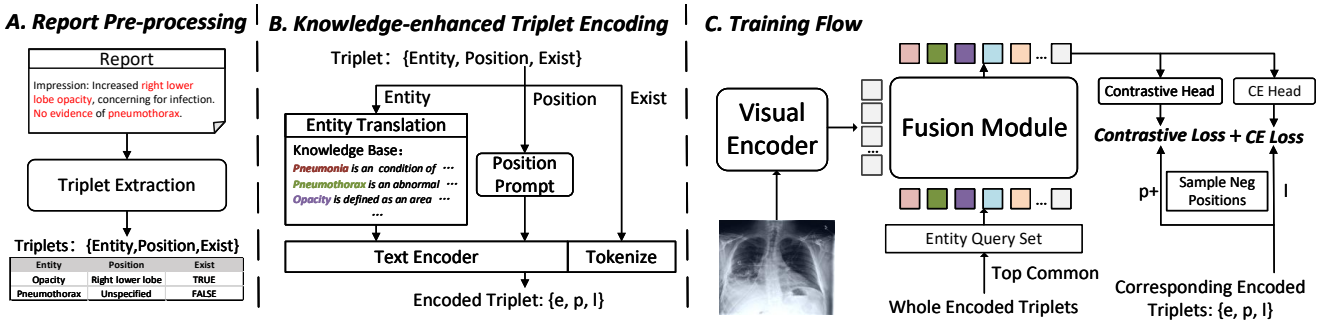

Figure 2: The whole framework of our method. We first pre-process the report into triplets leveraging triplet extraction module. Then we encode the extracted triplets and it is worth emphasizing that we translate the entities into detailed descriptions during encoding, by querying the medical knowledge base. Finally we change the training flow with triplets, *i.e.*, we query a transformer-based fusion module at entity level, which provides more detailed supervision signals.

3. Method

In this section, we start by describing the considered problem scenario in Sec. 3.1, followed by our report pre-process operation with triplet extraction in Sec. 3.2. Then we introduce our proposed knowledge-enhanced architecture in Sec. 3.3, including, visual encoder, knowledge-enhanced triplet encoder, and the fusion module for aligning visual-language signals. In Sec. 3.4, we describe the training procedure with the paired image-reports sourced from the daily routine X-ray scans and, in Sec.3.5, we introduce the procedure for inference.

3.1. Problem Scenario

Assuming we are given a training set with N samples, *i.e.*, $\mathcal{D}_{\text{train}} = \{(\mathcal{X}_1, \mathcal{T}_1), \ldots, (\mathcal{X}_N, \mathcal{T}_N)\}$, 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 (seen or unseen during training):

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

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 introduce our report pre-process operation with triplet extraction.

3.2. Report Pre-processing

To start with, we propose to pre-process medical reports with a **Triplet Extraction** module by removing the unnecessary complexity from language grammar. 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.

We condense the original reports with an off-shelf medical Named Entity Recognition (NER) method, namely RadGraph [29, 67], transforming reports into a set of triplets, as shown in Figure 2A. In detail, the medical key words can be extracted and classified as "entity" or "position" with the NER module. "Entity" refers to some clinical observations, like "Opacity". "Position" refers to the anatomical body part that occurs in a radiology report, like "right lower lobe". Besides, the NER module will also provide an "exist" label to conclude whether an entity is claimed to be exist, absent or uncertain in reports. Based on this, we can use a set of triplets, *i.e.*, $\{\text{entity}, \text{position}, \text{exist}\}$, to reformulate the sentence in reports, for example, the triplet $\{\text{Opacity}, \text{Right lower lobe}, \text{True}\}$ represents "It is true that there is opacity located at right lower lobe". **Note that**, the triplets with a specific "position" are not always termed as True in "exist" as radiologists may point out entities absent at some specific position.

Therefore, given a report $\mathcal{T}$ with multiple sentences, $\mathcal{T} = \{s_1, s_2, ..., s_M\}$, the extraction module independently operates on each of the sentences, and construct a number of triplets from the report:

$$\Phi_{\text{ex}}(s_j) = \{\text{entity}_n, \text{position}_n, \text{exist}_n\}, n \in [0, t_j], \quad (2)$$

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. After the triplet extraction, each report is equal to a set of triplets.

Discussion. In contrast to natural texts, information in medical reports tends to be more condensed, with radiologists pointing out the existence of abnormality and their positions

in the image. Meanwhile, medical terminologies tend to be professional, and within certain vocabulary (mostly listed in UMLS [5]), specially designed NER methods [29] demonstrate great performance on reports. Therefore, adopting the **Triplet Extraction** operation in medical VLP can avoid unnecessary complexity from understanding grammar, while still retaining the useful information in reports.

3.3. Architecture

In this section, we detail our proposed framework, consisting of three components, namely, visual encoding, knowledge-enhanced triplet encoding, and fusion module, as shown in Fig. 2B and Fig. 2C.

3.3.1 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:

$$\mathcal{V} = \Phi_{\text{visual}}(\mathcal{X}) \in \mathbb{R}^{h\times w\times d}, \tag{3}$$

h, w, d refer to the height, width, and feature dimension of the output, in our case, we adopt ResNet-50 as the visual backbone, and take the output from the 4th residual block. Note that, we make the architectural choice for fair comparison with existing work [68, 47, 25, 6], while other visual backbones, *e.g.*, ViT [17], can equally be applied.

3.3.2 Knowledge-enhanced Triplet Encoding

The goal of this module is to encode the triplets extracted from reports by incorporating medical domain knowledge as shown in Fig.2B.

Given a triplet as $\{\text{entity}, \text{position}, \text{exist}\}$, it is easy to code the "exist" as it only has three outcomes. We use $l \in \{0, 1, -1\}$ to tokenize it, 1 for True, 0 for False, -1 for uncertain. For the "entity" words, we translate them into **detailed descriptions** by querying some easy-access medical knowledge bases [1][2], *e.g.*, Description(["Pneumonia"])="It is a condition of the lung primarily ... present with opacities and pleural effusion ...". Despite its simplicity, converting the entities into descriptions is crucial for more reliable and zero-shot diagnosis, as it further decomposes the professional medical entities into basic attributes that are shared by different diseases, encouraging the model to capture a deep understanding of the visual evidence. For the "position" words, we use a prompt as "It is located at {position}" to form a sentence. Finally, we use ClinicalBERT [3] as a pre-trained text encoder, to compute the embedding for the "entity" and "position", and then adopt a linear MLP to project the embedding to desired dimensions:

$$e = \Phi_{\text{textual}}(\text{Description}(\{\text{entity}\})) \in \mathbb{R}^d, \tag{4}$$

$$p = \Phi_{\text{textual}}(\text{"It is located at } \{\text{position}\}\text{"}) \in \mathbb{R}^{d'}. \tag{5}$$

Each triplet has now been embedded into $\{e, p, l\}$.

Discussion. The extracted entities are medical terminologies that are only understandable to audiences with a medical background, while enriching them with detailed descriptions helps the model to capture a deep understanding of visual evidence for diseases. Such patterns can be generalized across diseases, as many attribute descriptions tend to be shared, enabling the model to build implicit relationships on seen classes and understand descriptions for unseen ones.

3.3.3 Fusion Module

With the triplets from reports, we can supervise the model **on the entity level** instead of the entire report level. The "position" and "exist" parts in triplets can be naturally seen as more fine-grained supervision labels. Specifically, we adopt a Transformer-based architecture, use the embedding of entities as query, iteratively attending the image embeddings, and output exist and position predictions of entities.

In detail, we select the top $|Q|$ most commonly appearing entities' embeddings in all training reports, to form an entity query set $Q = \{e_1, e_2, ..., e_{|Q|}\}$. The details of the entity query set is provided in the supplementary material (Sec. A). Then Q will be passed into a fusion module with the image representation $\mathcal{V}$ for alignment. The fusion module consists of multiple Transformer Decoder layers, with Q as Query, and $\mathcal{V}$ as Key and Value. The outputs are further fed into two MLPs, independently infer the existence of the entity and the entity's 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 predicted position for all entities. **Note that**, $\hat{m} \in \mathbb{R}^{H\times W}$ denotes the average of the cross-attention maps sourced from Transformer layers and is up-sampled to the size of input image with nearest interpolation. $\hat{m}$ is used for grounding at inference, as it naturally acts as a segmentation heatmap. During training, we will not directly calculate any loss on it.

Discussion. Adopting Transformer decoder enables to compute correspondences between entities and images at patch-level. Consequently, image features $\mathcal{V}$ are more suitable for downstream segmentation tasks and the average of cross-attention maps in each layers can be used directly for **zero-shot** grounding, providing explainable for diagnosis. Besides, the default self-attention on queries in Transformer structure can also build relationships across entities.

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

[2] UMLS [5] https://www.nlm.nih.gov/research/umls/

3.4. Training

Given a set of encoded triplets $\{e, p, l\}$ extracted form the pairing reports $\mathcal{T}$, we can compute training loss on the output of fusion module. For the existence prediction $\hat{s}$, we use binary cross-entropy with the corresponding "exist" labels l, and if l is -1 we just pass this label, denoted as $\mathcal{L}_{\text{cls}}$. To supervise the position prediction for each entity query, we adopt contrastive learning. We form a position set with top $|P|$ common position embeddings as a position set, $P = \{p_1, p_2, \cdots, p_{|P|}\}$, randomly sample M negative position embeddings from it, and use the corresponding position embedding p from triplets as positive:

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

where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors and $\mathcal{I}(\cdot, \cdot)$ is a random index sampling function. The position embeddings are un-normalized in calculation. **Note that**, some entities may not be mentioned in the report and thus, we can not find corresponding labels in triplets. We simply ignore the corresponding predictions while computing loss.

The final loss is the sum of the two:

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

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.

Discussion. In contrast to the existing approaches [68] that align images with entire reports, our training paradigm with triplets provides supervision at a more fine-grained entity level, rather than the global alignment between image and reports as has often done in existing approaches.

3.5. Inference

At inference time, given a test image, we can directly infer the existence of certain entities/disease, and ground their visual evidence. In particular, for the entities that have appeared in the entity query set Q, we simply adopt the corresponding elements from Q, while for those unseen ones, we replace the entity with a brief description provided by the user, and treat it as an extra query added to entity query set Q, **resembling zero-shot inference**. The existence output $\hat{s}$ can be directly applied for classification, the average cross-attention $\hat{m}$ between the target entity and the visual features are used for grounding.

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 [31, 22] 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 [58] 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 [51] 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 [25, 6]. 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 [48] and COVID Rural [54, 15] 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 [7] 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

This section describes the implementation for architectures. In **Pre-training**, the triplets extraction module and text encoders used in triplets encoding are all fixed, while the visual encoder and fusion module are trained end-to-end on the image-text pairs. In **Fine-tuning**, we adopt ResNet50 [24] initialized with image encoder for classification, and ResUNet [16] initialize its encoder with our pre-trained image encoder for segmentation. More details about exact values of different parameters and training progress can be found in supplementary material (Sec. B)

Dataset	RSNA Pneumonia			SIIM-ACR Pneumothorax			ChestX-ray14		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [68]	0.8042	0.5842	0.7611	0.6431	0.4329	0.5700	0.6101	0.1628	0.7102
GLoRIA [25]	0.7145	0.4901	0.7129	0.5342	0.3823	0.4047	0.6610	0.1732	0.7700
BioViL [6]	0.8280	0.5833	0.7669	0.7079	0.4855	0.6909	0.6912	0.1931	0.7916
CheXzero [56]	0.8579	0.6211	0.7942	0.6879	0.4704	0.5466	0.7296	0.2141	0.8278
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 [68]	0.6159	0.7057	0.6113	0.5208	0.6902	0.5266
GLoRIA [25]	0.6319	0.6938	0.5710	0.6659	0.7007	0.6083
BioViL [6]	0.6137	0.6958	0.5461	0.5382	0.6910	0.5375
CheXzero [56]	0.6462	0.7369	0.6629	0.6667	0.6400	0.6578
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 description.

4.4. Baselines

We compare with various existing state-of-the-art medical image-text pre-train methods, namely, ConVIRT [68], GLoRIA [25], BioViL [6] and CheXzero [56]. 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 authors. For zero-shot setting, we use the prompt as mentioned by BioViL [6] and compare to the very recent method (CheXzero [56]) that has shown to have better zero-shot diagnosis ability than radiologists. For fine-tuning, we all use the same setting as described in Sec. 4.3.

4.5. Metrics

AUC, F1 and ACC are measured for classification tasks. 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 Evaluation

In this section, we compare our method with other state-of-the-art methods under zero-shot setting, on classification and grounding. Due to the space limitation, we include the entire ablation study in the supplementary material (Sec. C), 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

Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑
GLoRIA [25]	0.7607	0.8330	0.1621	0.2182	0.3468
BioViL [6]	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 [25]	0.0651	0.2377	0.0585
BioViL [6]	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 and CheXzero 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 [25]	0.0364	0.2906	0.1073	0.0645	0.1141	0.2727	0.2821	0.1336	0.0596	0.1075
BioViL [6]	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 for entity encoding while *"Covid-19 Description"* refers to replace the name "Covid-19" with its description. Our method can achieve better performance on different metrics.

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 (Sec. D) for detailed comparison of 14 diseases.

Unseen Diseases. Here, we are considering a strict setting for openset classification, in particular, we use covid-19 to evaluate the systems. 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, our proposed approach, after introducing medical knowledge, *i.e.*, using entity descriptions, can understand the complex medical entity descriptions unseen in the training set, and significantly boost the performance from 0.66 to 0.74 on AUC and from 0.59 to 0.70 on ACC, demonstrating entity translation is vital for unseen diseases.

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 treated as seen, and "Covid-19" is treated as unseen. Due to the space limitation, we include visualization results in supplementary material (Sec. E).

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 [6], 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.

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 [68]	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 [25]	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 [6]	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 [68]	0.5706	0.6491	0.7201	0.5406	0.6121	0.7352	0.1995	0.2724	0.3737
GLoRIA [25]	0.6555	0.6907	0.7328	0.5673	0.5778	0.7694	0.1889	0.2809	0.3869
BioViL [6]	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 [68]	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 [25]	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 [6]	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 average scores are reported.

5.2. Fine-tuning Evaluation

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.

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 [68, 25, 6]. 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 (Sec. D) 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 triplet extraction module to extract useful medical-related triplets as more useful supervision signals, simplifying complex raw reports with minimal information loss. *Second*, we translate the entities in extracted triplets 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. *Additionally*, with fine-tuning, our method still outperforms state-of-the-art methods significantly, showing the superiority of our method.

Acknowledgement

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), STCSM (No. 22511106101, No. 18DZ2270700, No. 21DZ1100100), 111 plan (No. BP0719010), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

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

[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.

[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] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022.

[11] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020.

[12] 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.

[13] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[14] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. *arXiv preprint arXiv:2210.12186*, 2022.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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`.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665, 2022.

[37] 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.

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[39] 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.

[40] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 13753–13762, 2021.

[41] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022.

[57] 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.

[58] 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.

[59] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 655–664. Springer, 2022.

[60] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[61] 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.

[62] 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.

[63] 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.

[64] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510, 2022.

[65] 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.

[66] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[67] 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.

[68] 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`.