PRIOR: Prototype Representation Joint Learning from Medical Images and Reports

Pujin Cheng [1,2], Li Lin [1,3], Junyan Lyu[1,4], Yijin Huang [1,5],
Wenhan Luo [6], Xiaoying Tang [1,2,✉]
[1]Department of Electronic and Electrical Engineering, Southern University of Science and Technology
[2]Jiaxing Research Institute, Southern University of Science and Technology
[3]Department of Electrical and Electronic Engineering, The University of Hong Kong
[4]Queensland Brain Institute, The University of Queensland
[5]School of Biomedical Engineering, University of British Columbia
[6]Shenzhen Campus of Sun Yat-sen University

Abstract

Contrastive learning based vision-language joint pre-training has emerged as a successful representation learning strategy. In this paper, we present a prototype representation learning framework incorporating both global and local alignment between medical images and reports. In contrast to standard global multi-modality alignment methods, we employ a local alignment module for fine-grained representation. Furthermore, a cross-modality conditional reconstruction module is designed to interchange information across modalities in the training phase by reconstructing masked images and reports. For reconstructing long reports, a sentence-wise prototype memory bank is constructed, enabling the network to focus on low-level localized visual and high-level clinical linguistic features. Additionally, a non-auto-regressive generation paradigm is proposed for reconstructing non-sequential reports. Experimental results on five downstream tasks, including supervised classification, zero-shot classification, image-to-text retrieval, semantic segmentation, and object detection, show the proposed method outperforms other state-of-the-art methods across multiple datasets and under different dataset size settings. The code is available at https://github.com/QtacierP/PRIOR.

1. Introduction

Powered by large-scale labeled natural image datasets, deep learning has achieved great success in computer vision [17, 12, 41, 42, 51]. However, annotating medical images is extremely expensive and labor-intensive [4, 45]. A practical approach is to first pre-train a model on a large-scale labeled natural image dataset like ImageNet [11], and then fine-tune it on the downstream medical image dataset with limited annotation [39, 37, 58]. This approach may nevertheless fail to achieve generalized performance due to the domain gap between natural images and medical images [44, 33]. To effectively inherit representation from images of the same domain, self-supervised learning (SSL) methods [5, 16, 6, 8] have been proposed through pre-training on unlabeled datasets. However, it has been suggested that the performance gain of pre-training on unlabeled medical images is relatively limited compared to ImageNet initialization [57]. There are potentially two reasons: 1) The sample size of medical images, even unlabeled, is still quite limited compared to that of the ImageNet dataset; 2) Medical images often exhibit high inter-class similarity.

Recently, vision-language pre-training (VLP) has shown natural language supervision can effectively transfer linguistic information to visual representation via well-designed proxy tasks such as contrastive learning [38, 26, 54, 31, 25] and generative reconstruction [30, 55]. VLP may particularly work for medical image analysis since medical reports are highly likely to be accessible in most situations. However, jointly pre-training medical images and reports is still challenging. First, there are typically multiple sentences in a medical report, and the textual information is highly complex [53, 27, 24]. Second, most descriptions in a medical report are exclusively related to specific sub-regions in the corresponding medical image [32]. Most existing VLP methods tend to ignore fine-grained representation and may fail to transfer to locality-aware downstream tasks such as semantic segmentation and object detection. Several methods have employed local alignment losses through contrastive tasks [18, 32, 50]. For exam-

✉ Corresponding author: <tangxy@sustech.edu.cn>.

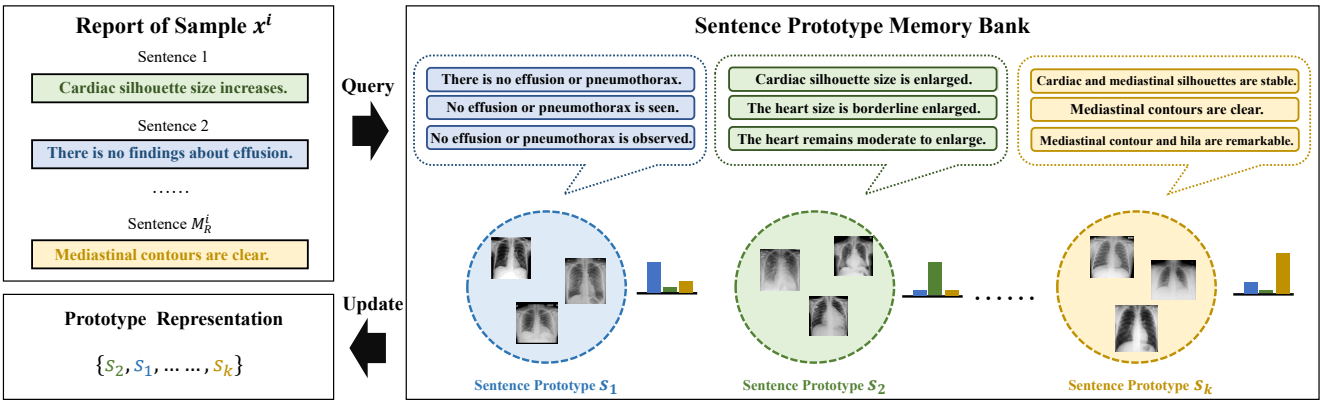


Figure 1. Illustration of the proposed sentence-wise prototype memory bank. The prototype embedding can group sentences sharing similar information. Each sentence representation is updated to the nearest prototype after querying.

ple, Huang *et al*. design a localized feature representation framework, but it still falls into the global alignment category [18]. Wang *et al*. propose a local contrastive loss to align locality-aware information [50], but a vanilla contrastive loss may easily ignore the similarity among nearby sub-regions due to their spatial locations and the overlapped sliding windows in convolution. To address this issue, Muller *et al*. employ positiveness probability sampling to avoid selecting a nearby sub-region as the negative sample [32]. However, this strategy is computationally heavy and time-consuming. Moreover, since contrastive learning is a discriminative SSL paradigm mainly focusing on high-level features, all those aforementioned methods tend to overlook low-level features, such as lesion boundaries in images and symptom descriptions in reports, which are nevertheless highly crucial for downstream medical image analysis tasks.

In this paper, we present a **P**rototype **R**epresentation framework via joint global and local alignment between medical **I**mages and rep**OR**ts (PRIOR), wherein we effectively combine contrastive learning and cross-modality conditional reconstruction. We consider sentence in reports and sub-region in images as the elementary local representation units, and the global representation is obtained via attention pooling over localized features. We propose a cross-modality alignment module to align representation between images and reports from both global and local views. To further learn locality-aware and fine-grained information, we utilize an encoder-decoder architecture to maximize the conditional mutual information between paired images and reports. The reconstruction decoder aims to reconstruct the masked image given the report and generate the report's representation given the image. We make use of the prior information that descriptions of most medical reports essentially can be summarized by multiple structured labels [19]. That is, sentence-level feature representation can be approximated by prototype categorization without any need to accurately retain redundant information such as syntax. As shown in Figure 1, each sentence in a report is discretely embedded as prototype representation. In this way, the sentence-level representation learning process can be treated as a classification-like task. Different from the image caption task which predicts each word auto-regressively [52, 29], sentences in a medical report are usually non-sequential. Inspired by [3], we use a parallel decoder to reconstruct sentence prototype embedding via bipartite matching. We successfully demonstrate the effectiveness of PRIOR on three challenging datasets for five downstream tasks, including supervised classification, zero-shot classification, image-to-text retrieval, semantic segmentation, and object detection.

Our main contributions are four-fold:

- We propose a novel cross-modality alignment module utilizing both global and local information to capture fine-grained features. Compared with previous works, our proposed alignment is more effectively operated on local representation.
- Considering the writing paradigm of a medical report, we present a prototype memory bank for report's sentence embeddings. Such discrete representation guides the linguistic features to further focus on high-level representation that links tightly to medical images.
- We leverage a conditional reconstruction task for both vision and language representation, which further facilitates cross-modality feature interaction and explores more structural and causal representation.
- PRIOR outperforms existing state-of-the-art (SOTA) methods on five tasks, including supervised classification, zero-shot classification, image-to-text retrieval, semantic segmentation, and object detection.

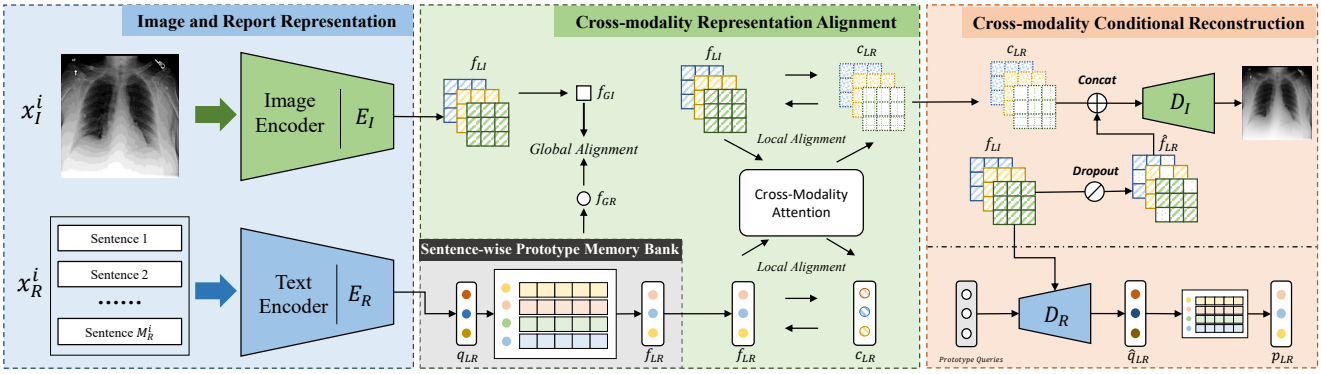


Figure 2. The overall framework of the proposed PRIOR. Given a pair of medical image and report, two independent encoders first encode each modality into a common embedding space. Then, the cross-modality alignment module aligns both global and local information between the two modalities. Finally, the cross-modality conditional reconstruction module reconstructs the masked image given the report and generates the sentence prototypes given the image.

2. Related Work

Self-Supervised Learning. Self-Supervised Learning (SSL) addresses the dilemma of limited annotated data. There are two main categories: (1) Contrastive SSL methods make use of the invariability of the same image under different augmented views. SimCLR [5] and MoCo [16, 6] attempt to maintain the invariability by spreading the representation of different images (negative samples) apart and clustering the representation of the same image (positive sample) together. Recent works have shown specific metric learning frameworks without any negative sample can outperform contrastive learning. Chen *et al*. propose a simple siamese representation learning framework with low memory and computation cost [7]. Several other works also identify that SSL with no image discrimination could obtain comparable performance to contrastive SSL [14]. (2) Generative SSL methods aim to learn data distribution with latent space embedding. Many of them rely on well-designed proxy tasks, such as colorization [23], image reconstruction [35, 46], inpainting [36], and so on. Recently, He *et al*. present a simple masked image auto-encoder with SOTA performance by using vision transformers [15].

Vision-Language Pre-training. Vision-Language Pre-training (VLP) is a specific type of SSL, wherein the self-supervised task usually relies on cross-modality interactions between visual and linguistic information. One of the most representative methods is CLIP [38]. It shows that contrastive learning can transfer rich semantic information from language supervision to visual representation [13, 9]. However, CLIP only learns coarse-grained representation and ignores localized information. Recently, many works have tried to build fine-grained relationships between natural images and text captions. For example, GLIP converts object detection into a phrase grounding task [26]. Yao *et al*. propose a VLP framework named FLIP involving fine-grained interactions, leveraging localized representation between visual and textual tokens [54]. However, none of them are specifically designed for pre-training medical images and reports. Furthermore, GLIP needs annotated bounding boxes for local alignment. And FLIP only focuses on the local alignment between coarse-grained sub-regions and words, which tends to ignore low-level information like textual structure in images and sentence-wise relation in reports. To capture low-level representation, CoCa adopts an encoder-decoder architecture for both contrastive learning and conditional generation [55], but it still cannot deal with the complex structures in medical reports and the low-level information in medical images.

Medical Image and Report Pre-training. In recent years, models that jointly pre-train vision and language for medical utility have been explored. ConVIRT performs global alignment through a contrastive loss [57]. On top of ConVIRT, GLoRIA [18], LoVT [32] and MGCA [50] all propose their own local alignment mechanisms. GLoRIA and MGCA consider each keyword in a report as the elementary local representation unit, but they do not consider any context across sentences. LoVT aligns sub-regions at the sentence level. However, it still abandons low-level features such as the shape of lesions and detailed symptom descriptions. Inspired by CoCa, we address the issues of existing medical image and report pre-training methods by introducing a conditional reconstruction task combined with prototype representation learning.

Prototype Representation Learning. The goal of prototype representation learning is to cluster similar units into a single embedding. Van *et al*. show that images could be represented discretely [47]. And Ramesh *et al*. prove that prototype learning works well on cross-modality interaction tasks [40]. For the medical domain, Chen *et al*. design a vision-language memory bank for radiology report generation [10]. Unlike report generation, foundation vision-

language models still need to preserve low-level visual information during the pre-training phase, since many downstream tasks involve boundaries and lesions. Here, we apply prototype learning to medical reports' sentence representation, converting a continuous sentence embedding space into a categorical distribution, all the while ensuring the preservation of structural integrity and fidelity across both high-level and fine-grained perspectives.

3. Methodology

VLP aims to learn a joint distribution $P(X_I, X_R)$ over a group of medical image $X_I = \{x_I^1, ..., x_I^N\}$ and report $X_R = \{x_R^1, ..., x_R^N\}$ pairs. Each sample x^i consists of an image x_I^i and the corresponding report x_R^i.

3.1. Framework Overview

The overall framework is shown in Figure 2. Given a pair of an image x_I^i and a report x_R^i, we first input them into two separate encoders. Then, each sentence embedding is updated by a learnable sentence-wise prototype memory bank (SPB) for the final linguistic representation. In addition to standard global contrastive learning, a local alignment module (LAM) is introduced to align local representation between sub-regions and sentences, aided by a cross-modality attention mechanism. Unlike the common Softmax based attention module, our proposed Sigmoid based LAM makes use of the common sense that not all sentences nor sub-regions are meaningful for cross-modality interaction. A cross-modality conditional reconstruction (CCR) module is designed to further leverage fine-grained representation. We reconstruct the masked image and the sentence prototypes as guided by the cross-modality representation.

3.2. Image and Report Representation

The image encoder E_I encodes each image x_I^i into M_I sub-regional representations, formulated as $f_{LI}^{i,v} \in \mathbb{R}^{M_I \times C_I}$, where v is the spatial location index of each sub-region. The global image representation f_{GI}^i is obtained by self-attention pooling over localized features [38].

Similarly, the text encoder E_R first encodes each report x_R^i into token-wise representations. After that, self-attention pooling is employed to derive M_R^i sentence-wise representations $q_{LR}^{i,u}$ over all tokens in the same sentence, where u indexes sentences and M_R^i is the total number of sentences in the report. Note that all these localized features will be updated by a sentence-wise memory bank that will be described in the next subsection. Different from previous works that consider the [CLS] token as the global representation, we gather sentence-wise features via an additional self-attention pooling operation to serve as the global representation f_{GR}^i.

3.3. Sentence Prototype Memory Bank

Normally, each sentence in a medical report identifies an observation, including symptoms, locations of lesions, etc. As such, a medical report's sentences can be treated as different representations of prototype features. Therefore, before going through the cross-modality LAM, we update the report's representation via a learnable memory bank. The proposed SPB $S \in \mathbb{R}^{K \times D}$ consists of K representation prototypes, where D is the dimension of the common embedding space. The operation of querying the sentence-wise prototype representation is formulated as

$$f_{LR}^{i,u} = s_k, \quad k = \arg\max_j \frac{q_{LR}^{i,u} \cdot s_j^T}{\|q_{LR}^{i,u}\| \|s_j\|}, \tag{1}$$

where $q_{LR}^{i,u}$ is the u-th sentence-wise representation of the i-th report, and s_j denotes the j-th prototype in the memory bank. We update the sentence-wise representation through querying SPB. Since Eq. (1) is discrete and not differentiable, we employ the Gumbel-softmax reparameterization trick [20] based on the querying distribution p_K over K cells. To push the localized representation towards the central prototype, an L1 loss is applied to explicitly update the memory bank, namely

$$\mathcal{L}_{proto} = \frac{1}{\sum_{i=1}^N M_R^i} \sum_{i=1}^N \sum_{u=1}^{M_R^i} \|f_{LR}^{i,u} - q_{LR}^{i,u}\|_1. \tag{2}$$

3.4. Cross-modality Representation Alignment

Paired image and report usually describe similar semantic information. For further cross-modality alignment, both global and local representation is projected into a common embedding space with dimension D via four independent MLPs. To explicitly align the global representation between the paired image and report, we maximize their mutual information. Specifically, we employ the InfoNCE loss [34] to estimate the lower bound of the mutual information. The global report-to-image alignment loss is defined as

$$\mathcal{L}_g^{I \leftarrow R} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(f_{GI}^i \cdot {f_{GR}^i}^T / \tau_1)}{\sum_{j=1}^B \exp(f_{GI}^i \cdot {f_{GR}^j}^T / \tau_1)}, \tag{3}$$

where B is the batch size and τ_1 is the temperature parameter. Similarly, the global image-to-report alignment loss is formulated as

$$\mathcal{L}_g^{R \leftarrow I} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(f_{GI}^i \cdot {f_{GR}^i}^T / \tau_1)}{\sum_{j=1}^B \exp(f_{GI}^j \cdot {f_{GR}^i}^T / \tau_1)}. \tag{4}$$

Global alignment focuses on discriminative high-level features. However, medical image analysis tasks are generally highly sensitive to fine-grained low-level information like lesion boundaries. As such, we design an LAM.

The first step of LAM is to generate cross-modality attention representation. Specifically, sub-region/sentence representation is respectively learned by sub-region-wise/sentence-wise weighted sum. Given that not all sub-regions/sentences contain meaningful semantic information, we do not use the Softmax function which regards each localized unit as contributing constantly to the cross-modality representation (*i.e.*, sum of weights is 1). The local report-to-image attention-based representation is formulated as

$$c_{LI}^{i,v} = \sum_{k=1}^{M_R^i} \sigma \left(\frac{Q^I f_{LI}^{i,v} \cdot K^I f_{LR}^{i,k\,T}}{\sqrt{D}} \right) \cdot V^I f_{LR}^{i,k}, \quad (5)$$

where Q^I, K^I and V^I are learnable matrices, and σ denotes the Sigmoid function. Similarly, the local image-to-report attention-based representation is

$$c_{LR}^{i,u} = \sum_{k=1}^{M_I} \sigma \left(\frac{Q^R f_{LR}^{i,u} \cdot K^R f_{LI}^{i,k\,T}}{\sqrt{D}} \right) \cdot V^R f_{LI}^{i,k}. \quad (6)$$

After obtaining the cross-modality attention representation, we perform local alignment for localized cross-modality information interaction. For local image-to-report alignment, the cross-modality representation and the localized linguistic representation of the same sentence are regarded as a positive pair, while those from different sentences are negative samples. We observe that different reports may share similar sentence representation. Therefore, for local alignment we use only sentences from the same report as the negative samples. As such, the image-to-report local alignment loss is defined as

$$\mathcal{L}_l^{R \leftarrow I} = -\frac{1}{\sum_{i=1}^{B} M_R^i} \sum_{i=1}^{B} \sum_{u=1}^{M_R^i} \log \frac{\exp(f_{LR}^{i,u} \cdot c_{LR}^{i,u\,T} / \tau_2)}{\sum_{k=1}^{M_R^i} \exp(f_{LR}^{i,u} \cdot c_{LR}^{i,k\,T} / \tau_2)}, \quad (7)$$

where τ_2 is the temperature parameter. For local report-to-image alignment, we conjecture that nearby sub-regions contain similar information induced by spatial structure and overlapping convolutions. It may cause feature collapse if we consider a positive-like sample as a negative one. Therefore, we apply the alignment methods proposed in [7], utilizing neither negative pairs in contrastive loss nor prior sampling strategies in [32]. The report-to-image local alignment loss is based on the cosine similarity and gets combined with asymmetrical projection, namely

$$\begin{aligned} \mathcal{L}_l^{I \leftarrow R} = &-\frac{1}{BM_I} \sum_{i=1}^{B} \sum_{v=1}^{M_I} \frac{1}{2} \textbf{sim} \left(h(f_{LI}^{i,v}), S(c_{LI}^{i,v}) \right) \\ &+ \frac{1}{2} \textbf{sim} \left(h(c_{LI}^{i,v}), S(f_{LI}^{i,v}) \right), \end{aligned} \quad (8)$$

where h is an independent MLP head, S is the stop-gradient operation, and **sim** denotes the cosine similarity. Finally, the total cross-modality representation alignment loss is

$$\mathcal{L}_{align} = \mathcal{L}_g^{I \leftarrow R} + \mathcal{L}_g^{R \leftarrow I} + \mathcal{L}_l^{I \leftarrow R} + \mathcal{L}_l^{R \leftarrow I}. \quad (9)$$

3.5. Cross-modality Conditional Reconstruction

Both global and local alignments concentrate on high-level features but are likely to ignore low-level information. To address this issue, we introduce a cross-modality conditional reconstruction task aiming to recover low-level information based on cross-modality interaction.

We consider two reconstruction tasks: masked image reconstruction and sentence prototype generation. We randomly zero out 50% image features via dropout. Then we concatenate the masked image features $\hat{f}_{LI}^i$ and the unmasked cross-modality features c_{LI}^i. To reconstruct the image, a lightweight decoder D_I is applied. The reconstruction loss of the masked image is presented as

$$\mathcal{L}_{ir} = \frac{1}{NM_I} \sum_{i=1}^{N} \| D_I \left(\textbf{cat}(\hat{f}_{LI}^i, c_{LI}^i) \right) - x_I^i \|_1, \quad (10)$$

where **cat** is the concatenation operation, $\hat{f}_{LI}^i$ represents the masked feature map, and c_{LI}^i denotes the cross-modality feature map.

For reconstructing the report information, we follow an encoder-decoder architecture and apply a cross-attention mechanism [49] to interact with visual features. Inspired by [3], parallel decoding is used to generate the sentence-wise prototypes. We employ U learnable embeddings (we call prototype queries) to generate different predicted prototypes in the decoder. If the number of sentences M_R^i is smaller than U, we pad the sentence embeddings with zero prototype $\varnothing$. Since the sentences in a report are not necessarily sequential, which means the decoding process should be permutation invariant, we apply the Hungarian algorithm [22] to identify the optimal prediction-label pair $\hat{\sigma}$ in the matching space $\mathfrak{S}_U$. We consider the L1 distance as the matching cost to perform bipartite matching

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_U}{\arg\min} \sum_{i=0}^{N} \sum_{u=0}^{M_R^i} \| s^{i,u} - \hat{q}_{\sigma(i,u)} \|_1, \quad (11)$$

where $s^{i,u}$ is the ground truth prototype and $\hat{q}_{\sigma(i,u)}$ is the predicted querying embedding that matches index $\sigma(i,u)$. After deriving the optimal matching pairs, we apply Eq. (1) to obtain the predicted prototype $\hat{s}_{\sigma(i)}$ based on the predicted distribution q_K.

The loss function for sentence prototype prediction has three components. The first term is the L1 distance between

the predicted querying embedding and the target prototype

$$\mathcal{L}_q = \frac{1}{\sum_{i=1}^{N} M_R^i} \sum_{i=0}^{N} \sum_{u=0}^{M_R^i} \|s^{i,u} - \hat{q}_{\sigma(i,u)}\|_1. \quad (12)$$

The second term focuses on distribution consistency, namely the KL divergence between the querying distribution p_K and the predicted distribution q_K

$$\mathcal{L}_{kl} = D_{KL}(q_k \| p_k). \quad (13)$$

The last term is similar to that in the global alignment. We first obtain the predicted global report representation p_{GR}^i via self-attention pooling over predicted prototypes p_{LR}. Then a global prediction alignment loss aiming to align the predicted global representation and the realistic global report representation is formulated as follows

$$\mathcal{L}_{gpa} = -\sum_{i=1}^{B} \log \frac{\exp(f_{GR}^i \cdot {p_{GR}^i}^T / \tau_3)}{\sum_{j=1}^{B} \exp(f_{GR}^j \cdot {p_{GR}^i}^T / \tau_3)}. \quad (14)$$

Finally, the overall loss for the cross-modality conditional reconstruction is

$$\mathcal{L}_{recon} = \mathcal{L}_{ir} + \mathcal{L}_q + \mathcal{L}_{kl} + \mathcal{L}_{gpa}. \quad (15)$$

3.6. Loss

The overall loss function of the proposed PRIOR contains the following terms: the cross-modality representation alignment loss $\mathcal{L}_{align}$, the SPB updating loss $\mathcal{L}_{proto}$ and the cross-modality conditional reconstruction loss $\mathcal{L}_{recon}$. We optimize these terms jointly via a weighted sum

$$\mathcal{L} = \mathcal{L}_{align} + \lambda_{proto}\mathcal{L}_{proto} + \lambda_{recon}\mathcal{L}_{recon}, \quad (16)$$

where λ_{proto} and λ_{recon} are hyper-parameters.

4. Experiments

4.1. Datasets

MIMIC-CXR [21]. We pre-train all SSL and VLP methods on version 2 of MIMIC-CXR, which contains 377110 X-ray chest images and 227827 corresponding reports. Because downstream tasks all employ frontal-view images, we remove all lateral-view images. Additionally, we abandon short reports containing fewer than four sentences, ending up with 182475 image-report pairs.

RSNA Pneumonia Detection [43]. The RSNA Pneumonia Detection dataset contains more than 30000 X-ray chest images. All data have image-level pneumonia classification and bounding-box pneumonia detection labels. Therefore, we utilize the RSNA dataset for two downstream tasks: supervised classification and object detection. For the classification task, we split the dataset by 60%/20%/20% for training/validation/testing. For the object detection task, we only use the pneumonia samples (a total of 6012) through five-fold cross-validation.

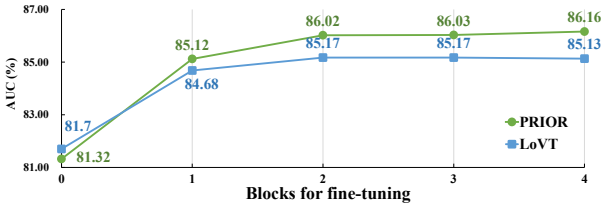


Figure 3. Partial fine-tuning results on 1% CheXpert. The number of blocks for fine-tuning increases from left to right. Fine-tuning with 0 block is equivalent to linear evaluation, while fine-tuning with 4 blocks is equivalent to full fine-tuning.

SIIM-ACR Pneumothorax. SIIM-ACR Pneumothorax contains 12954 X-ray chest images, together with image-level pneumothorax annotation and pixel-level segmentation mask if pneumothorax exists. We use them for downstream supervised classification and semantic segmentation. For the classification task, we split the dataset by 60%/20%/20% for training/validation/testing. For the segmentation task, we only use the pneumothorax samples (a total of 2669) through five-fold cross-validation.

CheXpert. [19] CheXpert contains more than 220000 X-ray chest images. Each image is labeled for five independent diseases. And thus the downstream task for this dataset is multi-label classification. We split the training set by 80%/20% for training/validation and use the official validation set for testing. Similar to that in GLoRIA [18], we create a sub-set (named CheXpert 5x200) from the training set of CheXpert for zero-shot classification and image-to-text retrieval, including 200 exclusively positive samples from each disease.

4.2. Implementation Details

Following previous works, we use the BERT model pre-trained by ClinicalBERT [1]. For image feature encoding, we apply ResNet50 [17] as the backbone. We resize all images into 224×224, and all reports are padded or truncated to have a fixed length of 256 tokens. We choose 768 to be the common latent dimension D. The SPB size K is set as 512. The temperature coefficients τ_1, τ_2, and τ_3 are all set as 0.01. The loss weight coefficients are set as $\lambda_{recon} = 1$ and $\lambda_{proto} = 10$. The overall framework is trained with a batch size of 128 for 100 epochs. The learning rate is initialized as $1e-5$ and decays following a cosine policy.

4.3. Evaluation Experiments

We compare our method with several widely-used initialization methods, including three image-level SSL methods (MoCo, MoCoV2, SimCLR), five medical VLP methods (ConVIRT, GLoRIA, BioViL, LoVT, MGCA), and ImageNet pre-training. All SSL and VLP methods are re-

Table 1. Supervised classification results by fine-tuning on three downstream datasets. All methods are trained on different portions of the training set from 1% to 100% and evaluated by AUC-ROC. Each value is the average of five runs. The best results are highlighted in bold and red, and the second-best results are highlighted in blue. Underline indicates that there is no statistically significant difference between the PRIOR and the best other method.

Init. Methods	RSNA			SIIM			CheXpert		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init.	56.42 ± 2.79	57.62 ± 3.27	83.63 ± 0.27	55.41 ± 1.72	57.83 ± 0.60	80.63 ± 0.55	52.96 ± 3.52	73.37 ± 0.68	75.12 ± 0.35
ImageNet Init.	80.27 ± 0.71	81.85 ± 0.46	88.01 ± 0.13	78.33 ± 1.79	82.88 ± 0.37	88.41 ± 0.41	77.57 ± 1.13	82.94 ± 0.91	87.08 ± 0.37
MoCo [16]	82.33 ± 0.47	85.22 ± 0.11	87.90 ± 0.09	75.49 ± 0.29	81.01 ± 0.67	88.43 ± 0.24	78.00 ± 0.62	86.27 ± 0.30	87.24 ± 0.05
MoCoV2 [6]	83.07 ± 0.49	85.88 ± 0.28	88.60 ± 0.07	77.10 ± 0.49	81.12 ± 0.64	90.72 ± 0.21	79.64 ± 0.53	86.04 ± 0.23	87.44 ± 0.27
SimCLR [5]	80.18 ± 2.78	84.60 ± 0.12	88.07 ± 0.11	74.97 ± 2.17	83.21 ± 0.49	88.72 ± 0.28	67.41 ± 2.74	86.74 ± 0.36	87.97 ± 0.22
ConVIRT [57]	84.17 ± 0.77	86.92 ± 0.13	88.74 ± 0.36	84.17 ± 0.77	85.66 ± 0.45	91.50 ± 0.08	85.02 ± 0.28	87.58 ± 0.53	88.21 ± 0.46
GLoRIA [18]	84.12 ± 0.47	86.83 ± 0.53	89.13 ± 0.12	85.05 ± 1.62	88.51 ± 0.78	92.11 ± 0.18	83.61 ± 0.52	87.40 ± 0.39	88.34 ± 0.12
BioViL [2]	81.95 ± 0.29	85.37 ± 0.61	88.62 ± 0.14	79.89 ± 1.06	81.62 ± 2.37	90.48 ± 0.14	80.77 ± 1.26	87.56 ± 0.16	88.41 ± 0.51
LoVT [32]	85.51 ± 0.36	86.53 ± 0.54	89.27 ± 0.13	85.47 ± 0.85	88.50 ± 0.76	92.16 ± 0.36	85.13 ± 0.48	88.05 ± 0.15	88.27 ± 0.07
MGCA [50]	**85.80 ± 0.68**	**87.66 ± 0.21**	**89.30 ± 0.16**	86.10 ± 0.31	**89.68 ± 0.09**	92.04 ± 0.09	85.63 ± 0.33	87.65 ± 0.33	88.30 ± 1.48
PRIOR (ours)	85.74 ± 0.36	87.08 ± 0.19	89.22 ± 0.17	**87.27 ± 0.39**	89.13 ± 0.11	**92.39 ± 0.23**	**86.16 ± 0.64**	**88.31 ± 0.20**	**88.61 ± 0.29**

trained on the MIMIC-CXR dataset with their official implementations and default hyper-parameters. ConVIRT applies only a global alignment loss, and may not perform well on locality-aware downstream tasks. GLoRIA, LoVT and MGCA are three SOTA VLP methods targeting medical images and reports with well-designed local alignment losses, While BioViL designs an improved language model for more high level clinic features.

All compared methods are evaluated on five downstream tasks, including supervised classification, zero-shot classification, image-to-text retrieval, semantic segmentation, and object detection. For zero-shot classification and image-to-text retrieval, we compare with GLoRIA and ConVIRT which present prompt engineering in their original papers. Although most of the existing methods perform linear evaluation, similarly to [15], we conduct a partial fine-tuning evaluation on CheXpert with 1% training data, updating only the last several residual blocks in ResNet50. We observe that the linear evaluation performance is not strictly keeping in line with the fine-tuning performance, as clearly shown in Figure 3. Additionally, if we unfreeze only the last block, the performance significantly increases, which agrees with previously-reported results [15]. Linear evaluation cannot well deal with non-linear features which are nevertheless critical for downstream tasks. Therefore, we only report the fine-tuning results.

4.4. Results on Classification

We conduct supervised classification experiments on RSNA, SIIM, and CheXpert respectively. AUC-ROC is employed as the evaluation metric. Table 1 shows that all VLP methods outperform image-level SSL methods and ImageNet initialization. Collectively, PRIOR achieves the best results, especially on SIIM and CheXpert with different training data sizes. Even though ConVIRT, GLoRIA, BioViL and LoVT have leveraged semantic representation from large-scale report supervision, they may have ignored the clinical paradigm relation between images and reports, the relation of which is successfully modeled by our proposed SPB. MGCA conducts disease-level alignment and performs well on the RSNA dataset, but it neglects sentence-level clinical information which is nevertheless crucial for complex downstream tasks like multi-label classification on CheXpert.

To analyze the quality of the cross-modality representation, we measure the zero-shot classification performance on CheXpert 5x200. We use models pre-trained on MIMIC-CXR to predict the labels of CheXpert. We compare PRIOR with two VLP methods, namely ConVIRT and GLoRIA, on the zero-shot classification task. For generalization, we employ the simplest prompt engineering strategy without any heavy ad-hoc design. The prediction is given by measuring the similarly between the predicted image and prompts. We take the sum of the global alignment loss and the cross-modality local attention as the similarity. As shown in Table 2, neither ConVIRT nor GLoRIA can capture correlative semantic information with simple prompt engineering. PRIOR attains the best zero-shot classification results, suggesting it can effectively transfer knowledge from linguistic supervision to vision domain.

4.5. Image-to-text Retrieval

We conduct image-to-text retrieval experiments on the CheXpert 5x200 dataset. Since the reports in CheXpert are inaccessible, we randomly select 1000 reports from MIMIC-CXR with exclusive 200 samples for each of 5 diseases. We measure the performance using Precision@K to check if the retrieved report matches the query image label. Our results in Table 3 demonstrate that PRIOR significantly outperforms both GLoRIA and ConVIRT.

4.6. Results on Segmentation and Detection

We examine the contribution of the local representation through the semantic segmentation task. We use U-Net [42] as the segmentation architecture. Table 4 presents the results of all methods on the SIIM dataset. Since GLoRIA,

Table 2. Zero-shot classification results on CheXpert 5x200.

VLP Methods	CheXpert Zero-shot Classification		
	Accuracy	F1-Score	Precision
ConVIRT [57]	21.30	19.03	17.65
GLoRIA [18]	23.20	15.79	**44.01**
PRIOR (ours)	**34.90**	**30.56**	35.88

Table 3. Image-to-text retrieval results on CheXpert 5x200.

VLP Methods	CheXpert Image-to-text Retrieval			
	Prec @ 1	Prec @ 2	Prec @ 5	Prec @ 10
ConVIRT [57]	20.3	19.8	19.7	19.9
GLoRIA [18]	29.3	29.0	27.8	26.8
PRIOR(ours)	**40.2**	**39.6**	**39.3**	**38.0**

Table 4. Semantic segmentation results by fine-tuning on SIIM. All methods are trained on different portions of the training set from 1% to 100% and evaluated by Dice. Underline indicates that there is no statistically significant difference between the PRIOR and the best other method.

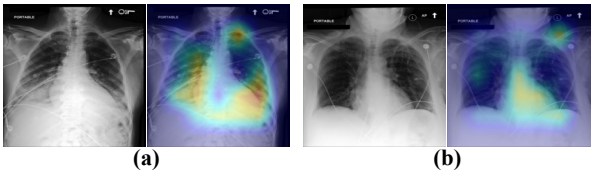
Init. Methods	SIIM Segmentation		
	1%	10%	100%
Random Init.	6.59 ± 0.40	16.61 ± 1.44	38.25 ± 0.90
ImageNet Init.	14.92 ± 3.38	25.26 ± 3.04	45.07 ± 0.87
MoCo [16]	17.66 ± 2.14	27.81 ± 2.19	41.18 ± 1.53
MoCoV2 [6]	18.19 ± 3.35	28.57 ± 2.92	44.01 ± 1.36
SimCLR [5]	16.55 ± 2.27	23.36 ± 0.66	40.62 ± 0.69
ConVIRT [57]	18.48 ± 2.43	27.32 ± 2.60	41.72 ± 0.69
GLoRIA [18]	18.78 ± 2.85	33.55 ± 2.14	45.46 ± 1.69
BioViL [2]	18.13 ± 1.67	27.78 ± 1.10	45.54 ± 1.14
LoVT [32]	18.81 ± 1.04	32.68 ± 0.95	44.65 ± 1.36
MGCA [50]	18.84 ± 2.37	33.54 ± 1.89	45.39 ± 1.35
PRIOR (ours)	**20.43 ± 2.20**	**34.81 ± 1.60**	**46.01 ± 1.03**

LoVT, MGCA and the proposed PRIOR consider localized feature representation during the pre-training phase, all of them outperform other methods. Among all the five medical image and report VLP methods, PRIOR achieves the most superior performance, especially with limited training data. A plausible reason is that PRIOR learns detailed locality-aware information via conditional reconstruction.

To further evaluate the effectiveness of our proposed method, we conduct object detection on the RSNA dataset. We use Faster R-CNN [41] as the detection architecture. Table 5 presents the performance comparison results on RSNA. We observe that some SSL methods outperform VLP methods, and we conjecture it is because medical reports may be incapable of precisely providing pneumonia information, which is consistent with the original motivation of building the RSNA dataset [43]. With that being said, our proposed PRIOR still outperforms all other methods, successfully establishing its effectiveness.

4.7. Ablation Study

This section presents the ablation study results of PRIOR under the 1% SIIM dataset setting for supervised classification and semantic segmentation. We compare the performance of the following variants: (1) Baseline, which only uses global alignment; (2) Baseline+LAM, which uses both global alignment and local alignment; (3) Baseline+LAM+CCR, which uses global alignment, local alignment and cross-modality conditional reconstruction; (4) Baseline+LAM+CCR+SPB (proposed PRIOR), which uses global alignment, local alignment, cross-modality conditional reconstruction, and sentence prototype memory bank. Note that variant (3) reconstructs only the masked image because the sentence-level representation is not categorized and unavailable without SPB. Table 6 tabulates the ablation study results, demonstrating all components are essential for extracting both global and local features.



(a) **(b)**

Figure 4. Representative cross-modality attention maps. (a) The related sentence is "*increased bibasilar opacities are combination of increased bilateral pleural effusions and bibasilar atelectasis*". (b) The related sentence is "*unchanged normal size of the cardiac silhouette* ".

4.8. Visualization of Cross-modality Attention

We visualize the cross-modality attention map to show the effectiveness and interpretability of the proposed PRIOR. Given an image and a corresponding sentence, we first obtain the cross-modality attention map by Eq. (5). Then, we visualize the attention map by resizing it to the same size as the image and superimposing it on top of the image. Figure 4 presents two examples to illustrate that the proposed attention module can detect correlative regions given a sentence. For example, the heatmap in Figure 4a highlights the abnormal regions precisely, aligning well with the report. Figure 4b also correctly detects the heart location.

4.9. Visualization of Feature Representation

To assess the quality of the feature representation from the proposed method, we utilize t-SNE [48] to visualize the high-level embeddings from the last layer of the image encoder on CheXpert 5x200. Apparently, results from the proposed method exhibit a clustering pattern more evidently, while results from other methods are more isotropic and homogeneous and fail to distinguish different categories.

4.10. Compared to generalized medical model

Recently, notable contributions have emerged in the form of substantial multi-modal medical datasets [56, 28]. One may wonder whether a domain-generalized large-scale dataset can bring significant improvements over a small-scale but single-modality dataset. In this context, we com-

Table 5. Object detection results by fine-tuning on RSNA. All methods are trained on different portions of the training set from 1% to 100% and evaluated by mAP(0.5:0.95). - means the method's mAP is lower than 0.1. Underline indicates that there is no statistically significant difference between the PRIOR and the best other method.

Init. Methods	RSNA Object Detection		
	1%	10%	100%
Random Init.	-	2.66 ± 1.34	18.99 ± 1.36
ImageNet Init.	-	7.85 ± 2.68	21.16 ± 1.94
MoCo [16]	-	16.25 ± 1.96	21.00 ± 2.00
MoCoV2 [6]	0.19 ± 0.22	18.45 ± 1.75	21.65 ± 2.10
SimCLR [5]	-	14.00 ± 1.96	21.76 ± 1.93
ConVIRT [57]	-	18.80 ± 2.79	21.28 ± 1.65
GLoRIA [18]	-	17.95 ± 1.37	21.37 ± 1.07
BioViL [2]	-	19.11 ± 1.13	19.50 ± 0.97
LoVT [32]	-	17.46 ± 2.21	21.80 ± 2.75
MGCA [50]	-	19.10 ± 1.83	21.33 ± 1.63
PRIOR (ours)	**0.20 ± 0.27**	**19.61 ± 1.93**	**22.20 ± 1.58**

Table 6. Ablation results of the proposed method on 1% SIIM. LAM denotes the local alignment module, CCR denotes the cross-modality conditional reconstruction and SPB denotes the sentence prototype memory bank.

Components			1% SIIM Classification	1% SIIM Segmentation
LAM	CCR	SPB		
			82.01 ± 2.21	17.88 ± 1.24
✓			84.86 ± 0.65	18.70 ± 2.33
✓	✓		85.80 ± 0.86	19.33 ± 2.22
✓		✓	86.81 ± 0.47	19.87 ± 2.03
✓	✓	✓	**87.27 ± 0.39**	**20.43 ± 2.20**

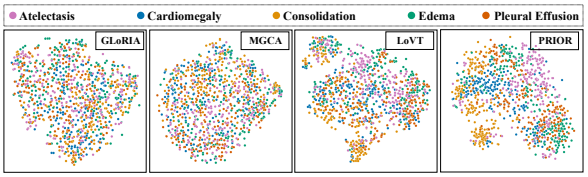


Figure 5. t-SNE visualization of the high-level embeddings from the last layer of the image encoder on CheXpert 5x200.

pare our proposed PRIOR with one of the large-scale models, namely BioMedCLIP [56]. Since BioMedCLIP is pre-trained on a million-scale dataset with a ViT backbone, it is unfair to compare PRIOR with it. Still, our PRIOR achieves comparative or even superior performance; e.g. 89.22% vs. **89.54%** on RSNA classification, **88.61%** vs. 87.97% on CheXpert classification and **30.56%** vs. 22.78% (F1-Score) on zero-shot CheXpert 5x200 classification (PRIOR vs. BioMedCLIP). Notwithstanding the comprehensive clinical insights harnessed by large-scale multi-modal datasets, there remain instances wherein a model specialized for a particular modality maintains a competitive edge in handling intricate scenarios.

5. Conclusion

In this paper, we have explored fine-grained representation learning from paired medical images and reports. We present a cross-modality alignment strategy by learning from both global and local information. A sentence-wise prototype memory bank is proposed to translate clinical descriptions into categorical representations and cluster similar clinical conditions. Based on the prototype memory bank, a cross-modality conditional reconstruction module is designed to reconstruct masked images and missing sentence prototypes through cross-modality interaction, facilitating low-level representation learning which is crucial for locality-aware downstream tasks. Extensive experiments on five downstream tasks demonstrate the effectiveness of the proposed method.

6. Acknowledgements

This study was supported by the Shenzhen Basic Research Program (JCYJ20190809120205578); the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX20210609103056042; JSGG20220831093004008); the Shenzhen Basic Research Program (JCYJ20200925153847004); the Shenzhen Science and Technology Innovation Committee (KCXFZ2020122117340001); the Guangdong Basic and Applied Basic Research Foundation (2023A1515012839).

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019. 6

[2] 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. Springer, 2022. 7, 8, 9

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5

[4] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3, 7, 8, 9

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3, 7, 8, 9

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. 3, 5

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 1

[9] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, Florence, Italy, July 2019. Association for Computational Linguistics. 3

[10] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[12] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1

[13] Han Fang, Pengfei Xiong, Luhui Xu, and Wenhan Luo. Transferring image-clip to video-text retrieval via temporal relations. *IEEE Transactions on Multimedia*, 2022. 3

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3, 7

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 3, 7, 8, 9

[17] 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. 1, 6

[18] 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. 1, 2, 3, 6, 7, 8, 9

[19] 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. 2, 6

[20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[21] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 6

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[23] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 3

[24] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673, 2019. 1

[25] 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

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 3

[27] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018. 1

[28] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023. 8

[29] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 2

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[31] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 1

[32] 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. 1, 2, 3, 5, 7, 8, 9

[33] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V Le, and Ruoming Pang. Domain adaptive transfer learning with specialist models. *arXiv preprint arXiv:1811.07056*, 2018. 1

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[35] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 3

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[37] Yadunath Pathak, Prashant Kumar Shukla, Akhilesh Tiwari, Shalini Stalin, and Saurabh Singh. Deep transfer learning based classification model for covid-19 disease. *Irbm*, 2020. 1

[38] 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, 3, 4

[39] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019. 1

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 8

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 7

[43] 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. 6, 8

[44] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021. 1

[45] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020. 1

[46] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2549–2559, 2018. 3

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[50] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 7, 8, 9

[51] Yizhi Wang, Guo Pu, Wenhan Luo, Yexin Wang, Pengfei Xiong, Hongwen Kang, and Zhouhui Lian. Aesthetic text logo synthesis via content-aware layout inferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2436–2445, 2022. 1

[52] Yiyu Wang, Jungang Xu, and Yingfei Sun. End-to-end transformer based model for image captioning. *arXiv preprint arXiv:2203.15350*, 2022. 2

[53] Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. *arXiv preprint arXiv:2106.06471*, 2021. 1

[54] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1, 3

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 3

[56] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 8, 9

[57] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 1, 3, 7, 8, 9

[58] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 1