MLIP: Enhancing Medical Visual Representation with Divergence Encoder and Knowledge-guided Contrastive Learning

Zhe Li [1], Laurence T. Yang [1,2,*], Bocheng Ren [1], Xin Nie [1], Zhangyang Gao [3], Cheng Tan [3], Stan Z. Li [3]
[1] Huazhong University of Science and Technology
[2] Zhengzhou University
[3] AI Lab, Research Center for Industries of the Future, Westlake University
[*]keycharon0122@gmail.com, ltyang@ieee.org, bc.Revincent@gmail.com, niexin@hust.edu.cn,
{gaozhangyang,tancheng,stan.zq.li}@westlake.edu.cn

Abstract

The scarcity of annotated data has sparked significant interest in unsupervised pre-training methods that leverage medical reports as auxiliary signals for medical visual representation learning. However, existing research overlooks the multi-granularity nature of medical visual representation and lacks suitable contrastive learning techniques to improve the models' generalizability across different granularities, leading to the underutilization of image-text information. To address this, we propose MLIP, a novel framework leveraging domain-specific medical knowledge as guiding signals to integrate language information into the visual domain through image-text contrastive learning. Our model includes global contrastive learning with our designed divergence encoder, local token-knowledge-patch alignment contrastive learning, and knowledge-guided category-level contrastive learning with expert knowledge. Experimental evaluations reveal the efficacy of our model in enhancing transfer performance for tasks such as image classification, object detection, and semantic segmentation. Notably, MLIP surpasses state-of-the-art methods even with limited annotated data, highlighting the potential of multimodal pre-training in advancing medical representation learning.[1]

1. Introduction

Representation learning for medical radiographs has gained significant attention recently, owing to the availability of abundant annotated data. Numerous approaches [20, 23, 46, 48] have employed deep learning in a supervised manner to learn representations for downstream tasks. However, the acquisition of large-scale annotated data is time-consuming and costly. unsupervised pre-training methods have emerged as a promising alternative. These methods, which do not rely on annotated data, harness medical reports as ancillary signals that provide targeted supervision for visual representation learning. By incorporating language information, these models can acquire more universal visual representations that are transferable to downstream tasks and capable of domain transfer.

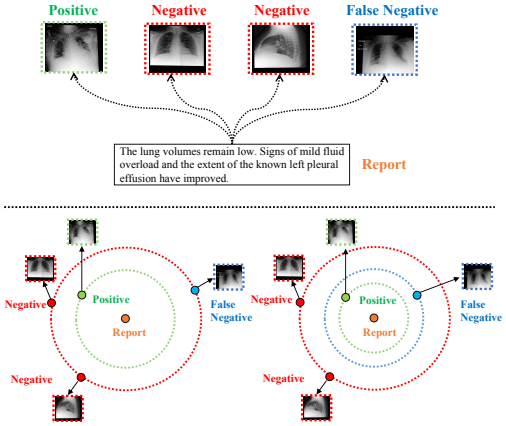

Figure 1. **Detailed illustration of false negatives in medical image-text.** Conventional approaches consider false negative samples as negatives that are distant from positive samples in the lower left corner. In contrast, in the lower right corner, our proposed method distinguishes false negatives from negatives, effectively bringing them closer to positives.

There are three mainstream paradigms in visual representation learning. Masked image modeling [29, 38, 60] follows ***mask-and-predict*** paradigm, randomly masking some patches and predicting missing information. Multimodal contrastive learning [10, 32, 64, 65] conducts ***embed-and-compare*** proxy tasks to maximize the mutual information between medical images and reports through

[*]Corresponding Author.
[1]Codes are available at **https://github.com/gentlefress/MLIP**

image-text contrastive learning. Multi-view self-supervised learning [9, 12, 13, 28] adopts an ***augment-and-compare*** paradigm, where an input image is randomly transformed into two augmented views and compare the two distinct views in the representation space.

However, the fact that pathological features only occupy a small part of a radiograph means that a significant portion of the information may not be relevant for our analysis, decreasing the utilization of medical image-text data. Moreover, due to the unique nature of medical image-report compared to general text-image pairs, different symptoms may correspond to the same disease, and traditional contrastive learning will mistake samples that are not in the same batch as negative samples even if they are very close in the semantic space. In Fig 1, we purpose to differentiate between false negative and negative samples and further reduce the distance between false negative and positive samples.

Driven by the revelation from [33, 39, 56], we design a knowledge-guided ***align-and-compare*** framework to capture multi-grained semantic information and to accurately align each image's pathology with the corresponding medical term [33, 36, 37]. We introduce a knowledge-guided medical multimodal pre-trained model, dubbed MLIP, to explore the inherent multi-granularity cross-modal correspondence for enhancing the generalizability of visual representation. Specifically, we employ a combination of three distinct image-text contrastive learning methods to embed language into vision at different granularity and utilize two proxy tasks to establish the match between vision and language. Our model exploits multi-level correspondences between medical radiographs and reports to enhance generalized medical visual representation with contrastive learning. Our approach demonstrates state-of-the-art performance in image classification, object detection, and semantic segmentation, even when working with limited annotated data.

The key contributions are summarized as follows:

- We introduce two dynamically updated **divergence encoders** for data augmentation, aiming to increase the number of samples and thus enhance the generalization ability of the model.
- We propose to leverage cross-modal attention-based **token-knowledge-patch** alignment and incorporate contrastive learning to facilitate the exploration of local representations.
- We propose a **knowledge-guided prototype clustering** contrastive learning approach, which focuses on conducting contrastive learning at the category level rather than the individual samples.
- We pre-train MLIP on the MIMIC-CXR dataset [35], evaluating the learned representations on seven downstream datasets. Experimental results demonstrate the superiority of our model over state-of-the-art methods, even with 1% and 10% training data.

2. Related Work

2.1. Text-guided Medical Visual Representations Learning

Medical reports are pivotal in unsupervised medical visual representation learning, with two primary methods dominating the field. The first method involves extracting disease labels from radiology reports using manually designed rules [34, 35], followed by pre-training image models for downstream tasks. However, defining the rules requires considerable human effort and domain expertise. On the other hand, the second method adopts image-text contrastive learning methods to integrate text and vision in an unsupervised manner [20, 32, 33, 56, 64]. These methods have been shown remarkable performance in diverse downstream tasks, including medical object detection [4], image classification [33, 64], and semantic segmentation [64]. However, they have not effectively explored visual representations at different granularities and rely on partial semantic information.

To address these limitations, MGCA [56] proposes to leverage multiple visual features at different granularities during the pre-training phase, enhancing the performance of models in downstream tasks. However, it overlooks the challenging sample issue in medical radiology. In this work, we propose a divergence encoder that manually updates its parameters based on the similarity between the output features and those of a common encoder. By increasing divergence between the two encoders, we enhance feature diversity and train the model to discriminate among similar samples effectively.

2.2. Knowledge-guided Pre-training

To enhance the model's knowledge and understanding ability by leveraging a broader background, numerous vision-and-language pre-training methods have been devised to incorporate domain-specific knowledge. These methods can be categorized into four distinct knowledge-guided schemes: embedding combination [66], data structure compatibility [26, 42], knowledge supervision [58], and neural-symbolic methods [2]. For instance, ERNIE-ViL [62] introduces a vision and language alignment technique by utilizing a scene graph extracted from the input text. Similarly, KB-VLP [11] incorporates object tags from images and knowledge graph embeddings from texts to enhance the acquisition of knowledge-aware representations. ARL [15] utilizes expert knowledge as an intermediate medium to align images and reports. Additionally, a recent study [45] proposes the automatic generation of visual and textual prompts, injecting expert medical knowledge into the prompt for pre-training.

In contrast to existing works, we propose an alignment method that leverages domain-specific knowledge as an in-

termediate mediator for aligning texts and images, along with a knowledge-guided prototype clustering contrastive learning. This approach integrates expert domain knowledge derived from the Unified Medical Language System (UMLS) [6]. By incorporating UMLS knowledge into both vision and language modalities, our approach leverages knowledge as a medium to achieve improved alignment between images and text, facilitating more effective clustering of image-text pairs. Importantly, our method effectively mitigates the influence of disease-level false negatives without relying on object detectors or scene graph parsers.

3. Proposed Approach

In this section, we present our approach for learning effective medical visual representations using medical reports. We utilize a knowledge-guided *align-and-compare* scheme, as depicted in Figure 2, to match and align modalities and compare them in the representation space. Our method comprises four key components: 1) global image-text contrastive learning; 2) local token-knowledge-patch alignment contrastive learning; 3) knowledge-guided category-level contrastive learning; and 4) proxy tasks to ensure matching and prevent shortcut exploitation by the network. We discuss each component in detail in the following subsections and provide an overview of the overall training objective.

3.1. Problem Setup

Recently, it has been demonstrated in [33, 56] that learning medical visual representation learning without labels can achieve competitive performance. In this study, we follow the setting in [56], given a training set of N medical image-report pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1,...,N}$, we use an image encoder f_v and a text encoder f_t encode $\mathcal{D}$ to a global feature set $\mathcal{E}_{il} = \{(v_i, t_i) | v_i = f_v(x_i), t_i = f_t(y_i)\}_{i=1,...,N}$, and a local feature set $\mathcal{E}_{tl} = \{(\mathcal{P}_i, \mathcal{S}_i)\}_{i=1,...,N}$, where $\mathcal{S}_i = \{s_i^1, s_i^2, ..., s_i^V\} \in \mathbb{R}^{V\times d}$ and $\mathcal{P}_i = \{p_i^1, p_i^2, ..., p_i^{M^2}\} \in \mathbb{R}^{M^2\times d}$. V denotes the length of the sentence and M^2 denotes the number of image patches.

Furthermore, we incorporate expert knowledge into our model by constructing an extracted knowledge graph, as described in [15]. This knowledge graph is denoted as $\mathcal{G} = \{(he_i, re_i, ta_i)\}_{i=1}^{N_\mathcal{G}}$, where $N_\mathcal{G}$ represents the number of graph triples, and he_i, re_i, and ta_i correspond to the head entity, relation, and tail entity, respectively. The inclusion of this expert knowledge enhances the model's understanding and reasoning capabilities, enabling more informed alignment and representation learning.

3.2. Global Image-text Contrastive Learning

To pull correct samples closer and push random samples apart in the latent space, we follow [31, 52], present a comprehensive discussion on global image-text contrastive learning by maximizing mutual information $\mathcal{I}(X, Y)$ between the vision element X and the language component Y:

$$\mathcal{I}(X, Y) = \sum_{y\in Y}\sum_{x\in X} P(x, y) \log \frac{P(x|y)}{P(x)}. \quad (1)$$

Eq.1 suggests that the fraction $\frac{P(x|y)}{P(x)}$ collapses to zero when x and y are incompatible with each other. Therefore, we hypothesize that $\frac{P(x|y)}{P(x)}$ is proportional to the similarity between x and y. Further, the maximization of mutual information corresponds to the maximization of the similarity $\text{sim}(x, y)$ between x and y, which can be represented as:

$$\mathcal{I}(v, t) \propto \mathcal{I}(X, Y) \propto \text{sim}(x, y) \propto \text{sim}(v, t). \quad (2)$$

Specifically, inspired by [12], we firstly utilize two projection layers h_v and h_t to map v_i and t_i into a normalized shared feature space, yielding $v_i^* \in \mathbb{R}^d$ and $t_i^* \in \mathbb{R}^d$, respectively. Then, we apply the dot product to model the similarity between v_i^* and t_i^*. To obtain more effective features, we perform Self-Attention [54] and LayerNorm [3] on features:

$$v_i^* = \text{LN}(\text{SA}\{h_v(v_i)\}); \quad (3a)$$
$$t_i^* = \text{LN}(\text{SA}\{h_t(t_i)\}), \quad (3b)$$
$$\text{sim}\{v_i^*, t_i^*\} = v_i^* {t_i}^T, \quad (3c)$$

where SA denotes Self-Attention module and LN denotes LayerNorm module.

We optimize this process via image-text contrastive loss based on InfoNCE loss [53], which are designed to maximize the mutual information between the correct image-text pairs in the latent space:

$$\mathcal{L}_{v2t}^{il}(v_i, t_i) = -\log\left(\frac{\phi_{il}(v_i, t_i)}{\sum_{k=1}^{B}\phi_{il}(v_i, t_k)}\right), \quad (4a)$$
$$\mathcal{L}_{t2v}^{il}(v_i, t_i) = -\log\left(\frac{\phi_{il}(v_i, t_i)}{\sum_{k=1}^{B}\phi_{il}(v_k, t_i)}\right), \quad (4b)$$

where $\phi_{il}(v_i, t_i) = \exp(\frac{\text{sim}(v_i^*, t_i^*)}{\tau_1})$, B is the batch size and τ_1 is the global temperature hyper-parameter.

Directly optimizing $\mathcal{I}(v, t)$ is a challenging task. As an alternative, [53] has proposed an alternative method to optimize the lower bound of mutual information:

$$\mathcal{I}(v, t) \geq \log N^{'} - \mathcal{L}^{\text{NCE}}(v, t), \quad (5)$$

where $N^{'}$ is the number of negative samples. In Eq.5, minimizing $\mathcal{L}^{\text{NCE}}(v, t)$ is equivalent to maximizing the lower bound of the mutual information between the medical image and the corresponding report.

To increase the number of samples and enhance the feature diversity, we perform a divergence encoder to achieve

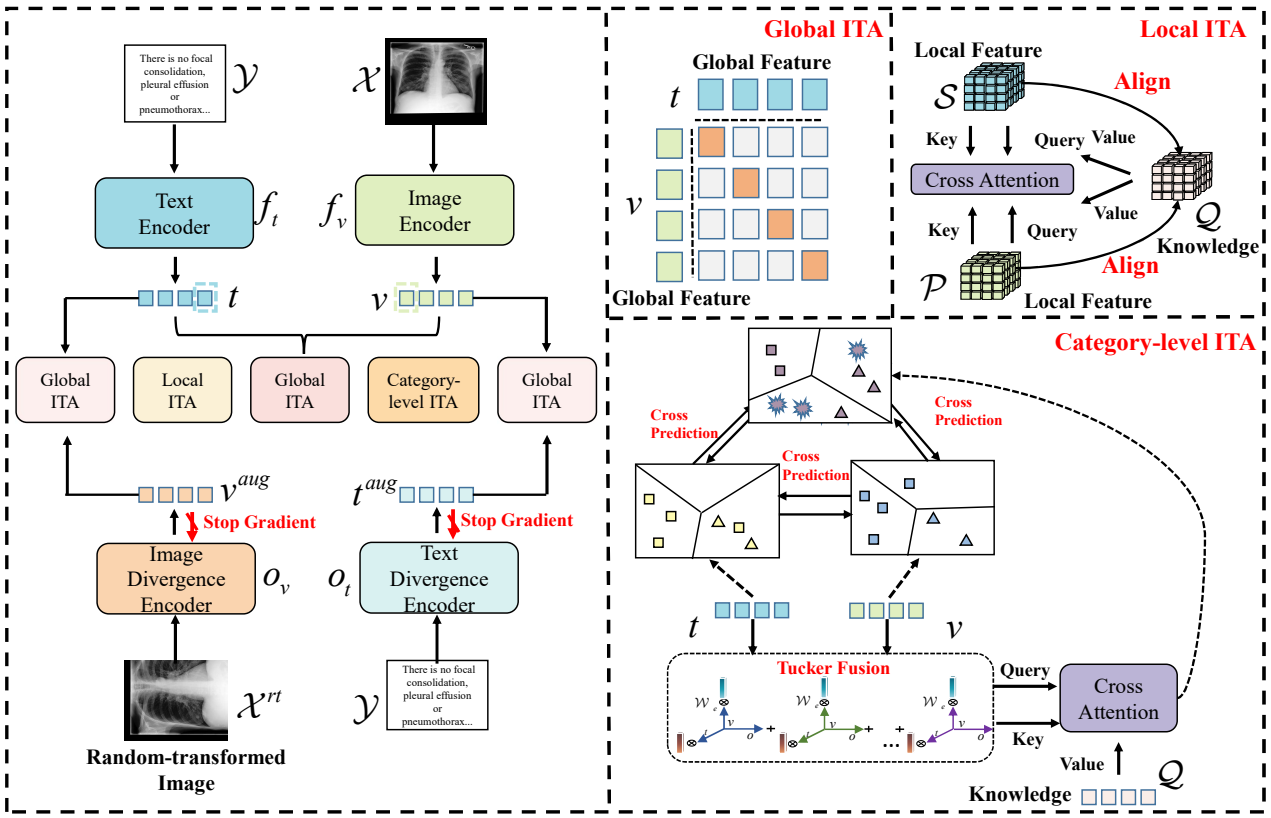

Figure 2. **The framework of our MLIP.** Our model architecture employs global, local, and category-level image-text contrastive learning. Given medical images and reports as inputs, we extract global features and local features for each modality using image and text encoders. We leverage global features for global image-text contrastive learning, while the local features are aligned with domain-specific knowledge from UMLS to achieve fine-grained image-text alignment. Through tucker fusion and cross-modal attention mechanisms, we combine the image, text, and knowledge representations, facilitating category-level prototype contrastive learning. Furthermore, to enhance feature diversity, we introduce a divergence encoder as a data augmentation strategy, generating similar yet distinct features. This enables global contrastive learning between images and augmented text, as well as between text and augmented images.

data augmentation and extend the gap between samples. We define image divergence encoder o_v and text divergence encoder o_t, initialized by f_v and f_t, respectively. Then we obtain features incrementally differentiated from v_i and t_i:

$$v_i^{aug} = o_v(x_i^{rt}); t_i^{aug} = o_t(y_i), \tag{6}$$

where x_i^{rt} denotes randomly transformed images. We manually update divergence encoders' parameters instead of relying on backpropagation:

$$\theta_{o_t} = s_t * \theta_{f_t} + (1 - s_t) * \theta_{o_t}, \tag{7a}$$

$$\theta_{o_v} = s_v * \theta_{f_v} + (1 - s_v) * \theta_{o_v}, \tag{7b}$$

where $s_t = \text{cosine}(t_i, t_i^{aug})$ and $s_v = \text{cosine}(v_i, v_i^{aug})$, and $\theta_{o_t}, \theta_{o_v}, \theta_{f_t}, \theta_{f_v}$ are the parameters of o_t, o_v, f_t, f_v, respectively. In this way, as the $s_v(s_t)$ increases, we aim to retain fewer parameters from $o_v(o_t)$ and incorporate more parameters from $f_v(f_t)$, in order to generate more diverse features. Then we use Eq.4a, 4b to compute $\mathcal{L}_{v2a}^{il}$ and $\mathcal{L}_{avt}^{il}$.

We compute the objective $\mathcal{L}_{il}$ as the average of the four loss values:

$$\begin{aligned}\mathcal{L}_{il} = &\frac{1}{2N}\sum_{i=1}^{N}(\mathcal{L}_{v2t}^{il}(v_i, t_i) + \mathcal{L}_{t2v}^{il}(v_i, t_i)) \\ &+ \frac{\lambda_0}{2N}\sum_{i=1}^{N}(\mathcal{L}_{v2a}^{il}(v_i, t_i^{aug}) + \mathcal{L}_{avt}^{il}(v_i^{aug}, t_i)),\end{aligned} \tag{8}$$

where N is the total number of samples and λ_0 denotes the weight for augmented image-text contrastive learning.

3.3. Local Token-knowledge-patch Alignment Contrastive Learning

In medical images, pathologies are often visually subtle and occupy a small fraction of the overall image, while only a few disease-related tags in the associated report accurately depict the critical medical condition. Given this observation, we employ a local image-text contrastive learning method to maximize the mutual information between local features and achieve cross-modal alignment between

images and texts, inspired by [18, 56].

However, traditional token-patch alignment contrastive learning is utilizing the local features of the image and text to compute the attention matrix, and then perform contrastive learning after aligning the images and texts. Since medical radiology is highly professional and there is a certain bias between different datasets, we regard professional knowledge from the UMLS [6] as a medium between vision and language. To achieve more accurate token-patch alignment, we align the knowledge with radiographs and reports.

Similar to global feature, we apply Self-Attention and LayerNorm module on every features:

$$p_i = \text{LN}(\text{SA}\{h_v(p_i)\}); s_i = \text{LN}(\text{SA}\{h_t(s_i)\}). \quad (9)$$

We apply the knowledge representation learning algorithm TransE [7] to the knowledge graph $\mathcal{G}$ to obtain entity embeddings. Subsequently, we utilize the Graph Attention Network [55] to capture local information in the graph neighborhood for each node. This allows us to obtain knowledge representations, denoted as $\{e_i\}_{i=1}^{N_e} \in \mathbb{R}^{N_e \times d_e}$, where d_e represents the feature dimension and N_e denotes the number of entity.

We adopt cross-modal attention mechanism [14, 44] to explore the matching between knowledge and image:

$$attn_{j,k}^{vk} = \text{softmax}(\frac{(Qp_i^j)^T(Ke_i^k)}{\sqrt{d}}), \quad (10a)$$

$$zv_i^j = \sum\nolimits_{k=1}^{N} attn_{j,k}^{vk}(Ve_i^k), \quad (10b)$$

where $Q, K, V \in \mathbb{R}^{d \times d}$ are trainable matrices. e_i is mapped to $\mathbb{R}^{M^2 \times d}$. zv_i^j is cross-modal knowledge embedding corresponding to p_i^j.

Lying in the purpose of maximizing the lower bound of mutual information, we leverage InfoNCE loss [53] to pull p_i^j and zv_i^j closer and push p_i^j and other cross-modal knowledge embeddings apart. However, given that irrelevant information only occupies a vast majority of medical images, we employ w_i^j to balance the weights of different patches. The loss $\mathcal{L}_{v2t}^{tl}$ is designed symmetrically as:

$$\mathcal{L}_{v2t}^{tl} = -\frac{1}{2NM^2}\sum_{i=1}^{N}\sum_{j=1}^{M^2} w_i^j(\log\frac{\phi_{tl}(p_i^j, zv_i^j)}{\sum_{k=1}^{M^2}\phi_{tl}(p_i^j, zv_i^k)} + \log\frac{\phi_{tl}(zv_i^j, p_i^j)}{\sum_{k=1}^{M^2}\phi_{tl}(zv_i^k, p_i^j)}), \quad (11)$$

where $\phi_{tl}(p_i^j, zv_i^j) = \exp(\frac{\text{sim}(p_i^j, zv_i^j)}{\tau_2})$, τ_2 is the local temperature hyper-parameter. To establish the correlation between the j-th visual patch and the [CLS] token, we assign the weight w_i^j using the last-layer attention mechanism averaged across multiple heads.

Similarly, for the j-th text token, we calculate corresponding cross-modal knowledge embedding zt_i^j and construct local contrastive loss $\mathcal{L}_{t2v}^{tl}$ to maximize the lower bound of mutual information between s_i^j and zt_i^j. The objective $\mathcal{L}_{tl}$ can be defined as the average of these two losses:

$$\mathcal{L}_{tl} = \frac{1}{2}(\mathcal{L}_{v2t}^{tl} + \mathcal{L}_{t2v}^{tl}). \quad (12)$$

3.4. Knowledge-guided Category-level Contrastive Learning

For a given radiograph-report pair, traditional contrastive learning approaches treat other radiograph-report pairs within the same batch as negative samples. However, in the context of category-level analysis, samples that belong to different batches but exhibit highly similar semantics should be considered positive samples. In our approach, we aim to select representative samples in each iteration, emphasizing their ability to capture meaningful disease-related information. In the medical domain, expert knowledge plays a crucial role in representation learning. We purpose to bridge the gap between the vast knowledge learned from general visual and textual data and its effective application in the intricate realm of medical radiology. Therefore, we incorporate expert knowledge from UMLS [6] as an auxiliary signal. Drawing inspiration from [8, 45], we propose a knowledge-guided clustering-based approach to improve the efficacy of learned representations. We bring together highly similar samples with high-level semantics, even when originating from different batches, and ensure their proximity in the feature space, rather than increasing their distance from one another.

Motivated by [41], we realize to filter out irrelevant information and explore more fine-grained relations between images and text. To achieve this, we employ a mechanism that identifies the most relevant topic in a given context. Specifically, we utilize v_i^* to find the most relevant topic in t_i^*, resulting in $\dot{t}_i$. Then, we use $\dot{t}_i$ to find the relevant topic in v_i^*, leading to $\dot{v}_i$. The process is mathematically defined as follows:

$$\dot{t}_i = \text{LN}(\text{softmax}(\frac{v_i^{*T}t_i^*}{\sqrt{d}})t_i^*); \dot{v}_i = \text{LN}(\text{softmax}(\frac{v_i^{*T}v_i^*}{\sqrt{d}})\dot{t}_i), \quad (13)$$

then we utilize tucker fusion [5] to seamlessly integrate visual and textual features, further fuse with knowledge representations:

$$\mathcal{Q} = ((\mathcal{T}_c \times_1 \dot{v}_i) \times_2 \dot{t}_i) \times_3 \mathcal{W}_o, \quad (14)$$

where $\mathcal{W}_o$ represents a mapping matrix which is trainable and maps fused features to a certain dimensional space, and $\mathcal{T}_c$ denotes the core tensor.

To further integrate knowledge with modality-specific features, we employ a linear mapping layer to project the

knowledge representation e_i into a d-dimensional space and incorporate it with fused features using cross-modal attention, thereby facilitating the fusion of information across modalities:

$$vkt_i = \text{SA}(\text{softmax}(\frac{\mathcal{Q}^T e_i}{\tau_3}) \cdot e_i), \tag{15}$$

where τ_3 is the temperature hyper-parameter we set to scale the attention.

For image-text features pair $(\dot{v_i}, \dot{t_i})$ and knowledge-fused features, we apply the iterative Sinkhorn-Knopp clustering algorithm [19] to generate a cluster assignment code $u^{vkt,i} \in \mathbb{R}^C$, by assigning vkt_i to C clusters separately. To facilitate this, we introduce a set $\mathcal{J} = j_1, ..., j_C$ that contains C trainable cross-modal prototypes, where each prototype $j_c \in \mathbb{R}^d$. We calculate the visual softmax probability $p^{v,i}$ by computing the cosine similarity between the visual feature vector $\dot{v_i}$ and all cross-modal prototypes in $\mathcal{J}$. Similarly, the textual softmax probability $p^{t,i}$ is obtained by measuring the cosine similarity between the textual feature vector $\dot{t_i}$ and all cross-modal prototypes in $\mathcal{J}$:

$$p_c^{v,i} = \frac{\exp(\dot{v_i}^T j_c/\tau_4)}{\sum_l \exp(\dot{v_i}^T j_l/\tau_4)}; p_c^{t,i} = \frac{\exp(\dot{t_i}^T j_c/\tau_4)}{\sum_l \exp(\dot{t_i}^T j_l/\tau_4)}, \tag{16}$$

where τ_4 is a category-level temperature hyper-parameter and c denotes the c-th element of the vector.

To enable knowledge-guided category-level contrastive learning, we employ $u^{vkt,i}$ as the pseudo-label for training $\dot{t_i}$ and $\dot{v_i}$. This allows the three features to interact in the latent space and guide the shifting of positive and negative samples with the assistance of domain-specific knowledge. The objective loss $\mathcal{L}_{cl}$ is formulated as follows:

$$\mathcal{L}_{cl} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{c=1}^{C} (u_c^{vkt,i} \log p_c^{v,i} + u_c^{vkt,i} \log p_c^{t,i}). \tag{17}$$

3.5. Image-text Matching and Text Swapping

In order to identify the alignment between radiographs and their corresponding reports, we propose two pretext tasks aimed at bridging the semantic divide between visual and linguistic information within the feature space: 1) computing relevance scores between image patch and contextualized sentence to evaluate the degree of correlation between the image and text elements; 2) randomly substituting medical reports corresponding to the image with a predetermined probability, improving the discriminative ability on mismatched samples of the model.

We assume that the text features t and image features v have been normalized. Therefore, we construct the similarity between the two modalities as a relevance score:$r(v,t) = v^T \cdot t$, subsequently, we randomly select another image $v^{'}$ and obtain its corresponding relevance score $r(v^{'}, t)$. To ensure that the difference between $r(v,t)$ and $r(v^{'}, t)$ is greater than a pre-specified margin $\mathcal{G}$, we utilize the hinge loss function to compute image-text match loss:

$$\mathcal{L}_{itm} = \max(0, \mathcal{G} - r(v,t) + r(v^{'}, t)). \tag{18}$$

Similarly, we propose a text swapping task, which involves randomly replacing text with a predefined probability γ. We employ a bidirectional similarity Hinge loss to penalize the model for insufficient discriminative ability. This task aims to enhance the model's ability to distinguish between different reports. We employ a cross-modal attention mechanism to fuse the text and image modalities, then compute the relevance score by performing a weighted summation of the similarity between the fused representation and the original text-image pair. Our objective is to ensure that this score exceeds the score obtained after replacing the text by a margin $\mathcal{G}'$:

$$r_{ts}(v,t) = v^T \cdot t + \alpha \cdot \text{CA}(v,t)^T \cdot \text{CA}(t,v), \tag{19a}$$

$$r_{ts}(v,t^{'}) = v^T \cdot t + \alpha \cdot \text{CA}(v,t^{'})^T \cdot \text{CA}(t^{'},v), \tag{19b}$$

$$\mathcal{L}_{ts} = \max(0, \mathcal{G}' - r_{ts}(v,t) + r_{ts}(v,t^{'})), \tag{19c}$$

where $\text{CA(x,y)} = \text{softmax}(\frac{x^T \cdot y}{\sqrt{d}}) \cdot y$. Through these two designed proxy tasks, we compute the image-text matching loss $\mathcal{L}_{itm}$ and the text swapping loss $\mathcal{L}_{ts}$. These losses quantify the model's ability to accurately match radiographs to their appropriate reports, thereby providing a measurable objective for the optimization process.

3.6. Overall Objective

Our training approach involves joint optimization of the five losses, aiming to promote the acquisition of effective and generalizable medical image representations by the network. The overall training objective can be expressed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{il} + \lambda_2 \mathcal{L}_{tl} + \lambda_3 \mathcal{L}_{cl} + \lambda_4 \mathcal{L}_{itm} + \lambda_5 \mathcal{L}_{ts}, \tag{20}$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are hyper-parameters employed to balance the weights associated with each respective loss.

4. Experiments

4.1. Pre-training Dataset and Implementation Details

Our MLIP framework is initially pre-trained on the MIMIC-CXR 2.0.0 dataset [35], with data consistency ensured through preprocessing methods from [64]. Lateral views are excluded from the dataset as downstream datasets only include frontal-view chest images. Inspired by [56], we extract impression and finding sections from free-text reports, providing comprehensive descriptions of medical diseases.

We filter out empty or short reports, resulting in approximately 217,000 image-text pairs. Details about our implementation can be found in the supplementary 6.1.

4.2. Downstream Tasks

Medical Object Detection. We assess the capability of our pre-trained image encoder for medical object detection on the **RSNA** Pneumonia dataset [50] (stage 2 version) and the **Object CXR** dataset [30]. The detection performance is evaluated using the YOLOv3 [25] frozen setting, where the pre-trained ResNet-50 [27] image encoder acts as a fixed backbone for YOLOv3. In this configuration, only the classification layers are fine-tuned. To evaluate the efficiency of data utilization, we conduct experiments in the zero-shot scenario and further fine-tune the model using 1%, 10%, and 100% of the available training data. Evaluation is performed using the Mean Average Precision (mAP) metric, computed with IOU thresholds ranging from 0.4 to 0.75.

Method	RSNA (mAP)				Object CXR (mAP)			
	Zero-shot	1%	10%	100%	Zero-shot	1%	10%	100%
Random Init	∼	1.0	4.0	8.9	∼	∼	∼	4.4
ImageNet Init	∼	3.6	8.0	15.7	∼	∼	8.6	15.9
ConVIRT [64]	3.7	8.2	15.6	17.9	∼	∼	8.6	15.9
GLoRIA-CheXpert [33]	4.4	9.8	14.8	18.8	∼	∼	10.6	15.6
GLoRIA-MIMIC [33]	6.2	10.3	15.6	23.1	∼	∼	8.9	16.6
MGCA [56]	7.8	12.9	16.8	24.9	∼	∼	12.1	19.2
M-FLAG [40]	8.6	13.7	17.5	25.4	∼	∼	13.6	19.5
PRIOR [16]	10.7	15.6	18.5	25.2	1.4	2.9	15.2	19.8
MLIP (Ours)	12.3	17.2	19.1	25.8	2.7	4.6	17.4	20.2

Table 1. **Fine-tuned results (mAP [%]) of object detection with 1%, 10%, and 100% of the available training data in RSNA and Object CXR.** ∼ means mAP is smaller than 1%.

Medical Semantic Segmentation. We evaluate the performance of our model for medical semantic segmentation on the **SIIM** Pneumothorax dataset [63] and the **RSNA** Pneumonia dataset [50]. Following the methodology presented in [33], we adopt the fine-tuning protocol of U-Net [48] to assess the segmentation task. Specifically, we utilize the pre-trained ResNet-50 image encoder as a fixed backbone for the U-Net architecture and train the decoder component using varying proportions of the available training data (1%, 10%, and 100%). We also evaluate our model in the zero-shot scenario. To evaluate the quality of segmentation, we compute Dice scores [59] as the chosen metric for performance assessment.

Medical Image Classification. We perform medical image classification on the **RSNA** Pneumonia dataset [50], **COVIDx** dataset [57], and **CheXpert** dataset [34]. To evaluate the transferability of our pre-trained image encoder, we adopt the Linear Classification setting following the methodology proposed in prior work [33, 56]. This involves freezing the pre-trained ViT-B/16 [21] or ResNet-50 image encoder and training only a linear classification head for the downstream classification task. Additionally, to assess data efficiency, we conduct experiments in the zero-shot scenario and evaluate the model using 1%, 10%, and 100% of the training data for each classification dataset. The evaluation metrics used are the area under the receiver operating characteristic (ROC) curve (AUROC) for RSNA and CheXpert, and accuracy (ACC) for COVIDx-v6, consistent with the evaluation criteria outlined in [64]. More details and experiment can be found in the supplementary 6.2 and 6.3.

Method	RSNA (Dice)				SIIM (Dice)			
	Zero-shot	1%	10%	100%	Zero-shot	1%	10%	100%
Random Init	3.9	6.9	10.6	18.5	∼	9.0	28.6	54.3
ImageNet Init	17.6	34.8	39.9	64.0	2.2	10.2	35.5	63.5
ConVIRT [64]	23.3	55.0	67.4	67.5	11.7	25.0	43.2	59.9
GLoRIA-CheXpert [33]	32.0	59.3	67.5	67.8	19.8	35.8	46.9	63.4
GLoRIA-MIMIC [33]	34.6	60.8	68.2	67.6	21.0	37.6	56.4	64.0
MGCA [56]	34.9	63.0	68.3	69.8	33.5	49.7	59.3	64.2
M-FLAG [40]	40.7	64.6	69.7	70.5	37.2	52.5	61.2	64.8
PRIOR [16]	41.8	66.4	68.3	72.7	38.6	51.2	59.7	66.3
MLIP (Ours)	44.3	67.7	68.8	73.5	40.2	51.6	60.8	68.1

Table 2. **Semantic segmentation results (Dice [%]) achieved on the SIIM and RSNA datasets.** Each dataset is fine-tuned using 1%, 10%, and 100% of the available training data. The best results obtained for each setting are highlighted in red, while the suboptimal results are highlighted in blue.

4.3. Results

Results on Medical Object Detection. We evaluate the ResNet-50-YOLOv3 architecture on the RSNA and Object CXR datasets. Our results, presented in Table 1, demonstrate a significant improvement over ConVIRT [64], GLoRIA [33], MGCA [56], M-FLAG [40] and PRIOR [16]. Notably, our method achieves superior performance using only 1% of the data, surpassing alternative approaches that require 10% or even 100% of the data for fine-tuning.

Results on Medical Semantic Segmentation. In Table 2, we present the semantic segmentation results (Dice [%]) achieved on the SIIM and RSNA datasets using the ResNet-50-U-Net architecture. MLIP leverages contrastive learning and category-level approaches to achieve remarkable performance improvements, consistently obtaining the best results in various settings, as highlighted in red. Specifically, MLIP outperforms the MGCA [56] by 4.7% on the RSNA dataset and 1.9% on the SIIM dataset when fine-tuned with only 1% of the training data. Moreover, MLIP achieved state-of-the-art results in zero-shot scenarios.

Results on Medical Image Classification. Table 3 shows the medical linear classification results on RSNA and COVIDx datasets. We divide existing pre-trained methods into two categories: pre-trained on CheXpert [34] and pre-trained on MIMIC-CXR[35]. The results of other approaches are from original papers, and we refer to [56],

Method	CheXpert (AUC)				RSNA (AUC)				COVIDx (ACC)			
	Zero-shot	1%	10%	100%	Zero-shot	1%	10%	100%	Zero-shot	1%	10%	100%
Random Init	-	56.1	62.6	65.7	-	58.9	69.4	74.1	-	50.5	60.3	70.0
ImageNet Init	-	74.4	79.7	81.4	-	74.9	74.5	76.3	-	64.8	78.8	86.3
pre-trained on CheXpert												
DSVE [22]	26.6	50.1	51.0	51.5	18.7	49.7	52.1	57.8	-	-	-	-
VSE++ [24]	27.3	50.3	51.2	52.4	19.1	49.4	57.2	67.9	-	-	-	-
GLoRIA [33]	50.4	86.6	87.8	88.1	39.2	86.1	88.0	88.6	20.9	67.3	77.8	89.0
pre-trained on MIMIC-CXR												
Caption-Transformer [17]	42.2	77.2	82.6	83.9	-	-	-	-	-	-	-	-
Caption-LSTM [61]	45.6	85.2	85.3	86.2	-	-	-	-	-	-	-	-
Contrastive-Binary [51]	46.8	84.5	85.6	85.8	-	-	-	-	-	-	-	-
ConVIRT [64]	47.6	85.9	86.8	87.3	34.7	77.4	80.1	81.3	17.8	72.5	82.5	92.0
GLoRIA-MIMIC [33]	51.7	87.1	88.7	88.0	40.6	86.6	89.2	90.4	22.1	67.3	81.5	88.6
MGCA (ResNet-50) [56]	50.2	87.6	88.0	88.2	41.0	88.6	89.1	89.9	24.5	72.0	83.5	90.5
M-FLAG (ResNet-50) [40]	55.9	87.8	88.4	88.6	41.8	88.8	89.4	90.2	25.4	72.2	84.1	90.7
PRIOR (ResNet-50) [16]	56.3	87.6	88.6	88.8	42.4	88.9	89.5	90.5	25.9	72.3	84.7	91.0
MLIP (Ours, ResNet-50)	56.9	87.8	88.7	88.9	42.9	88.8	89.6	90.6	26.3	73.0	85.0	90.8
MGCA (ViT-B/16) [56]	50.0	88.8	89.1	89.7	39.2	89.1	89.9	90.8	33.2	74.8	84.8	92.3
MLIP (Ours, ViT-B/16)	57.0	89.0	89.4	90.0	53.0	89.3	90.0	90.8	34.8	75.3	86.3	92.5

Table 3. **Image classification results in zero-shot scenarios and fine-tuning with 1%, 10%, and 100% of the training data in CheXpert, RSNA and COVIDx.** The evaluation metric used is AUC [%] for CheXpert and RSNA, and ACC [%] for COVIDx. The best results achieved for each setting are highlighted in red, while the suboptimal results are highlighted in blue.

pre-train GLoRIA with MIMIC-CXR datasets. We evaluate these approaches in the zero-shot scenario and with 1%, 10% and 100% of the data for fine-tuning, the results all outperform the SOTA. For a fair comparison, we pre-train our model with ResNet-50 and ViT-B/16 architecture. Except for the ViT-B/16 architecture, which yields comparable results to MGCA when fine-tuning is conducted using 100% of the available data, all others achieve better performance than the same architecture.

4.4. Ablation Study

Table 4 presents ablation results on semantic segmentation for both RSNA and SIIM datasets. We observe that leveraging knowledge as an intermediate medium for aligning image-text pairs in contrastive learning substantially enhances the model's performance. Moreover, category-level contrastive learning aids in mitigating false negatives, thereby improving the model's generalization. Global contrastive learning acts as a performance lower bound, complementing local and category-level approaches and yielding promising outcomes. Other ablation studies can be found in the supplementary 6.4.

4.5. Visualization

To further understand the inner workings of MLIP, we present learned local correspondences between radiographs and medical reports in the form of heatmaps and showcase

Tasks Setting			RSNA (Dice)			SIIM (Dice)		
Global ITA	Local ITA	Category-level ITA	1%	10%	100%	1%	10%	100%
	✓	✓	57.4	66.3	71.7	49.3	56.7	64.6
✓		✓	60.6	68.1	70.4	47.0	48.8	66.4
✓	✓		64.7	68.2	73.3	50.0	51.3	67.7
✓	✓	✓	67.7	68.8	73.5	51.6	60.8	68.1

Table 4. **Results of ablation study on proxy tasks for the semantic segmentation task.** Global ITA's pivotal role is evident, which can be attributed to the role of the divergence encoder.

the performance of MLIP on downstream tasks (semantic segmentation and object detection) in the supplementary 6.5. The visual evidence supports that MLIP excels in fine-grained feature extraction, boosting accuracy.

5. Conclusion

In this study, we propose MLIP, a novel medical visual representation learning framework that integrates language information into the visual domain. By introducing a divergence encoder to enhance representations and handle difficult samples, along with a language-knowledge-image alignment method guided by domain expertise, we alleviate false negative issue and imprecise alignment issue in other models. Experimental results demonstrate the effectiveness of MLIP on multiple datasets, even in zero-shot scenarios and with limited annotated data. Our proposed divergence encoder and knowledge-assisted alignment approach have broader applicability.

References

[1] 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. 1

[2] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *ICML*, pages 279–290. PMLR, 2020. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Michael Baumgartner, Paul F Jäger, Fabian Isensee, and Klaus H Maier-Hein. nndetection: a self-configuring method for medical object detection. In *MICCAI*, pages 530–539. Springer, 2021. 2

[5] Hedi Ben-Younes, Rmi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, pages 2612–2620, 2017. 5

[6] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *NAR*, 32 (suppl_1):D267–D270, 2004. 3, 5

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013. 5

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NIPS*, 33:9912–9924, 2020. 5

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[10] 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 *MICCAI*, pages 529–539. Springer, 2020. 1

[11] Kezhen Chen, Qiuyuan Huang, Yonatan Bisk, Daniel McDuff, and Jianfeng Gao. Kb-vlp: Knowledge based vision and language pretraining. In *ICML*, page 2021, 2021. 2

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 2, 3

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2

[14] 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 *ECCV*, pages 104–120. Springer, 2020. 5

[15] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *ACM MM*, pages 5152–5161, 2022. 2, 3

[16] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21361–21371, 2023. 7, 8, 2

[17] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020. 8

[18] Wanyun Cui, Guangyu Zheng, and Wei Wang. Unsupervised natural language inference via decoupled multimodal contrastive learning. In *EMNLP*, pages 5511–5520, 2020. 5

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS*, 26, 2013. 6

[20] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24 (9):1342–1350, 2018. 1, 2

[21] 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. 7, 1

[22] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *CVPR*, pages 3984–3993, 2018. 8

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

[24] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 8

[25] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *CVPR*, pages 1–6. Springer Berlin/Heidelberg, Germany, 2018. 7

[26] Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models. In *EMNLP*, pages 2281–2290, 2020. 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7, 1

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1

[30] J Healthcare. Object-cxr-automatic detection of foreign objects on chest x-rays, 2020. 7, 2

[31] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[32] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv e-prints*, pages arXiv–1811, 2018. 1, 2

[33] 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 *ICCV*, pages 3942–3951, 2021. 2, 3, 7, 8, 1

[34] 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 *AAAI*, pages 590–597, 2019. 2, 7, 1

[35] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2, 6, 7, 1

[36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 2

[37] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019. 2

[38] Zhe Li, Zhangyang Gao, Cheng Tan, Stan Z Li, and Laurence T Yang. General point model with autoencoding and autoregressive. *arXiv preprint arXiv:2310.16861*, 2023. 1

[39] Zhe Li, T. Yang Laurence, Xin Nie, BoCheng Ren, and Xianjun Deng. Enhancing sentence representation with visually-supervised multimodal pre-training. In *ACM MM'23*, 2023. 2

[40] Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2023. 7, 8

[41] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, pages 13753–13762, 2021. 5

[42] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908, 2020. 2

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[44] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *NIPS*, 29, 2016. 5

[45] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022. 2, 5

[46] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018. 1

[47] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1, 7, 2

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[50] 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): e180041, 2019. 7, 2

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019. 8

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794. Springer, 2020. 3

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 3, 5

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017. 3

[55] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. 1050(20):10–48550, 2017. 5

[56] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NIPS*. 2, 3, 5, 6, 7, 8, 1

[57] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):1–12, 2020. 7

[58] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *TACL*, 9:176–194, 2021. 2

[59] Zhaobin Wang, E Wang, and Ying Zhu. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review*, 53:5637–5674, 2020. 7

[60] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 1

[61] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057. PMLR, 2015. 8

[62] 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 *AAAI*, pages 3208–3216, 2021. 2

[63] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. 7, 2

[64] 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 *MLHC*, pages 2–25. PMLR, 2022. 1, 2, 6, 7, 8

[65] Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *MICCAI*, pages 320–328. Springer, 2017. 1

[66] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. In *ACL*, pages 1441–1451, 2019. 2

[67] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

MLIP: Enhancing Medical Visual Representation with Divergence Encoder and Knowledge-guided Contrastive Learning

Supplementary Material

6. Implementation Details

6.1. Implementation of Pre-traing

Config	Value
optimizer	AdamW [43]
learning rate	2e-5
weight decay	0.05
learning rate schedule	cosine
warmingup epochs	20
initial learning rate	1e-8
batch size	190
feature dim	128
τ_1	0.1
τ_2	0.07
τ_3	0.2
$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$	1

Table 5. **Experiment setting for MLIP pre-training.**

Architecture and Experiment Settings: Our model consists of a text encoder and an image encoder. Following [33], for the text encoder f_t, we adopt the 6-layer BioClinicalBERT [1] model from the HuggingFace library. To ensure consistency with [56], we either truncate each report or pad it with [PAD] tokens to maintain a fixed length of 112 text tokens. The text features are represented by the embedding of the [CLS] token in the final layer, resulting in a 768-d vector. Additionally, the word token-level features $\mathcal{S} \in \mathbb{R}^{768\times112}$ are represented by the embeddings of individual word tokens in the last layer.

As for the image encoder f_v, we utilize both ResNet50 [27] and ViT-16 [21], which is initialized with weights pretrained on the ImageNet-1k [49]. We divide the radiograph into patches of size 16×16, resulting in 196 visual tokens for each image. To process these patches, we prepend a learnable embedding (referred to as the [CLS] token embedding) to the sequence of embedded patches and input them into the vision transformer. The embedding of the [CLS] token in the last layer represents the image-level feature, resulting in a 768-d vector. Additionally, the embeddings of the individual patches in the last layer represent the visual patch-level features $\mathcal{P} \in \mathbb{R}^{768\times196}$.

For each modality, we design two divergence encoders o_t and o_v, whose parameters are initialized with the f_t, f_v. We train our model 50 epochs on 4 pieces of RTX 3090 GPUs. Table 5 shows the hyper-parameters during our pre-training.

Data Preprocessing: Following [56], we employ the JPG version of the MIMIC-CXR dataset [35] for pre-training our MGCA. For each image, we resize the larger dimension to 256 while padding zeros on the smaller side, resulting in a standardized image size of 256 × 256. During training, we randomly crop a 224 × 224 region from the image, normalize it to the range$(0, 1)$, and pass it through the image encoder. The medical reports are sourced from the original MIMIC-CXR dataset. Consistent with the approach in [56], we extract alphanumeric character sequences while discarding all other characters and symbols. Moreover, we exclude reports containing fewer than 3 tokens. To tokenize each report, we employ the WordPiece tokenizer implemented by BioClinicalBERT [1].

6.2. Implementation of Downstream tasks

Classification: For the fine-tuning of the CheXpert dataset [34], we exclusively utilize a batch size of 96. However, for the remaining linear classification settings, we employ a batch size of 48. Similar to the image preprocessing approach for the MIMIC-CXR dataset, we resize the larger dimension to 256 while padding zeros on the smaller side, resulting in a standardized image size of 256×256. Subsequently, for training, we randomly crop the image to 224×224, while for validation and testing, we perform a centered crop. The cropped image is then normalized to the range$(0, 1)$ before being inputted into the classifier.

In the linear classification setting, we freeze the pretrained image encoder (either ResNet-50 or ViT-B/16) and solely train the classification head, which is initialized with randomized weights. We utilize the AdamW optimizer [43] with a learning rate of 5e-4 and a weight decay of 1e-6. The image classifier is fine-tuned for 50 epochs, and early stopping is employed if the validation loss does not decrease for 10 consecutive runs. The checkpoint model with the lowest validation loss is saved for testing purposes.

Object Detection: In our object detection experiments, we follow a specific setup. However, for fine-tuning the

RSNA [50] and Object-CXR datasets [30], we only fine-tune 1% of the data using a batch size of 8. For other object detection tasks, we utilize a batch size of 16. We do not employ any data augmentation techniques. Our preprocessing involves resizing each image to 224×224 and normalizing the pixel values to the range$(0, 1)$. The resulting images are then fed into the object detection model.

Following [56], we utilize the AdamW optimizer with a learning rate of 5e-4 and weight decay of 1e-6. We do not employ a learning rate scheduler in our training process. In our experiments, we replace the Darknet-53 backbone with a pre-trained ResNet-50 model. Prior to training, we freeze the image encoder and randomly initialize the remaining layers. For predicting bounding boxes, we extract three-stage features from the $2nd$, $3rd$, and $4th$ bottleneck building blocks. The anchors used in our experiments are consistent with those described in the original paper [47], but we rescale them based on the input image size of 224×224.

We fine-tune the object detection model for 50 epochs and employ early stopping if the validation loss does not decrease for 10 consecutive runs. Finally, we save the checkpoint model with the lowest validation loss for testing purposes.

Semantic Segmentation: We assess the segmentation performance of our MLIP on two datasets: the SIIM Pneumothorax [63] Dataset and the RSNA Pneumonia dataset [50]. We follow the data preprocessing procedure outlined in [56]. For the RSNA dataset, we generate pneumonia region masks based on the provided bounding boxes. Specifically, we resize both the images and masks to a size of 512×512. To augment the training set, we employ the ShiftScaleRotate function from the albumentations Python library [2], which applies random affine transformations such as translation, scaling, and rotation. The specific augmentation parameters are as follows: a rotation limit of 10, a scale limit of 0.1, and an augmentation probability of 0.5. After augmentation, we normalize the images to the range$(0, 1)$ before feeding them into the semantic segmentation model.

We adopt the U-Net [48] architecture with a ResNet-50 encoder implemented by the Sementation-Models-PyTorch library [3] to evaluate the semantic segmentation performance of the pre-trained ResNet-50 model. During training, we utilize the AdamW optimizer with a learning rate of 5e-4 and weight decay of 1e-6. Following the approach in [33], we employ a combined loss function consisting of $\alpha\times$ FocalLoss and DiceLoss, where α is set to 10. We fine-tune the semantic segmentation model for 50 epochs and apply early stopping if the validation loss does not decrease for 10 consecutive runs. The checkpoint model with the lowest validation loss is saved for testing.

[2]https://albumentations.ai/

[3]https://github.com/qubvel/segmentation_models.pytorch

Methods	CheXpert Image-to-text Retrieval			
	Prec@1	Prec@2	Prec@5	Prec@10
ConVIRT [64]	20.3	19.8	19.7	19.9
GLoRIA [33]	29.3	29.0	27.8	26.8
PRIOR [16]	40.2	39.6	39.3	38.0
MLIP (Ours)	41.7	40.3	39.0	39.4

Table 6. **Image-to-text retrieval results on CheXpert 5x200.**

To assess the semantic segmentation performance of the pre-trained ViT-B/16 model, we employ the SETR-PUP (progressive upsample) architecture introduced in [67], replacing the encoder with the pre-trained ViT. Our implementation is based on this repository [4]. Unlike the ResNet-50 variant, we resize each image to a size of 224×224 before feeding it into the SETR-PUP model. In this setup, we freeze the pre-trained image encoder and only train the decoder portion. The loss function and other training hyperparameters remain the same as in the ResNet-50 U-Net fine-tuning setting.

6.3. Image-text Retrieval

We perform image-to-text retrieval experiments on the CheXpert 5x200 dataset. Due to the unavailability of reports in CheXpert, we randomly choose 1000 reports from MIMIC-CXR, with 200 samples exclusively assigned to each of the 5 diseases. We evaluate the performance using Precision@K metric to assess the matching between the retrieved report and the query image label. The results presented in Table 6 clearly indicate that MLIP outperforms both GLoRIA and PRIOR with a significant margin.

6.4. Ablation Study

Table 7 presents the ablation results of our main contributions on the object detection task. Our proposed divergence encoder enhances feature diversity and enables the model to better adapt to challenging samples. With the assistance of expert knowledge, the alignment between medical images and medical reports becomes more efficient. Lastly, the proxy tasks designed in our approach strengthen the model's ability to discriminate negative samples. Additionally, we conduct ablation study on five losses, as shown in Table 8.

6.5. Visualization

Visualization of Activated Regions: We introduce a re-weighting mechanism to assess the importance of visual tokens in generating image-level representations. To provide visual evidence of this process, Figure 4 depicts the weights assigned to the visual tokens by the ViT model. It is important to note that these weights are obtained by averaging the

[4]https://github.com/fuying-wang/MGCA

Tasks Setting			RSNA (mAP)			Object CXR (mAP)		
DE	KA	TS+ITM	1%	10%	100%	1%	10%	100%
✗			11.7	13.4	19.8	1.2	12.4	16.1
	✗		13.2	15.6	21.6	2.8	15.3	17.9
		✗	16.2	17.9	23.5	3.7	16.8	18.8
✔	✔	✔	17.2	19.1	25.8	4.6	17.4	20.2

Table 7. **Results of ablation study on main contributions for the object detection task,** including divergence encoder (DE), knowledge augmentation (KA) and text-swapping + image-text matching (TS+ITM).

Loss Setting					RSNA (Dice)			SIIM (Dice)		
$\mathcal{L}_{il}$	$\mathcal{L}_{tl}$	$\mathcal{L}_{cl}$	$\mathcal{L}_{itm}$	$\mathcal{L}_{ts}$	1%	10%	100%	1%	10%	100%
✗					57.4	66.3	71.7	49.3	56.7	64.6
	✗				60.6	68.1	70.4	47.0	48.8	66.4
		✗			64.7	68.2	73.3	50.0	51.3	67.7
			✗		65.8	68.3	73.0	50.7	53.4	66.9
				✗	66.0	68.5	73.2	50.9	53.7	67.2
✔	✔	✔	✔	✔	67.7	68.8	73.5	51.6	60.8	68.1

Table 8. **Results of ablation study on five losses for the semantic segmentation task.**

attention weights from the last layer of ViT across multiple heads. The highlighted pixels in the figure indicate regions with relatively high attention weights. This visualization demonstrates that ViT has the capability to automatically learn and allocate attention to critical regions by aligning cross-modal instance-level representations.

Visualization of Downstream Tasks Effects: We conduct visual effect evaluations for downstream tasks in Figure 5, specifically object detection and semantic segmentation. Drawing upon domain expertise, we conclude that our model achieves competitive performance in lesion segmentation and lesion detection.

Visualization of Important Words: Figure 3 presents three instances of radiology reports. In these reports, the top 5 words with the highest weights, indicating their importance, are highlighted in red font. It is important to note that these weights are obtained by averaging the attention weights from BERT's last layer across multiple heads. Upon analysis, we observe that the majority of the highlighted words (such as pneumothorax, bilateral infiltrates, lung collapse) are closely associated with the patients' medical conditions. This observation suggests that the BERT effectively learns to prioritize disease-related words during the alignment of cross-modal instance-wise embeddings.

presence of bilateral infiltrates with consolidation in the lower lobes. This pattern suggests a possible diagnosis of pneumonia. Additionally, there is evidence of a small right-sided pleural effusion. Further evaluation and appropriate management are recommended.

right-sided consolidation in the middle lobe. No evidence of pleural effusion or pneumothorax is noted. The consolidation is suggestive of pneumonia. Antibiotic therapy and close monitoring are recommended for the patient.

bilateral pleural effusions with associated volume loss in the lung bases. No evidence of pulmonary infiltrates or pneumothorax is observed. Further evaluation with ultrasound or CT scan is recommended to determine the underlying cause of the effusions.

presence of a large right-sided pneumothorax with lung collapse. Immediate intervention, such as the insertion of a chest tube, is necessary to relieve the pneumothorax and reinflate the lung. Urgent attention is recommended

Figure 3. **Visualization on the importance of words according to the attention weights learned by text encoder.** Words with top 5 highest weights are highlighted by red font.

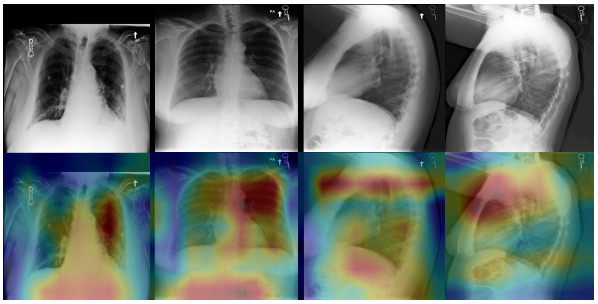

Figure 4. The visualization of the activated visual tokens in our ViT. The highlighted pixels in the visualization correspond to regions that have been identified as important by the model. These regions have been learned through the training process, where the model dynamically assigns attention to specific visual tokens based on their relevance to the task.

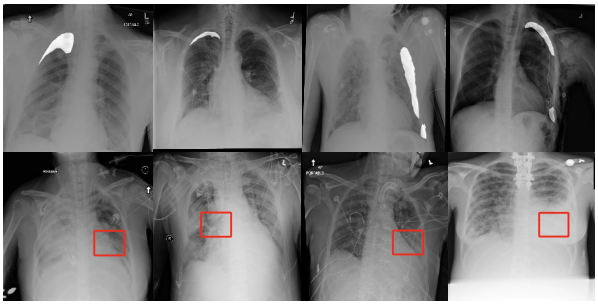

Figure 5. Visualization results for downstream tasks. The top row showcases the results of the semantic segmentation task, while the bottom row displays the results of the object detection task.